nature portfolio

Peer Review File

Distinct hydrologic response patterns and trends worldwide revealed by physics-embedded learning

Corresponding Author: Dr Chaopeng Shen

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
The text describes a novel hybrid hydrological model, δHBV2, that combines conceptual modeling (HBV) with neural networks and differentiable routing. It's designed to learn from thousands of sites and produce high-resolution outputs. The planned extensive evaluation against established Global Water Models (GWMs) across diverse river systems (mixed-anthropogenic, natural, and small basins) is also a significant aspect, setting the stage for important comparisons and potential new insights.

This work has high potential significance for hydrology. δHBV2 offers a hybrid approach that bridges process-based and data-driven modeling, allowing for both physical grounding and powerful data learning. Its scalability and high-resolution output address critical modern hydrological needs. The direct comparison with GWMs, a first for this model, is crucial for validating its performance and positioning it as a potentially more flexible and data-driven alternative in the field. The model itself, δHBV2, is presented as original, having been recently introduced in Song et al.

The methodology appears conceptually sound and robust, aligning with current advanced hydrological modeling trends. The hybrid, differentiable, multiscale approach, combined with extensive training and rigorous evaluation against benchmarks, goes beyond state of the art. However, the paper lacks somehow sufficient detail for reproduction. Specifics on the neural network architecture, differentiable HBV/Muskingum-Cunge implementations, data preprocessing, and training/testing splits would be essential for replication.

I find it difficult to see progress in hydrological pattern understanding and insights if as training inputs, model outputs are used, because as I understand it, this is the case here, albeit the paper shows considerable advances in modeling and is worth considering for publication, after revisions.

Also, it is still true that even this hybrid approach seems to have a lot of limitations in understanding patterns in areas where there have been considerable human activities and pressure on the water cycle. Thus, there is still much room for improvement.

In my opinion, the authors need to explain these two last points more in detail and outline their implications.

(Remarks on code availability)
NA

Reviewer #2

(Remarks to the Author)
This is an important paper that defines the state-of-the art in global-scale hydrologic modeling. The model, grounded in an interpretable physical conceptualization, leverages AI and big data to learn spatially explicit parameters. Its performance exceeds that of other global physical model and deep-learning benchmarks. It even performs reasonably well in arid regions, which have typically been challenging for hydrologic model performance. Because of the model's physical interpretability, the results can be interrogated to show new trends in hydrological signatures that have not been recognized

before.
Overall, the methodology is sound and the presentation is polished. I have a few comments, intended to make the paper more accessible to the broad audience of Nature Communications.
The biggest shortcoming was some gaps in the methods description. First, a clearer explanation of the LSTM that learns parameter values in the differentiable model is needed. The Supplementary Text refers to “real parameter values,” but how are these determined? In general, I am unclear how the time series of parameter values that the LSTM is trained on is generated. The text indicates that sequence-to-sequence training is used, but additional details such as sequence length would be helpful.
Second, the Methods lacks a description of the benchmark models to which the performance of the differentiable HBV models are compared. The strong performance (though still less than that of the differentiable HBV model) of the GWM 0 model benchmark is intriguing, and I was hoping to find out more about how it was bias-corrected/post-processed but disappointed that the Methods or supplemental text did not contain this information. Some description of the LSTM benchmark would also be appreciated.
Third, the trends and changes observed in the model results were dominantly attributed to climate change, which is probably generally true. However, the possible role of land-use change or other major changes such as dam building were not discussed (see, for example, the second minor comment below). This seems like an important oversight in the discussion overall.

MINOR/LINE-BY-LINE COMMENTS
• Regarding the sentence: “Where ET/P increases substantially, it could be a sign of insufficient water for the ecosystem.” I don’t agree with this statement. An increase in this ratio reflects more water use over time but not insufficient water. Rather, a leveling off would reflect the transition to water-limited conditions.
• Regarding the text: “However, central-western USA, southern Africa, the south fringe of the Sahara Desert, central Asia, and Ethiopia have seen large declining (more negative) trends, highlighting increasing streamflow variability. These changes are regional but can be substantial; we believe they may be caused by changes in precipitation intensities.” Land-use change can be another major contributor to these observations; I was surprised this wasn’t mentioned here.

(Remarks on code availability)
Though I did not try to run the code, I did look at the repository and found it to be fairly straightforward and consistent with my expectations.

Reviewer #3

(Remarks to the Author)
I provide a review of the manuscript entitled “Distinct hydrologic response pattern and trends worldwide revealed by physics-embedded learning”. he manuscript describes the development of a physics-based machine learning hydrological model integrated with a differentiable routing model to improve the simulation of spatio-temporal streamflow dynamics globally and characterise hydrological signatures (including elasticity, flashiness and temporal autocorrelation). This new model outperforms other global water models in its ability to partition precipitation into different water components and to identify the trends and integrated hydrological signatures of these components. Much of the model scheme is detailed, as is the observational data used to train and test the new model (δHBV2). However, significant improvements are needed, such as improving the readability of the text, reducing the intensive use of brackets and better organising the technical parts of the manuscript for a general readership. I also have some concerns about the methodology. In short, I cannot recommend direct publication in Nature Communications in its current form.
1. The numbers of pages are missing. Hence, I can only comment by referring specific sentences.
2. Main concern: In the abstract, baseflow fraction is highlighted, while the method is falling short, especially the parameterization uncertainty and the comparison between simulated and observed baseflow. Also, in Fig. S3, the global map of simulated baseflow have much larger values compared to an existing global map of baseflow index from ensemble observations and metrics. See Xie 2024 (Xie, J., Liu, X., Jasechko, S. et al. Majority of global river flow sustained by groundwater. Nat. Geosci. 17, 770–777 (2024).).
3. In terms of long-term water balance, is it calculated based on a part of full periods of datasets or not? Also, I found the authors constrain the routing model with MERIT unit basins on the sum of all MERIT basins belonging to one drainage area. Does this mean the water balance is constrained at the drainage area level while not at its finer or coarser levels?
4. Can the authors better introduce different Global water models? Many models are mentioned in the first line in the second page, while other names are presented in Fig. 1. They need to summarize their corresponding spatio-temporal resolutions as well, to see if the comparison with δHBV1 and δHBV2 is fair.
5. I highly recommend including a schematic figure to better illustrate which hydrological processes can be advanced by different parts of model schemes in the introduction in the last paragraph in page 3 but also heavy technical text elsewhere.
6. Can the authors discuss the computational feasibility of large-scale simulations which is not yet discussed?
7. I found it’s difficult to change between different training and testing datasets. Why not filter consistent basin for different model training and testing?
8. Fig. 1 (d-e) show trends of water components from 2001-2020. It confuses me that why 2020 is used while other parts of the manuscript have 2015? And is the result from δHBV2?
9. Fig. 1 (c) at which temporal interval the ratio between A/P is calculated?
10. Presenting ET/P trends come with a surprise. Some discussions seem to come from the trends of Q/P but not ET/P (e.g. “Papua New Guinea have seen increasing blue water while Central Europe and subtropical and midlatitude regions in South America have seen increasing green water”), while Q/P is not presented and is not simply opposite to ET/P. Same issue for “Where blue-water fraction increases and baseflow ratio decreases in tandem”.

11. A few suspectable statements, such as “these changes are correlated with trends in precipitation (supplementary figure s2)”. Can you report correlation coefficient?
12. Main concern: “Since δHBV2 conceptual model backbone is not fundamentally different from other models, the parameterization through big-data appears to have greatly improved the sensitivity to decadal-scale changes…” The new model involves new observations and a differentiable routing model to refine the upstream basins’ simulation and aggregate it to a drainage area. To do so, MERIT high resolution DEM data are used. Can the authors further elaborate how this largely improve the overall model performance? And assuming, if using the same data for regionalized parameterization of the old model scheme, can that already provide a good model performance compared to traditional global water models?
13. In semi-arid regions, summer elasticity ε is high. This is not only because low flow is driven by storage-dependent groundwater releases which are controlled by precipitation accumulation in prior months, but also because the semi-arid regions have higher evapotranspiration.
14. Winter ε is low. But I don’t agree with the current explanation that “outcome of more linear responses after thresholds are fulfilled” which is also very abstract for understanding. I think it is likely related to snow formation during winter.
15. Figure 4: Is there any meaning to arrange rivers clockwise?
16. Section Daily streamflow variability and trends: too many brackets intervening reading.
17. Main concern: Text “Both δHBV2 and LST thus represent state-of-the-art, allowing them to accurately capture the baseflow-surface runoff partitioning and flow variability” and Figure 5: why LSTM alone is mentioned here, is it related to δHBV1 or δHBV2 or an independent training? Also, if a LSTM model scheme directly pursue a similar simulation accuracy as δHBV2, why need the large effort on δHBV2? I encourage the authors to clarify their model novelty better.
18. Figure 5: different datasets should also be explained directly in the main text, as different model performance is well related to different datasets. The authors need to explain why different datasets provide different model performance as well.
19. Section Global datasets: I suggest adding a data table with spatio-temporal information.
20. Sentence “Such a scale also allows nonlinear and threshold behaviors to better manifest, e.g., concentrated rainfall in a small mountainous region can lead to substantial saturation and runoff but…”: check the grammar
21. Ensure the abbreviations are clearly introduced throughout the manuscript.

