Dear Editors and Reviewers,

We sincerely thank the two reviewers for their insightful and constructive comments on our manuscript, "A physically guided deep learning reconstruction of terrestrial water storage anomalies at 0.1° across China" (essd-2026-282). To address these comments and further improve the quality and scientific value of our dataset, we plan to make substantial revisions to the manuscript, mainly including: (1) updating the input datasets and extending the reconstructed TWSA time series to the more recent years until 2025; (2) refining the deep learning framework and training strategy to enhance model generalization and reduce mascon boundary artifacts; (3) improving uncertainty quantification through more robust estimation of model uncertainty and the incorporation of input-related inference uncertainty; and (4) refining the evaluation strategy to enable more appropriate comparisons with existing products and improve the assessment of reconstruction reliability.

Given the additional experiments and analyses required, at this stage we provide our revision plans of addressing each comment. Please see our point-by-point response below, with comments in blue and our response in black. Once these additional analyses have been completed, we will submit a fully revised manuscript, together with a detailed description of the changes made in response to the reviewers' comments.

Reviewer 1

Overall comment: In this study, Li et al. (2026) applied a deep learning model to downscale the JPL mascon to 0.1 degrees, incorporating spatial details from a hydrological model (PCR-GLOBWB). Evaluations across different scales and comparisons with two other publicly available products show that the proposed products perform better. However, after a detailed read, several major points need to be carefully reconsidered and revised before this manuscript can be published in ESSD. One general comment on the dataset is that the "main product" described by the authors is only available through 2019 (about 7 years ago). For a data description paper being considered in 2026, the data record is a bit out of date, and I highly recommend that the authors extend it so that the community can benefit more from it. There are several other issues regarding method definition, description, and product evaluations. Please refer to my detailed comments below.

Response to the overall comment: We appreciate the reviewer's constructive comments for improving our study, which will help us further improve the quality and scientific value of our dataset. Based on the comments, we plan to incorporate the latest available input datasets, extend the temporal coverage of the reconstructed TWSA dataset to more recent years to 2025, improve uncertainty quantification through additional training runs and feature perturbations, and refine several technical aspect of the deep learning framework. Given the substantial amount of additional experimental work required, at this stage we provide our point-by-point responses and planned revisions below. A fully revised manuscript, together with a detailed description of the changes made, will be submitted once all additional experiments and analyses have been completed.

Major comments

1. Line 190: The referenced glacier products are only available annually. How did you use them for your monthly downscaling? Please mention it clearly. On line 205, you mentioned that they provide the monthly products, but the link does not refer to them. Please correct.

Response to comment 1: We first explain how the monthly glacier mass change time series was derived from the referenced glacier product (Hugonnet et al. 2021) for the original study period (2002–2019), and then explain how we will refer to the methods to generate monthly glacier mass change time series extending to 2025 in the revised manuscript.

First, the repository of the referenced glacier product (accessed at https://doi.org/10.6096/13) distributes two distinct families of data, including (1) Elevation-change maps at 100 m posting, which are available only as multi-year rates (i.e., 5-, 10- and 20-year periods), and (2) Cumulative time series of elevation, volume and mass change, aggregated globally, per RGI region, per glacier, and per tile of 0.5°, 1°, 2° and 4° tilings. The latter one (cumulative time series) is provided at a monthly time step over 2000–2019, which was generated using Gaussian process regression to interpolate multi-temporal glacier

elevation observations into continuous monthly time series, followed by the conversion of elevation changes into glacier volume and mass changes (Hugonnet et al. 2021).

In our previous manuscript, we obtained the monthly time series of cumulative glacier volume changes relative to 2000 at per tile of 0.5° from Hugonnet et al. (2021). We assumed spatially uniform glacier elevation changes within each 0.5° grid cell, and divided the volume changes by the corresponding glacier areas from RGI 6.0 to derive glacier thickness changes. These thickness changes were subsequently converted into liquid water equivalent (LWE) using a glacier ice density of 850 kg m^{-3} and a water density of 1000 kg m^{-3}. The resulting LWE changes were then weighted by the glacier-covered fraction of each 0.1° grid cell to obtain gridded glacier water storage changes. Finally, we calculated glacier water storage anomalies relative to the 2004–2009 mean baseline to ensure consistency with GRACE and GRACE-FO TWSA.

