

1 **To:** The Editor and Referees of Earth System Science Data (ESSD)
2 **Subject:** Point-by-Point Response to the Referees' Comments
3 **Manuscript Title:** High-resolution global groundwater storage anomalies: A 1 km downscaled dataset

4 Dear Editor and Referees,

5 Thank you for your careful evaluation of our manuscript and for the constructive comments provided
6 during the interactive discussion. We have revised the manuscript accordingly. The main revisions
7 include:
8 (i) expanding the in-situ validation framework to include spatially blocked groundwater-monitoring
9 evaluation, aquifer-scale temporal comparisons, monthly STL-based validation, and regional literature
10 benchmarks;
11 (ii) conducting controlled source-sensitivity experiments using multiple GRACE- and GHM-derived
12 target products;
13 (iii) replacing the previous regional uniform specific-yield parameterization with a scale-dependent
14 effective specific yield S_{yeff} and evaluating its uncertainty;
15 (iv) providing additional details on the SSA gap-filling procedure, GLDAS processing, and the
16 temporal coverage of the product;
17 (v) revising the interpretation of kilometre-scale spatial patterns and residuals relative to the
18 GHM-simulated background to avoid unsupported causal attribution; and
19 (vi) improving the consistency of abbreviations, notation, terminology, and manuscript wording.

20 For each referee comment, the response is organized in the following sequence: (1) Referee Comment,
21 (2) Author's Response, and (3) Author's Changes in the Manuscript. **Referee comments are shown in**
22 **bold black text**, **our responses in blue text**, and *specific manuscript changes in red italic text*.

23 A marked-up version of the revised manuscript showing all changes is appended to this response file, as
24 requested by the editor. All page and line numbers refer to the clean revised manuscript.

25

26

Point-by-point response to Referee 1 Comments

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.

Reviewer comments are shown in **black/bold**; our response follows in **blue normal text**; and specific revisions to the manuscript are summarized as *italicized red normal 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 could be extended when a temporally homogeneous and globally consistent set of high-resolution predictor datasets becomes available for the post-2020 period.

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.

Author's changes in manuscript:

[p. 8, lines 154-157] Sect. 2.1 (Data Sources), third paragraph: added explicit statement that the temporal coverage is limited to April 2002–December 2020 due to predictor availability.

[pp. 8-9, Table 1 (caption at line 158)] Table 1: temporal coverage column updated to reflect the actual end date of each predictor (mostly Dec 2020).

[p. 9, lines 159-179; pp. 13-16, lines 286-351] Sect. 2.1.1 and 2.2: clarified that the full GRACE/GRACE-FO record through November 2025 is used only for retrospective SSA reconstruction and does not extend the temporal coverage of the 1 km product.

[pp. 1-2, lines 16-39; pp. 58-59, lines 1317-1348] Abstract and Conclusions: temporal scope stated explicitly as April 2002–December 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 spatially independent groundwater-monitoring evaluation, aquifer-scale temporal comparisons, monthly seasonal analysis, regional benchmarking against published studies, cross-scale consistency assessment, and comparisons with global hydrological model simulations.

To provide an observation-based assessment independent of existing GWSA products, we introduced three in situ groundwater-level datasets. The first is the global groundwater-monitoring network compiled by Jasechko et al. (2024), which contains annual groundwater-depth records from 152,069 monitoring wells spanning 1801–2022. The second is the Global-Scale Time Series Dataset for Groundwater Studies within the Earth System (GROW; B  thge et al., 2026), which includes more than 200,000 groundwater time series from 55 countries, with daily, monthly, and annual records. The third is the China Groundwater Level Yearbook for Geo-environmental Monitoring, which provides daily records from January 2005 to December 2017 for 1,074 georeferenced stations across 31 provincial-level regions, including 266 stations with continuous 13-year observations. These datasets were processed and spatially matched to the WhyMAP aquifer framework for the independent validation analyses described below.

2.1 Spatially Independent In-situ Evaluation Using Global Groundwater-Monitoring Observations

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, S_{yeff} , whereas the remaining blocks were held out for independent evaluation. All wells belonging to the same block remained in the same subset. S_{yeff} 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. 4 were calculated exclusively from held-out grid comparisons.

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

2.2 Aquifer-scale In-situ Time-series Comparisons

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 S_y 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 relatively strong temporal agreement in several climate-dominated aquifer systems, including the Stampriet-Kalahari Basin ($CC = 0.82$), the Cuvelai-Upper Zambezi Basin ($CC = 0.72$) and Russian Platform Basins ($CC = 0.67$), while weaker or decoupled temporal behavior occurs in some intensively managed or hydrogeologically complex aquifers.

2.3 Monthly-scale Validation of Seasonal, Interannual, and Long-term Variability

In addition, we added a monthly-scale validation of the reconstructed GWSA against processed in situ groundwater-level observations. To directly assess the seasonal fluctuations highlighted by the reviewer, we additionally calculated and compared the seasonal contribution of the reconstructed and observed series. The data sources and preprocessing procedures are described in Sect. 2.1.5. Five representative validation regions were selected for monthly analysis. The analysis included 55 GROW stations with records covering April 2002–December 2020 and 87 stations from the China Groundwater Level Yearbook with records covering 2005–2017.

For each aquifer, station observations were aggregated using the median method. The reconstructed GWSA and groundwater-level series were then standardized to remove dimensional differences. STL decomposition was applied to separate the original, long-term trend, seasonal, and residual components. Pearson correlation coefficients were calculated for the original, trend, and seasonal components, and the seasonal contribution was calculated for both series. These procedures and the corresponding time-series plots are presented in the revised Sect. 3.2.2 and Fig. 9.

Across the five aquifers, the mean correlation coefficients for the original, trend, and seasonal components were 0.51, 0.55, and 0.39, respectively. The reconstruction therefore agreed slightly better with long-term variability than with monthly or seasonal fluctuations. The mean seasonal contribution was approximately 0.39 for the reconstruction and 0.48 for the in situ observations. For AQU 13, the reconstructed and observed seasonal contributions were relatively close for both the High Plains Aquifer and the North China Plain. The difference was also small for the Floridan Aquifer in AQU 24. In contrast, the observed seasonal contributions reached 0.74 and 0.72 in the Apennine and Scandinavian aquifer systems in AQU 34, respectively, compared with 0.49 and 0.58 for the reconstructed series. The reconstruction therefore showed closer seasonal agreement in the larger aquifer systems examined here, but tended to underestimate seasonal contributions in some local and shallow aquifers.

The High Plains Aquifer and North China Plain had original-series correlations of 0.67 and 0.82, respectively. For the High Plains Aquifer, the reconstruction captured the transition from early depletion to later recovery, and its seasonal contribution was close to that of the in situ observations (0.27 and 0.23). However, the near-zero seasonal correlation indicated a possible phase offset or lag in the seasonal cycle. For the North China Plain, the reconstruction reproduced the persistent groundwater deficit and showed similar seasonal contributions to the observations (0.22 and 0.26), consistent with a relatively coherent response across the alluvial plain. In contrast, the Apennine and Scandinavian aquifer systems showed weaker correspondence in long-term or monthly behavior,

which may reflect local recharge contrasts, snowmelt and freeze–thaw effects, shallow groundwater responses, sparse monitoring, and hydrogeological heterogeneity.

This additional analysis directly evaluates seasonal fluctuations, interannual variability, and long-term changes beyond correlation with an existing GWSA product. Nevertheless, because groundwater-level observations and storage anomalies have different physical dimensions, the monthly comparison was conducted using standardized series and should be interpreted as an assessment of temporal co-variation rather than a direct validation of absolute storage magnitude. The revised manuscript now states this limitation explicitly.

2.4 Regional Benchmarking against Published Groundwater-storage Studies

To address specifically the requested comparison with previous regional studies, we have now completed a four-region literature benchmark and incorporated it into Sects. 2.5.3 and 3.2.3, new Fig. 10, and Table 4. The regions are the North China Plain, northwestern India, the Sacramento–San Joaquin–Tulare domain in California, and the High Plains Aquifer. Monthly numerical data from Zhang et al. (2023) and Liu et al. (2022) were used for the North China Plain and California. Where a directly accessible numerical series was unavailable, we digitized the published monthly curve from Rodell et al. (2009) for northwestern India and annual well-based curves from Rateb et al. (2020) for the High Plains. The study-area definitions and approximate boundary matching are described explicitly; the High Plains footprint follows the 39-cell rule of Breña-Naranjo et al. (2014).

Regional product values were area-weighted, converted to millimetres, and compared without the 3-year smoothing used in Figs. 6–8. Each pair was independently centered over the same baseline, and missing published values were not interpolated. Multiple archive epochs in the same calendar month were averaged for California (January 2012 and April 2015). The evaluation covers temporal variability (r and seasonal amplitude), magnitude (RMSE, Bias, and SD ratio), and long-term change (Theil–Sen slopes, slope differences, sign agreement, and separate exact-window OLS benchmarks). Reconstructed product months are retained at their stated dates and identified in Fig. 10. All four primary time-series articles, and the additional trend benchmarks from Famiglietti et al. (2011) and Breña-Naranjo et al. (2014), are now cited in the manuscript.

The results are region-dependent. The North China Plain has the strongest agreement ($r = 0.963$; $RMSE = 32.0$ mm), followed by California ($r = 0.789$; $RMSE = 77.3$ mm) and northwestern India ($r = 0.731$; $RMSE = 66.4$ mm). Product/reference SD ratios are 0.912, 0.673, and 0.593, respectively.