(Remarks on code availability)

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
In my opinion, the authors addressed all comments in a satisfactory manner, and I'd be happy to see this manuscript published.

(Remarks on code availability)

Reviewer #2

(Remarks to the Author)
I continue to be highly supportive of this paper and believe it represents a significant advancement of broad interest and that it generates novel insight on global changes in water distribution. I am quite satisfied with the authors’ responses to all of the review comments received. I had a few new insights when reading the manuscript (which do not detract from my supportiveness of the paper) and two minor editorial comments below.

It occurred to me when rereading the manuscript that, despite the very large improvements, an overall average R-squared of 0.43 in simulated vs. observed trends still leaves much room for improvement. I was a little surprised by the spatial distribution of some of the positive and negative trends and found myself wondering which ones were reliable (especially given that some of these results depart from what has been previously reported, as the authors allude to with regards to previous Budyko analyses). Could the authors consider using, for example, stippling (as done by the IPCC) to demarcate the reliability of trends in different regions, either based on the outcomes of the temporal test (for Q) or R-squared of trends? They would need to pick thresholds in performance measures to determine which regions should be stippled, but I do think this would lead to a more honest assessment.

It also occurs to me that there may be confusion triggered by the use of “green water fraction” to exclusively mean evapotranspiration fluxes. Many people will understand “green water” to indicate soil water. I suggest a slight change in terminology to “green water fluxes” (while retaining parenthetic references to evapotranspiration to distinguish these from downward or lateral fluxes) to ameliorate this potential source of confusion.

Line 137-138: Change to “Due to the recent timescale of its introduction…”
Line 716: tested --> test

(Remarks on code availability)

Reviewer #3

(Remarks to the Author)
The authors have well addressed my concerns and questions. The code and guidelines look clean and are easy to follow. I'm happy to recommend a direct publication.

(Remarks on code availability)

Many thanks to you for handling our manuscript and to the reviewers for their constructive comments. In the last round of comments, reviewers generally gave positive feedback on our modeling progress, but demanded clarifications, debated some of our arguments, asked about the discrepancy in baseflow ratio from literature studies, and requested revision in writing style. No major new experiment was requested. We have completed a round of revision. In terms of new figures, we (i) have provided new results showing how ET/P is negatively correlated with Precipitation to support our argument (Figure S7 on page 6 of this doc), (ii) explained how our simulated baseflow recession behaviors agree with the observed while offering new insight a at the local scale (new Figure S6 on page 8 of this document), (iii) updated model sketch (revised Figure S1 on page 12 of this doc), and (iv) added local runoff to precipitation ratio (Q'/P, revised Figure S4 on page 14 of this doc). We have revised the manuscript's writing extensively, both in Results and Methods, to make clarifications and improve the flow. We believe the revised manuscript is much improved. We thank you again for your processing. Below, in curly brackets {} we refer to the line numbers in the clean revised manuscript without track change.

Chaopeng

Reviewer #1

The text describes a novel hybrid hydrological model, δHBV2, that combines conceptual modeling (HBV) with neural networks and differentiable routing. It's designed to learn from thousands of sites and produce high-resolution outputs. The planned extensive evaluation against established Global Water Models (GWMs) across diverse river systems (mixed-anthropogenic, natural, and small basins) is also a significant aspect, setting the stage for important comparisons and potential new insights.

This work has high potential significance for hydrology. δHBV2 offers a hybrid approach that bridges process-based and data-driven modeling, allowing for both physical grounding and powerful data learning. Its scalability and high-resolution output address critical modern hydrological needs. The direct comparison with GWMs, a first for this model, is crucial for validating its performance and positioning it as a potentially more flexible and data-driven alternative in the field. The model itself, δHBV2, is presented as original, having been recently introduced in Song et al.

The methodology appears conceptually sound and robust, aligning with current advanced hydrological modeling trends. The hybrid, differentiable, multiscale approach, combined with extensive training and rigorous evaluation against benchmarks, goes beyond state of the art. However, the paper lacks somehow sufficient detail for reproduction. Specifics on the neural network architecture, differentiable HBV/Muskingum-Cunge implementations, data preprocessing, and training/testing splits would be essential for replication.

Thanks for your suggestion, the structures of all neural networks and hyperparameters have been mentioned in the supplementary text S2-S5. The input data preprocessing has also been provided in supplementary text S3. The implementations of HBV are provided in supplementary Table S7. Model training and testing split has been provided in the *Model training and evaluation* in *Methods* section.

1. I find it difficult to see progress in hydrological pattern understanding and insights if as training inputs, model outputs are used, because as I understand it, this is the case here, albeit the paper shows considerable advances in modeling and is worth considering for publication, after revisions.

Thanks for your effort in understanding our manuscript.
We need to clarify that at no point we used any model outputs as inputs to δHBV2. The NN module embedded inside the workflow is not directly trained. The training target is observed discharge. However, it is worthwhile to add this sentence to the Discussion "*We also emphasize that we neither need nor have ground truth data to directly supervise the neural network outputs (i.e., the physical parameters). Gradient information is propagated backward from the loss function defined between simulated and observed streamflow, through the process-based model, to update the neural networks in this end-to-end framework.*" {line 650}.

Although we did not modify the structure of the hydrological model, parameterization remains a persistent challenge in global-scale simulations. Few observations are incorporated into traditional global-scale hydrologic model parameterization. Even in the case of GWM0, which performs limited calibration, the process is conducted at the annual scale. In differentiable modeling, parameterization is replaced with a neural network trained to learn the general relationship between inputs and physical parameters. With extensive global streamflow data, parameterization becomes more effective and the accuracy of streamflow simulation is greatly improved, especially at this high resolution.

2. Also, it is still true that even this hybrid approach seems to have a lot of limitations in understanding patterns in areas where there have been considerable human activities and pressure on the water cycle. Thus, there is still much room for improvement.
In my opinion, the authors need to explain these two last points more in detail and outline their implications.

Agreed. We have indeed documented the limitations in a specific section, particularly around human activities. For example:
"*While land use and human water use could have contributed to the shifts in water balance, these two processes are not explicitly simulated by the current version of the model (see Limitations).*" {line 245}

"*Examined regionally, δHBV2 tends to perform well in boreal or northern midlatitude rivers, but tends to be challenged in river basins where human water use is significant, ...*"{line 313}

Nevertheless, we believe even currently, the model has some capacity to implicitly adapt to human water use through big-data learning. The learned physical parameters may partially compensate for the lack of reservoirs. This can be evident from the fact that the drop in performance from natural to human-dominated large rivers are even less drastic than that of GWM #2, a model known for human water use modules. With respect to the larger reservoirs, we definitely need a more capable reservoir and water use module, both of which are now in development.

To clarify this, we added: *"Overall, due to structural limitations in both the HBV and Muskingum-Cunge models, δHBV2 has limited capacity to represent human impacts such as reservoirs, land use changes, and human water use, although physical parameters learned from observational data can partially compensate for some of the resulting errors."*{line 448}

Reviewer #2

This is an important paper that defines the state-of-the art in global-scale hydrologic modeling. The model, grounded in an interpretable physical conceptualization, leverages AI and big data to learn spatially explicit parameters. Its performance exceeds that of other global physical model and deep-learning benchmarks. It even performs reasonably well in arid regions, which have typically been challenging for hydrologic model performance. Because of the model's physical interpretability, the results can be interrogated to show new trends in hydrological signatures that have not been recognized before.
Overall, the methodology is sound and the presentation is polished. I have a few comments, intended to make the paper more accessible to the broad audience of Nature Communications.