Second, in the revised manuscript, we plan to extend the monthly glacier mass change time series to 2025, consistent with the extended TWSA reconstruction period. Because the glacier dataset of Hugonnet et al. (2021) only covers 2000–2019, we will reconstruct the entire 2000–2025 series using annual glacier mass change estimates at the 0.5° spatial resolution from the World Glacier Monitoring Service (WGMS) (Dussaillant et al. 2025). The monthly time series from Hugonnet et al. (2021) will be used to derive the climatological seasonal cycle over 2000–2019, which guides the temporal disaggregation of WGMS annual estimates into monthly glacier mass changes. Specifically, based on the climatological seasonal cycle from Hugonnet et al. (2021), we will determine the relative contributions of individual months to glacier mass changes and use these seasonal patterns to distribute the annual mass change estimates from WGMS into monthly increments, while ensuring consistency with the WGMS annual totals. This approach aims to preserve the seasonal characteristics derived from Hugonnet et al. (2021) while constraining the annual mass changes using WGMS estimates.

After obtaining the monthly glacier mass changes at 0.5° resolution for 2002–2025, we will apply the same spatial disaggregation procedure used for the original 2002–2019 dataset. Specifically, assuming uniform glacier surface elevation changes within each 0.5° grid cell, we will convert the glacier mass changes into water-equivalent thickness changes and then derive glacier water storage changes at 0.1° resolution using the glacier-covered fraction of each grid cell.

We acknowledge that the assumptions of spatially uniform glacier elevation changes within each 0.5° grid cell and the use of a climatological seasonal cycle to distribute annual mass changes could introduce uncertainties. Nevertheless, these approximations provide a practical approach of incorporating high-resolution glacier mass changes at monthly scales, given the limited availability of temporally continuous glacier observations. We will acknowledge these assumptions and their associated uncertainties in the revised manuscript.

2. Line 228 and the title: As you are writing a paper with a strong focus on DL, I highly recommend avoiding the term "physics-informed," as it causes considerable confusion with PINN.

Response to comment 2: We appreciate this comment and will avoid the term "physics-informed" in the title and throughout the manuscript. In addition, in the revised version we define 0.1° high-resolution TWSA from the combination of PCR GLOBWB2 and satellite-retrieved glacier mass changes as the "bottom-up TWSA ($TWSA_{BU}$)", which indicates the sum of different water storage components and avoid the confusion with Physics-Informed Neural Network (PINN).

3. Line 243: Yes, accumulated meteorological fluxes are typically comparable to TWS changes. However, the errors are also accumulated. This is the reason that comparisons between GRACE data and P, ET, R are usually in the differential domain. Can you comment on this? And please clarify what n is in your case?

Response to comment 3: We retain accumulated meteorological fluxes as predictors because of their typical comparability to TWS changes and the assistance in improving the performance of TWSA reconstruction indicated by previous studies (Wang et al. 2024). "n" in previous Line 243 refers to the number of months in the study period, which equals 213 months in the previous study period (Apr 2002–Dec 2019) and 285 months in the revised data period (Apr 2002 to Dec 2005).

To account for the potential accumulation of errors in these predictors, we will specifically estimate the uncertainties associated with individual input channels and quantify their propagation into the final high-resolution TWSA reconstruction. This will help us examine whether accumulated input errors substantially affect the reliability of the reconstructed TWSA. Further details on uncertainty estimation are provided in our response to comment 7.

4. Major issues with Section 3 Methodology

On line 302, you clearly mention that your CNN takes inputs with a spatial size of 3 degrees, but Fig. 2 clearly does not show a 3-degree size; rather, it shows around 15 degrees (5 JPL mascons). Please clarify and keep it consistent.

