

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.

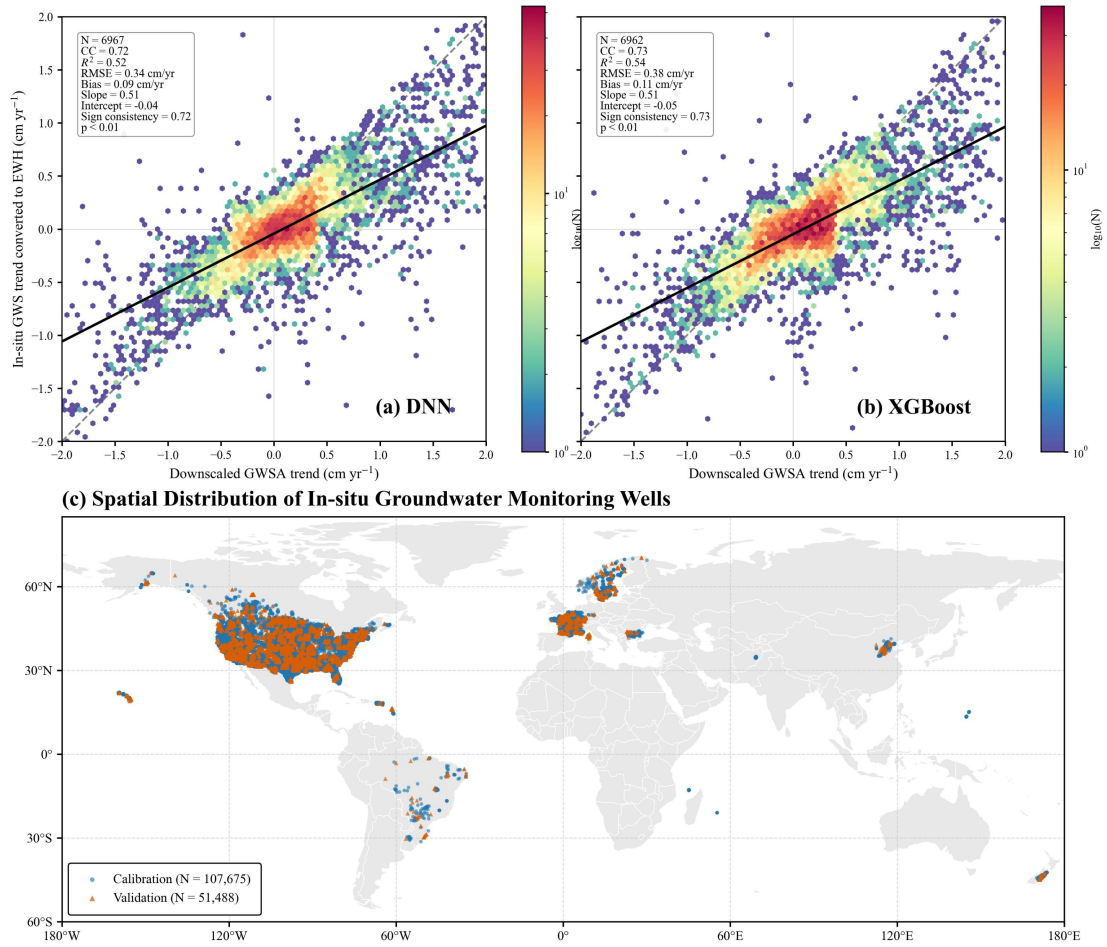


Figure 4. Spatially independent, quality-controlled in-situ evaluation of long-term downscaled groundwater-storage trends and spatial distribution of the monitoring network.

(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.

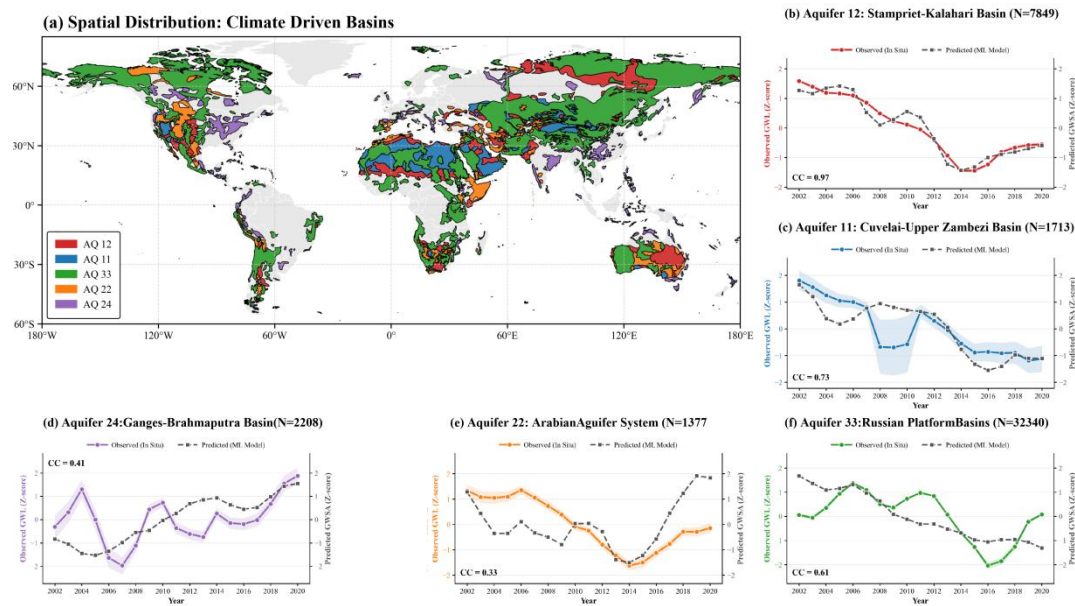


Figure 6. Spatial distribution and interannual variations of GWL and downscaled GWSA for climate-driven/natural-dominated major aquifer basins.