Thank you for your comments, and we have gone through your suggestions one by one.

1. The biggest shortcoming was some gaps in the methods description. First, a clearer explanation of the LSTM that learns parameter values in the differentiable model is needed. The Supplementary Text refers to "real parameter values," but how are these determined? In general, I am unclear how the time series of parameter values that the LSTM is trained on is generated. The text indicates that sequence-to-sequence training is used, but additional details such as sequence length would be helpful.

Thanks for asking. We neither have nor need ground truth for the physical parameters. The beauty of this workflow is that the network is indirectly supervised, and the only loss is calculated on the simulated streamflow and backpropagated through the model calculations. This

is the soul of differentiable modeling which enables neural networks and physical calculations to be trained together. To clarify this, we added the following text: "*We also emphasize that we neither need nor have ground truth data to directly supervise the neural network outputs (i.e., the physical parameters). Gradient information is propagated backward from the loss function defined between simulated and observed streamflow, through the process-based model and physical parameters, to update the neural networks in this end-to-end framework".* {line 650}

As for the problem of sequence length, we revised the sentence as:*"We trained our model with a sequence of 365 days following a 365-day warmup period, and static parameters were derived from the last day of LSTM outputs."* {line 575}

2.Second, the Methods lacks a description of the benchmark models to which the performance of the differentiable HBV models are compared. The strong performance (though still less than that of the differentiable HBV model) of the GWM 0 model benchmark is intriguing, and I was hoping to find out more about how it was bias-corrected/post-processed but disappointed that the Methods or supplemental text did not contain this information. Some description of the LSTM benchmark would also be appreciated.

We anonymized the GWMs to avoid distraction from the main content of the paper. Upon the request of the reviewer, we added this description about how GWM0 performs bias correction in the Supplementary Information (Text S6):

GWM0 takes a two step-wise method to calibrate and bias-correct the rainfall-runoff process. The rainfall-runoff equation in the GWM0 can be described as:

$$R_l = P_{eff}\left(\frac{S_s}{S_{smax}}\right)^{\gamma_l}$$

The rainfall-runoff exponent, denoted as γ_l*, is the only calibratable parameter ranging from 0.3 to 3. This parameter is calibrated with the objective to allow annual mean flow bias to fall within 1% of observation. The calibration is carried out in around 1,300 large basins around the world. All grid points within the calibrated ones will get the same value, while grid cells outside will be extrapolated via a linear regression method with mean annual temperature, mean available soil water capacity, fraction of open water bodies, mean basin land surface slope, fraction of permanent snow and ice, and the aquifer-related groundwater recharge factor as the regressors. If values at the extremes (i.e.,* γ_l *= 0.3 or 3) still result in model underestimation or overestimation; it would indicate systematic bias. In that case, a post-simulation linear correction factor was fitted and applied to all grid cells within this basin to achieve the abovementioned 1% bias target. If this correction alone still does not meet the 1% bias criterion, a second adjustment is applied directly to the simulated discharge only at the gauging stations. This secondary incrementally increases or decreases discharge along the flow path from upstream cells to the gauging stations without altering the runoff fields.*

We also add some description in the main to facilitate the reader:
"Among these models, GWM0 is unique in calibration and applying a two-stage bias-correction strategy. First, an area-based correction coefficient is uniformly applied to all grid cells within a basin to adjust runoff, aiming to match observed streamflow at the annual scale within a 1% margin. If this adjustment is insufficient, a second-stage correction is applied directly to the simulated streamflow at target gages, without further modifying the runoff (detailed description can be found in Supplementary Text S6)." {line 745}

3.Third, the trends and changes observed in the model results were dominantly attributed to climate change, which is probably generally true. However, the possible role of land-use change or other major changes such as dam building were not discussed (see, for example, the second minor comment below). This seems like an important oversight in the discussion overall.

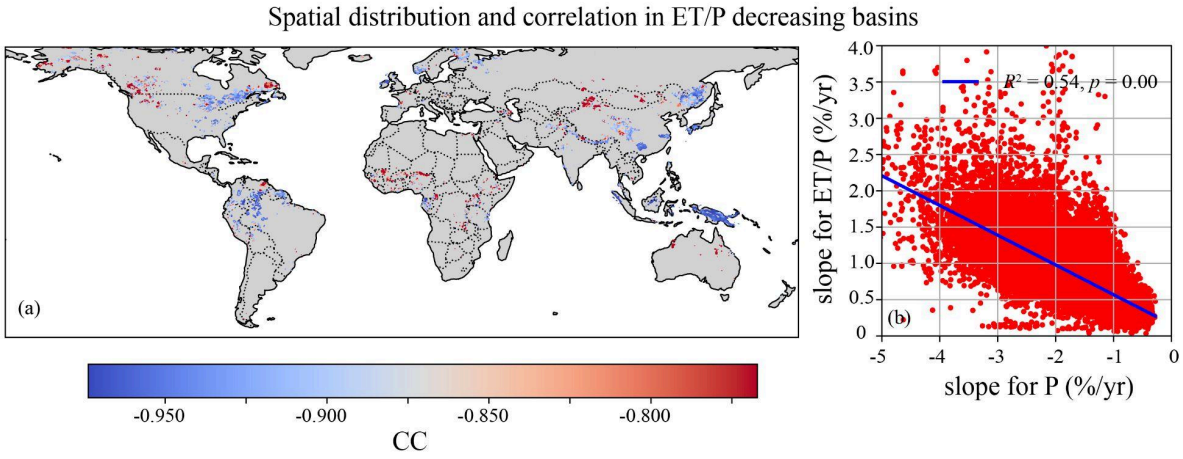
Good question. While we agree both climate change and landuse change both alter hydrology, the predominant effect captured by this model is climate change because land use was used as a static input. We added this sentence for clarification: "*In our training or forward simulations, land-use inputs are regarded as static, so the model's long-term output shifts mainly reflect changes in the climate inputs. It is possible that the dynamic parameters produced by the NNs jointly trained on streamflow data, can capture some interannual covariation of vegetation-related characteristics due to climate shifts, but we do not expect this to be a major effect. The impacts of land use change, e.g., in central Asia and Ethiopia, are not explicitly simulated, which should be considered in future efforts.* " {line 453}

MINOR/LINE-BY-LINE COMMENTS

4. Regarding the sentence: "Where ET/P increases substantially, it could be a sign of insufficient water for the ecosystem." I don't agree with this statement. An increase in this ratio reflects more water use over time but not insufficient water. Rather, a leveling off would reflect the transition to water-limited conditions.

We have revised this part as "*Where model-diagnosed MERIT-basin-scale ET/P increases substantially, it often results from precipitation declines. In fact, we find negative correlation between changes in ET/P and precipitation in space and time (Supplementary Figure S7). When annual precipitation declines, there is less excess water that can overcome the storage thresholds to become runoff, so the more precipitated water exits as green water, and blue water drops disproportionately.The prominent shifts toward green water which occur in Germany, central Siberia, southern Brazil, central Chile, the Congo basin, and northern Australia are due to decline precipitation as discussed above (Supplementary Figure S4) and suggest a disproportionate decrease in streamflow available for human use in these regions.*"

To clarify, we wrote this to describe the *model-diagnosed ET vs P dynamics*, which does not explicitly simulate human water use, because correlating the map with rainfall changes suggests regions with ET/P change are mostly where precipitation is reduced. What we plotted is local hydrologic flux for the MERIT basins. The figure below shows this --- the temporal correlation between annual precipitation and ET/P is mostly negative, and the spatial correlation in their change trends are also negative. While we agree water use could be part of it, the model is unlikely to respond to water use increase because we do not have that term in the model and ET is calculated based on energy balance and energy-water co-limitations. At this scale at which this plot is made (MERIT basins), it is unlikely water use constitutes a major signal worldwide (it might for those large reservoirs). It is possible that the model internally compensate for human water use that is correlated with energy input, but the figure below undoubtedly shows simulated ET/P is responding to precipitation changes in an inverse manner, both in time and in space.



Revised Figure S7. (a) The temporal correlation coefficient between annual ET/P and precipitation anomalies, for basins where ET/P has a significant negative trend (same locations as Figure 1d), are all negative. (b) The spatial correlation between precipitation change trend in 2001-2020 with the ET/P trend (from simulations) are also negative, for basins where the ET/P trend slope ranges from 0 to 4 and exhibits a statistically significant trend (p-value < 0.05).