Response to comment 4: In the revised manuscript, we will enlarge the patch size to include multiple JPL mascons within a patch as suggested in comment 8, and then revise Fig. 2 accordingly to ensure consistency between the methodological description and the figure.

5. Section 3.1 Training strategy

Major issue of this section: The whole methodology, from model architecture, training design (especially the loss function), to the uncertainty quantification approach, is too similar to the one proposed by Gou and Soja 2024, which the authors also referred to in the introduction section. It is fine to further improve the methods and/or apply them to another set of data/target regions, but you need to:

(1) Properly credit the previous developments. You should clearly mention that your method is based on the pipeline introduced by Gou and Soja (2024) and that you made further modifications in your methodology section.

(2) For the modification part: You need to highlight which part of the method you have changed and why you made the choice. How does it compare to the previously introduced method?

(3) The uncertainty quantification part is certainly not fully completed. Various studies have already mentioned this issue. The authors need to revise it to provide more realistic uncertainty information. See the comments below.

Response to comment 5: We appreciate the reviewer's constructive suggestions and will address each of the three points separately below.

First, we credit the previous developments and will clearly mention that our method is based on the pipeline introduced by Gou and Soja (2024), particularly in the Method part.

Second, we will clarify the improvement of our method in the following two aspects.

(1) Reconstruction of the gap period between GRACE and GRACE-FO mission. Gou and Soja (2024) used GRACE retrieved and model simulated TWSA as both training labels and input features for inference. Consequently, their framework cannot reconstruct TWSA during periods without GRACE observations, such as the gap period between GRACE and GRACE-FO missions (July 2017–May 2018). In contrast, we revise the framework to use GRACE observations and bottom-up TWSA only as labels to train the deep learning model parameters, without incorporating them into the inference process. This improvement enables our model to reconstruct temporally continuous high-resolution TWSA, including periods without GRACE observations.

(2) Incorporation of glacier mass changes. The fine-scale spatial information in Gou and Soja (2024) was provided only by hydrological model simulations from WaterGAP Global Hydrology Model (WGHM), which lacks the explicit module of glacier mass changes. To address this limitation, we construct bottom-up TWSA by combining hydrological simulations with satellite-derived glacier mass changes. This allows our framework to better represent TWSA variations in glacierized regions.

Third, we agree that our previous uncertainty estimation, which characterized the model uncertainty from three independent training runs, did not fully characterize the uncertainty of the reconstructed TWSA. We will therefore improve the uncertainty estimation by increasing the number of independent

training runs and incorporating input data perturbations to account for additional sources of uncertainty. Further details of uncertainty estimation are provided in our response to comment 7 below.

6. Line 374: You actually spent two paragraphs explaining how to determine the lambdas, and the conclusion is that they are not necessary at all. I do acknowledge your efforts in this hyperparameter tuning. Still, you should comment on why lambda=1 provides good results and, ultimately, may consider reducing it in the main text to provide more concise information.

Response to comment 6: In the revised manuscript, we will comment that lambda=1 provided a good balance between large-scale consistency with GRACE observations and the reliability of fine-scale spatial patterns. We will also streamline the discussion of hyperparameter tuning in the main text to provide more concise information.

7. Line 375: So you basically just get the ensemble of three trainings. So, it's just "half" of the deep ensembles (see Lakshminarayanan et al., 2017) and can therefore only capture epistemic uncertainties (another issue: 3 repeated runs might be too limited for approximating epistemic uncertainties). So how about aleatoric uncertainties? Actually, Gou and Soja 2024 also devote quite some text to discussing this issue and mention it as one of their main limitations. To my understanding, they tried to consider the aleatoric uncertainty to a certain extend. So, regarding the uncertainty quantification part, your method does not advance on the previous study, right?

Response to comment 7: To provide a more comprehensive uncertainty assessment, we plan to improve the uncertainty quantification in the following aspects.

(1) Epistemic uncertainty. We will increase the number of independent training runs from three to five to provide a more robust estimation of epistemic uncertainty. The pixel-level epistemic uncertainty will be estimated as the standard deviation of predictions from these independent runs.

