To: The Editor and Reviewer 1, *Earth System Science Data (ESSD)*

Subject: Point-by-point response to Reviewer 1 Comments

Manuscript Title: High-resolution global groundwater storage anomalies: A 1 km downscaled dataset

Dear Editor and Reviewer 1,

We would like to express our sincere gratitude for your thorough and highly constructive evaluation of our manuscript. Your insightful comments have been instrumental in improving the rigor, reliability, and clarity of our dataset and manuscript.

In response to your comments, we have substantially strengthened the manuscript by expanding the independent in-situ validation framework, adding representative aquifer-scale temporal comparisons, conducting controlled source-sensitivity experiments across multiple GRACE- and GHM-derived target products, revising the treatment and uncertainty analysis of the effective specific yield, and providing a more detailed evaluation of the SSA gap-filling procedure. We have also clarified the temporal scope of the 2002–2020 downscaled product and the retrospective use of the extended GRACE/GRACE-FO record.

Below, we provide a detailed, point-by-point response to your comments. The reviewer's comments are marked in **black bold**, and our responses are presented in blue/regular text.

Major Comments

1. As shown in Table 1, TWSA data extend through the end of 2025, whereas the other input data sets end in 2020. This five-year temporal mismatch requires justification. Why was the GWSA data set not updated through 2025? Other satellite-based products that are regularly updated and currently available should be considered to extend the temporal coverage of the input data.

Response:

We thank the reviewer for raising this important point regarding the temporal coverage of the dataset. We agree that extending the 1 km GWSA product beyond 2020 would be scientifically valuable, and we carefully evaluated the feasibility of such an extension during the revision.

Although the CSR GRACE/GRACE-FO mascon record used in this study is currently available through November 2025, the downscaled GWSA product is restricted to April 2002–December 2020 because the established production framework requires a temporally homogeneous set of global high-resolution dynamic predictors. In the present framework, several key 1 km or sub-kilometre hydroclimatic and remotely sensed predictor series used for model training and reconstruction terminate in 2020. Replacing these predictors after 2020 with alternative products having different spatial resolutions, retrieval algorithms, forcing data, or error structures would introduce a temporal discontinuity into the predictor space and would change the methodological basis of the published product.

We therefore chose to preserve temporal and methodological consistency rather than extend the dataset by combining heterogeneous predictor products. This temporal restriction has now been stated explicitly in Sect. 2.1 and Table 1 of the revised manuscript.

We nevertheless agree with the reviewer that extending the temporal coverage is an important future objective. Future versions of the dataset will be extended when a temporally homogeneous and globally consistent set of high-resolution predictor datasets can be assembled for the post-2020 period without altering the spatial and methodological basis of the present reconstruction.

We also clarify that the availability of GRACE/GRACE-FO observations through 2025 is used only to improve the retrospective SSA reconstruction of historical GRACE gaps, as described in our response to Comment 5. It does not provide the missing high-resolution predictor information required to produce a physically and statistically consistent 1 km downscaled product after 2020.

2. The validation is insufficient to establish the reliability of the proposed data set. First, the estimated GWSA should be evaluated against available groundwater-level observations from monitoring wells. Second, the authors should select representative groundwater basins and compare their GWSA estimates with those reported in previous regional studies, with particular attention to differences in temporal variability, magnitude, and long-term trends. Third, comparisons with simulations from existing continental- or global-scale land surface and groundwater models would provide an additional independent assessment. Validation should not be limited to correlation analysis against an existing GWSA product. For comparisons with monitoring-well observations, the authors should also present and interpret time-series plots to determine whether the estimated GWSA captures observed seasonal fluctuations, interannual variability, and long-term trends.

Response:

We fully agree that the reliability of the dataset should not be established solely through correlation with the original GRACE-derived GWSA field. Following the reviewer's suggestion, we substantially expanded the validation framework to include independent groundwater-monitoring observations, representative aquifer-scale temporal comparisons, cross-scale consistency assessment, and comparisons with global hydrological model simulations.

First, we incorporated a global groundwater-monitoring database containing 152,069 wells, from which records satisfying the temporal coverage and quality-control requirements were retained for evaluation. To reduce the influence of spatial clustering, the monitoring network was partitioned into 1° spatial blocks. Approximately 70% of the blocks were used for calibration of the scale-dependent effective specific yield, Syeff, whereas the remaining blocks were held out for independent evaluation. All wells belonging to the same block remained in the same subset. Syeff was estimated only from calibration blocks and transferred to held-out blocks without refitting. The final comparison was then performed on common 0.25° grids, and all reported statistics in revised Fig. 5 were calculated exclusively from held-out grid comparisons.

This spatially independent design provides the principal in-situ assessment of long-term groundwater-storage trend magnitude and spatial covariance. For the XGBoost product, the held-out evaluation yields $CC = 0.734$, $R^2 = 0.538$, $RMSE = 0.378$ cm yr^{-1}, and Bias = 0.106 cm yr^{-1}. Corresponding DNN statistics are also reported to demonstrate the robustness of the evaluation across model architectures.

Second, following the reviewer's suggestion, we added representative aquifer-scale time-series comparisons in Figs. 6–8 to assess interannual variability and longer-term temporal co-variation between groundwater observations and the downscaled GWSA. Because groundwater level and equivalent-water-height storage have different physical dimensions and a uniform local Sy is not available for every aquifer, these comparisons are intentionally treated as conversion-free assessments of temporal co-variation rather than direct tests of absolute storage magnitude. Both series were referenced consistently to the 2004–2009 baseline, standardized using Z-scores, and subjected to the same 3-year rolling-average treatment. The results show strong temporal agreement in several climate-dominated aquifer systems, including the Stampriet-Kalahari Basin (R = 0.82) and Russian Platform Basins (R = 0.67), while weaker or decoupled temporal behavior occurs in some intensively managed or hydrogeologically complex aquifers.