(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

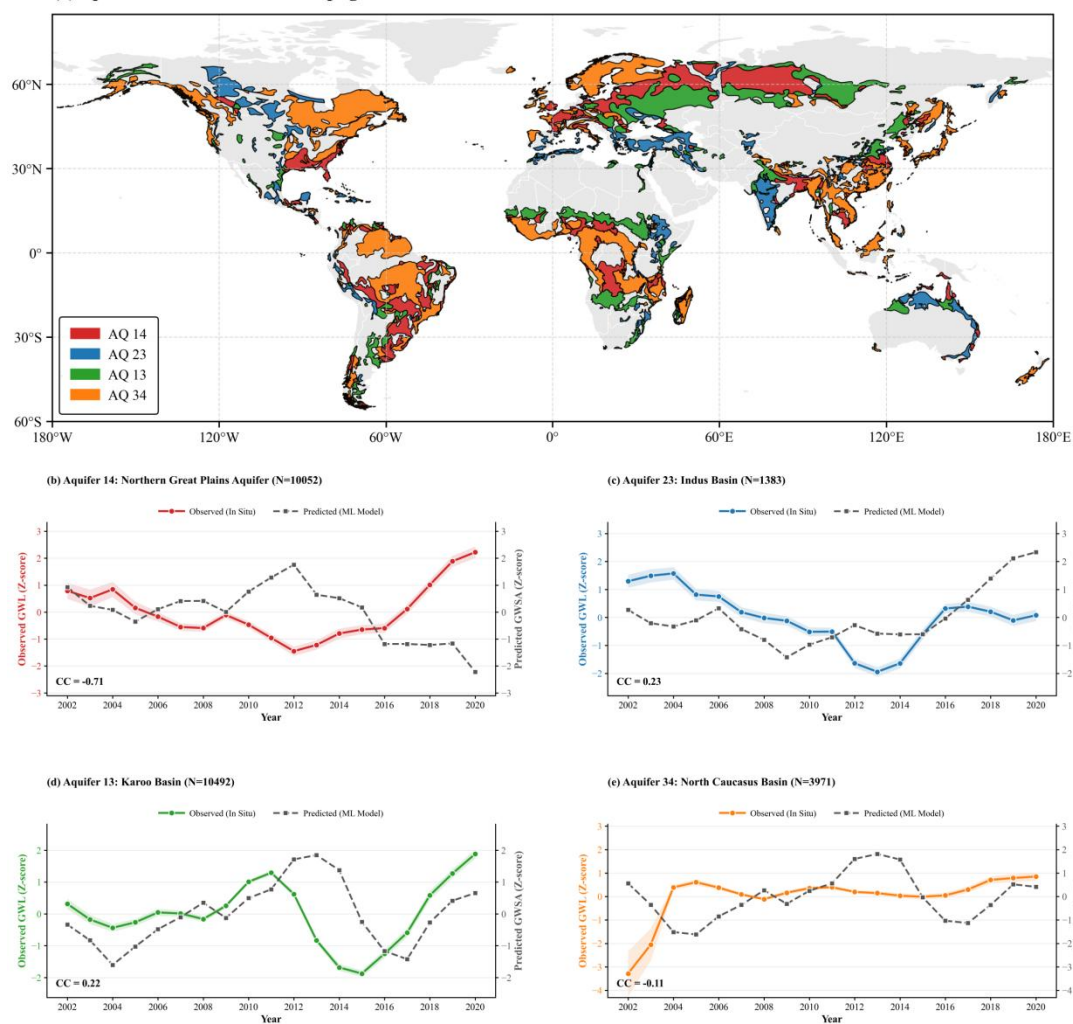


Figure 7. Spatial distribution and interannual variations of groundwater dynamics for anthropogenic-dominated major aquifer basins.

(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.

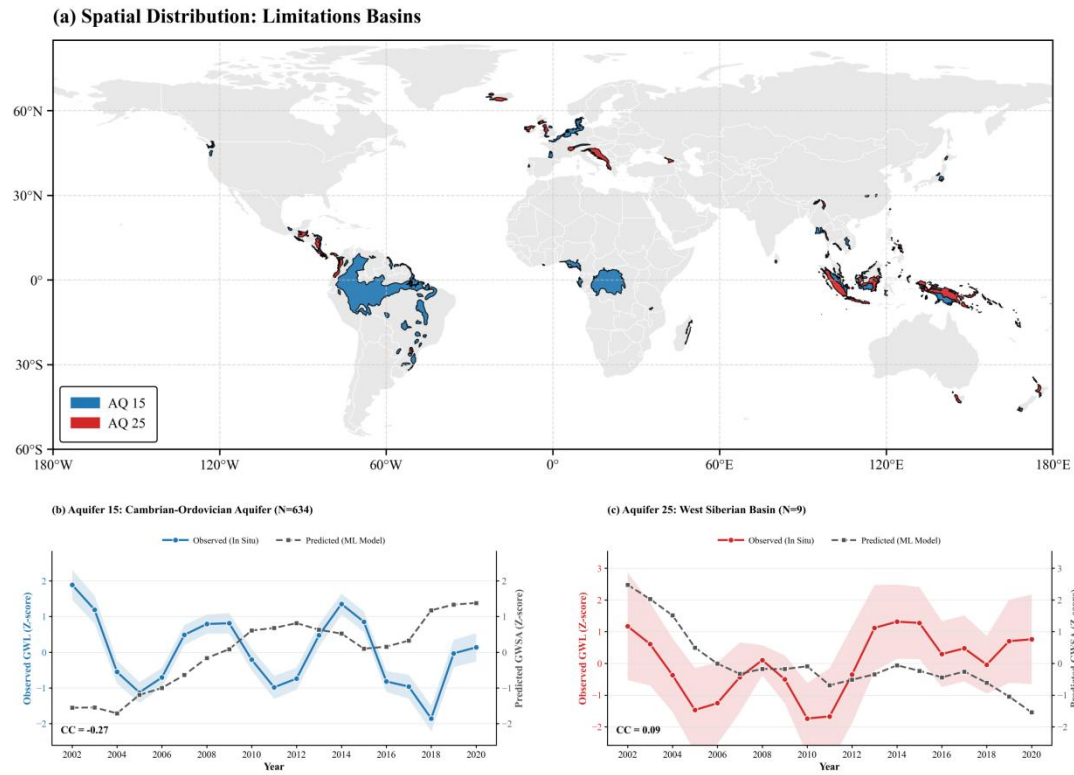


Figure 8. Spatial distribution and interannual variations of groundwater anomalies for aquifers under physical limitations or specific climatic-hydrogeological constraints.

(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.