(2) Aleatoric uncertainty. In the revised manuscript, we will incorporate feature-related aleatoric uncertainty by quantifying the propagation of input data uncertainties into the reconstructed TWSA through Monte Carlo-based perturbations. Although this approach does not capture all sources of aleatoric uncertainty, it provides a practical approximation given the limited availability of uncertainty information for the input datasets.

Specifically, our deep learning framework uses 12 input channels in inference, including four static and eight dynamic predictors. The dynamic predictors are derived from ERA5-Land and MODIS NDVI. For ERA5-Land variables, we plan to use the standard deviation across the 10 ensemble members of the corresponding ERA5 variables provided by ECMWF as an approximation of their uncertainties, because the ensemble-based uncertainty information is only directly available for ERA5 reanalysis. For MODIS NDVI, we will adopt the nominal uncertainty of approximately 0.025 under clear-sky conditions, as reported in the official MODIS Vegetation Indices (MOD13 Series) Product User Guide.

We will then randomly perturb each of the eight dynamic input channels according to its estimated uncertainty to assess how input data errors propagate into the reconstructed TWSA in the model inference. The four static predictors will remain unperturbed because they primarily represent relatively stable spatial background characteristics, and reliable spatially explicit uncertainty distributions are not readily available. We will then generate 30 sets of jointly perturbed dynamic inputs through Monte Carlo sampling for each of the five independently trained U-Net models, yielding 150 reconstructed TWSA predictions in total. For each model, the standard deviation across its 30 predictions will be used to estimate the inference uncertainty associated with input data errors, as an approximation of the input-related aleatoric uncertainty.

(3) Total uncertainty. We plan to estimate the total uncertainty of the reconstructed TWSA using a variance decomposition framework, combining the between-model variability from independent training runs (epistemic uncertainty) and the within-model variability arising from Monte Carlo input perturbations (input-related aleatoric uncertainty).

(4) Comparison with previous studies and limitations. Compared with Gou and Soja (2024), who considered aleatoric uncertainty in only one input feature (GRACE) during inference, our proposed

approach will account for uncertainties in all eight dynamic predictors through joint Monte Carlo perturbations, providing a more comprehensive assessment of input-related uncertainty propagation.

Nevertheless, we acknowledge that this approach does not capture all sources of uncertainty. For example, uncertainties associated with the training labels are not explicitly propagated through the U-Net training process. Consequently, the resulting total uncertainty estimates may still be underestimated. We will explicitly discuss these limitations in the revised manuscript.

8. Fig. 3: The validation AE loss is around double the training AE loss. This is a risky sign, indicating that the model cannot maintain agreement with GRACE data. I guess the issue is associated with the small patch size and the choice of 3-deg JPLM. The sharp mascon boundary may have a remarkable negative impact in this case; see your Fig. 7.

Response to comment 8: In the revised framework, we will increase the patch size to include multiple GRACE mascons within each patch, thereby reducing the influence of sharp mascon boundaries on the reconstructed TWSA. In addition, we will implement an early stopping strategy to mitigate potential overfitting and improve the model's generalization performance. Specifically, training will be stopped when the validation loss shows an increase, and the model checkpoint with the lowest validation loss will be retained for inference.

9. Fig. 7: Both of the maps show clear “mascon-like” shapes. These shapes indicate a crucial point: the downscaled products may not overcome the limitation imposed by a 3-degree mascon boundary. It is a rather fatal problem for a product that is argued to have 0.1-degree resolution. Please comment on it.

Response to comment 9: In the revised framework, we will increase the patch size and adopt sliding-window sampling and inference strategies to mitigate the influence of the 3° mascon boundaries on the reconstructed TWSA. Furthermore, we will quantitatively assess potential mascon boundary artifacts by calculating the ratio of TWSA spatial gradients across mascon boundaries to those within mascons. This analysis will help determine whether the revised framework effectively reduces the 3° mascon boundary artifacts.