We acknowledge, however, that the 3-year smoothing suppresses the monthly seasonal cycle and therefore does not constitute a direct evaluation of the seasonal fluctuations specifically requested by the reviewer. To address this aspect more explicitly, we are currently extending the analysis to several representative aquifers in China for which sufficiently continuous monthly groundwater-level observations are available. The additional comparison will use unsmoothed monthly records to evaluate the correspondence between observed groundwater-level variability and the downscaled GWSA, with particular attention to the timing and magnitude of seasonal fluctuations as well as their interannual evolution. These additional monthly-scale comparisons will be incorporated into the subsequent revised manuscript and will complement the existing aquifer-scale comparisons in Figs. 6–8.

Third, we added comparisons with GHM-simulated background variability. In the revised Sect. 3.6, this analysis is explicitly framed as "Residual Signals Not Explained by the Modeled Natural Background." The residuals are not interpreted as direct anthropogenic pumping estimates. They may contain contributions from human influence, GHM structural and forcing errors, GRACE component-separation uncertainty, and downscaling uncertainty. Their purpose is therefore diagnostic: persistent spatial departures can identify regions requiring further investigation with independent pumping, irrigation, hydrogeological, or other high-resolution observations.

Together, the spatially blocked in-situ evaluation, aquifer-scale standardized time-series comparisons, cross-scale consistency assessment, and GHM-based diagnostic analysis provide complementary tests of the reconstructed product. The current analyses evaluate long-term trend magnitude and interannual temporal co-variation, while the additional unsmoothed monthly comparisons for selected aquifers in China will specifically address seasonal-scale groundwater variability.

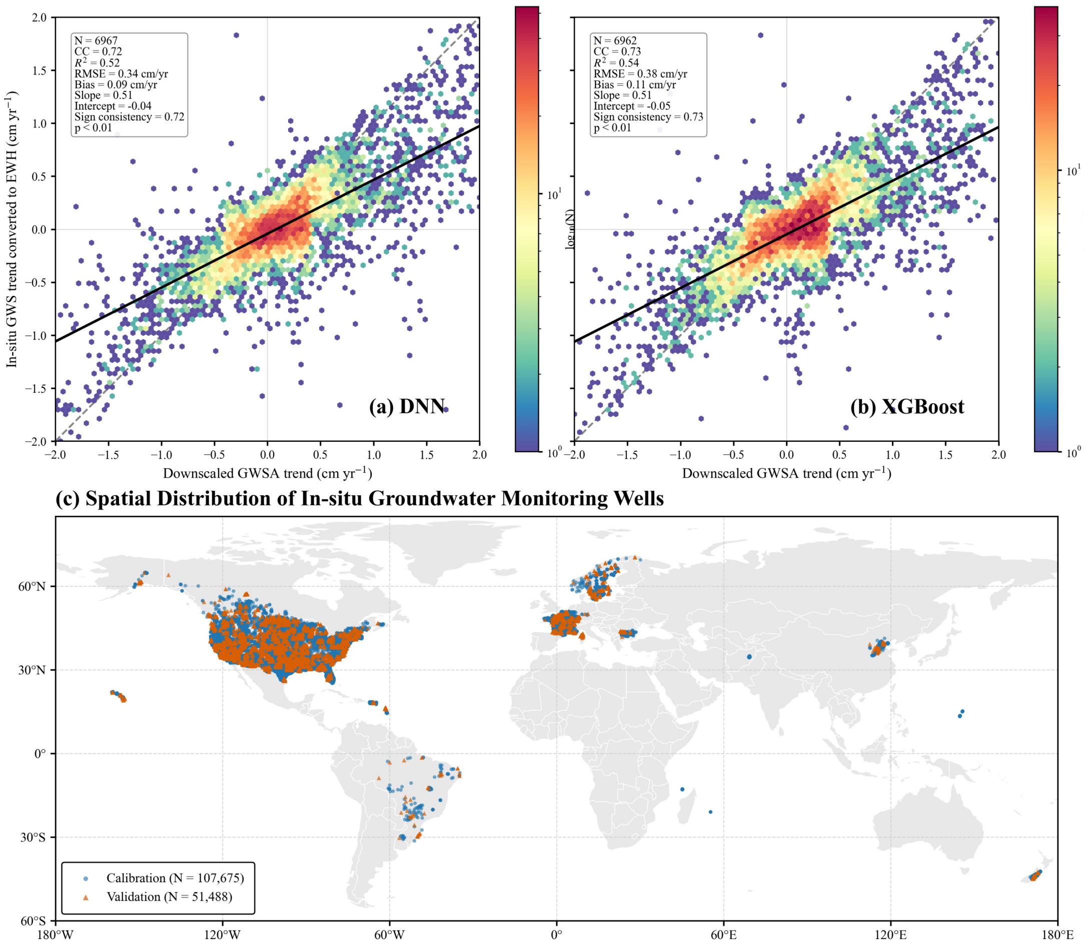

Figure 5. Spatially independent, quality-controlled in-situ evaluation of long-term downscaled groundwater-storage trends and spatial distribution of the monitoring network. (a) Held-out 0.25° grid comparison between DNN-derived GWSA trends and groundwater-level trends converted to equivalent water height using S_{yeff} estimated only from calibration spatial blocks and blindly transferred to the held-out blocks. (b) Corresponding held-out comparison for XGBoost. The network was partitioned into 1° spatial blocks; approximately 70% were calibration blocks and 30% were held-out validation blocks, with all wells from a block assigned to one subset. Neither held-out wells nor their blocks contributed to S_{yeff} estimation or transfer selection. Colors indicate the logarithmic density of valid held-out grid comparisons; the gray dashed line is the 1:1 relationship and the solid black line is the fitted regression. N and all reported statistics (*CC*, R^2, *RMSE*, *Bias*, slope, intercept, trend-sign consistency, and significance) are calculated exclusively from held-out 0.25° grids. (c) Spatial distribution of groundwater-monitoring wells and the 1° spatial-block partition used for calibration and validation.