Product/reference Sen slopes are $-17.71/-20.27$ mm yr^{-1} in the North China Plain, $-6.96/-15.80$ mm yr^{-1} in California, and $-22.96/-39.47$ mm yr^{-1} in northwestern India. California seasonal amplitudes are similar (82.0/82.7 mm), whereas product seasonality is stronger in the North China Plain (23.2/8.9 mm) and northwestern India (63.9/23.5 mm). These comparisons therefore reveal amplitude discrepancies despite positive temporal correlations.

The High Plains comparison reveals an important limitation. The annual comparison gives $r = 0.526$ and an effectively neutral product Sen slope (-0.10 mm yr^{-1} ; 95% confidence interval -5.10 to $+5.41$ mm yr^{-1}), compared with -9.28 mm yr^{-1} for the digitized well-based composite. Over the exact March 2003–February 2013 window of Breña-Naranjo et al. (2014), the product OLS trend is $+4.50$ mm yr^{-1} ,

opposite to the published $-27.6 \pm 3.3 \text{ mm yr}^{-1}$. The product also underestimates depletion in the exact northwestern India and early California windows. These disagreements and the incomplete match to Liu et al.'s full 2003–2021 record are now discussed in Sect. 3.2.3 and Sect. 4.3.

We retain an explicit distinction between literature benchmarking and independent monitoring-well validation. The three monthly regional references share GRACE information with our target; the High Plains reference is well-based but annual and approximately combined from digitized subregional curves. Consequently, the new analysis directly addresses regional variability, magnitude, and published trends, but it does not constitute independent monthly well validation of the seasonal cycle. The 3-year-smoothed well comparisons likewise cannot establish monthly seasonal skill. This remaining limitation is now stated explicitly, and the conclusions are restricted to the temporal scales and reference types actually evaluated.

Regional references added or explicitly linked to the new comparison

Zhang, J., Hu, L., Sun, J., and Wang, D.: Reconstructing groundwater storage changes in the North China Plain using a numerical model and GRACE data, *Remote Sens.*, 15, 3264, <https://doi.org/10.3390/rs15133264>, 2023.

Rodell, M., Velicogna, I., and Famiglietti, J. S.: Satellite-based estimates of groundwater depletion in India, *Nature*, 460, 999–1002, <https://doi.org/10.1038/nature08238>, 2009.

Liu, P.-W., et al.: Groundwater depletion in California's Central Valley accelerates during megadrought, *Nat. Commun.*, 13, 7825, <https://doi.org/10.1038/s41467-022-35582-x>, 2022.

Rateb, A., et al.: Comparison of groundwater storage changes from GRACE satellites with monitoring and modeling of major U.S. aquifers, *Water Resour. Res.*, 56, e2020WR027556, <https://doi.org/10.1029/2020WR027556>, 2020.

Famiglietti, J. S., et al.: Satellites measure recent rates of groundwater depletion in California's Central Valley, *Geophys. Res. Lett.*, 38, L03403, <https://doi.org/10.1029/2010GL046442>, 2011.

Breña-Naranjo, J. A., Kendall, A. D., and Hyndman, D. W.: Improved methods for satellite-based groundwater storage estimates: A decade of monitoring the High Plains Aquifer from space and ground observations, *Geophys. Res. Lett.*, 41, 6167–6173, <https://doi.org/10.1002/2014GL061213>, 2014.

2.5 Comparison with Global Hydrological Model Simulations

Finally, we added comparisons with GHM simulations as an additional model-based diagnostic. In the revised Sect. 3.6, this analysis is explicitly framed as “Residual Signals Not Explained by the GHM-Simulated 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.

Taken together, the spatially blocked in-situ evaluation, aquifer-scale standardized time-series comparisons, monthly STL-based validation, unsmoothed regional literature benchmarks, cross-scale consistency assessment, and GHM diagnostics provide complementary evidence across multiple temporal and spatial scales. The revised manuscript identifies where temporal variability is reproduced, where seasonal or storage-change amplitudes are attenuated, and where a regional trend

is not reproduced, without claiming uniformly high agreement or independent 1 km observational resolution.

Author's changes in manuscript:

[pp. 12-13, lines 254-285] New Sect. 2.1.5: describes the three in-situ groundwater-level datasets and their preprocessing.

[pp. 21-24, lines 487-569] Sect. 2.5 (In situ observational validation): restructured into three subsections — 2.5.1 S_{yeff} correction and sensitivity analysis, 2.5.2 Conversion-free validation, 2.5.3 Regional comparison with published estimates.

[pp. 33-37, lines 757-820] Sect. 3.2.1: new conversion-free aquifer-scale temporal comparisons (Figs. 6–8).

[pp. 37-40, lines 821-883] Sect. 3.2.2: new monthly STL-based validation (Fig. 9).

[pp. 40-42, lines 884-926] Sect. 3.2.3: new regional literature benchmarks (Fig. 10, Table 4).

[pp. 51-53, lines 1122-1160] Sect. 3.6: new GHM residual diagnostic (Fig. 16).

[pp. 56-58, lines 1236-1314] Sect. 4.3: expanded discussion of validation limitations, regional discrepancies, and scale mismatch.

[pp. 30-31, lines 688-700] Fig. 4: replaced previous binned validation with unbinned held-out 0.25° grid comparisons for DNN and XGBoost.

[p. 31, lines 712-715] Table 3: added sensitivity of in-situ evaluation to $\pm 15\%$ and $\pm 30\%$ perturbations in S_{yeff} .

[p. 41, lines 909-915] Table 4: added regional agreement metrics with published estimates.

[pp. 1-2, lines 16-39; pp. 58-59, lines 1317-1348] Abstract and Conclusions: updated to reflect the expanded validation framework and its findings.

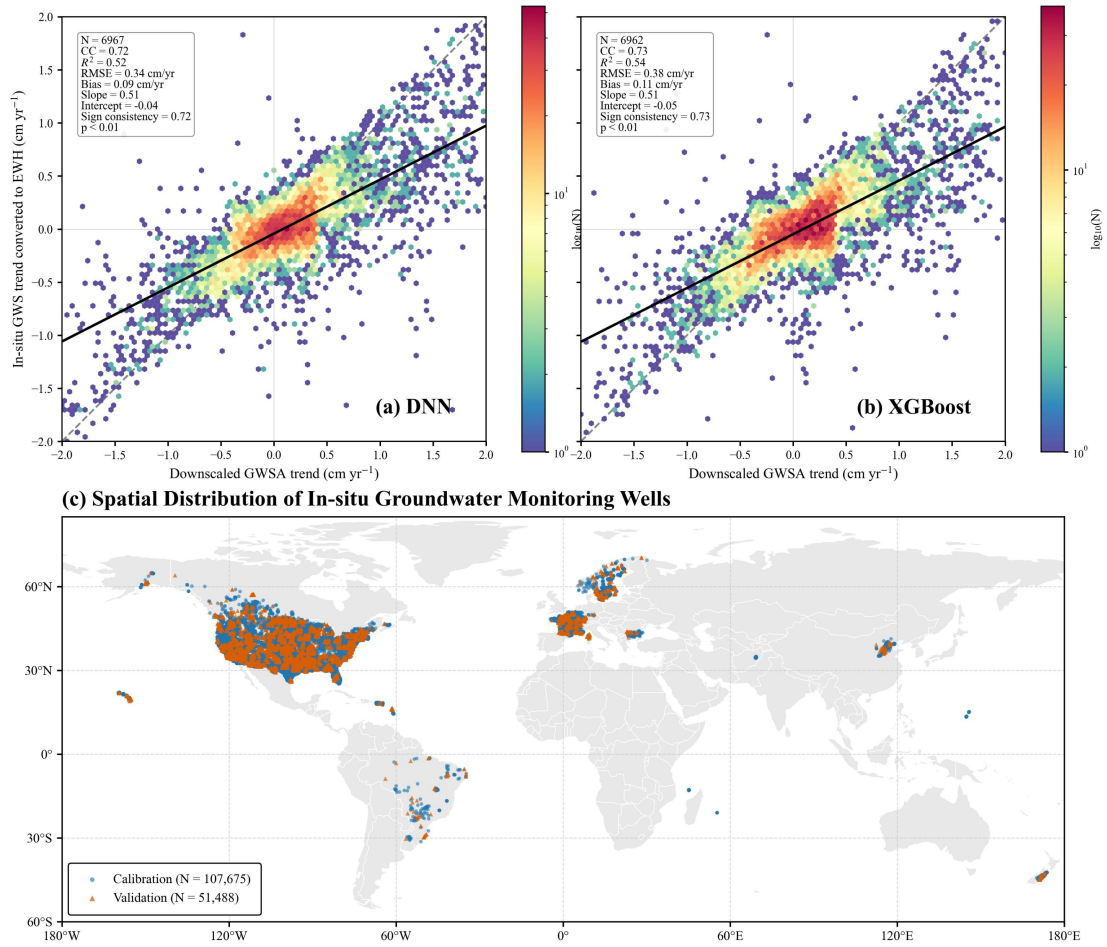


Figure 4. Spatially independent, quality-controlled in-situ evaluation of long-term downscaled groundwater-storage trends and spatial distribution of the monitoring network.

(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.