We have incorporated this figure as Figure S7 in the revised Supplementary Information. And refer to this figure from the main text.

5. Regarding the text: "However, central-western USA, southern Africa, the south fringe of the Sahara Desert, central Asia, and Ethiopia have seen large declining (more negative) trends, highlighting increasing streamflow variability. These changes are regional but can be substantial;

we believe they may be caused by changes in precipitation intensities." Land-use change can be another major contributor to these observations; I was surprised this wasn't mentioned here.

Thanks for your suggestions, similar to the question #3, we add the sentence in line 402.
"*In our training or forward simulations, land-use inputs are regarded as static, so the model's long-term output shifts mainly reflect changes in the climate inputs. It is possible that the dynamic parameters produced by the NNs jointly trained on streamflow data can capture some interannual covariation of vegetation-related characteristics due to climate shifts, but we do not expect this to be a major effect. The impacts of land use change, e.g., in central Asia and Ethiopia, are not explicitly simulated, which should be considered in future efforts.(line 454)* "

Reviewer #3

I provide a review of the manuscript entitled "Distinct hydrologic response pattern and trends worldwide revealed by physics-embedded learning". The manuscript describes the development of a physics-based machine learning hydrological model integrated with a differentiable routing model to improve the simulation of spatio-temporal streamflow dynamics globally and characterise hydrological signatures (including elasticity, flashiness and temporal autocorrelation). This new model outperforms other global water models in its ability to partition precipitation into different water components and to identify the trends and integrated hydrological signatures of these components. Much of the model scheme is detailed, as is the observational data used to train and test the new model (δHBV2). However, significant improvements are needed, such as improving the readability of the text, reducing the intensive use of brackets and better organising the technical parts of the manuscript for a general readership. I also have some concerns about the methodology. In short, I cannot recommend direct publication in Nature Communications in its current form.

Thanks for your suggestions, and we have gone through all your suggestions and please see the detailed revision as follows. We have indeed reduced the use of parenthesis throughout the manuscript.

1. The numbers of pages are missing. Hence, I can only comment by referring to specific sentences.

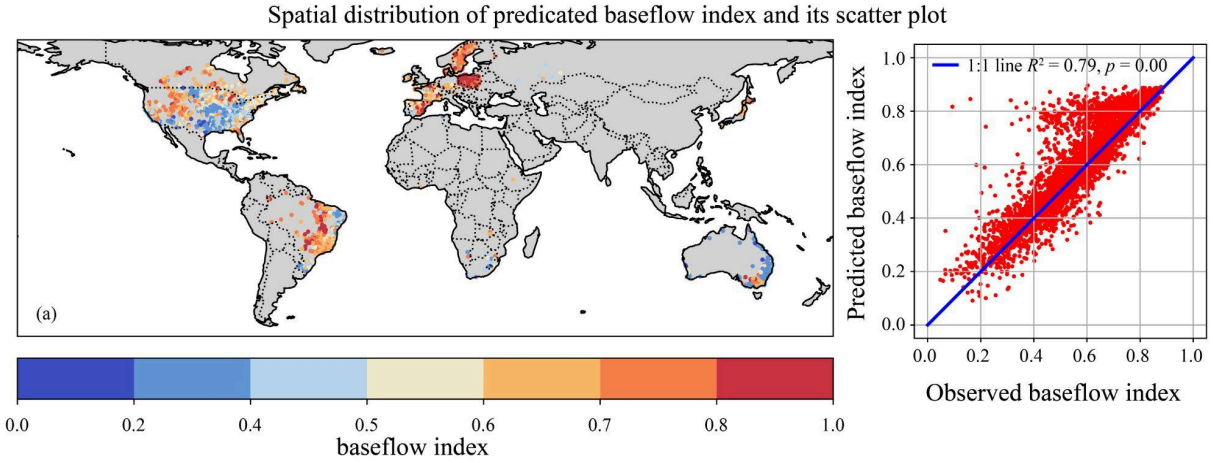
We added the number of pages and lines. Apologies for some software hiccups in the first submission.

2. Main concern: In the abstract, baseflow fraction is highlighted, while the method is falling short, especially the parameterization uncertainty and the comparison between simulated and observed baseflow. Also, in Fig. S3, the global map of simulated baseflow have much larger

values compared to an existing global map of baseflow index from ensemble observations and metrics. See Xie 2024 (Xie, J., Liu, X., Jasechko, S. et al. Majority of global river flow sustained by groundwater. Nat. Geosci. 17, 770–777 (2024).).

Thank you for this note. We provide a table below that breaks down the differences. We would like to clarify that our baseflow ratio is a diagnostic variable at MERIT-basin scale defined as the model-simulated ratio of local baseflow to total runoff output. In contrast, Xie et al., employed statistical baseflow separation methods to estimate the baseflow index at the gauge level, some of which are very large. The definition and value of the baseflow therein depends on some algorithmic parameters used for baseflow separation. Furthermore, it is well known that hydrologic fluxes are scale-dependent [1,2]. As a result, it is not surprising that the fluxes are different.

	Xie et al., 2024	This paper (δHBV2)
Conceptual difference	Baseflow separation applied at streamflow gages.	Local baseflow: total runoff ratio from model-estimated fluxes.
Scale difference	Mostly much larger basins	Median 37 km^2 basins
Dependence	On baseflow separation algorithms and parameters	On δHBV2 structure and parameters trained on streamflow data.



Revised Figure S6. Gauge-based baseflow ratio calculated using our simulated daily streamflow and an ensemble of 12 baseflow-separation methods on the training stations from 1980 to 2015. (a) map of the baseflow index for the training stations (b) 1 vs. 1 plot of baseflow ratios calculated based on our simulated flows and observed flows.

In fact, if we apply the same baseflow separation algorithms on our simulated flow at the streamflow gages, we would have obtained very similar baseflow ratios as what Xie et al., 2024

obtained (Figure S6), which suggest our overall simulated behavior is similar. This means the local baseflow index in our map provides new and local-scale diagnostic information that was not available before, while the overall simulation is consistent with observations.

This proves that the simulated baseflow recession behavior is very similar to the observed one when the same analysis tool is used. The seemingly higher value of baseflow ratios than the literature arose from the two abovementioned differences.

We put the above figure in Supplementary Figure S6 and add the following description.
"*On a side note, the baseflow ratios in Supplementary Figure S5 are noticeably higher than those from gage-based separation methods*[55] *due to conceptual and scale differences, and our simulated baseflow behavior is very similar to the observations if the same analysis method is applied (Supplementary Text S1 and Figure S6).*" {line 228}

What's more, we also add the Text S1 for detailed description:
"There are two reasons why our baseflow ratio in Supplementary Figure S5 appear larger than what was presented in the literature. First, prior studies applied baseflow separation to observed daily streamflow, whereas we calculate the ratio by dividing simulated baseflow flux by total simulated runoff, so the concepts are different. Second, our analysis is at the MERIT basin scale, much smaller than typical gaged basins, and hydrologic fluxes are known to be scale-dependent. Finally, if the same baseflow separation method is applied to our simulated streamflow at gage locations, the resulting ratios closely matched those from observed data (Supplementary Figure S6), confirming that the model reproduces realistic baseflow recession behaviors. Our approach thus offers complementary diagnostic insight, and future work could assess which method better reflects actual groundwater contributions."

3. In terms of long-term water balance, is it calculated based on a part of full periods of datasets or not? Also, I found the authors constrain the routing model with MERIT unit basins on the sum of all MERIT basins belonging to one drainage area. Does this mean the water balance is constrained at the drainage area level while not at its finer or coarser levels?

We evaluated the long-term water balance from 2001-2020 from the annual-scale, as annotated on the figure and Figure 1 caption.

Regarding mass balance, we added this explanation *"Since NNs in δHBV2 only provides the parameters, a mass balance between all hydrologic fluxes and states is strictly enforced throughout the model at the MERIT-level basins by the HBV modules and the routing scheme. Training the aggregated model's behavior on streamflow data at the gages will indirectly condition the model's behavior at smaller MERIT basin. In addition, since there are gages of varying catchment sizes, with some basins <100 km2, a mean area around 500 km2, and a*

maximum area around 50,000km2, we can constrain the model's behavior at different spatial scales. "{line 664}

4. Can the authors better introduce different Global water models? Many models are mentioned in the first line in the second page, while other names are presented in Fig. 1. They need to summarize their corresponding spatio-temporal resolutions as well, to see if the comparison with δHBV1 and δHBV2 is fair.