10. Problem with evaluations and discussions.

Line 695: Xiong et al. (2025) used SHC-based products and did not provide anything based on the mascon product. Therefore, all the following comparisons that use the JPL mascon as the ground truth are unfair and cannot yield a convincing conclusion.

Response to comment 10: We agree with this comment, and will use JPL-SHC products to evaluate the ability of the TWSA dataset developed by Xiong et al. (2025) to maintain consistency with GRACE observations at coarse resolutions.

11. Line 720: Did you also examine if the correlations are statistically significant? And although groundwater dominates NCP. You should also carefully consider other components, especially root-zone soil moisture.

Response to comment 11: We will add statistical significance tests for the correlations between observed groundwater level variations and downscaled TWSA changes. In addition, we plan to combine groundwater storage variations derived from in-situ well observations with root-zone soil moisture variations from ERA5-Land to estimate subsurface water storage anomalies, providing a more representative basis for comparison with the downscaled TWSA.

Minor comments

12. Line 74: The study by Vishwakarma et al. (2021) does not belong to ML methods.

Response to comment 12: We will remove this study in the Introduction of Machine Learning methods.

13. Fig. 1: This figure does not show the necessary information clearly. For example, the positions of wells are not readable.

Response to comment 13: We will add an enlarged inset of the North China Plain to the main map of China in Fig. 1, with larger markers and higher-contrast colours to improve the visibility of the in-situ groundwater well locations.

14. Line 172: Is it RL06.3 or? Please specify.

Response to comment 14: We will update the GRACE dataset to the latest version (RL06.3) in the revised manuscript and explicitly specify the data version in the text.

15. Line 324: Then how did you handle the ocean pixels? Did you remove them or mask them out?

Response to comment 15: We mask out ocean pixels so that they are excluded from the loss calculation. We will clarify this in the revised manuscript.

16. Line 330: But the problem is that the same location may also be used for training in another month. As you highlighted, you want to learn the spatial details, so it's not really a fair "training-validation" strategy, right?

Response to comment 16: We agree with the reviewer that randomly splitting training and validation samples spatially within each month may lead to potential temporal information sharing. Therefore, in the revised manuscript, we will follow previous studies (Arshad et al. 2024, Gou and Soja 2024, Wang et al. 2024) and adopt a temporal splitting strategy, dividing the available months into training (70%), validation (15%), and independent test (15%) sets to provide a more reliable assessment of the model's generalization capability.

17. Line 334: Please first cite Adam, then report more details on your LR scheduling so that the audience can fully reproduce it.

Response to comment 17: We will cite Adam (Kingma and Ba 2014), and report more details on the learning rate scheduling, which is set with an initial learning rate of 0.1, decayed by a factor of 0.998 per epoch.

18. Line 398: Well, you can argue that using ERA5L here is for consistency. But then, you can't judge whether your method is better than their other/original GRACE methods anymore, since you explicitly incorporate this information. They no longer provide independent evaluation.

Response to comment 18: In the revised manuscript, we will clarify that ERA5L is used to evaluate internal model consistency rather than provide an independent validation. We will not use ERA5L estimates to evaluate other downscaling products or original GRACE products.

19. Fig 3: It's hard to gain valuable information from this figure as the point density is missing. Please consider including this density information so that the distributions of the points can be better analyzed.

Response to comment 19: We will use density scatter plots to illustrate the comparisons among downscaled TWSA, bottom-up TWSA, and GRACE TWSA, allowing for a clearer visualization of the point distributions.

20. Line 451: I doubt that the issue is due to land-ocean leakages. You used the CRI version, which explicitly accounts for the ocean-land leakage issue, right? So, it should be fine for your case, given the current accuracy.

Response to comment 20: We agree with the reviewer that the used CRI version of the JPL mascon product accounts for land-ocean leakage. We will therefore remove the attribution of the relatively poor reconstruction performance in coastal regions to GRACE land-ocean leakage in the revised manuscript.