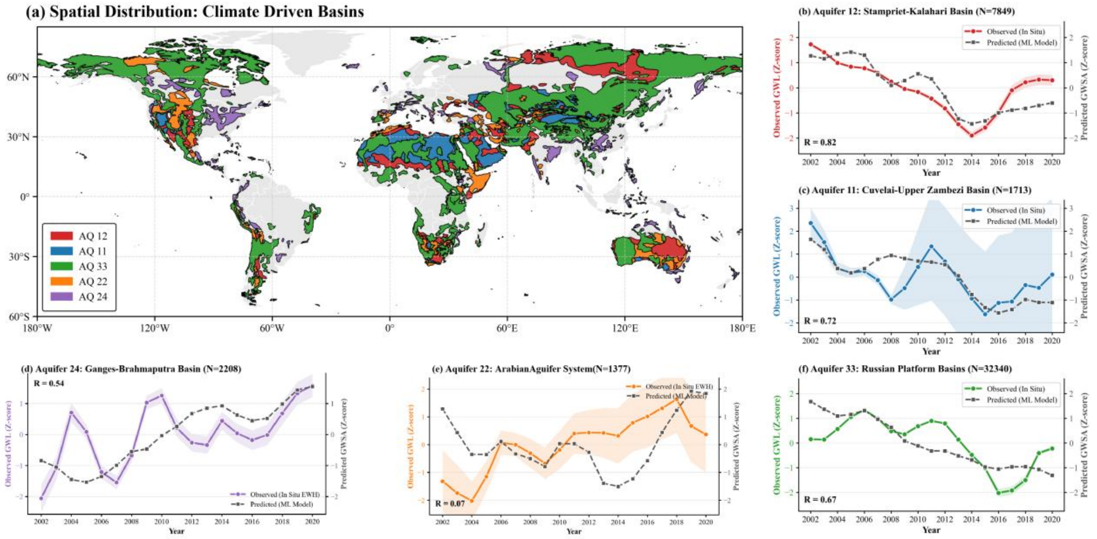

Figure 6. Spatial distribution and interannual variations of GWL and downscaled GWSA for climate-driven/natural-dominated major aquifer basins. (a) Spatial locations of selected aquifers with specific IDs (AQ 12: Stampriet-Kalahari Basin, AQ 11: Cuvelai-Upper Zambezi Basin, AQ 33: Russian Platform Basins, AQ 22: Arabian Aquifer System, and AQ 24: Ganges-Brahmaputra Basin). (b–f) Temporal comparisons between in situ observed GWL (solid colored curves with standard error shading) and machine learning-predicted downscaled GWSA (gray dashed curves) from 2002 to 2020. To ensure temporal and dimensional consistency, both time series were anchored to a 2004–2009 baseline, smoothed via a 3-year rolling average to suppress high-frequency noise, and standardized using *Z*-scores. *N* represents the number of valid monitoring wells, and *R* denotes the *Pearson CC* between the standardized observations and predictions.