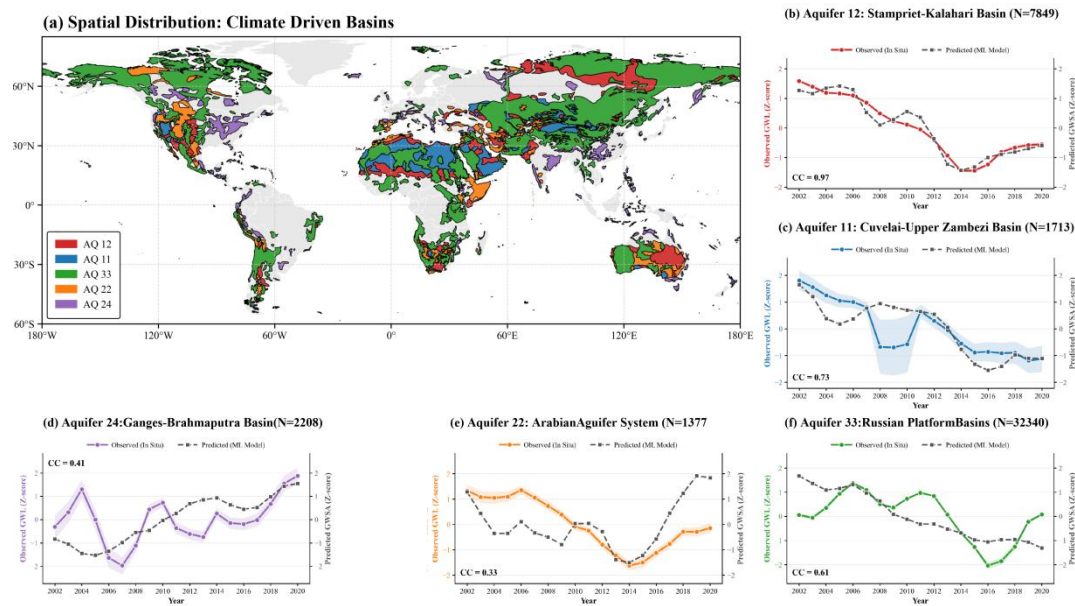


Figure 6. Spatial distribution and interannual variations of GWL and downscaled GWSA for climate-driven/natural-dominated major aquifer basins.

(a) Spatial locations of selected aquifers with specific IDs (AQ 12: Stampriet-Kalahari Basin, AQ 11: Cuvetlai-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 CC denotes the Pearson correlation coefficient between the standardized observations and predictions.

(a) Spatial Distribution: Anthropogenic Basins

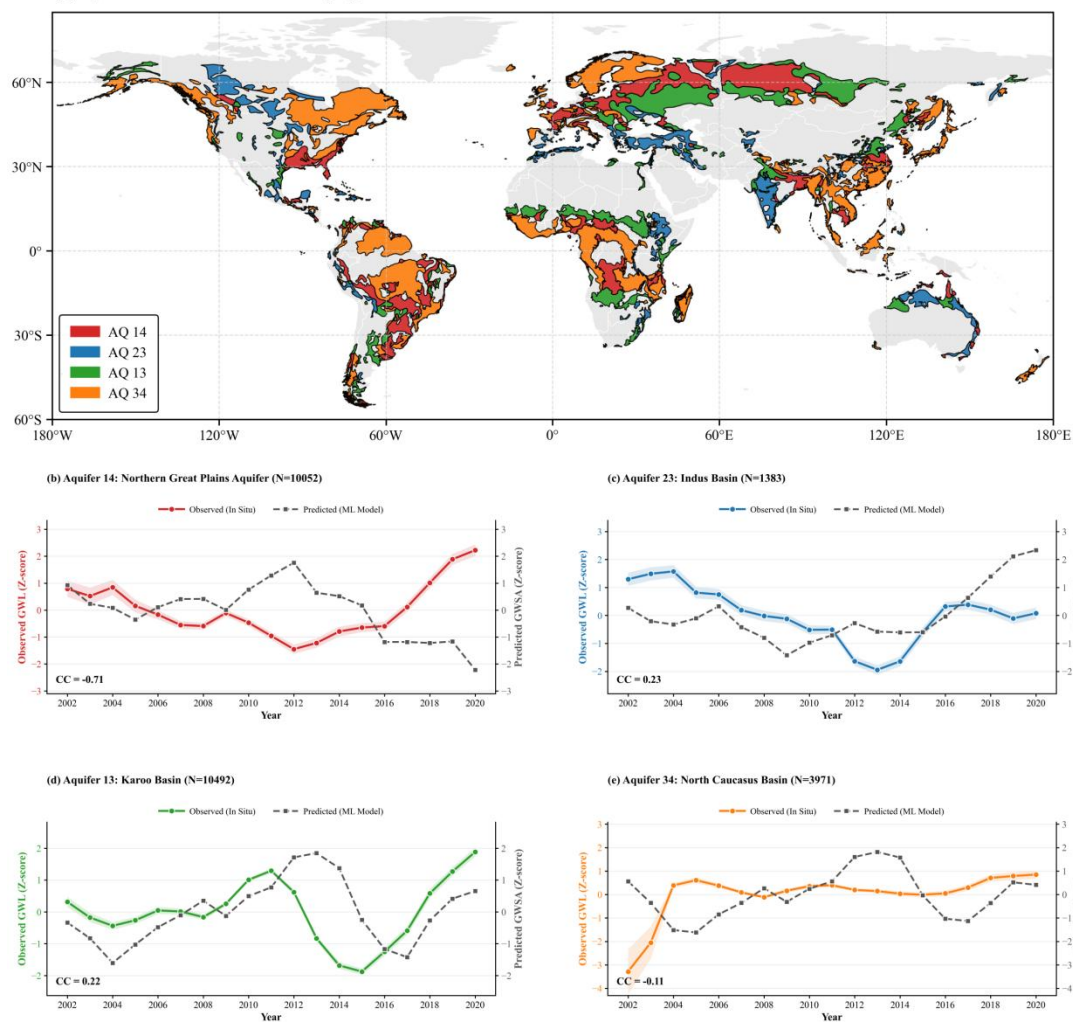


Figure 7. Spatial distribution and interannual variations of groundwater dynamics for anthropogenic-dominated major aquifer basins.

(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 CC indicates the corresponding Pearson correlation coefficient for each basin.

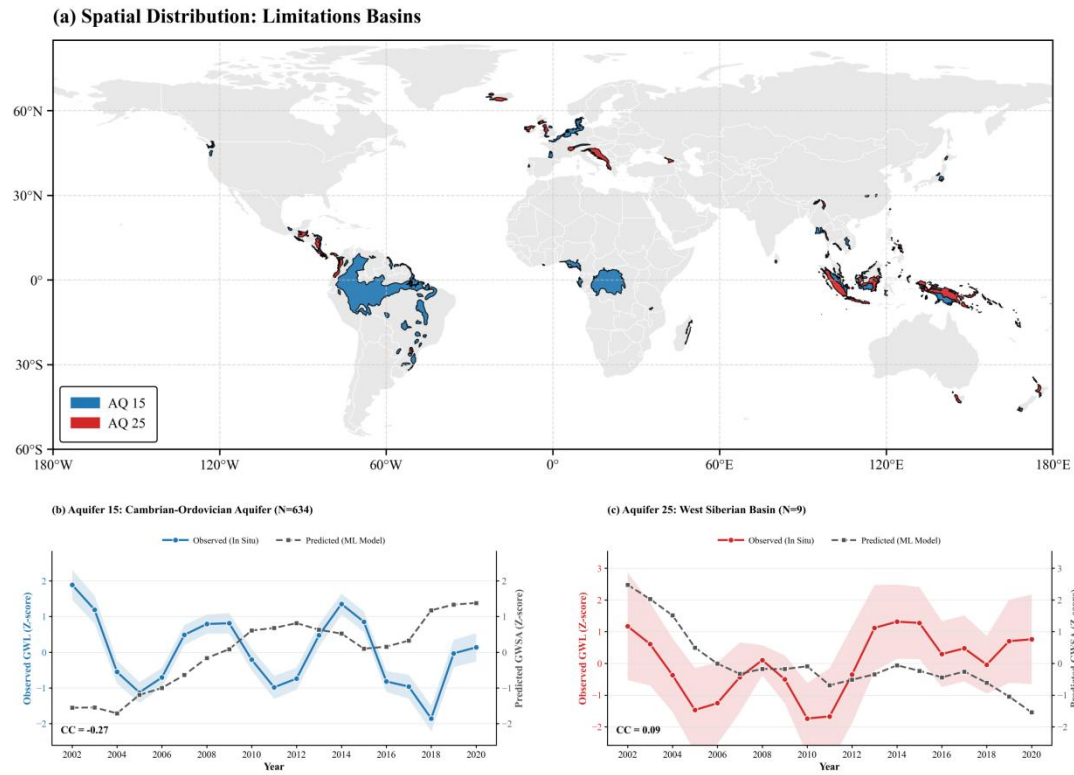


Figure 8. Spatial distribution and interannual variations of groundwater anomalies for aquifers under physical limitations or specific climatic-hydrogeological constraints.

(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 CC denotes the Pearson correlation coefficient between the in situ records and model estimates.