21. Fig. 5: Didn't you mention that you chose Hydro Basin L5 (Line 403)? They are certainly not in this figure.

Response to comment 21: We apologize for the lack of clarity in the original manuscript. We first evaluate the consistency between downscaled TWSA and GRACE observations at large basin scales, for which China is divided into nine major river basin regions, as shown in previous Fig. 5. In contrast, the HydroBASINS Level 5 catchments mentioned in previous Line 403 represent small to medium catchments beyond the GRACE solution, which are used to assess catchment-scale water balance in Section 4.2 (Fig. 8 in the original manuscript). We will clarify the distinction between these two sets of basin boundaries in the revised manuscript.

22. About Section 4.3: I'm a bit confused by the design of this manuscript: You explicitly exclude the highly irrigated glacierized regions in your previous evaluations. But now you're even focusing on them and providing quite a lot of discussion. They actually contradict each other. Could you please explain your intention a bit more? Thanks!

Response to comment 22: We appreciate this comment and provide the following clarification. To assess the reliability of fine-scale spatial information in the reconstructed TWSA beyond the original GRACE resolution, we first evaluate water balance closure across small to medium catchments, excluding highly irrigated and glacierized regions. Highly irrigated catchments are strongly affected by human activities, limiting the applicability of natural water balance assumptions, while glacierized catchments are characterized by complex cryospheric processes and large uncertainties in hydrological flux estimates. Therefore, these regions were excluded from the water balance evaluation in Section 4.2.

In Section 4.3, we further evaluate the reconstructed TWSA in highly irrigated and glacierized regions, where observations of dominant water storage components provide a more suitable basis for assessing the reliability of TWSA variations. Specifically, TWS variations in highly irrigated regions are largely dominated by groundwater changes, which can be evaluated using groundwater level observations from in-situ wells. In high mountain regions, where glacier mass changes play a dominant role in TWS variations, we use glacier elevation changes derived from satellite altimetry for comparison. These targeted evaluations complement the catchment-scale water balance assessment in Section 4.2 by examining the reconstructed TWSA from a different perspective. We will clarify the distinct objectives and rationale of these evaluation approaches in the revised manuscript.

23. Fig. 9: Again, the readability of these figures is not good. And please choose different color maps for the absolute correlation and the differences, respectively.

Response to comment 23: We will revise Fig. 9 by using different colormaps for absolute correlations and correlation differences, enlarging the plots, and enhancing the visual contrast to improve readability.

24. Line 709: But Gou and Soja 2024 clearly mentioned that their uncertainty information is underestimated. So being close to theirs is not a good sign for your product. Instead, it would be a meaningful contribution if you could provide more realistic uncertainty information.

Response to comment 24: We agree with the reviewer and will not use the similarity to Gou and Soja (2024) to support the reliability of our uncertainty estimates. In addition, while Gou and Soja (2024) considered uncertainty in only one input feature during model inference, we plan to introduce Monte Carlo-based joint perturbations of all eight dynamic input channels to better characterize aleatoric uncertainty. Further details on the proposed improvements to uncertainty quantification are provided in our response to comment 7.

References:

Arshad, A., et al. (2024). Downscaled-GRACE Data Reveal Anthropogenic and Climate-Induced Water Storage Decline Across the Indus Basin. *Water Resources Research* 60(7): e2023WR035882.

Dussaillant, I., et al. (2025). Annual mass change of the world's glaciers from 1976 to 2024 by temporal downscaling of satellite data with in situ observations. *Earth System Science Data* 17(5): 1977–2006.

Gou, J. Y. and B. Soja (2024). Global high-resolution total water storage anomalies from self-supervised data assimilation using deep learning algorithms. *Nature Water* 2(2): 139–150.

Hugonnet, R., et al. (2021). Accelerated global glacier mass loss in the early twenty-first century. *Nature* 592(7856): 726–731.

Kingma, D. P. and J. Ba (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv*:1412.6980.

Wang, Y. M., et al. (2024). Spatial downscaling of GRACE-derived groundwater storage changes across diverse climates and human interventions with Random Forests. *Journal of Hydrology* 640: 131708.