Thank you for your suggestions. We have added descriptions of the different models in the Methods section. In response to Reviewer #2's comment, we have also included an explanation of the bias-correction method used in GWM#0 (new Supplementary Text S6, pasted in page 4 of this doc above). However, due to the number of models involved and the scope of our paper, we cannot describe all the physics inside each other model. The following shows the new sentence (line 708):
"Global Water Models (GWMs)—including global hydrological models, land surface models, and dynamic global vegetation models—simulate the terrestrial water cycle at the global scale. In this study, we compared our results with six established GWMs: WaterGAP2, DBH, H08, LPJmL, MATSIRO, and PCR-GLOBWB. The 0.5° simulations, forced by GSWP3 atmospheric data, were obtained from the ISMIP2a protocol. Readers are referred to Müller Schmied et al[79]. for more information about these models. Due to the resolution discrepancy, we only compared to GWMs at the major river outlets, whose catchments are distinctly larger than the gridsize of GWMs, at monthly, annual scales or decadal scales." {line 738} As stated, we want to make fair comparisons in our work, and most existing GWMs operate at a spatial resolution of 0.5°, although some groups have developed higher-resolution versions (e.g., 5 arc-minutes), which are not yet publicly available. Therefore, when comparing our model with traditional GWMs #0-#5, we only made comparisons at the monthly or annual scale, and restricted the evaluation to large rivers (typically >100000 km²), where the effect of spatial resolution is less pronounced. Later discussion also reinforces this: "*Since these 61 large river stations were not used in training the model and much larger than the models' resolution, δHBV2's high performance is due to the collective knowledge gained from the numerous small basins used for big-data training, which is crucial for capturing such shifts.(line 188)"*

We adopted evaluation metrics such as elasticity based on the same forcing used in traditional GWMs (e.g., GSWP3), to ensure consistency. For daily-scale comparisons, we did not use 0.5° GWMs as baselines. Instead, we compared our model against GLOFAS, a finer-resolution (0.1°) model that is calibrated with observed data. We add the following explanations for clarification: "*To ensure the reliability and fairness of the elasticity calculation, We uniformly used GSWP3 as the precipitation dataset and applied the F-test, retaining only results with p-values below 0.1. ...*" {line 781} and "*To further validate the relevance of high-resolution δHBV2 for local water management, we compared its daily streamflow simulations in small-to-medium basins (<50,000*

km²) with LSTM, lumped δHBV, and a widely-used operational global-scale product, GloFAS, as previous GWMs lack sufficient spatial resolution for this scale. " {line 382}

We also revise the description about different datasets used in comparison to make sure it was conducted in a fair way.

"The resolution of MERIT basins is much finer than the ~0.5 degree grid resolution of GWMs provided in the Inter-Sectoral Impact Model Intercomparison Project (ISIMIP). As a result, there is a discrepancy in the catchment area represented in the models, with MERIT having better topological correctness. Due to the relatively lower spatial resolution, GWMs were only benchmarked on large rivers at monthly or annual temporal scales to ensure a fair comparison. The first set of comparisons thus focus on annual and monthly evaluations on 33 large rivers benchmarked in the literature[52], which have major human influences including reservoirs and water withdrawals. To evaluate the models under more natural conditions, we further added 28 large rivers (sometimes using upstream gages of rivers among the first 33) as a second comparison set.

To benchmark δHBV2 against δHBV1, LSTM, and GloFAS, we compiled a third comparison set of thousands of smaller gages. GloFAS may still have catchment-area discrepancy with our data-driven models, so we restricted the comparison with GloFAS to gages where the catchment-area discrepancy between all models and the gage-registered area was less than 20%. To separately demonstrate the models' temporal, spatial, and overall generalizability, we used three small-gage datasets (ds0-ds2). ds0 contains 4746 training basins for the temporal test. ds1 contains 2,509 basins for the prediction in ungauged basins (PUB). None of the basins in ds1 were used in the training of δHBV2 or the calibration of GloFAS. ds2 contains all small 5,558 basins from ds0 and ds1, excluding those with a large area discrepancy to enable a fair comparison between δHBV2 and GloFAS." {line 522}

5. I highly recommend including a schematic figure to better illustrate which hydrological processes can be advanced by different parts of model schemes in the introduction in the last paragraph in page 3 but also heavy technical text elsewhere.

Thanks. We have added the following Figure to illustrate the main components of the model.

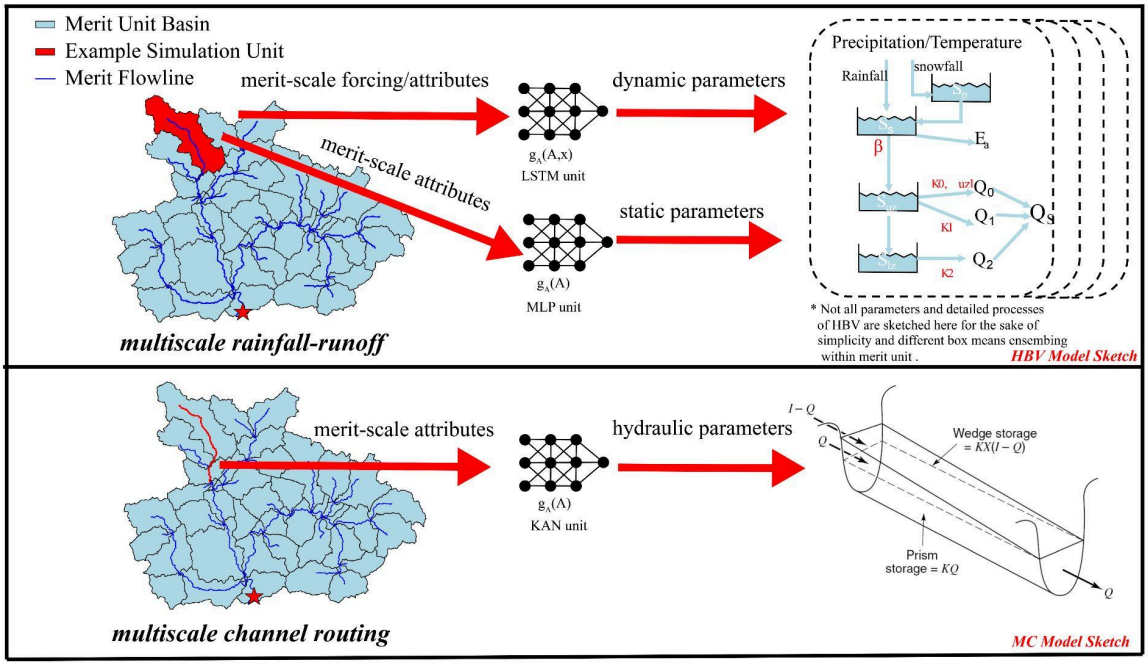


Figure S1. The sketch of the differentiable model framework

The performance improvements enacted by difference choices were broken down in two earlier papers[3,4]. We added the following descriptions into Methods to explain:
"*Based on our earlier benchmarks, a model with the simple HBV structure and NN-based differentiable parameter learning can roughly equal the performance of locally-calibrated operational hydrologic models[42], achieving around a median NSE of 0.64 on the CAMELS dataset. However, its regionalized parameterization offers benefits in applications to ungauged basins and lower parameter equifinality[72]. Adding multiple hydrologic response units to implicitly represent heterogeneity within each basin largely elevates the performance to around 0.71. Setting two or three parameters as dynamic and adding a capillary flux further elevates it to 0.75[48], which is nearly equivalent to state-of-the-art LSTM on small basins while performing better for extremes. The recent addition of multiscale learning and differentiable Muskingum-Cunge routing further elevates performance in arid regions and major rivers.*
" {line 655}

6. Can the authors discuss the computational feasibility of large-scale simulations which is not yet discussed?

Thanks for your suggestion, we add the following description:
"Our highly efficient system completes a 10-year global rainfall–runoff simulation in 7 hours on a single NVIDIA A100 GPU, which are reduced to 2 hours and 2.5 days when using 4 GPUs."{line 640} and Despite the current degree of parallelization, there is still considerable room for further acceleration.

7. I found it's difficult to change between different training and testing datasets. Why not filter consistent basins for different model training and testing?