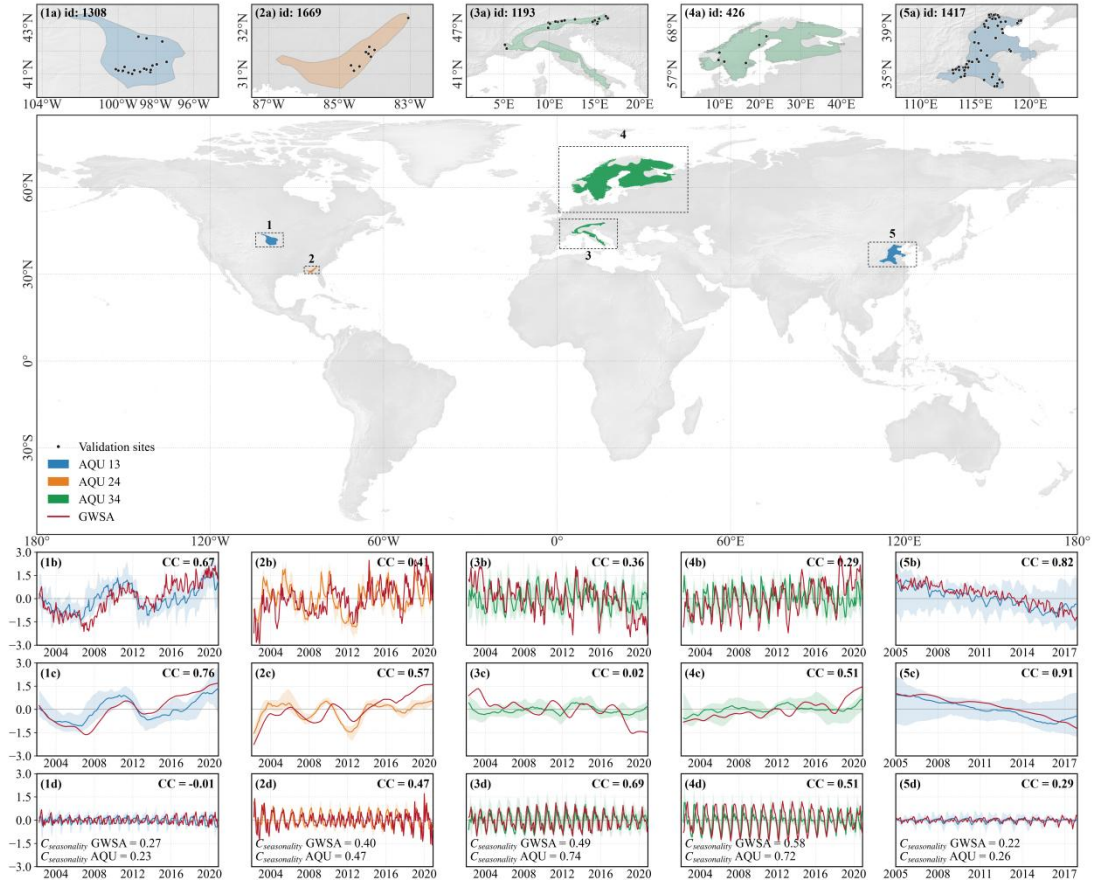


Figure 9. Spatial distributions and monthly STL decompositions of in situ groundwater-level observations and reconstructed GWSA across five representative validation regions.

(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 $C_{\text{seasonality}}$ denotes the seasonal contribution.

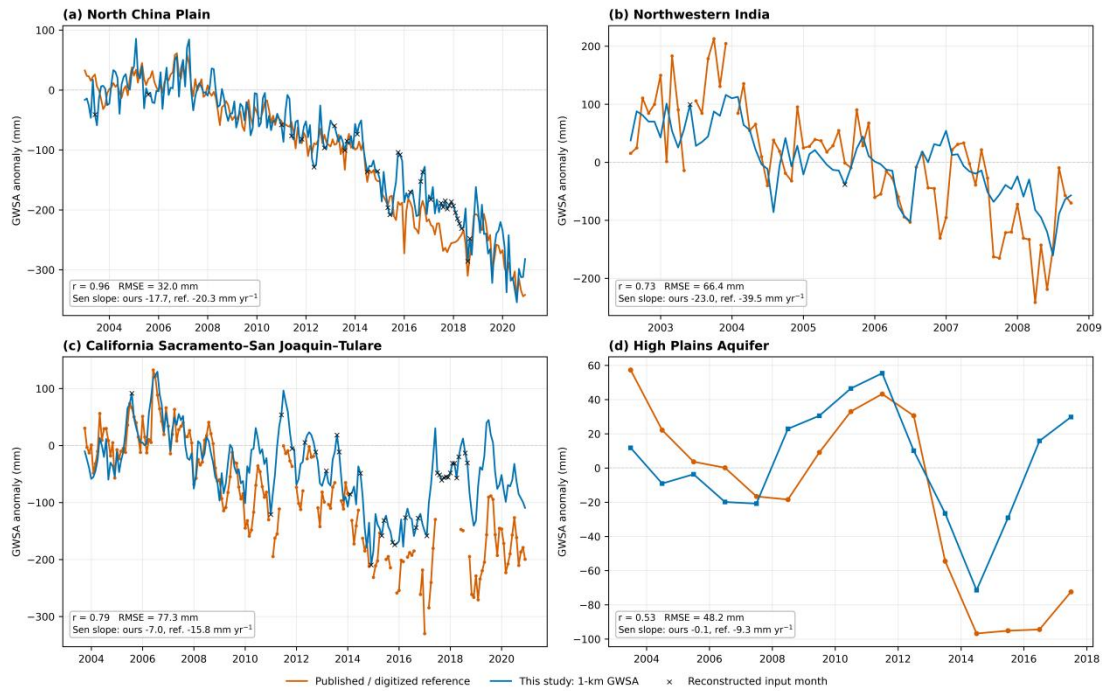


Figure 10. Regional evaluation of the 1 km GWSA reconstruction against published groundwater-storage estimates.

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.

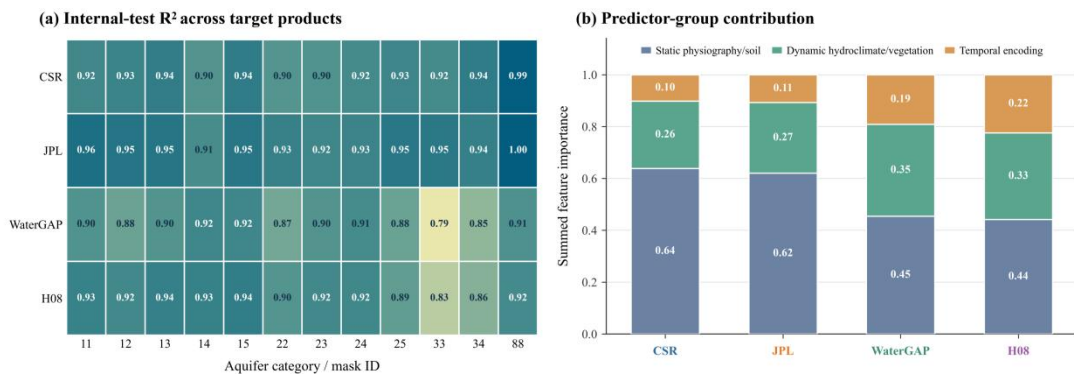


Figure 5. Controlled sensitivity of the aquifer-stratified XGBoost downscaling workflow to alternative target products.

(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.