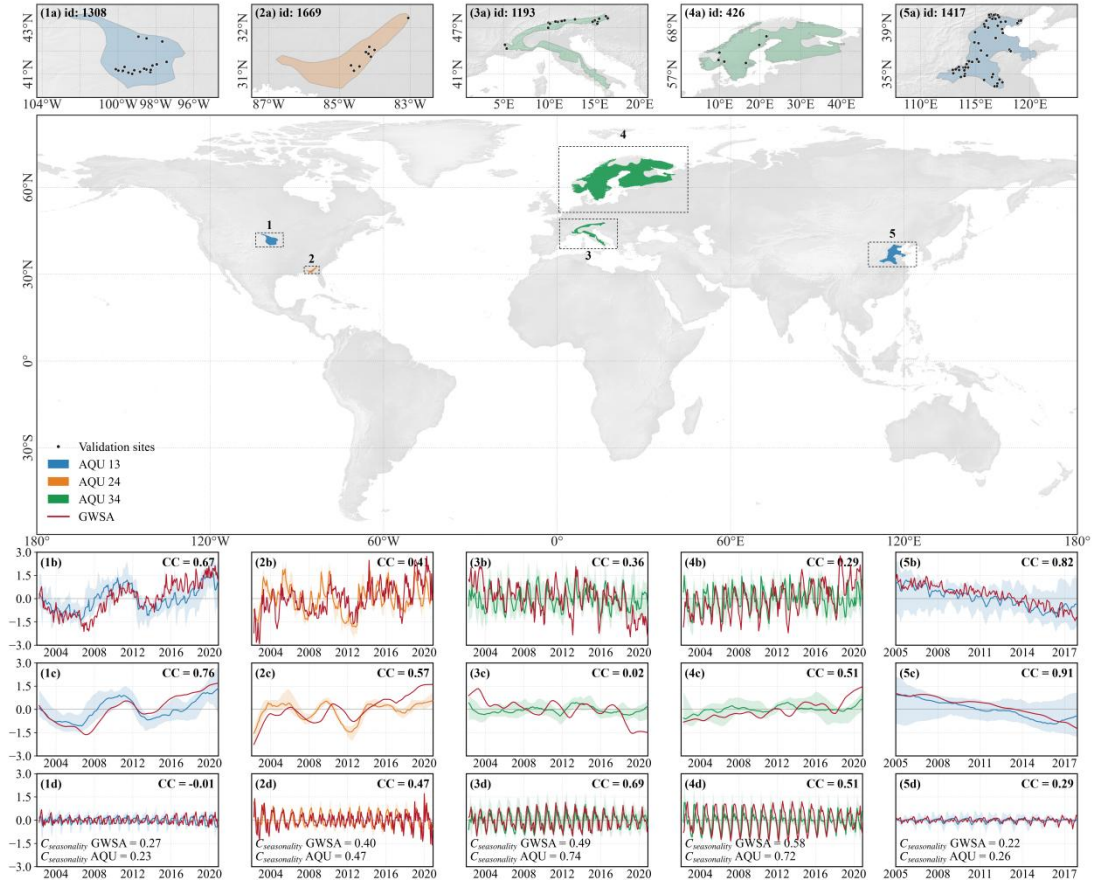


Figure 9. Spatial distributions and monthly STL decompositions of in situ groundwater-level observations and reconstructed GWSA across five representative validation regions.

(a) shows the extents of the five validation aquifers and the corresponding in situ monitoring sites, including the High Plains Aquifer (AQU 13; ID 1308), the Floridan Aquifer (AQU 24; ID 1669), the Apennine Peninsula aquifer systems (AQU 34; ID 1193), the Scandinavian Peninsula aquifer systems (AQU 34; ID 426), and the North China Plain (AQU 13; ID 1417). (b), (c), and (d) show the standardized monthly series, trend components, and seasonal components, respectively. Shaded areas indicate the 10th–90th percentile ranges, CC denotes the Pearson correlation coefficient between the in situ observations and the reconstructed results, and $C_{seasonality}$ denotes the seasonal contribution.

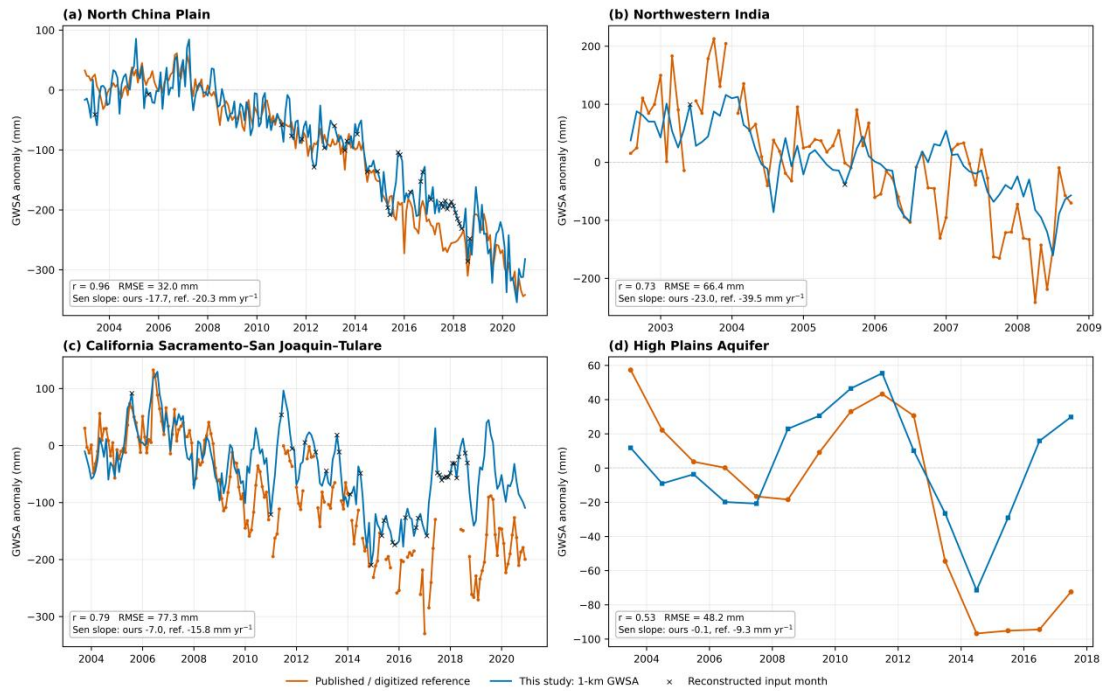


Figure 10. Regional evaluation of the 1 km GWSA reconstruction against published groundwater-storage estimates.

Monthly comparisons are shown for (a) NCP (Zhang et al., 2023), (b) northwestern India (Rodell et al., 2009), and (c) California (Liu et al., 2022); (d) shows the annual HPA well-based composite digitized from Rateb et al. (2020). Blue denotes this study and orange the reference; small orange markers retain isolated available samples beside gaps. Black crosses identify reconstructed input months. Each pair is centered over a common baseline; gaps in published monthly data remain unfilled. Insets report r , RMSE, and paired-series Sen slopes. The annual HPA series cannot validate monthly seasonality, and its full-period sign agreement should not obscure the opposite trends in the separate 2003–2013 benchmark.

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 same 19 predictors, aquifer masks, development-sample size and sampling procedure,

random 70%/30% internal split, and hyperparameter-search protocol were used for all four target configurations to make the source-sensitivity experiments methodologically comparable.

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. 5 and Sect. 3.1.3, 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.

Author's changes in manuscript:

[p. 7, lines 133-138; p. 9, lines 174-179; pp. 10-11, lines 202-212] Sect. 2.1 and 2.1.1–2.1.3: added descriptions of the JPL, H08, and WaterGAP datasets used for the sensitivity experiment.

[p. 32, lines 721-736] New Sect. 3.1.3: reports the controlled sensitivity experiment and feature-importance analysis (Fig. 5).

[p. 33, lines 738-745] Fig. 5: new figure showing internal-test R^2 across 12 aquifer categories and aggregated feature-importance profiles for the four target products.

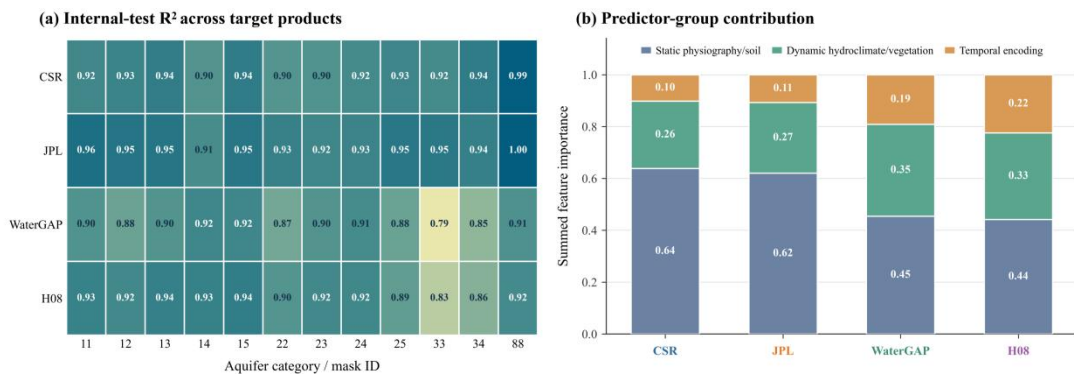


Figure 5. Controlled sensitivity of the aquifer-stratified XGBoost downscaling workflow to alternative target products.

(a) Internal-test R^2 for CSR, JPL, WaterGAP, and H08 across 12 aquifer categories and masks. (b) Aggregated feature importance for static physiography/soil, dynamic hydroclimate/vegetation, and

temporal encoding. All experiments used the same predictor database and random 70%/30% internal split. Because target amplitudes differ, cross-product comparison emphasizes dimensionless R^2 . These statistics characterize within-distribution reconstruction and source sensitivity; they are not temporally blocked forecasting 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 S_y , the authors should evaluate the sensitivity of their results to the selected S_y 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 S_{yeff} by perturbing the baseline coefficients by $\pm 15\%$ and $\pm 30\%$. For the XGBoost product, the baseline evaluation yielded $R^2 = 0.538$, $CC = 0.734$, $RMSE = 0.378 \text{ cm yr}^{-1}$, and $Bias = 0.106 \text{ cm yr}^{-1}$. Under the perturbed scenarios, R^2 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 S_{yeff} 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. 3.1.2 and Sect. 4.3 of the revised manuscript.