Thanks for your suggestions. These arrangements were made to ensure fair and meaningful comparisons. We prepared different evaluation datasets based on the resolution and nature of the models being compared. For comparisons with global water models (GWMs) that operate at relatively coarse resolution, we focused on water balance evaluation at the scale of large river systems and accordingly constructed two datasets: a mixed large-river dataset and a natural large-river dataset.

For local-scale performance evaluation, we compared our model against a high-resolution GWM (GLOFAS) and prepared three small-to-medium-sized river basin datasets to demonstrate the models' temporal, spatial, and overall generalizability with temporal, spatial, and overall tests, respectively. We apologize for the added complexity, but there were many unavoidable constraints due to availability of other models' results for comparisons.

8. Fig. 1 (d-e) show trends of water components from 2001-2020. It confuses me that why 2020 is used while other parts of the manuscript have 2015? And is the result from δHBV2?

Due to the sparsity of observed streamflow data after 2016, we limited the training and validation against observations to the period ending in 2015. However, for applications focused on specific hydrological phenomena, we extended our model predictions through 2020. We also add this explanation in the note of Figure 1.
"note that: while the model was trained and validated against the observations up to 2015, it was used to simulate hydrologic dynamics through 2020 based on meteorological forcings."
{line 163}

9. Fig. 1 (c) at which temporal interval the ratio between A/P is calculated?

Good catch. Here we calculate the metric from 1981- 2000 on an annual scale. We added "annual" to Figure 1 caption.

10. Presenting ET/P trends come with a surprise. Some discussions seem to come from the trends of Q/P but not ET/P (e.g. "Papua New Guinea have seen increasing blue water while Central Europe and subtropical and midlatitude regions in South America have seen increasing green water"), while Q/P is not presented and is not simply opposite to ET/P. Same issue for "Where blue-water fraction increases and baseflow ratio decreases in tandem".

We added this one transition sentence "*Here we demonstrate the partitioning using evapotranspiration-to-precipitation ratio (ET/P) with local runoff ratio Q'/P provided in*

Supplementary Figure S3. It should be noted that local Q'/P and ET/P may not sum to one each year due to storage effects" {line 198}

It can be seen that the color scale indicates if blue or green water fraction increases. Even if the two do not add to 1.0 in each year, they are often fairly close because annual scale storage changes tend to be small compared to ET and Q'. We show the basin-local runoff Q'/P trend in a new Supplementary Figure.

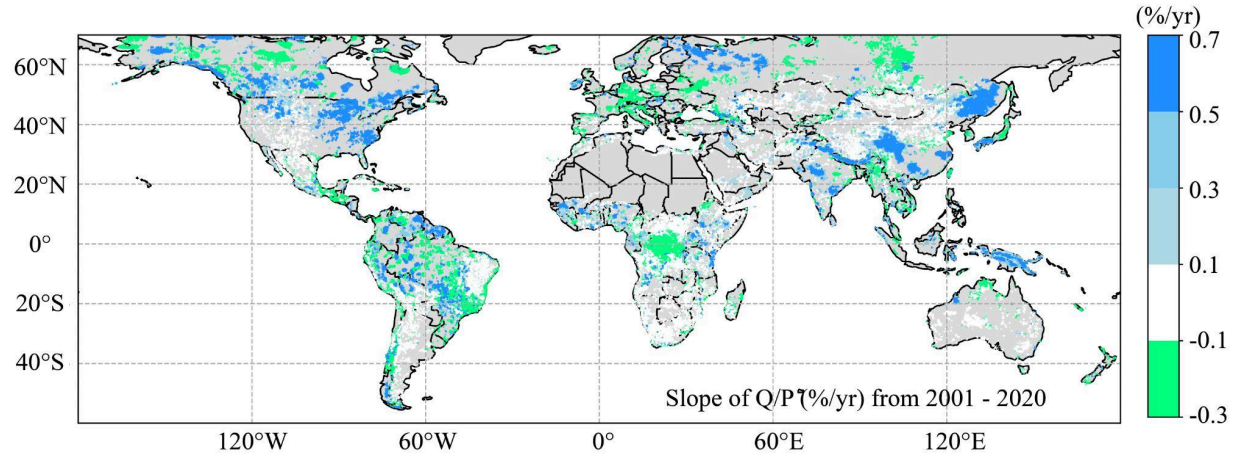


Figure S4. *Spatial trends in annual Q'/P ratios from 2001-2020, shown as percent change per year in basins with statistically significant trends (Mann-Kendall test, $p < 0.05$ colored)*

11. A few susceptible statements, such as "these changes are correlated with trends in precipitation (supplementary figure s2)". Can you report correlation coefficients?

Good suggestion. Here we further calculate the correlation of ET/P series against P series from 2001 to 2020. We can see that for most of the basins where ET/P declined, there is a negative correlation between ET/P and P. There is also a negative spatial correlation between their temporal trends. Please see the new Figure S6, pasted on Page 6 of this Response document.

12. Main concern: "Since δHBV2 conceptual model backbone is not fundamentally different from other models, the parameterization through big-data appears to have greatly improved the sensitivity to decadal-scale changes…" The new model involves new observations and a differentiable routing model to refine the upstream basins' simulation and aggregate it to a drainage area. To do so, MERIT high resolution DEM data are used. Can the authors further elaborate how this largely improve the overall model performance? And assuming, if using the same data for regionalized parameterization of the old model scheme, can that already provide a good model performance compared to traditional global water models?

The biggest difference is that previous models cannot exhaustively absorb information from thousands of basins to fine-tune the model parameterization, and we further improved

high-resolution performance via multiscale learning. The performance improvements enacted by difference choices were broken down in two earlier papers[3,4]. We added the following descriptions into Methods, which are better placed at a different spot to avoid interrupting the flow here:

"*Based on our earlier benchmarks, a model with the simple HBV structure and NN-based differentiable parameter learning could roughly equate the performance of locally-calibrated operational hydrologic models[42], achieving around a median NSE of 0.64 on the CAMELS dataset. However, its regionalized parameterization offers benefit in applications to ungauged basins and lower parameter equifinality[72]. Adding multiple hydrologic response units to implicitly represent heterogeneity within each basin largely elevates the performance, to around 0.71. Setting two or three parameters dynamic and adding a capillary flux further elevates it to 0.75[48], which is nearly equivalent to the state-of-the-art LSTM on small basins while performing better for extremes. The recent addition of differentiable Muskingum Cunge below further elevates performance in arid regions major rivers*." {line 674}

13. In semi-arid regions, summer elasticity ε is high. This is not only because low flow is driven by storage-dependent groundwater releases which are controlled by precipitation accumulation in prior months, but also because the semi-arid regions have higher evapotranspiration.

Thanks for your suggestions, we revised the sentence "*In these regions, summer ε is high due to a low runoff baseline caused by high evapotranspiration and the dominance of storage-dependent groundwater release driven by precipitation accumulated in previous months.*" {line 332}

14. Winter ε is low. But I don't agree with the current explanation that "outcome of more linear responses after thresholds are fulfilled (line 319) " which is also very abstract for understanding. I think it is likely related to snow formation during winter.

Thanks for your suggestion, we revised the sentence as *"Winter ε has a smaller range in values than summer ε, likely due to more linear hydrological responses once key thresholds—such as land abstraction—are exceeded, resulting in more proportional runoff response to precipitation."* {line 334}

15. Figure 4: Is there any meaning to arrange rivers clockwise?

By employing the clock diagram, we are able to more clearly identify which model performs best in different basins and where our model could perform better.

We have added the explanation to the figure caption:
"Monthly Nash-Sutcliffe Efficiency (NSE) scores (1981–2000) for global water models (GWMs) over (a) mixed-anthropogenic-impact and (b) natural basins. Rivers are arranged clockwise by δHBV2 performance (highest NSE on the outer ring, lowest at center) to show the best model over each basin. Marker shape and color represent different GWMs; river name color denotes

biome type. The figure format and basin selection follow conventions from previous studies. GWM0 includes post-simulation bias correction; others do not. Amur, Columbia, Danube, Irtysh, Missouri, Madera, OB', Orange, Tocantins, Yenisey are repeated as the points in the right panel are taken from more upstream, natural gages of these rivers."

16. Section Daily streamflow variability and trends: too many brackets intervening reading (from line 360).

Thanks for your suggestions, and we have revised our main text again and removed parentheses at appropriate places.