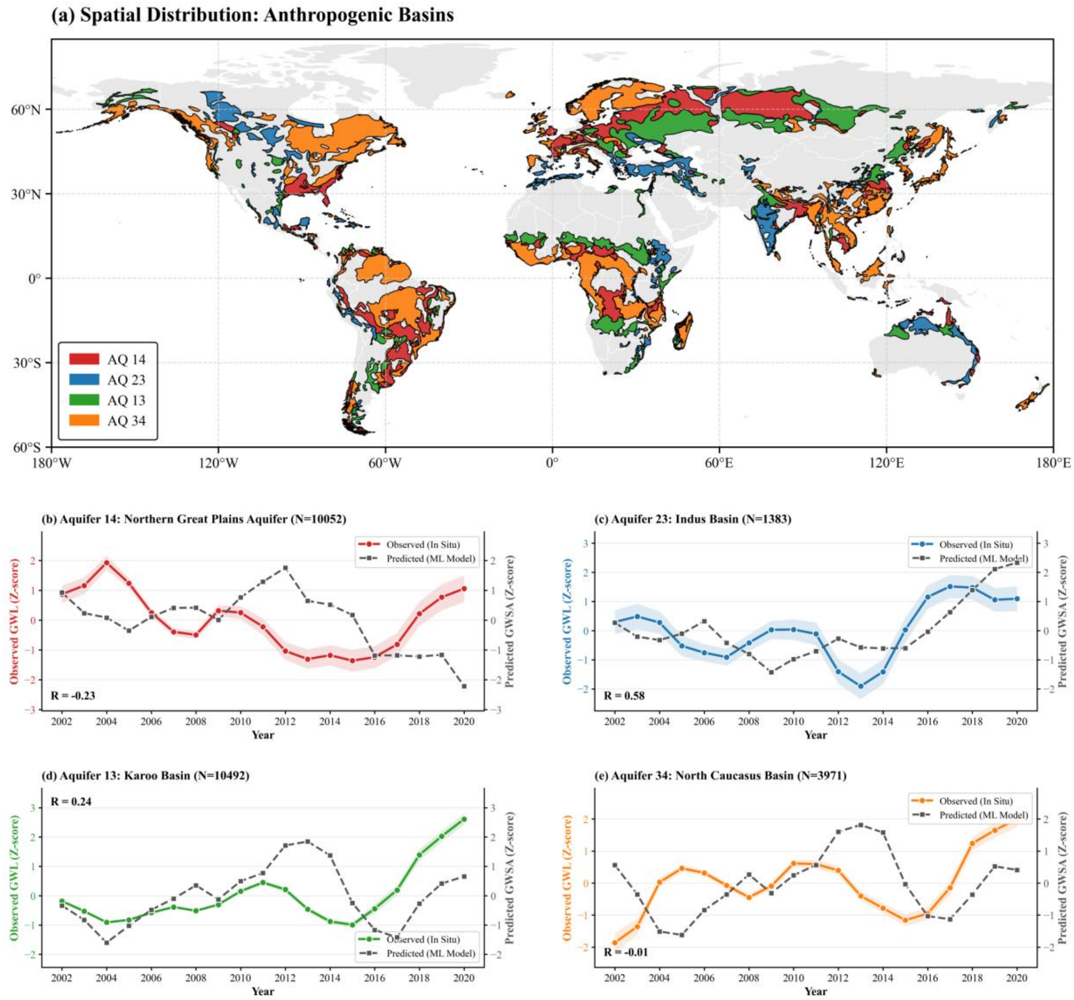

Figure 7. Spatial distribution and interannual variations of groundwater dynamics for anthropogenic-dominated major aquifer basins. (a) Spatial locations and identifiers of selected human-impacted aquifers (AQ 14: Northern Great Plains Aquifer, AQ 23: Indus Basin, AQ 13: Karoo Basin, and AQ 34: North Caucasus Basin). (b–e) Interannual time series of in situ observed GWL (colored solid lines with shading representing standard errors) versus machine learning-derived downscaled GWSA (gray dashed lines) for the period 2002–2020. Both sequences were adjusted to a 2004–2009 baseline, processed with a 3-year rolling average, and plotted in a Z-score standardized space to facilitate direct comparison of evolutionary rhythms. *N* indicates the total count of valid monitoring wells, and *R* indicates the corresponding *Pearson CC* for each basin.

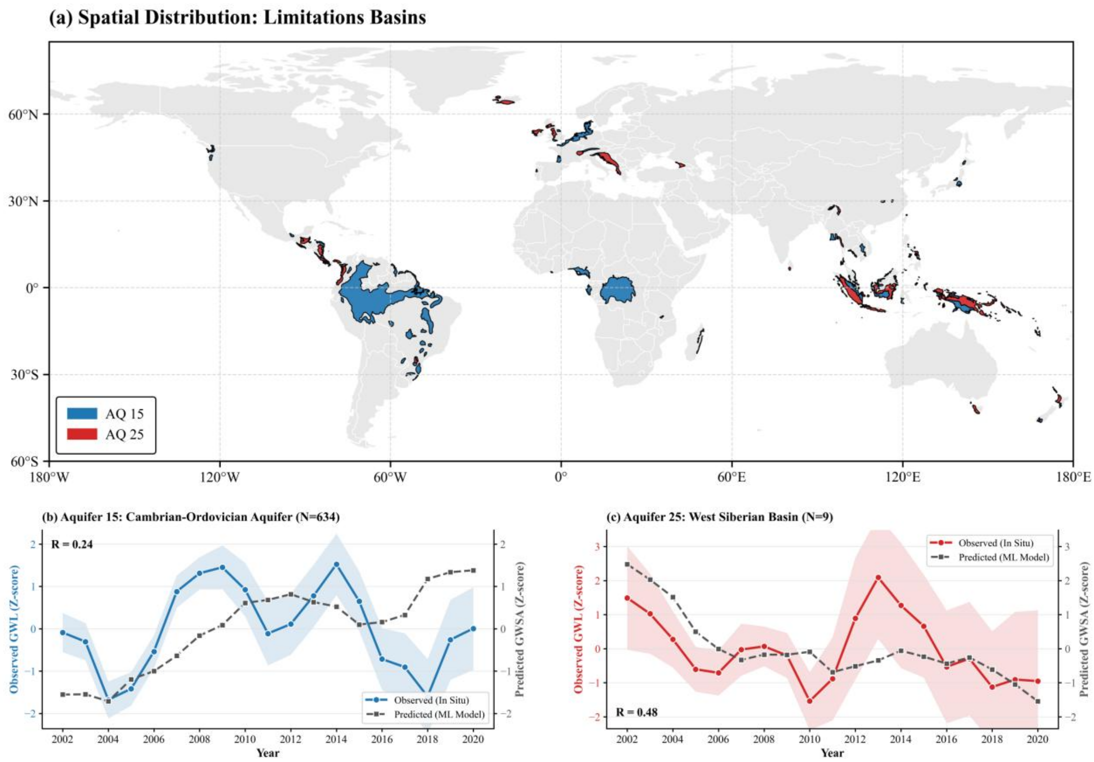

Figure 8. Spatial distribution and interannual variations of groundwater anomalies for aquifers under physical limitations or specific climatic-hydrogeological constraints. (a) Geographic locations of the selected aquifers (AQ 15: Cambrian-Ordovician Aquifer and AQ 25: West Siberian Basin). (b – c) Comparisons of in situ observed GWL (colored solid lines with standard error bands) and predicted downscaled GWSA (gray dashed lines) from 2002 to 2020. The time series reflect Z-score standardized values after applying a 2004 – 2009 baseline alignment and a 3-year rolling average. *N* represents the effective number of observation wells utilized, and *R* denotes the *Pearson CC* between the in situ records and model estimates.