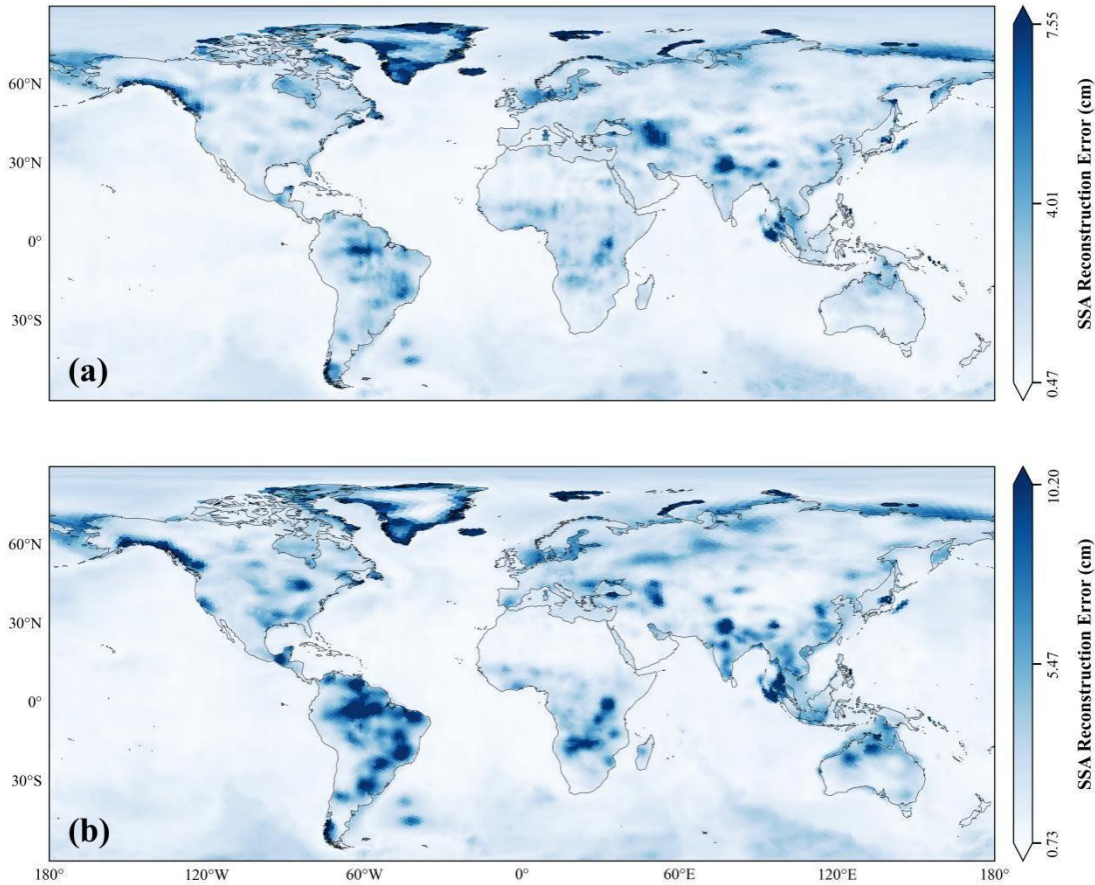


Figure 2. Evaluation of the reconstruction accuracy of Singular Spectrum Analysis (SSA).

(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

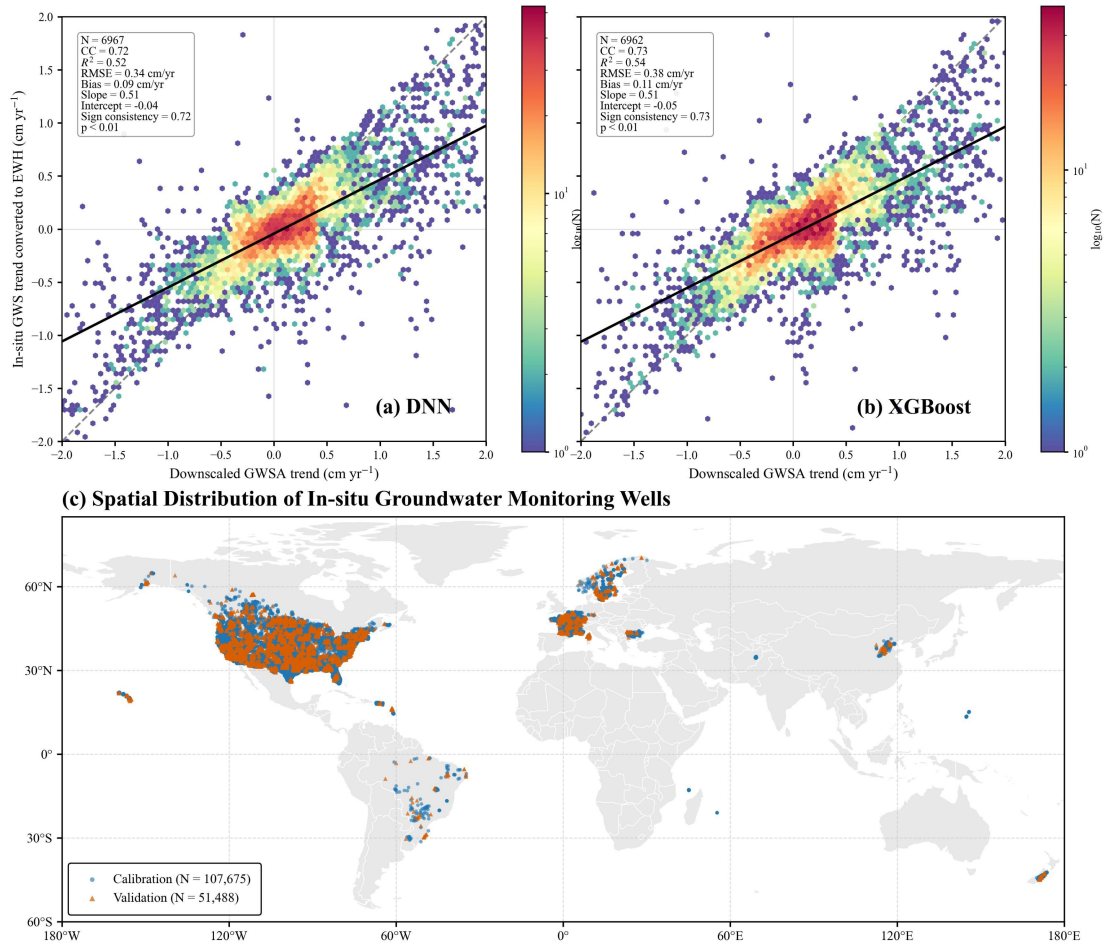


Figure 4. Spatially independent, quality-controlled in-situ evaluation of long-term downscaled groundwater-storage trends and spatial distribution of the monitoring network.

(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 , $RMSE$, $Bias$, $slope$, $intercept$, $trend\text{-}sign\text{ consistency}$, 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.