Accordingly, we have revised Sects. 2.5.1 to clarify the S_{yeff} conversion and sensitivity framework, Sect. 3.1.2 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.

Author's changes in manuscript:

[pp. 21-23, lines 487-532; p. 31, lines 712-715] Replaced the old regional uniform S_y table (former Table 2) with the scale-dependent S_{yeff} framework.

[pp. 22-23, lines 500-532] Sect. 2.5.1: new subsection describing the S_{yeff} calibration, spatial-block design, and conversion procedure.

[p. 31, lines 701-711] Sect. 3.1.2: reports the quantitative sensitivity results.

[p. 31, lines 712-715] Table 3: complete perturbation experiment for DNN and XGBoost.

[pp. 56-57, lines 1252-1265] Sect. 4.3: discussion of S_{yeff} uncertainty and its implications.

[pp. 30-31, lines 688-700] Fig. 4: spatial-block allocation and held-out validation design shown in panel (c).

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
DNN	-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
	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
XGBoost	-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
	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

The spatial-block allocation and held-out validation design are shown in revised Fig. 4, reproduced above in our response to Comment 2.

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.

Author's changes in manuscript:

[p. 9, lines 159-179; pp. 13-16, lines 286-351] Sect. 2.1.1 and 2.2: clarified that the full GRACE/GRACE-FO record through November 2025 is used only for retrospective SSA reconstruction and does not extend the temporal coverage of the 1 km product.

[pp. 26-27, lines 608-632] Sect. 3.1.1: reports the global and land-only error statistics for both SSA phases.

[pp. 27-28, lines 634-637] Fig. 2: new maps of SSA reconstruction error for the short-gap (a) and prolonged-gap (b) configurations.

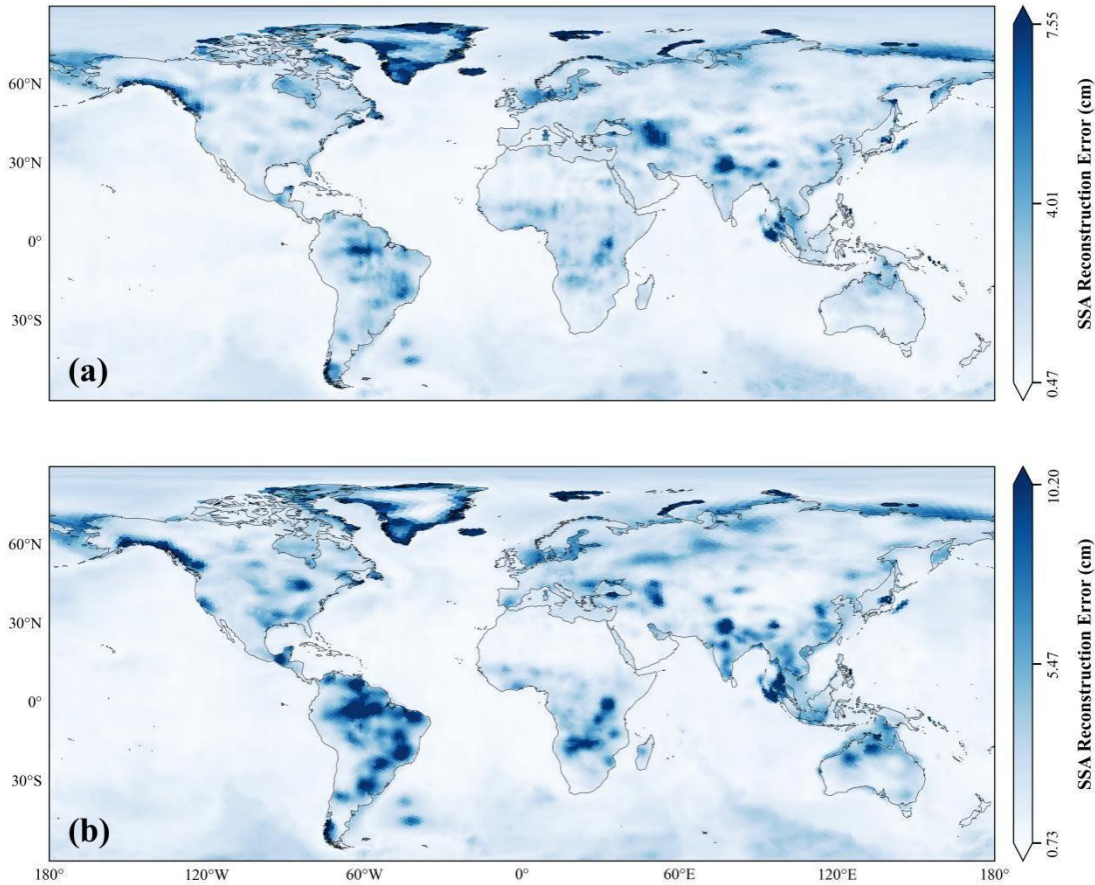


Figure 2. Evaluation of the reconstruction accuracy of Singular Spectrum Analysis (SSA).

(a) Reconstruction error in the SSA-filling-a phase (cm), quantified as the standard deviation of residuals between observed and reconstructed values. (b) Reconstruction error in the SSA-filling-b phase (cm), quantified as the minimum aggregated root mean square error (RMSE) from annual masking experiments.

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) , 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. 2.4.2. 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.

Author's changes in manuscript:

[pp. 1-2, lines 16-41] Abstract: abbreviation definitions reviewed and harmonised.

[pp. 7-58, lines 124-1314] Sects. 2.1–2.7, 3.1–3.6, 4, and all captions: repeated full definitions after first introduction removed; abbreviations used consistently.

[p. 12, line 252] Removed unnecessary one-time abbreviations (e.g., GEE is no longer abbreviated if used only once).

[representative captions: pp. 27-28, lines 634-637; p. 53, lines 1149-1160] Table and figure captions: statistical metrics, aquifer identifiers, units, and model terminology standardised.

Point-by-point response to Referee 2 Comments

Dear Editor and Reviewer 2,

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 revised the manuscript to clarify the component-separation limitations associated with dynamic surface-water storage, correct and document the monthly GLDAS processing procedure, strengthen the spatially independent in-situ validation and effective-specific-yield uncertainty analysis, and clarify the scope of the random 70%/30% model-development split. We have also replaced the previous binned validation with an unbinned common-grid evaluation and revised the interpretation of kilometre-scale spatial patterns and GHM residuals to avoid unsupported causal attribution.

Refer comments are shown in black/bold; our response follows in **blue normal text**; and specific revisions to the manuscript are summarized as *italicized red normal text*.

Major Comments

1. For Eq. (15), the manuscript states that Terrestrial Water Storage (TWS) represents the aggregate water reserves both above and beneath the land surface, encompassing surface water, soil moisture, groundwater, snow cover, ice, and vegetation water storage. However, Eq. (15) does not explicitly account for surface-water or ice storage anomalies. Although glacier/ice storage may reasonably be neglected in typical non-glaciated land areas, surface-water storage (including rivers, lakes, reservoirs, wetlands, and floodplains) cannot simply be ignored in a global-scale analysis. I am not convinced that applying a permanent-waterbody mask adequately resolves this issue. First, masking removes spatial pixels, whereas subtraction separates mass-storage components. Masking water-body pixels does not eliminate the gravity signal associated with temporal changes in lake or river storage, including potential leakage into adjacent land areas. Second, many important surface-water variations are associated with non-permanent water bodies, including seasonal floodplains, temporary inundation, reservoir fluctuations, and wetland-storage changes. Consequently, if surface-water storage increases but is not explicitly removed from TWSA, the inferred GWSA may be overestimated; conversely, decreases in surface-water storage could potentially be misinterpreted as groundwater depletion. I therefore suggest that the authors more explicitly discuss and, if possible, quantify the uncertainty introduced by the omission of dynamic surface-water storage. This issue may also be relevant to the interpretation of Fig. 5. In particular, could the dendritic pattern resembling a river network in the Amazon Basin partly result from residual surface-water signals that were not explicitly removed, rather than representing groundwater-storage variability alone? I believe this alternative explanation deserves further examination.

Response:

We thank the reviewer for raising this important issue. We agree with the physical distinction between spatial masking and explicit component subtraction. GRACE-observed terrestrial water storage integrates groundwater together with soil moisture, snow, canopy water, surface water, and other land-water storage components. Therefore, time-varying surface-water storage can propagate into a residual GRACE-derived GWSA estimate, particularly in regions characterized by large river, lake, reservoir, wetland, or floodplain-storage variations.

The formulation adopted in Eq. (15) follows the commonly used residual GRACE–land-surface-model framework, in which modeled soil-water storage, snow-water equivalent, and canopy-water-storage

anomalies are subtracted from GRACE TWSA to estimate groundwater-storage anomalies. We therefore clarify that the absence of an explicit dynamic surface-water term represents a known component-separation limitation of this residual framework rather than an assumption that surface-water variability is physically absent.

The permanent-waterbody and glacier masks used in the present study were intended as error-reduction measures rather than substitutes for dynamic component subtraction. They reduce contamination over the most evident permanent-water and ice domains, but they cannot remove time-varying storage in rivers, lakes, reservoirs, wetlands, and seasonal floodplains or GRACE leakage into adjacent land pixels. We have revised the manuscript to state this distinction explicitly.