Xiong, Y. H., et al. (2025). High-Resolution Terrestrial Water Storage Anomalies and Components in China From GRACE/GFO via Joint Inversion Downscaling. *Water Resources Research* **61**(7): e2024WR038996.

Referee 2

The authors developed a physically-guided deep learning framework based on a U-Net architecture to downscale GRACE/-FO Terrestrial Water Storage Anomalies (TWSA) from a native resolution of 3° to a high resolution of 0.1° across China. The incorporation of a dual-loss function—combining large-scale mass conservation from GRACE mascons and pixel-scale spatial structural constraints from hydrological simulations (PCR-GLOBWB 2)—is methodologically sound and represents an encouraging attempt to resolve cryospheric and anthropogenic signals. Overall, the manuscript is well-written, clearly organized, and the produced high-resolution dataset will be of great value to the regional hydrological community. To further strengthen the manuscript and improve the clarity, transparency, and robustness of the study, a few constructive suggestions are provided below for the authors' consideration.

Response to the overall comment: We thank the reviewer for recognizing the value of our study and for providing constructive suggestions to further improve our work. To address these comments and provide the community with a more up-to-date TWSA dataset, we plan to refine several technical details of the deep learning framework, incorporate the latest available input datasets, extend the temporal coverage of the reconstructed TWSA dataset to more recent years until 2025, and improve uncertainty quantification through additional independent training runs and Monte Carlo-based input perturbations. At this stage, we provide our point-by-point responses and planned revisions below. A fully revised manuscript, together with a detailed description of the changes made, will be submitted once the additional experiments and analyses have been completed.

1. Line 330-331. The authors partitioned the spatial domain into 70% training and 30% validation sets within each monthly time step rather than across the temporal axis during 2002–2019. While this spatial-within-month splitting strategy is reasonable for spatial reconstruction and inpainting tasks, it is suggested to enrich the discussion regarding potential temporal information sharing when evaluating long-term basin-scale time series (e.g., Fig. 6) and spatial metrics (e.g., Fig. 4). In addition, it would be beneficial to add an independent test set.

Response to comment 1: We agree that the previous strategy, which partitioned the spatial domain into the training and validation sets within each monthly time step, could cause temporal information sharing when evaluating basin-scale time series. In the revised version, we will extend the reconstructed TWSA dataset to 2025 and adopt a temporal splitting strategy to assess the model's temporal generalization capability, with an independent test set. Specifically, the entire time period (April 2002 to December 2025) will be split along the temporal axis into training (70%), validation (15%), and test sets (15%). We will separately evaluate the model performance on the training, validation, and test sets and present the corresponding results in the Supplementary Information to provide a more comprehensive assessment of the model's reliability.

2. Lines 375–381. It would be helpful to provide the explicit mathematical formula used to calculate the pixel-level uncertainty from the three independent training runs. Furthermore, while the substantial computational burden of training deep networks for 800 epochs is well appreciated, an ensemble size of N=3 is statistically on the smaller side for estimating standard deviations. If computational resources permit, increasing the ensemble size (e.g., to 5–10 runs) or providing a brief discussion on the variance stability with N=3 would make the uncertainty quantification even more convincing.

Response to comment 2: We thank the reviewer for this suggestion. We will increase the ensemble size from three to five independent training runs to improve the robustness of model uncertainty estimation. The pixel-level model uncertainty will be calculated as the standard deviation of predictions from these independent runs, and the corresponding mathematical formula will be explicitly provided in the revised manuscript as follows:

$$\sigma_{m,i,t} = \sqrt{\frac{1}{N-1}\sum_{n=1}^{N}(TWSA_{i,t}^{n} - \overline{TWSA_{i,t}})^2}$$

where σ represents the estimated uncertainty, and *TWSA* is the reconstructed results from the U-Net deep learning model. The subscript *m* represents the model uncertainty, whereas *i* and *t* denote the pixel

and month. The superscript n denotes the reconstructed TWSA from the n-th independent training run, and N is the total number of independent runs.