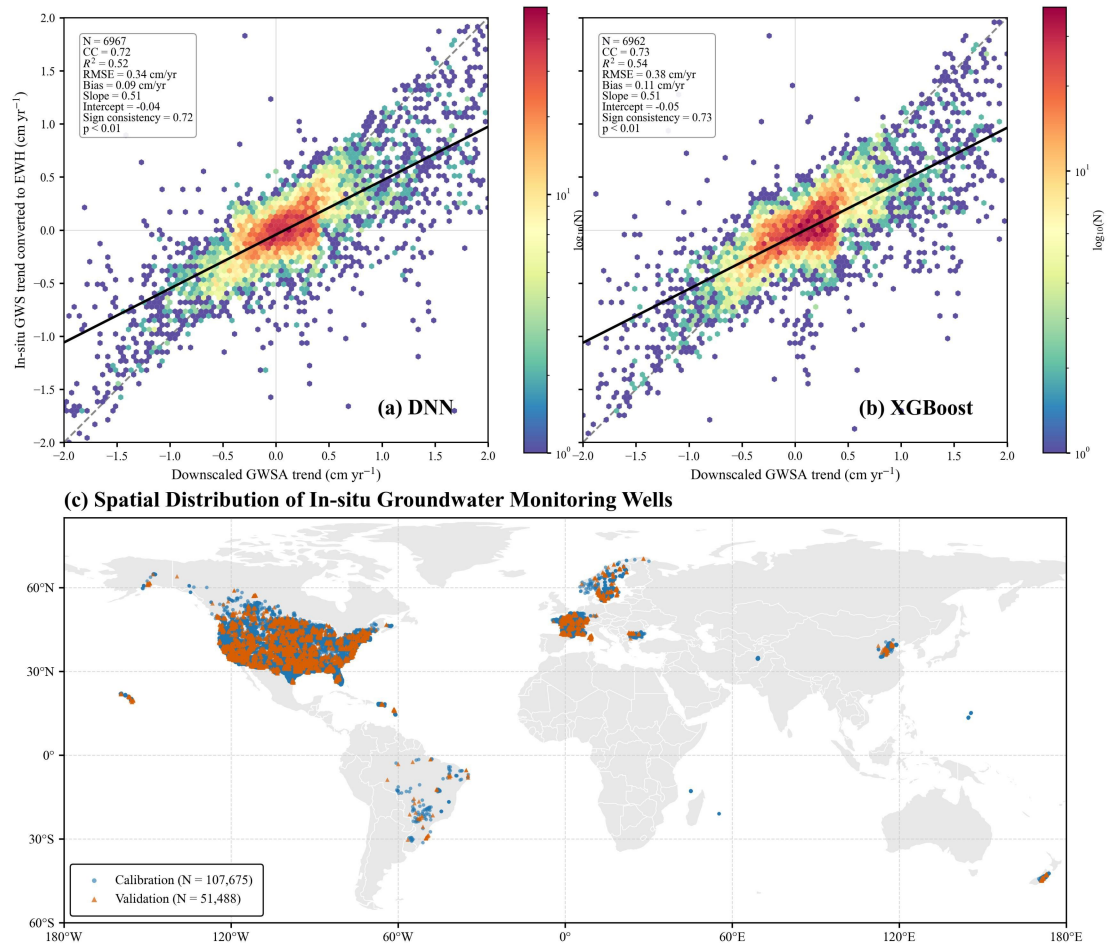


Figure 4. Spatially independent, quality-controlled in-situ evaluation of long-term downscaled groundwater-storage trends and spatial distribution of the monitoring network.

(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.

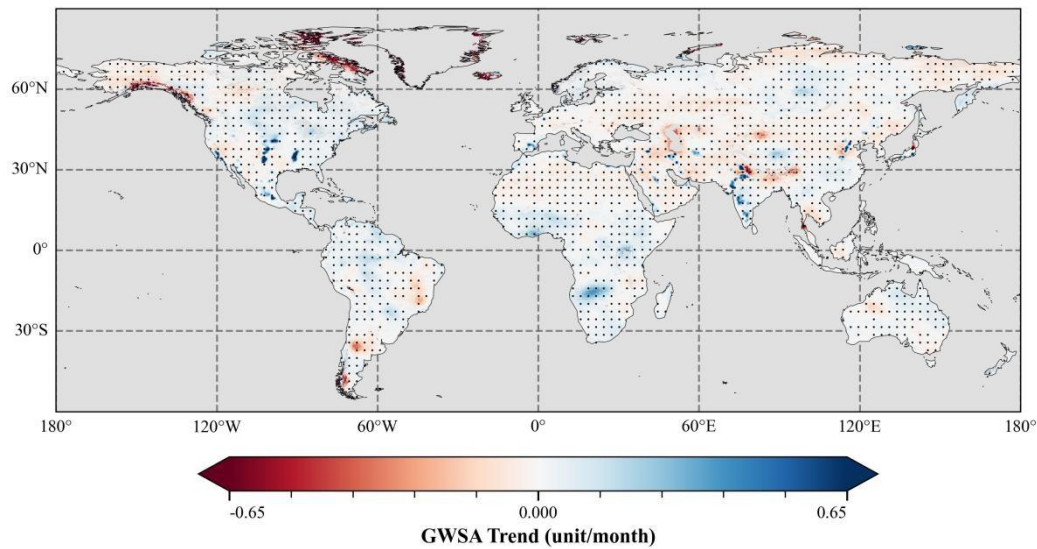


Figure 16. Residual signals not explained by the GHM-simulated background ($p < 0.01$).

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

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

Chen Zhang^{1,3,5,★}, Yifei Fan^{2,4,5,★}, Wenxia Han^{2,4,5,*}, Yingchun Ge^{1,3,5}, Jiabei Zhu^{1,3,5}, Tao Che^{1,3,5,*}

1 State Key Laboratory of Cryospheric Science and Frozen Soil Engineering, Northwest Institute of Eco-Environment and Resources, Chinese Academy of Sciences, Lanzhou 730000, China

2 Key Laboratory of Geology and Environment of Salt Lake, Qinghai Institute of Salt Lakes, Chinese Academy of Sciences, Xining, China

3 Heihe Remote Sensing Experimental Research Station, Key Laboratory of Remote Sensing of Gansu Province, Northwest Institute of Eco-Environment and Resources, Chinese Academy of Sciences, Lanzhou 730000, China

4 Key Laboratory of Green and High-End Utilization of Salt Lake Resources and Qinghai Provincial

5 University of Chinese Academy of Sciences, Beijing 100049, China

★ These authors contributed equally to this work.