A globally homogeneous monthly dynamic-surface-water correction was not implemented in the present production framework. We therefore do not introduce a numerical global correction that is unsupported by the revised dataset. Implementing such a correction consistently would require a temporally continuous and globally homogeneous surfacewater-storage dataset covering the full 2002–2020 period and representing rivers, lakes, reservoirs, wetlands, and floodplains under compatible anomaly definitions and spatial support. Introducing an incompletely constrained or heterogeneous correction during the present revision could itself create additional structural uncertainty and would require regeneration and revalidation of the complete GWSA dataset.

We agree particularly with the reviewer’s interpretation of the Amazon Basin. The dendritic pattern highlighted in the original figure, now shown in revised Fig. 11, cannot be treated as groundwater-only evidence. We have removed the previous wording suggesting that the model “successfully reconstructs” a groundwater drainage network. The revised manuscript describes these kilometre-scale features as predictor-informed spatial patterns under coarse-scale GRACE constraints and explicitly recognizes residual dynamic surface-water storage and GRACE leakage as plausible contributors in the Amazon and other floodplain-dominated regions.

The uncertainty and future-research sections now identify globally consistent observations or multi-model estimates of river, lake, reservoir, wetland, and floodplain storage as an important future constraint. Such data would allow explicit component subtraction and sensitivity or ensemble analyses in subsequent versions of the dataset.

Author's changes in manuscript:

[pp. 17-18, lines 379-395] Sect. 2.3 (GWSA Inversion and Masking Strategy): added explicit statement that Eq. (15) is a residual framework with a known limitation regarding dynamic surface-water storage, and that masking does not substitute for component subtraction.

[p. 18, lines 396-404] Sect. 2.3.1: clarified that the permanent-water mask reduces but does not eliminate leakage or temporal surface-water variations.

[pp. 43-44, lines 935-968] Sect. 3.3: explicitly added residual dynamic surface-water storage and GRACE leakage as plausible contributors to the Amazon dendritic pattern.

Fig. 11 caption: revised to describe the kilometre-scale patterns as predictor-informed under coarse-scale GRACE constraints and to avoid groundwater-only causal interpretation.

[p. 58, lines 1301-1307] Sect. 4.3 (last paragraph): added discussion of surface-water storage as a key future constraint for reducing component-separation uncertainty.

2. In line 165, the manuscript refers to the annual time-series data of SWS, SWE, and CWS. Is annual a typo, or were annually aggregated data actually used rather than monthly data? If annual values were indeed used together with monthly GRACE TWSA, this would introduce an important temporal-scale inconsistency. Monthly variations in soil moisture, snow water equivalent, and canopy water storage that should have been removed from TWSA would remain embedded in the inferred groundwater component. This could be particularly important in monsoon regions, snow-dominated regions, and other areas with strong seasonal landwater-storage variability. If so, the conclusion that the seasonal contribution exceeds 50%

across 52.3% of the global land area would also require reconsideration. If annual is simply a wording error and monthly GLDAS data were actually used, this should be corrected explicitly. It would also be helpful to provide the exact GLDASNoah variables used, their units, the treatment of the soil-moisture layers, and the temporal aggregation and anomaly-calculation procedures.

Response:

We thank the reviewer for identifying this ambiguity. The word “annual” in the previous manuscript was a wording error. The GLDAS Noah V2.1 components used together with monthly GRACE TWSA were monthly rather than annually aggregated. We have corrected this wording throughout the revised manuscript. Table 1 now identifies GLDAS-2.1 as a monthly dataset covering April 2002 to December 2020, consistent with the temporal resolution used for the GWSA inversion.

For the soil-water-storage term, SWS was calculated as the sum of the four Noah soil-moisture layers: SoilMoi0_10cm_inst (0–10 cm), SoilMoi10_40cm_inst (10–40 cm), SoilMoi40_100cm_inst (40–100 cm), and SoilMoi100_200cm_inst (100–200 cm). Snow water equivalent and canopy water storage were obtained from SWE_inst and CanopInt_inst, respectively. These GLDAS storage variables are provided in kg m⁻²; because 1 kg m⁻² of water is equivalent to 1 mm of water thickness, they were converted to centimetres of equivalent water thickness before component subtraction.

All GLDAS components were processed at the monthly scale. Monthly anomalies were calculated relative to their respective 2004–2009 mean values, consistent with the GRACE TWSA reference period, before applying Eq. (15).

$$\Delta GWS = \Delta TWS - (\Delta SWS + \Delta SWE + \Delta CWS) \quad (15)$$

We have revised Sect. 2.1.2 and the description accompanying Eq. (15) to make the soil-layer treatment, temporal resolution, unit harmonization, and anomaly calculation explicit. Because the actual calculation used monthly nongroundwater storage components, the seasonal-contribution analysis was not based on a mixture of monthly GRACE observations and annually aggregated GLDAS components. Consequently, the reported seasonal results do not require recalculation as a result of this wording correction.

Author's changes in manuscript:

[p. 10, lines 191-201] Sect. 2.1.2: corrected 'annual' to 'monthly' throughout; added exact Noah variable names (SoilMoi0_10cm_inst, SoilMoi10_40cm_inst, SoilMoi40_100cm_inst, SoilMoi100_200cm_inst, SWE_inst, CanopInt_inst), units (kg m⁻²), conversion to cm, soil-layer treatment, and anomaly-calculation procedure.

[pp. 8-9, Table 1 (caption at line 158)] Table 1: temporal resolution of GLDAS-2.1 confirmed as Monthly.

[p. 17, lines 359-369] Sect. 2.3 and Eq. (15): notation harmonised to use ΔSWS consistently for soil-water storage; units and baseline period (2004–2009) stated explicitly.

Minor comments

3. Line 337: Please clarify what is meant by the 'Warp function'.

Response:

We thank the reviewer for requesting clarification of the “Warp function.” We have revised the Methods to specify that this refers to the `gdal.Warp` geospatial processing operation used in Python for raster resampling and grid alignment.

In the present workflow, the 19 high-resolution predictor layers were mean-aggregated from the 1 km grid to the 0.25° modeling grid and aligned spatially with the GWSA target field before model training. The operation is therefore a geospatial preprocessing step used to ensure a common grid and spatial support; it is not an additional predictive or statistical model.

Author's changes in manuscript:

[p. 19, lines 423-428] Sect. 2.4, step (1): clarified that `gdal.Warp` is a Python geospatial resampling operation for mean aggregation and grid alignment.

4. Lines 368-370: The study uses a random 70%/30% split for model training and validation.

Because groundwater and hydroclimatic datasets generally exhibit substantial spatial and temporal autocorrelation, a random split may lead to overly optimistic estimates of model performance if nearby pixels or temporally adjacent observations occur in both the training and validation sets. I recommend supplementing the current validation with more stringent approaches, such as spatially blocked and temporally blocked validation.

Response:

We thank the reviewer for this important methodological comment. We agree that a random 70%/30% split does not remove spatial or temporal autocorrelation and can yield optimistic error estimates when neighboring grid cells or temporally adjacent records occur in both subsets. We have revised the manuscript so that these statistics are no longer described as evidence of spatial or temporal generalization.

Specifically, within each aquifer category, 200,000 grid-month records were randomly sampled from the complete April 2002–December 2020 pool. Seventy percent were used for model fitting and hyperparameter optimization, and 30% were withheld for internal testing. This ensures that the reported test records were not used to fit the development model, but it does not guarantee temporal separation. After architecture and hyperparameters were selected, the production model was refitted using all available 2002–2020 samples, consistent with the objective of reconstructing the complete historical record.

To address spatial dependence in the principal product evaluation, we added a spatially independent in-situ validation. The groundwater-monitoring network was partitioned into 1° blocks, all wells in a block remained in the same subset, S_{eff} was estimated only from calibration blocks, and final statistics were calculated only from held-out 0.25° grids.

This block design provides a spatially independent observation-based assessment of final product reliability; it does not constitute a spatially blocked retraining test of the downscaling model. The random machine-learning split is retained only for internal model development.

We did not implement a chronological 2002–2009 training / 2010–2020 testing experiment. Such a design would test transfer to a later hydroclimatic regime and would therefore address forecasting or temporal extrapolation rather than the stated objective of a retrospective, observation-constrained reconstruction of all monthly conditions during 2002–2020. It would also require a separate temporally segmented model and could not replace the full-period production model without changing

the product definition. Accordingly, we now state explicitly that the dataset is not a post-2020 forecasting product and that the random-split metrics do not quantify future-period predictive skill.

Comparable retrospective GRACE downscaling work has used a 70%/30% grid-based training/testing split rather than a chronological forecast split (Arshad et al., 2024, Water Resources Research, 60, e2023WR035882, <https://doi.org/10.1029/2023WR035882>). We cite this precedent only to explain the reconstruction-oriented protocol, not to claim that random sampling eliminates dependence. Indeed, structured-data validation literature warns that random cross-validation can underestimate predictive error when spatial or temporal dependence is present (Roberts et al., 2017, Ecography, 40, 913–929, <https://doi.org/10.1111/ecog.02881>). The revised Sects. 2.4, 2.5, and 4.3 now make this limitation and scope explicit.

Author's changes in manuscript:

[p. 19, lines 429-435] Sect. 2.4, step (2): revised description of the 70 %/30 % split to clarify its purpose (internal model development and hyperparameter selection) and its limitations regarding spatial/temporal autocorrelation.

[pp. 21-23, lines 487-532] Sect. 2.5: spatially independent block validation design described and used for the principal observation-based assessment.

[p. 57, lines 1272-1278] Sect. 4.3: added explicit statement that the random-split metrics do not quantify future-period predictive skill and that the dataset is not a forecasting product.