3. Products derived from different data sources should be systematically compared and cross-validated. The manuscript should quantify the consistency and discrepancies among these products and discuss how differences in data sources propagate into the final GWSA estimates.

Response:

We thank the reviewer for emphasizing source-related uncertainty. In response, we have now completed a controlled target-product sensitivity experiment rather than leaving this comparison only as future work.

The aquifer-stratified XGBoost workflow was rerun for four target configurations: CSR- and JPL-derived GRACE/GRACE-FO GWSA fields and groundwater-storage fields from WaterGAP and H08. The 19 predictors, aquifer masks, 200,000-sample development pool, random 70%/30% internal split, and hyperparameter-search protocol were held constant so that differences could be attributed primarily to the target product.

Across the 12 aquifer categories and masks, mean internal-test R^2 was 0.928 for CSR, 0.946 for JPL, 0.888 for WaterGAP, and 0.908 for H08. JPL achieved the highest R^2 in 10 categories, while CSR and H08 each ranked first in one. The corresponding R^2 ranges were 0.900–0.994, 0.915–0.996, 0.794–0.924, and 0.828–0.941, respectively. Because the products have different target amplitudes, we use dimensionless R^2 for the cross-product comparison and retain RMSE and MAE for within-product interpretation only.

The feature-importance analysis provides an additional diagnostic of uncertainty propagation. CSR and JPL assigned 63.8% and 62.0% of total importance to static physiography and soil attributes, whereas WaterGAP and H08 assigned 45.4% and 44.2%. Temporal encoding contributed 10.2% and 10.7% for CSR and JPL but 19.1% and 22.5% for WaterGAP and H08. The complete importance profiles were highly consistent within the GRACE-derived pair (Spearman $\rho = 0.954$) and within the GHM pair ($\rho = 0.909$), but cross-family correlations were lower ($\rho = 0.382$–0.425).

These results, now reported in revised Fig. 3 and Sect. 3.1.2, show both robustness and source dependence: the workflow achieves high within-product reconstruction skill across all four targets, but the balance between static heterogeneity and temporal dynamics changes with the target product. We explicitly describe these random-split statistics as internal source-sensitivity diagnostics rather than independent or temporally blocked validation.

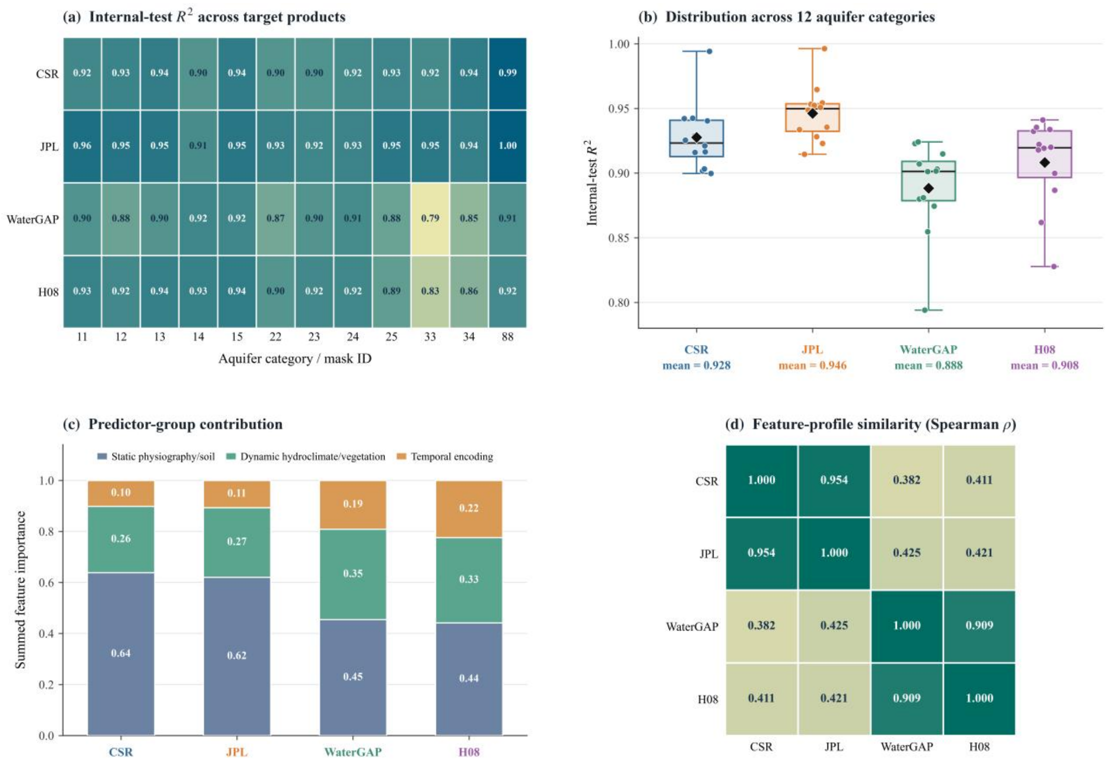

Figure 3. Controlled target-product sensitivity of the aquifer-stratified XGBoost workflow. (a) Internal-test R^2 across aquifer categories for CSR, JPL, WaterGAP, and H08. (b) R^2 distributions and means. (c) Predictor-group importance. (d) Spearman similarity of the complete feature-importance profiles. Identical predictors and model-development procedures were used in all experiments; the results quantify within-distribution source sensitivity and are not temporal-forecast tests.