Correspondence to: Wenxia Han (wenxia_han@163.com) and Tao Che (chetao@lzb.ac.cn)

Abstract. ~~The global: Groundwater~~ depletion ~~of groundwater~~ poses a ~~significant~~ challenge to water security. ~~However, but~~ the coarse spatial resolution of Gravity Recovery and Climate Experiment (GRACE ~~satellite~~) observations ~~obscures fine-scale hydrological dynamics and~~ limits the ~~separation of~~ ~~localized anthropogenic extraction signals from large-scale climatic forcing. To address this~~ ~~limitation~~ characterization of groundwater variability at regional and local scales. Here, we present a ~~high-resolution (1 km), continuous~~ monthly ~~dataset of~~ global groundwater storage ~~anomalies~~ anomaly (GWSA) dataset covering ~~the period from April~~ 2002 ~~to December~~ 2020. The dataset ~~is was~~ generated using ~~a production framework that integrates~~ Singular Spectrum Analysis (SSA) for temporal gap filling ~~and an aquifer stratified machine learning approach driven by~~ 19 high-resolution hydroclimatic and geophysical predictors. ~~To ensure robustness, and spatial consistency, multiple predictive approaches~~ aquifer-stratified machine-learning downscaling. XGBoost and deep neural networks (DNNs) were evaluated as ~~part of a quality control procedure, and the most stable model~~ candidate models, and XGBoost was selected for final dataset production based on ~~multi-scale validation. Cross-scale evaluation~~ shows that the downscaled dataset preserves the large-scale spatial patterns of the internal, cross-scale, and in-situ evaluations. When aggregated to the original GRACE observations scale, the 1 km product

agreed closely with ~~high agreement the original GWSA~~ ($R^2 = 0.972$, $RMSE = 2.10$ cm). Independent validation using 1,518 in situ evaluation used a global archive of 152,069 groundwater-monitoring wells, combined with a geographically stratified and a 1° spatial-block calibration-validation design. Effective specific yield matrix for dimensional conversion, further demonstrates the ability of the dataset (S_{yeff}) was estimated from calibration blocks and transferred to capture long-term groundwater variability across diverse hydrogeological conditions (held-out blocks before comparison at a common 0.25° grid. For XGBoost, the held-out evaluation yielded $R^2 = 0.44$, $p < 0.01$). The resulting 1 km dataset provides enhanced spatial detail and enables the identification of sharp nonlinear boundaries associated with intensive human pumping, as well as spatial polarization patterns in groundwater storage changes. This dataset offers a reliable, $RMSE = 0.378$ cm yr^{-1} , and $Bias = 0.106$ cm yr^{-1} . The association between modeled and observation-constrained resource-derived trends remained similar under $\pm 30\%$ perturbations in S_{yeff} , although amplitude-related metrics were more sensitive to the conversion coefficient. Standardized time-series comparisons and monthly evaluations of representative aquifers showed better agreement for long-term variability than for water resource assessment, hydrological modeling, and studies of coupled climate and human influences seasonal variability. Comparisons with published regional estimates showed close agreement in the North China Plain, weaker depletion magnitudes in northwestern India and California, and a trend-sign mismatch in the High Plains Aquifer during 2003–2013. Experiments using CSR, JPL, H08, and WaterGAP also showed that reconstruction performance and predictor importance depend on the target product. The dataset provides kilometre-scale estimates of groundwater-storage variability under the large-scale GRACE constraint. Local interpretation of the reconstructed spatial patterns requires additional groundwater—systems-level, abstraction, and hydrogeological observations.

Keywords: Groundwater storage anomaly; GRACE downscaling; Machine learning; High spatiotemporal resolution; ~~Human pumping characteristics~~ Product uncertainty

1. ~~1~~ Introduction

Global groundwater depletion presents a significant threat to water security. In the context of the accelerating global water cycle, deep aquifer systems are experiencing substantial dynamic evolution (Famiglietti, 2014; Taylor et al., 2013). As the world's largest accessible freshwater resource, groundwater plays a crucial role in mitigating the impacts of climate variability on ecosystems (Rodell et al., 2018; Wada et al., 2010; Weber et al., 2016). Over the past four decades, anthropogenic overextraction has led to the accelerated decline of groundwater levels in multiple regional aquifers globally, resulting in seawater intrusion, land subsidence, water quality deterioration, ecosystem degradation and reduced river discharge. Additionally, global climate change, especially the intensification of drought events, has decreased effective precipitation infiltration and groundwater recharge, while increasing surface evapotranspiration and the reliance of agricultural irrigation on groundwater, thus further exacerbating the depletion risk of groundwater systems. Several large-scale aquifer systems, such as those in the North China Plain, the U.S. High Plains (Ogallala) and northwestern India, are approaching critical depletion thresholds (Rodell et al., 2018; Scanlon et al., 2012). However, conventional groundwater monitoring mainly relies on observation-well networks, which are constrained by limited spatial coverage, time-series gaps caused by insufficient maintenance and barriers to data sharing, making spatiotemporally continuous monitoring challenging.

The Gravity Recovery and Climate Experiment (GRACE), launched in 2002, and its follow-on mission GRACE-FO have supported the continuous retrieval of groundwater storage anomalies (GWSA), emerging as an important approach for monitoring groundwater changes. Nevertheless, due to their coarse spatial resolution, with ~~the native~~ an effective spatial resolution ~~being only $3^\circ \times 3^\circ$ of several~~ hundred kilometres, existing data cannot resolve the significant spatial heterogeneity caused by complex topography, lithological heterogeneity and differences in aquifer structures. Moreover, local scale studies face substantial limitations, particularly the difficulty of establishing reliable connections between GRACE estimates and local ground-based observation sites, which further restricts the accuracy of groundwater-change assessments. Therefore, there is an urgent need to downscale GRACE-derived GWSA.

~~In Comparison to physically based downscaling,~~ Statistical downscaling has ~~attracted increasing~~ attention because of its simple computational framework and strong cross-regional transferability,

especially in dealing with nonlinear relationships. Its fundamental assumption is that the statistical relationships between predictor variables, such as hydrological, vegetation and climatic factors, and the target variable remain approximately consistent between the original resolution and the high resolution. On this basis, existing studies have developed a series of GWSA downscaling models, including been widely used to derive higher-resolution GWSA estimates from coarse-resolution satellite or model products. Existing approaches include partial least squares regression (PLS), artificial neural networks (ANN), ground-based scaling factors (SF), and the most widely used tree-based machine learning models, such as factor methods, random forests (RF), and gradient boosted trees Extreme Gradient Boosting (XGBoost). However, existing studies are often confined to most applications have been conducted at regional or basin scales (Miro and Famiglietti, 2018; Seyoum et al., 2019), which limits their use for global observation. Moreover, most groundwater assessment. In addition, many downscaled products still rely heavily on coarse-resolution global hydrological models with coarse spatial resolutions for model training development and lack constraints from independent in situ observational benchmarks groundwater observations (Scanlon et al., 2018). This dependence often causes the resulting products to represent only smoothed interpolations of low frequency background signals, with limited ability to capture the localized characteristics induced by human groundwater pumping. Several issues therefore remain for global high-resolution GWSA products. First, comparison with groundwater-monitoring data is affected by differences in spatial support and physical units between point-scale groundwater levels and grid-scale groundwater storage. Second, consistency with the original coarse-resolution target alone does not show whether the reconstructed data reproduce observed temporal variability, seasonal behavior, storage magnitude, and long-term trends. Third, downscaling results may depend on the target product used for model training, and differences among GRACE solutions and global hydrological models may propagate into the reconstructed spatial patterns and predictor relationships.