[pp. 63-64, lines 1502-1504] References: Roberts et al. (2017) added to the reference list and cited in Sect. 4.3 and included in the reference list. Arshad et al. (2024) is cited here only to provide context for the reconstruction-oriented protocol and is not used as evidence that random sampling removes spatial or temporal dependence. Arshad et al. (2024) is cited only in this response and therefore has no manuscript line reference.

5. Line 425: I agree with the comment from the first referee regarding the specific-yield data. Given that S_y can exhibit pronounced spatial heterogeneity, the use of broad regional averages may obscure important hydrogeological differences. It would be useful to discuss whether higher-resolution S_y information could improve the in-situ validation.

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 scaledependent effective specific yield, $S_{y\text{eff}}$, used as a level-to-storage conversion coefficient for harmonizing groundwater-level trends with equivalent-water-height groundwater-storage trends. $S_{y\text{eff}}$ 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 $S_{y\text{eff}}$ 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 $S_{y\text{eff}}$ by perturbing the baseline coefficients by $\pm 15\%$ and $\pm 30\%$. For the XGBoost product, the baseline evaluation yielded $R^2 = 0.538$, $CC = 0.734$, $RMSE = 0.378 \text{ cm yr}^{-1}$, and $Bias = 0.106 \text{ cm yr}^{-1}$. Under the perturbed scenarios, R^2 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 $S_{y,eff}$ 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. 3.1.2 and Sect. 4.3 of the revised manuscript.

Accordingly, we have revised Sects. 2.5.1 to clarify the $S_{y,eff}$ conversion and sensitivity framework, Sect. 3.1.2 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.

Author's changes in manuscript:

[pp. 21-23, lines 487-532; p. 31, lines 712-715] Replaced the old regional uniform S_y table (former Table 2) with the scale-dependent $S_{y,eff}$ framework.

[pp. 22-23, lines 500-532] Sect. 2.5.1: new subsection describing the $S_{y,eff}$ calibration, spatial-block design, and conversion procedure.

[p. 31, lines 701-711] Sect. 3.1.2: reports the quantitative sensitivity results.

[p. 31, lines 712-715] Table 3: complete perturbation experiment for DNN and XGBoost.

[pp. 56-57, lines 1252-1265] Sect. 4.3: discussion of $S_{y,eff}$ uncertainty and its implications.

[pp. 30-31, lines 688-700] Fig. 4: spatial-block allocation and held-out validation design shown in panel (c).

Table 3

Sensitivity of the global in-situ evaluation to perturbations in $S_{y,eff}$. 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	$S_{y,eff}$ perturbation	CC	R^2	RMSE(cm yr ⁻¹)	Bias(cm yr ⁻¹)	Slope	Trend-sign consistency
DNN	-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
	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
XGBoost	-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
	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

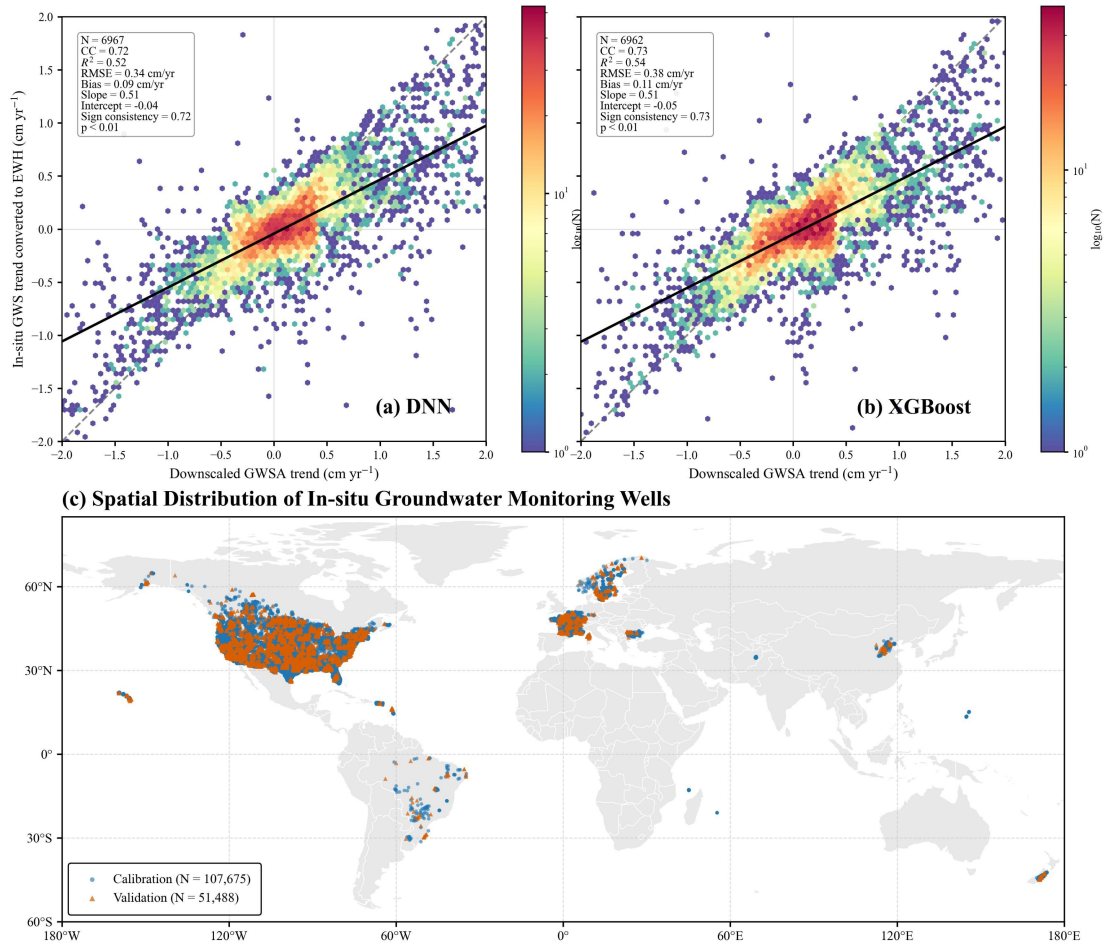


Figure 4. Spatially independent, quality-controlled in-situ evaluation of long-term downscaled groundwater-storage trends and spatial distribution of the monitoring network.

(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.

6. Lines 509-510: Statements such as 'the sharp boundaries caused by localized human pumping' should, in my view, be expressed more cautiously. The additional fine-scale spatial information in the 1-km product necessarily originates largely from the high-resolution predictor variables, because GRACE itself does not resolve groundwater-storage variability at 1-km scales. Therefore, the presence of sharp spatial boundaries in the downscaled product does not by itself demonstrate that these boundaries are directly caused by localized groundwater pumping. Independent high-resolution observations would be needed to support such an interpretation.

Response:

We agree and have removed direct causal wording. The revised manuscript consistently describes the kilometre-scale patterns as predictor-informed spatial heterogeneity under coarse-scale GRACE constraints. Agricultural or landmanagement boundaries are now described as spatial associations, not proof of pumping, and the text explicitly states that attribution requires independent high-resolution observations.

The Discussion further explains that XGBoost and DNN generate different spatial textures because the high-resolution information is model- and predictor-dependent. This model dependence is treated as structural uncertainty rather than as additional satellite resolution.

Author's changes in manuscript:

[pp. 42-43, lines 927-954] Sect. 3.3: revised interpretation of fine-scale patterns; removed wording implying direct causal attribution to pumping.

[pp. 48-50, lines 1047-1113] Sect. 3.5.1 and 3.5.2: depletion and accumulation patterns described as statistical features and spatial associations; causal attribution avoided.

[p. 44, lines 956-968] Fig. 11 caption: dendritic Amazon patterns and agricultural-boundary gradients described as predictor-informed spatial heterogeneity under coarse-scale GRACE constraints.

[p. 51, lines 1115-1121] Fig. 15 caption: sharp class boundaries treated as statistical features requiring independent verification, not as direct evidence of abrupt aquifer boundaries or overextraction.

7. Figure 4: The use of 1,518 monitoring wells is potentially an important strength of this study. However, the validation procedure involves several levels of aggregation, including aggregation into 15 bins. Such binning suppresses local variability and observational noise and may therefore increase the apparent agreement between modeled and observed trends. I recommend additionally reporting the unbinned station- or grid-level validation results, including R², RMSE, bias, and the corresponding scatter plot. A map showing the spatial pattern of the in-situ observations would also be helpful for assessing the geographical representativeness of the validation dataset.

Response:

We agree with the reviewer and have replaced the previous 15-bin presentation with an unbinned common-grid evaluation. In the revised Fig. 4, no additional binning according to modeled or observed trend magnitude is applied. Instead, the DNN and XGBoost products are evaluated directly against all retained held-out 0.25° grid comparisons after the common-grid aggregation and uniform quality-control procedure.

The 0.25° aggregation retained in the revised analysis serves a different purpose from the previous statistical binning. Individual groundwater wells represent point-scale hydraulic conditions, whereas the downscaled GWSA represents spatially integrated storage variations. Wells and model estimates must therefore first be compared on a common spatial support. Within each held-out 0.25° grid, the converted in-situ trends and downscaled GWSA trends are averaged independently. No subsequent grouping of those grid-level comparisons into trend intervals or statistical bins is performed. Thus, “unbinned” means removal of the former 15-bin trend aggregation, not removal of the physically required point-to-grid harmonization.

Revised Fig. 4 reports held-out 0.25° grid comparisons ($N = 6,967$ for DNN; $N = 6,962$ for XGBoost), including CC , R^2 , $RMSE$, $Bias$, regression slope and intercept, trend-sign consistency, and significance. N denotes retained grids, not individual wells. Panel (c) maps the monitoring network and calibration/validation allocation, making geographical representativeness explicit.