4. As shown in Table 2, assigning a uniform specific yield value to each of the seven regions may not adequately represent the spatial heterogeneity of aquifer properties. Because GWSA estimates derived from groundwater-level observations are sensitive to Sy, the authors should evaluate the sensitivity of their results to the selected Sy values and discuss the uncertainty introduced by this regional parameterization.

Response:

We thank the reviewer for highlighting the uncertainty associated with spatial variability in aquifer storage properties. We agree that assigning a single specific-yield value to broad hydrogeological regions cannot adequately represent the substantial heterogeneity of aquifer systems at the global scale.

In the revised manuscript, we therefore replaced the previous regionally uniform S_y parameterization with a scale-dependent effective specific yield, S_{yeff}, used as a level-to-storage conversion coefficient for harmonizing groundwater-level trends with equivalent-water-height groundwater-storage trends. S_{yeff} is treated as an effective observational conversion parameter rather than as a direct estimate of local aquifer-specific yield, and it is used only in the in-situ evaluation framework.

To reduce spatial dependence between coefficient estimation and validation, the global monitoring network was partitioned into 1° spatial blocks. Approximately 70% of the blocks were used for S_{yeff} calibration, whereas the remaining 30% were retained for validation. The calibrated coefficients were transferred to the validation locations without refitting, and the converted groundwater-storage trends were subsequently compared with the downscaled GWSA at a common 0.25° grid scale following the same aggregation and quality-control procedures described in the revised Methods.

We further quantified the sensitivity of the validation results to uncertainty in Syeff by perturbing the baseline coefficients by ±15% and ±30%. For the XGBoost product, the baseline evaluation yielded R2=0.538, CC = 0.734, RMSE = 0.378 cm yr^{-1}, and Bias = 0.106 cm yr^{-1}. Under the perturbed scenarios, R2 ranged from 0.477 to 0.520, CC from 0.691 to 0.721, RMSE from 0.359 to 0.408 cm yr^{-1}, and Bias from 0.104 to 0.107 cm yr^{-1}. The fitted regression slope showed a larger response, varying from 0.379 to 0.593, while trend-sign consistency remained between 0.696 and 0.743. DNN exhibited a comparable sensitivity pattern. These results indicate that uncertainty in Syeff primarily affects amplitude scaling, whereas the overall association structure and trend-direction agreement remain comparatively stable. These results are now reported explicitly in Table 3 and discussed in Sect. 3.1.3 and Sect. 4.3 of the revised manuscript.

Accordingly, we have revised Sects. 2.5.2–2.5.4 to clarify the S_{yeff} conversion and sensitivity framework, Sect. 3.1.3 to report the quantitative sensitivity results, Table 3 to provide the complete perturbation experiment, and Sect. 4.3 to discuss the associated uncertainty and its implications for the interpretation of the in-situ evaluation.

Table 3

Sensitivity of the global in-situ evaluation to perturbations in S_{yeff}. The 0% scenario represents the baseline configuration. All perturbation scenarios were evaluated using the same spatial aggregation, uncertainty treatment, quality-control procedure, and statistical metrics.

Model	S_{yeff} perturbation	*CC*	R^2	*RMSE (cm yr^{-1})*	*Bias (cm yr^{-1})*	*Slope*	*Trend-sign consistency*
	−30%	0.690	0.476	0.393	0.102	0.381	0.693
	−15%	0.719	0.517	0.377	0.100	0.446	0.708
DNN	0%	0.722	0.522	0.343	0.094	0.508	0.722
	+15%	0.714	0.510	0.348	0.098	0.556	0.730
	+30%	0.704	0.496	0.357	0.099	0.592	0.728
	−30%	0.691	0.477	0.408	0.105	0.379	0.696
	−15%	0.714	0.510	0.391	0.106	0.442	0.720
XGBoost	0%	0.734	0.538	0.378	0.106	0.506	0.734
	+15%	0.721	0.520	0.359	0.104	0.563	0.743
	+30%	0.705	0.497	0.369	0.107	0.593	0.735