17. Main concern: Text "Both δHBV2 and LSTM thus represent state-of-the-art, allowing them to accurately capture the baseflow-surface runoff partitioning and flow variability" and Figure 5: why LSTM alone is mentioned here, is it related to δHBV1 or δHBV2 or an independent training? Also, if a LSTM model scheme directly pursue a similar simulation accuracy as δHBV2, why need the large effort on δHBV2? I encourage the authors to clarify their model novelty better.

Here LSTM refers to an independent LSTM simulation as a benchmark. Revised to *"Both δHBV2 and LSTM thus represent the state of the art for smaller basins, allowing them to accurately capture the flow variability. However, LSTM cannot simulate diagnostic variables like ET, baseflow, recharge, soil moisture, and snowmelt for providing narratives to stakeholders, and can suffer more when extrapolating to data-scarce regions and unseen extreme events*[48]*. "* {line 395}

18. Figure 5: different datasets should also be explained directly in the main text, as different model performance is well related to different datasets. The authors need to explain why different datasets provide different model performance as well.

Thanks for your suggestion. This problem was similar to the question from Question #4, #7. We prepared different datasets to make the comparison fair on our end, and we were limited by the availability of other models and datasets for comparisons.

Revised sentence:
" Due to the relatively lower spatial resolution, GWMs were only benchmarked on large rivers at monthly or annual temporal scales to ensure a fair comparison. The first set of comparisons thus focus on annual and monthly evaluations on 33 large rivers benchmarked in the literature[52]*, which have major human influences including reservoirs and water withdrawals."* {line 525}

19. Section Global datasets: I suggest adding a data table with spatio-temporal information.

Thanks for your suggestion, since all forcing and attributes are in the 0.1 degree, so we add the following note in Table S6.
"*Note: All climate forcings are provided at 0.1° spatial and daily temporal resolution, and attribute data were static and resampled accordingly.* "

20. Sentence "(line 629) Such a scale also allows nonlinear and threshold behaviors to better manifest, e.g., concentrated rainfall in a small mountainous region can lead to substantial saturation and runoff but…":check the grammar

Thanks for your suggestion, we have revised it into "*Such a scale also allows nonlinear and threshold behaviors to better manifest. For example, concentrated rainfall in a small mountainous region can lead to substantial saturation and runoff, but if the same rainfall is spread across a large basin, then very little runoff would occur*" {line 635}

21. Ensure the abbreviations are clearly introduced throughout the manuscript.
Thanks for the suggestion, we have uniformed our model abbreviations in the manuscript.

1. Chaopeng, S., Yalan, S., Martyn, C. & Jiangtao, L. Prominent impacts of hydrologic scaling laws on climate risks. Preprint at https://doi.org/10.21203/rs.3.rs-4584048/v1 (2024).Dear Nature Communications editor(s),
2. Blöschl, G. & Sivapalan, M. Scale issues in hydrological modelling: A review. *Hydrol. Process.* **9**, 251–290 (1995).
3. Feng, D., Liu, J., Lawson, K. & Shen, C. Differentiable, Learnable, Regionalized Process‐Based Models With Multiphysical Outputs can Approach State‐Of‐The‐Art Hydrologic Prediction Accuracy. *Water Resour. Res.* **58**, e2022WR032404 (2022).
4. Song, Y., Bindas, T., Shen, C. & Ji, H. High-resolution national-scale water modeling is enhanced by multiscale differentiable physics-informed machine learning. *Water Resour. Res.* (2025) doi:10.1029/2024WR038928.

Dear Nature Communications editor,

Many thanks to the editor and reviewers for a fast turnaround. Only Reviewer #2 raised one question about presenting the model quality, besides some minor suggestions. We much appreciate that comment and have provided a new Revised Figure S2 to address this question. Stippling may not directly apply here due to resolution, not having a model ensemble, and the need to address two separate issues (significance of the trend vs. reliability of the model). Thus we choose to show maps of model performances for both large basins and upstream ones in revised Figure S2. The errors on the trend mainly arise from Africa and north-central Asian basins where training data is scarce, but some other basins are well captured despite data scarcity. We detailed this in the response below. We believe the revised version is now suitable for publication.

Chaopeng

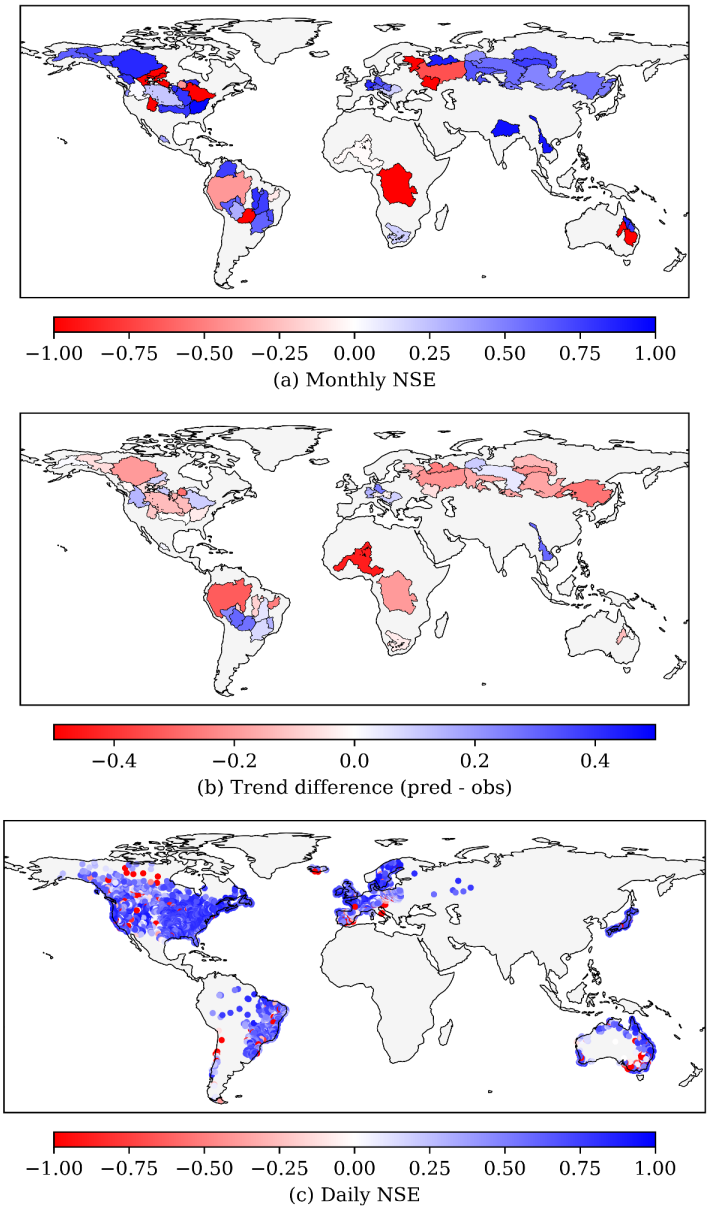
R2: It also occurs to me that there may be confusion triggered by the use of "green water fraction" to exclusively mean evapotranspiration fluxes. Many people will understand "green water" to indicate soil water. I suggest a slight change in terminology to "green water fluxes" (while retaining parenthetic references to evapotranspiration to distinguish these from downward or lateral fluxes) to ameliorate this potential source of confusion.

Thanks for the comment about "green water flux". We have added some "fluxes" where appropriate, as you suggested. For example, "In general, midlatitudes in North America and Asia and tropical areas like Central America and Papua New Guinea have seen decreasing green water fluxes " {line 208-210}, "When annual precipitation declines, there is less excess water that can overcome the storage thresholds to become runoff, so the more precipitated water exits as green water fluxes, and blue water drops disproportionately." {line 244-246}. "parts of Europe have experienced increases in green water fluxes". However, it would seem cumbersome to change "green-blue-water partitioning" in the subtitle or abstract. This phrase is used by lots of papers in the same meaning as ours, e.g., Stephens et al.[1], and since we defined it in the paper, it should be ok.

R2: It occurred to me when rereading the manuscript that, despite the very large improvements, an overall average R-squared of 0.44 in simulated vs. observed trends still leaves much room for improvement. I was a little surprised by the spatial distribution of some of the positive and negative trends and found myself wondering which ones were reliable (especially given that some of these results depart from what has been previously reported, as the authors allude to with regards to previous Budyko analyses). Could the authors consider using, for example, stippling (as done by the IPCC) to demarcate the reliability of trends in different regions, either based on the outcomes of the temporal test (for Q) or R-squared of trends? They would need to pick thresholds in performance measures to determine which regions should be stippled, but I do think this would lead to a more honest assessment.