As a complementary check, new Sects. 2.5.3 and 3.2.3, Fig. 10, and Table 4 compare regional time series with Zhang et al. (2023), Rodell et al. (2009), Liu et al. (2022), and Rateb et al. (2020). The revised discussion reports attenuated depletion in northwestern India and California and an opposite High Plains trend over the exact 2003–2013 benchmark of Breña-Naranjo et al. (2014). These literature comparisons do not alter the held-out statistics. Shared GRACE inputs, digitization, and the annual High Plains reference are explicitly distinguished from independent monthly well validation.

Author's changes in manuscript:

[pp. 30-31, lines 688-700] Fig. 4: replaced binned validation with unbinned held-out 0.25° grid comparisons; added panel (c) showing spatial distribution of monitoring wells and 1° block partition.

[p. 23, lines 523-532] Sect. 2.5.1: updated to describe the unbinned common-grid aggregation and quality-control procedure.

[pp. 29-30, lines 666-685] Sect. 3.1.2: reports unbinned statistics (CC, R^2 , RMSE, Bias, slope, intercept, trend-sign consistency) for DNN and XGBoost.

[p. 24, lines 555-569; pp. 40-42, lines 884-926] Sect. 2.5.3 and 3.2.3: new regional literature benchmarks (Fig. 10, Table 4) as a complementary, non-independent assessment of regional variability and trend magnitude.

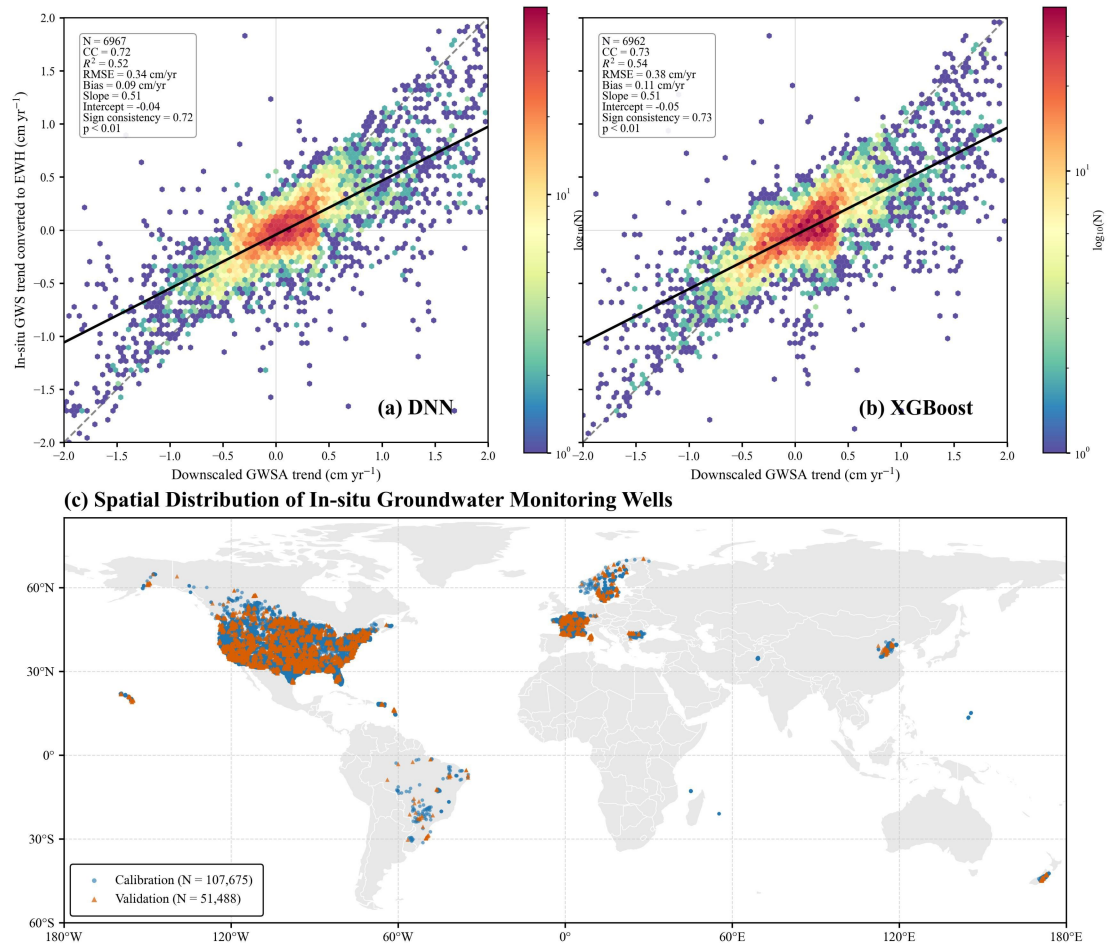


Figure 4. Spatially independent, quality-controlled in-situ evaluation of long-term downscaled groundwater-storage trends and spatial distribution of the monitoring network.

(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.

8. Section 3.5: The interpretation of the residuals as anthropogenic pumping signals should be more conservative. The inferred anthropogenic component is obtained by subtracting GHM-simulated background signals from the reconstructed GWSA. As the manuscript itself acknowledges, these residuals inevitably include structural errors from the GHMs, in addition to uncertainties in the GRACE-derived GWSA and the downscaling procedure. I therefore suggest referring to these residuals as signals consistent with anthropogenic groundwater depletion, or as depletion signals not explained by the GHM-simulated background, rather than treating them as direct estimates of groundwater pumping.

Response:

We agree and adopted this more conservative interpretation throughout. The section has been retitled “Residual Signals Not Explained by the GHM-Simulated Background.” Residuals are now described as diagnostic departures that may contain human influence, GHM structural and forcing errors, GRACE component-separation uncertainty, and downscaling uncertainty. They are not direct estimates of pumping or isolated anthropogenic extraction.

The revised text uses spatial association with irrigation or managed land only for hypothesis generation. It also states that interpolating a 0.5 degree GHM signal to 1 km does not create independent 1 km observations and that causal attribution requires follow-up high-resolution data.

Author's changes in manuscript:

[pp. 51-52, lines 1122-1147] Sect. 3.6: retitled 'Residual Signals Not Explained by the GHM-Simulated Background'; text revised to avoid treating residuals as direct pumping estimates.

[p. 53, lines 1149-1160] Fig. 16 caption: updated to state that residuals may incorporate GHM structural uncertainty, GRACE component-separation uncertainty, and downscaling uncertainty; 1 km detail is predictor-informed under coarse-scale GRACE constraints.

[p. 54, lines 1177-1186; pp. 57-58, lines 1288-1300] Sects. 4.1 and 4.3: residual interpretation revised to emphasise diagnostic/hypothesis-generating purpose rather than causal attribution.

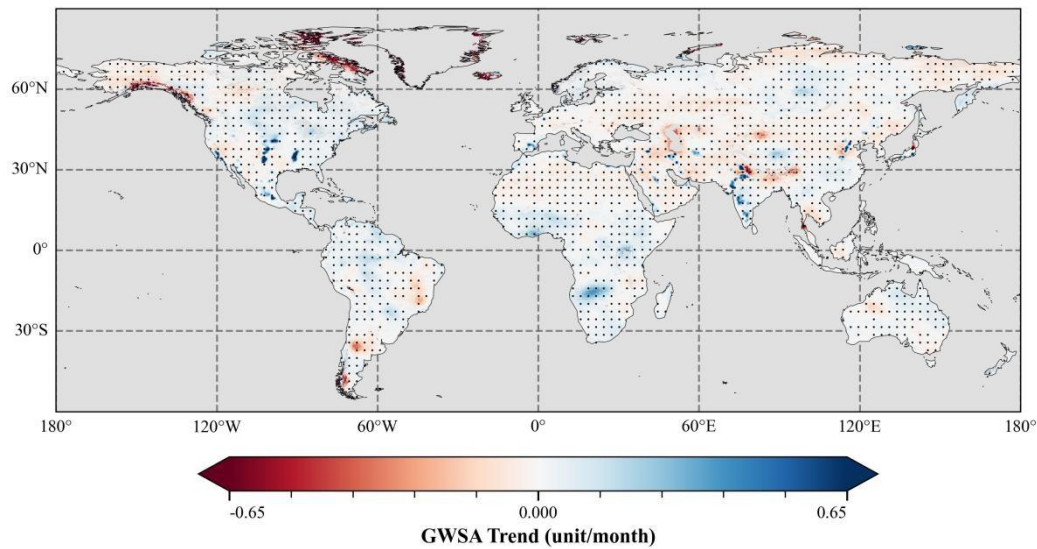


Figure 16. Residual signals not explained by the GHM-simulated background ($p < 0.01$).

The analysis bilinearly interpolates GHM-simulated background to 1 km and subtracts them from the 1 km downscaled GWSA reconstruction sequence. The residual sequences were first de-seasonalized using STL, after which long-term trends were estimated using the Theil–Sen method and their significance was assessed using the Mann–Kendall test. Red regions indicate negative residual depletion signals not explained by the GHM-simulated background; blue areas indicate positive residuals, which may reflect recharge, model bias, or other unresolved processes. Dense black dots denote residual trends passing the MK significance test ($p < 0.01$). Residuals are not direct estimates of pumping or isolated anthropogenic extraction: they may incorporate GHM structural uncertainty, uncertainty in GRACE-derived GWSA separation, and downscaling uncertainty. The 1 km detail is predictor-informed under coarse-scale GRACE constraints and requires independent high-resolution observations for causal attribution.

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