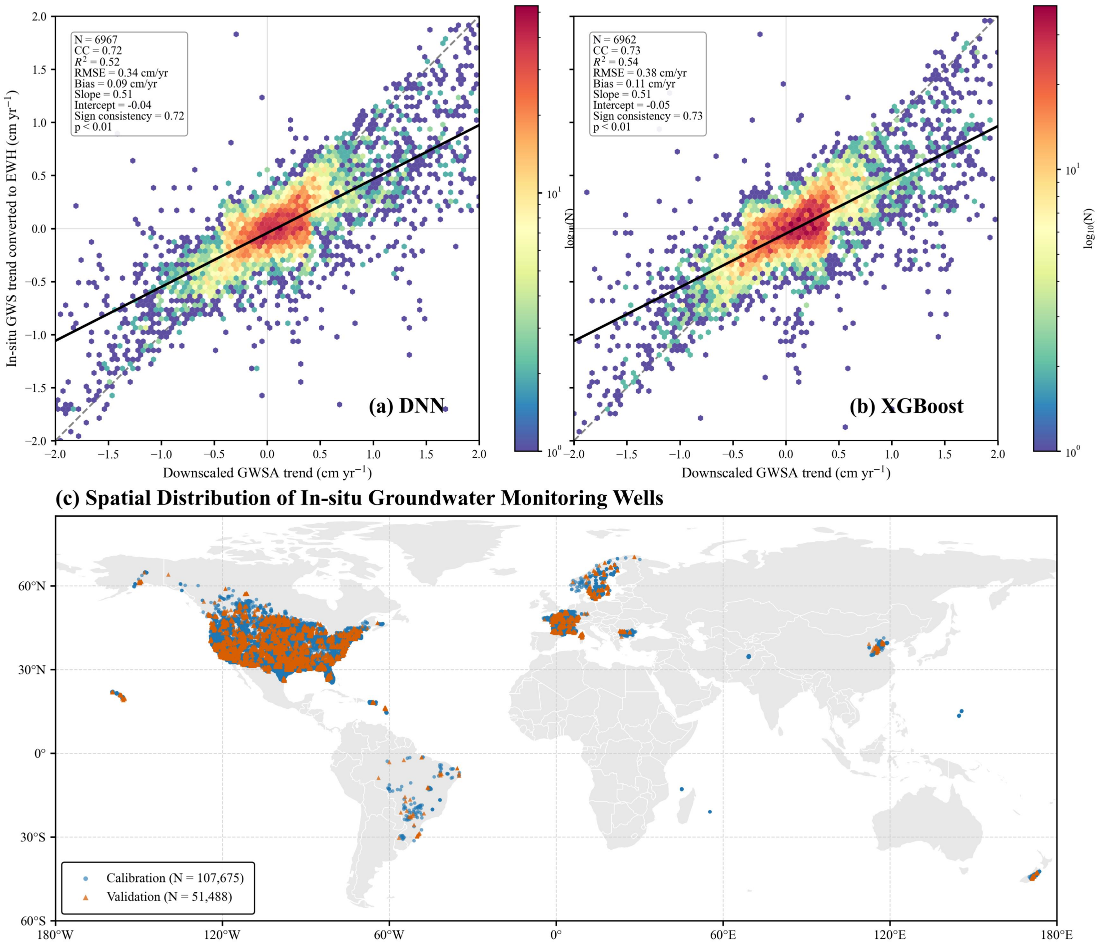

Figure 5. Spatially independent, quality-controlled in-situ evaluation of long-term downscaled groundwater-storage trends and spatial distribution of the monitoring network. (a) Held-out 0.25° grid comparison between DNN-derived GWSA trends and groundwater-level trends converted to equivalent water height using S_{yeff} estimated only from calibration spatial blocks and blindly transferred to the held-out blocks. (b) Corresponding held-out comparison for XGBoost. The network was partitioned into 1° spatial blocks; approximately 70% were calibration blocks and 30% were held-out validation

blocks, with all wells from a block assigned to one subset. Neither held-out wells nor their blocks contributed to Syeff estimation or transfer selection. Colors indicate the logarithmic density of valid held-out grid comparisons; the gray dashed line is the 1:1 relationship and the solid black line is the fitted regression. N and all reported statistics (CC, R2, RMSE, Bias, slope, intercept, trend-sign consistency, and significance) are calculated exclusively from held-out 0.25° grids. (c) Spatial distribution of groundwater-monitoring wells and the 1° spatial-block partition used for calibration and validation.

5. The manuscript should provide a more detailed description of how temporal data gaps in the GRACE and GRACE-FO records were handled. The authors should specify the gap-filling method, its underlying assumptions, and the uncertainty it introduces into the reconstructed GWSA time series.

Response:

We thank the reviewer for requesting additional clarification of the GRACE/GRACE-FO gap-filling procedure and its associated uncertainty. The revised manuscript now provides a detailed description of the two-stage Singular Spectrum Analysis (SSA) reconstruction framework and evaluates uncertainty separately for short and prolonged missing-data periods.

SSA-filling-a was used to reconstruct 22 discrete missing months occurring before June 2017. SSA-filling-b was used for the 11-month gap between the GRACE and GRACE-FO missions and two additional missing GRACE-FO months. Missing values were initialized as zero only as starting values for the iterative procedure; they were not retained as reconstructed observations. The final estimates were obtained through iterative singular-value decomposition of the trajectory matrix, modal selection, and convergence testing.

For the short-gap configuration, reconstruction uncertainty was quantified as the standard deviation of the differences between observed GRACE values and SSA-reconstructed values during available months. For the prolonged-gap configuration, annual withholding experiments were performed: each year from 2004 to 2015 was successively removed, reconstructed, and evaluated against the original observations using RMSE. The 12 experiments were then aggregated to determine the optimal SSA parameter configuration.

The revised results show a median short-gap reconstruction error of 1.156 cm. For the prolonged-gap configuration, the global median reconstruction error is 1.679 cm, increasing to 2.708 cm over land excluding Antarctica. The corresponding spatial distributions and percentile ranges are now reported in Fig. 2 and the revised Results section.

We further clarify that the complete GRACE/GRACE-FO record through November 2025 was used only because this is a retrospective reconstruction rather than a forecasting exercise. Post-2020

GRACE/GRACE-FO observations provide additional temporal information for stabilizing the SSA reconstruction of earlier missing months, but they are not used to generate a downscaled 1 km product after 2020. The temporal extent of the final downscaled product remains constrained by the availability of the high-resolution predictor datasets, as explained in our response to Comment 1.