Thanks for the idea about stippling and reliability. It seems the reviewer wants to have an idea where the model is trustworthy, similar to what is presented in the IPCC map (Figure R1 below). In page 4, we will discuss why stippling may not be the suitable approach. However, to address the model quality issue, we now add a figure (revised Supplementary Figure S2) that shows the map of (a) monthly NSE of large rivers, (b) difference in trend estimate between observed and simulation (containing all the pairs in Figure 1c and more); and (c) monthly NSE of small training and test stations (ds+ds1). With these three plots, readers could infer where the model is more trustworthy and how well it might do for both large-scale dynamics and upstream catchments. We added this note "The most challenging trends are those in Africa, e.g., Niger (GRDC 1834101) or Congo (GRDC 1147010), where the model could not capture the rising trend in Q/P due to a paucity of training data." {line 193-195}.

The following note is added to the Limitations section: "While we do not have an ensemble of models to produce a formal uncertainty estimate, we present maps of large basin and upstream-catchment evaluations (Supplementary Figure S2) as a gauge of model reliability at different scales. Some large basins, e.g., Ganges, Siberian, and southern Brazilian ones, are well simulated in terms of both NSE and long-term trend, despite not having any training basins. This suggests the model does possess some ability to generalize in space. Nevertheless, some poor-performing basins still tend to be due to a void of training stations, especially African and North-central Asian ones, where hydrologic dynamics may also be systematically different from the training basins. These regions are expected to improve when we learn from more data, including expanded training stations, new streamflow estimates from the Surface Water and Ocean Topography (SWOT) mission, and non-streamflow observations such as soil moisture. Future effort can use an ensemble based on different training data to assess uncertainty." {line 466 - 477}



Revised Supplementary Figure S2. (a) Monthly NSE for each large river basin, both natural and mixed; (b) The difference between estimated and observed streamflow trends for large rivers. Some basins in (a) are missing here because they do not have long enough record (10 years) to reliably estimate trend; (c) Daily NSE for the training and test small basins.

The value of 0.44 in Figure 1c may underplay the model's power because some of the large basins, e.g., Ganges, were removed from the evaluation due to not having long enough data (10 years) to establish a trustworthy trend estimate. The value is lowered mainly due to very different basin behaviors in Africa. We actually see that Ganges was simulated with a daily NSE of 0.93 in 7 years when it has data, and will likely have a decent trend estimate if the observation was long enough. Orinoco has a similar story. We add this note "Some well-simulated basins, e.g.,

Ganges, and Orinoco, were not represented in (c) as their records too short (<10 years) to estimate the trend, and they are expected to elevate the R^2 if data were available." {Figure 1 caption}.

As the reviewer commented, despite some data gaps, the model does present a major advance compared to established ones. This work also should not be perceived as a static end point. It demonstrates the potential of big data learning in global hydrology, and there are plenty of opportunities to improve it as indicated in the quoted text above. The work supporting IPCC was also not achieved in one paper.

Why we looked into it but could not produce a stippling figure:
(a) The IPCC map (below) was produced using model agreement. Here, we only have one high-resolution model and thus cannot produce such an uncertainty estimate. In addition, the point we are making is that big data learning allows our model to be quite different (and more accurate) than the established models. Thus, it seems that having a high model agreement with GWMs may not be a reliable error estimate, and comparisons with observations provide a more trustworthy benchmark. The computational expense to generate such an ensemble is high and it is out of the scope of this paper. We will need to wait follow-up work for formal uncertainty estimate. In fact, we have upcoming work with groundwater recharge that shows, although the ensemble of the global water models (GWMs) is indeed better than all of its members, our single model is closer to observations than the GWM ensemble. The power of big data learning cannot be understated!

(b) We have very high-resolution simulations (~36 km^2 median), much finer than the 0.5 degrees used earlier (~3100 km^2 near the equator). If we try to add stippling to Figure 1, as you can imagine, the results will be highly cluttered and messy to view.

(c) Figure 1+Figure S2 also allows the disentanglement of two issues: statistical trend significance of the trend and (b) model quality at small and large basin scales (a&c). Figure 1 shows only points with significant trends. Thus, we think adding the Supplementary Figure S2 is a better way to present model quality.

We hope you see that we much appreciate your opinion and have taken applicable measures to respond to the reliability issue. While the solution seems somewhat different from your suggestion, we hope you see that it is a suitable one for the issue. We look forward to building on top of this work and employing an appropriate approach to produce full uncertainty quantification in the future, one that is perhaps even more reliable than model agreement.

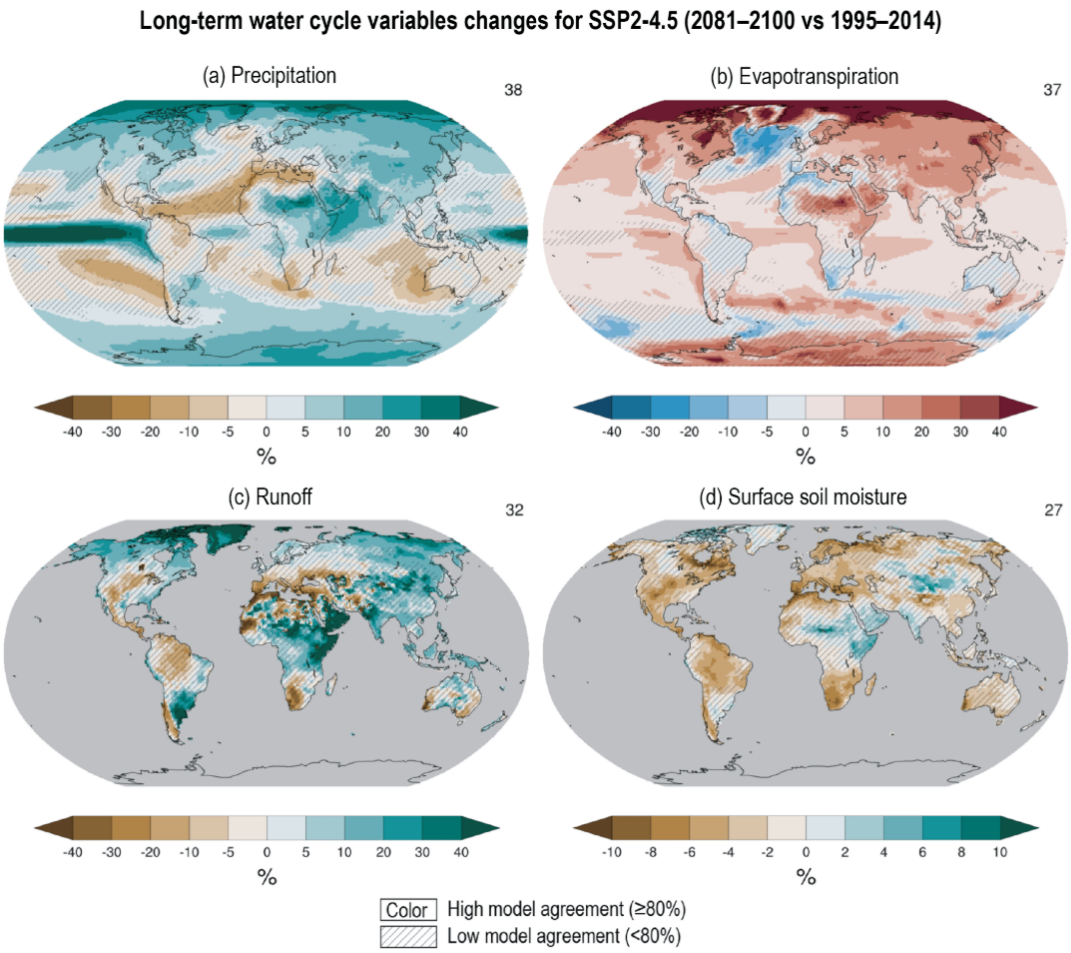


Figure R1. IPCC Stippling style.

Minor revision

1.Line 137-138: Change to "As it was only recently introduced in Song et al, …"

Thanks for your suggestion and we have revised as suggested.

2. Line 716: tested --> test

Thanks for your suggestion and we have revised as suggested.

1. Stephens, C. M. *et al.* Changes in Blue/Green Water Partitioning Under Severe Drought. *Water Resour. Res.* **59**, e2022WR033449 (2023).