3. Line 367. Could the authors provide more details on how the benchmark uncertainty value of "1.96 cm" for the JPL-M product was obtained? Specifically, it would be valuable to clarify whether this value represents the formal satellite measurement error provided in the JPL-M product or an empirical metric, so that readers can better understand the baseline comparison between this value and the AE loss of the downscaled TWSA.

Response to comment 3: The uncertainty for the JPL-M was obtained from the satellite measurement, provided with the GRACE JPL-M product. We used the median value of JPL-M uncertainty estimates across China as the regional benchmark of GRACE observations. This will be clearly clarified in the revised manuscript.

4. The definition and description of TWSA_PI could be made more consistent throughout the manuscript. In some parts, it is defined as the sum of PCR-GLOBWB 2 simulations and satellite-derived glacier mass changes, while in others, it is referred to mainly as simulations from PCR-GLOBWB 2. It is suggested to introduce a clear mathematical equation explicitly defining TWSA_PI and ensure that the terminology and notations remain unified across all sections.

Response to comment 4: We appreciate this comment. To avoid potential confusion with physics-informed neural networks (e.g., PINN), we will revise "physically informed TWSA ($TWSA_{PI}$)" to "bottom-up TWSA ($TWSA_{BU}$)". $TWSA_{BU}$ represents the sum of different water storage components, including surface (rivers, lakes, reservoirs, canopy, and snow) and subsurface (soil and groundwater) water storage simulated by PCR-GLOBWB 2 and glacier storage retrieved from satellite observations. This will be consistently clarified in the revised manuscript.

5. Line 185. It would be helpful to add a brief physical explanation for upscaling the JPL-M mascon product from its nominal 0.5° resolution back to its native 3° resolution (e.g., ensuring data independence by excluding land-grid scaling factors and respecting the effective spatial resolution of GRACE), explaining why using the 0.5° grid directly was avoided.

Response to comment 5: We will clarify that the motivation of upscale JPL-M from its nominal 0.5° resolution back to its native 3° resolution is to exclude land-grid scaling factors and respect the effective spatial resolution of GRACE. The 3° estimation does not include land-grid-scaling gain factors based on the CLM model, so that the data can provide observational information that is entirely independent of hydrological models.

6. Line 324: The bounding coordinates (15°–57°N, 69°–138°E) do not cover the entire territorial waters of China (such as the southernmost islands in the South China Sea). It is suggested to rephrase this as "...covering mainland China and surrounding regions..." to ensure geographic and cartographic precision.

Response to comment 6: We appreciate this comment and will correct this in the revised manuscript.

7. Lines 108, 178, and 550. The dates defining the gap period between GRACE and GRACE-FO conflict.

Response to comment 7: We will clarify that the gap period between the GRACE and GRACE-FO missions is from July 2017 to May 2018 and ensure consistency throughout the revised manuscript.

8. Figure 1: The locations of the in-situ groundwater wells in the North China Plain inset are difficult to distinguish. Please enlarge the marker size and/or select a higher-contrast color to improve legibility.

Response to comment 8: We will add an enlarged inset of the North China Plain to the main map of China in Figure 1, with larger markers and higher-contrast colors to improve the visibility of the in-situ groundwater well locations.

9. Figure 2: In the figure schematic and caption, define the abbreviations "AE" (Absolute Error) and "MAE" (Mean Absolute Error). Figure captions should be self-contained so readers can interpret the workflow without referring back and forth to the main text.

Response to comment 9: We will clarify the abbreviations in the caption of Figure 2 as follows: AE indicates the Absolute Error between averaged GRACE TWSAs and averaged predicted TWSAs over each patch, and MAE denotes the Mean Absolute Error between bottom-up TWSAs and predicted TWSAs at the pixel level.

10. Line 686. The sub-grid variability is displayed in the fourth row (d1–d3), not the “fourth column”. Please correct this typographical error.

Response to comment 10: We will correct this typo in the revised main text.

Again, we sincerely appreciate these constructive comments to further improve our work.