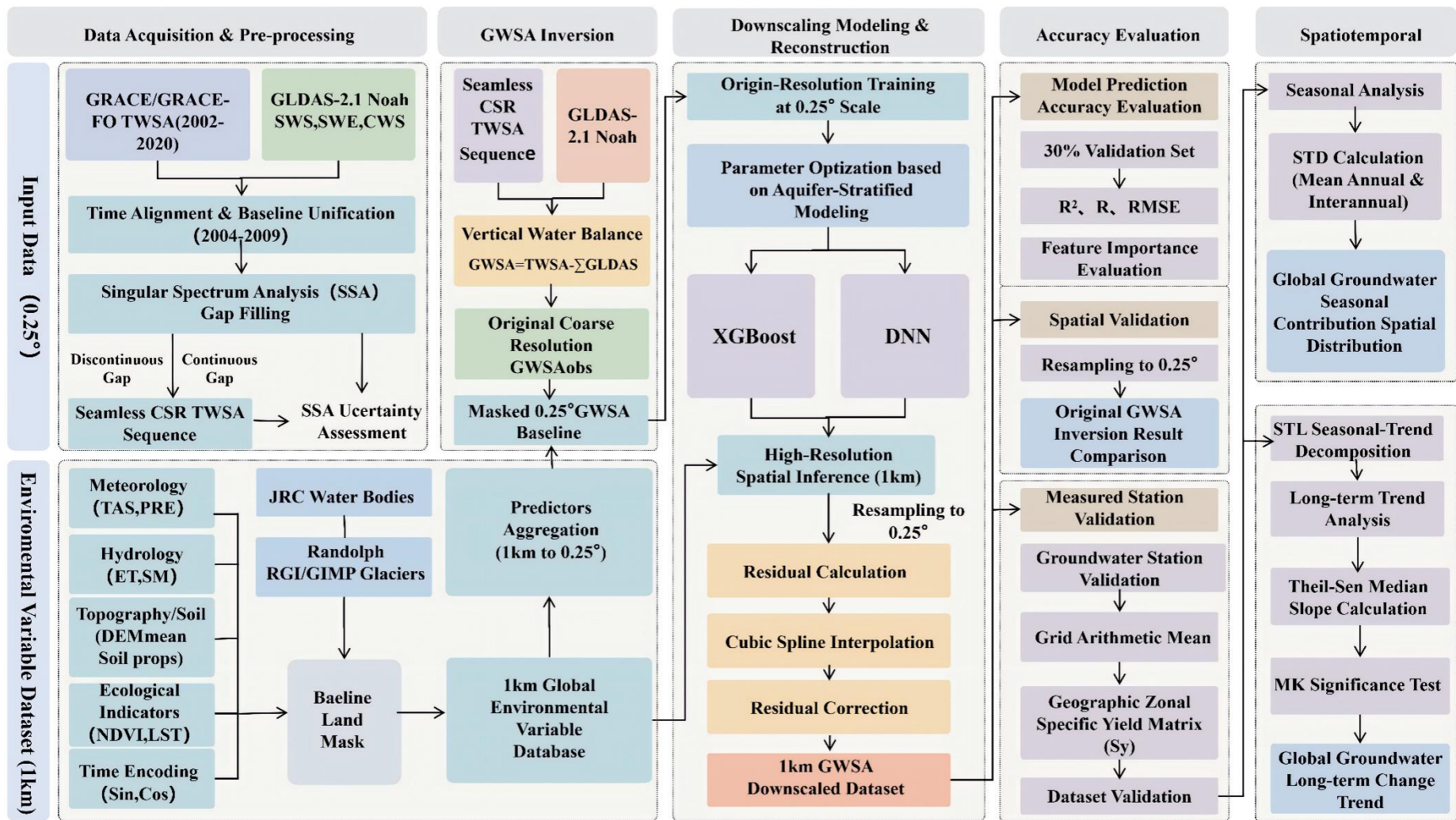
In this study, we developed a spatiotemporally seamless global 1 km monthly Groundwater Storage Anomaly (GWSA) dataset covering the period from April 2002 to December 2020. Our The framework integrates Singular Spectrum Analysis (combines SSA) for temporal gap filling (Yi and Sneeuw, 2021) and utilizes multisource 1 km with 19 high-resolution environmental covariates. To ensure the physical

consistency and spatial robustness of the dataset, we systematically evaluated representative machine learning (ML) architectures, including deep neural networks (DNN) and XGBoost, as a rigorous quality control measure as candidate downscaling models (Reichstein et al., 2019; Shen, 2018). We validated our dataset), and the production model was selected using a comprehensive global observation network of 1,518 in situ monitoring internal evaluation, cross-scale consistency, and independent groundwater observations. For the in-situ evaluation, a global groundwater-monitoring archive containing 152,069 wells (Jasechko et al., 2024) and cross-compared it with GHMs. The results indicate that the downscaled GWSA can provide a was divided into 1° spatial resolution capable of effectively identifying local agricultural boundaries, thereby bridging the spatial scale gap in global assessments blocks. Effective specific yield was estimated from calibration blocks and providing a robust, observation-constrained benchmark dataset for global water resource management transferred to held-out blocks before groundwater-level and the development of regional groundwater digital twins GWSA trends were compared on a common 0.25° grid. Standardized time-series comparisons were also conducted without specific-yield conversion, and monthly observations were used to examine long-term and seasonal variability in representative aquifers. The reconstructed GWSA was further compared with published estimates for the North China Plain, northwestern India, California, and the High Plains Aquifer. In addition, CSR, JPL, H08, and WaterGAP were used as alternative target products to examine the sensitivity of the downscaling results to the input data source.

2 Data and Methodology

To generate the 1 km global groundwater storage anomaly (GWSA) dataset, a systematic framework composed of five inter-related stages was devised (Fig. 1). Initially, multi-source data integration was implemented. This entailed filling the temporal gaps in 0.25° GRACE/GRACE-FO observations via Singular Spectrum Analysis (SSA) and constructing a 1 km environmental covariate database. Subsequently, coarse-resolution GWSA was inverted by means of a vertical water balance method and refined through water body and glacier masking. To tackle the spatial scale disparity, an aquifer-stratified downscaling model was employed. This model utilized tree-based (XGBoost) and continuous (DNN) architectures, followed by a residual correction mechanism based on cubic spline interpolation. The

reconstructed high-resolution product was then subjected to a stringent multi-scale accuracy assessment, including 0.25° internal consistency checks and independent validation ~~against 1,518~~using 152,069 in-situ groundwater monitoring stations ~~through a regional specific yield (S_r) matrix with annual~~observations and 140 stations with monthly observations. The finalized dataset was utilized to conduct spatiotemporal analysis, characterizing global groundwater dynamics through seasonal contribution analysis and long-term trend detection. The specific theoretical and operational details of each stage are elaborated in the following subsections.