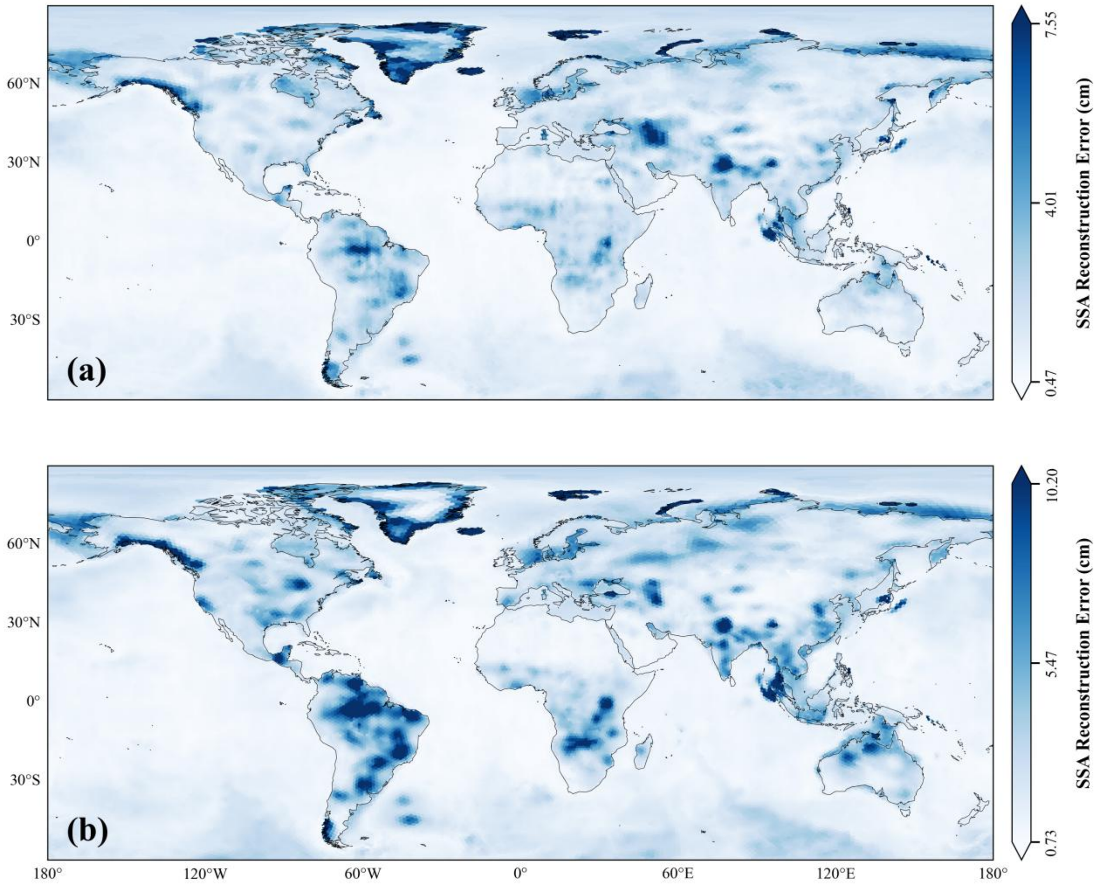

Figure 2. Evaluation of the reconstruction accuracy of SSA. (a) Reconstruction error in the SSA-filling-a stage (cm), quantified as the dispersion of residuals between reconstructed and observed values over all available months. (b) Reconstruction error in the SSA-filling-b stage (cm), quantified as the aggregated RMSE derived from repeated masking experiments in which the original observations for 2003–2016 were withheld one year at a time and then reconstructed.

6. Each abbreviation should be defined only at its first occurrence and used consistently thereafter. Repeatedly spelling out a term after its abbreviation has already been defined is unnecessary.

Response:

We thank the reviewer for pointing out the inconsistencies in abbreviation and notation usage. We have systematically reviewed the Abstract, main text, tables, and figure captions and revised the manuscript to improve the consistency and readability of abbreviations and technical notation.

In the revised manuscript, abbreviations are introduced at their first relevant occurrence and are subsequently used consistently without repeatedly spelling out the corresponding full terms. We also removed several unnecessary or one-time abbreviations where the full terminology was clearer. For example, major terms such as Gravity Recovery and Climate Experiment (GRACE), groundwater storage anomalies (GWSA), Singular Spectrum Analysis (SSA), groundwater level (GWL), Google Earth Engine (GEE), and the World-wide Hydrogeological Mapping and Assessment Programme (WHYMAP) are now explicitly defined at their first occurrence.

We further harmonized statistical and methodological terminology throughout the manuscript. Root mean square error (RMSE) is defined once in the Methods and used consistently thereafter, while mean absolute error (MAE), which is additionally reported in the aquifer-specific model comparison, is now explicitly defined in Sect. 2.5.1. We also revised repeated or potentially ambiguous shorthand expressions in the validation sections, including the terminology used for the effective specific-yield conversion and the standardized time-series comparison.

In addition, notation in the main text, tables, and figure captions has been checked for consistency so that statistical metrics, aquifer identifiers, units, and model-related terminology follow a uniform presentation throughout the revised manuscript. These revisions improve terminological consistency and avoid unnecessary repetition of terms after their abbreviations have already been introduced.

We appreciate the reviewer’s suggestion, which helped us substantially improve the clarity and consistency of the manuscript.

We hope that these extensive revisions address the reviewer’s concerns and substantially strengthen the validation, uncertainty characterization, and scientific positioning of our 1 km global dataset. Thank you once again for your time and highly valuable feedback.

Sincerely,

The Authors