149
150

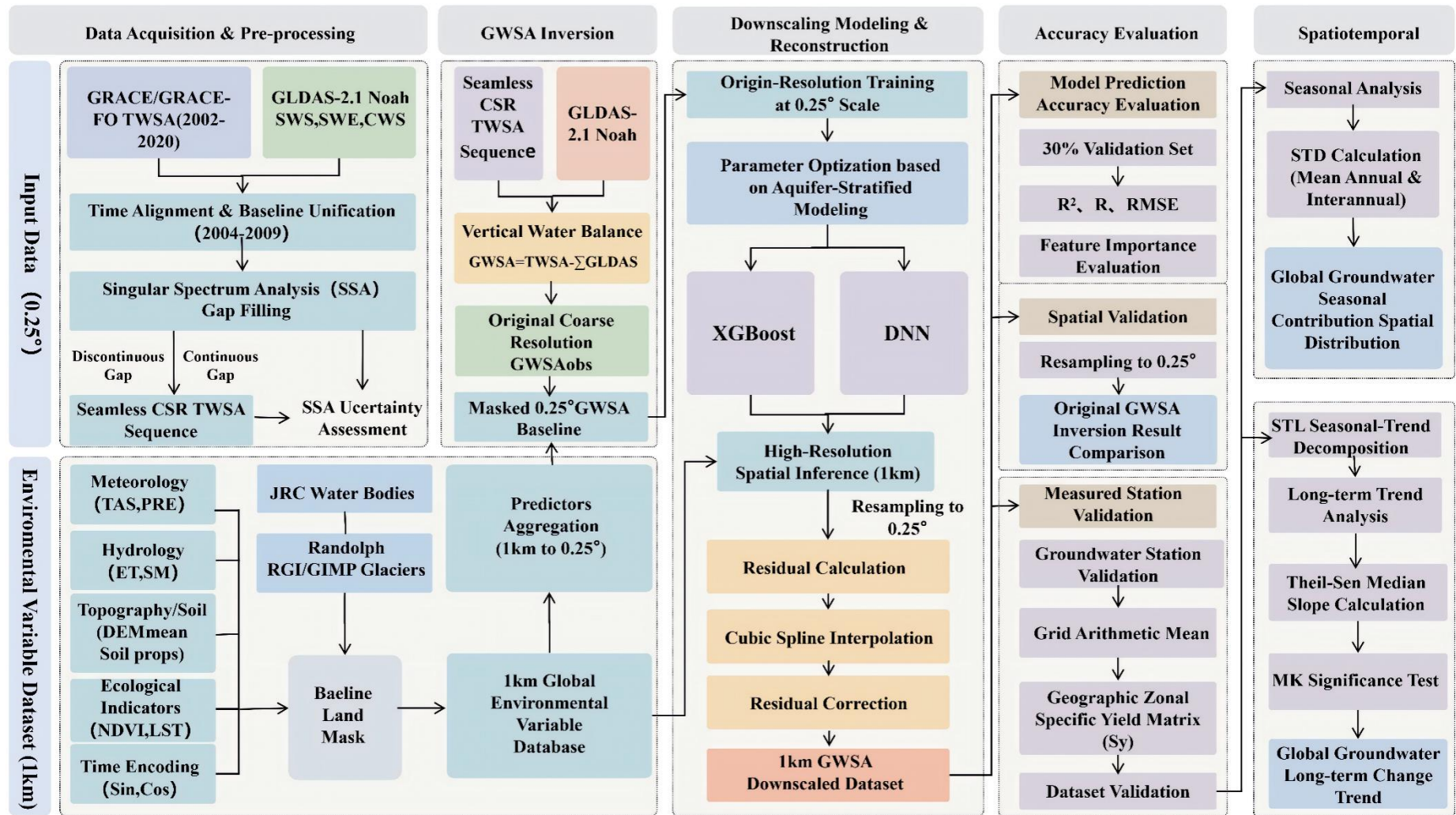


Figure 1. Methodological framework for 1 km GWSA inversion and downscaling.

The framework consists of five main modules: (1) Data Acquisition and Pre-processing of 0.25° input data and 1 km environmental variables; (2) GWSA Inversion based on vertical water balance; (3) Downscaling Modeling & Reconstruction via machine learning (XGBoost and DNN) and residual correction; (4) Accuracy Evaluation through model prediction metrics, spatial comparison, and measured station validation; and (5) Spatiotemporal analysis evaluating both seasonal variations and long-term change trends

2.1 Data Sources

All data supporting the findings of this study are publicly available. This research utilized the GRACE/GRACE-FO mass concentration (mascon) solutions (RL0603) provided by the University of Texas Space Research Center (CSR) and the GLDAS-2.1 product from the NASA Global Land Data Assimilation System Version 2 (GLDAS-2) as the primary target variables for the downscaling modeling process. Additionally, a spatiotemporal unified database of environmental variables with a global 1 km resolution was established in this study. This database encompasses meteorological, hydrological, topographic and soil, and biophysical variables, along with the cumulative fluxes of certain variables.

To assess the sensitivity of the GWSA downscaling framework to the data source, we introduced three additional datasets alongside the CSR GRACE/GRACE-FO and GLDAS reference data. The JPL RL06.3Mv04 CRI-filtered mascon product was used as an alternative satellite TWSA input to CSR, while the H08 and WaterGAP 2.2c groundwstor products were converted using the reference-period procedure and used as independent hydrological-model GWSA benchmarks for comparison with the satellite-derived GWSA.

The World-wide Hydrogeological Mapping and Assessment Programme (WHYMAP) provides a standardized framework for delineating major aquifer units and has been widely used in global groundwater assessments and model evaluations (Jasechko et al., 2024; Scanlon et al., 2018). In this study, aquifer boundaries were used as the common spatial framework for model development and validation. Separate downscaling models were developed for different aquifer categories, and the reconstructed GWSA was evaluated against observations within the corresponding aquifer units.

Meteorological variables include temperature (TAS) and precipitation (PRE); hydrological variables consist of soil moisture (SM) and actual surface evapotranspiration (ET); topographic and soil

attribute variables are composed of global mean elevation (DEM), proportion of clay particles (Clay), proportion of silt particles (Silt), proportion of sand particles (Sand), coarse fragments (CFVO), cation exchange capacity (CEC), soil pH, soil organic carbon content (SOC), and total nitrogen (TN); biophysical variables involve normalized vegetation index (NDVI) and land surface temperature (LST). These variables serve as crucial predictive factors for facilitating the refined mapping of global groundwater reserve anomalies. All data employed in model construction are listed, and detailed information on datasets is presented in Table 1.

Although GRACE/GRACE-FO mascon solutions are available through November 2025, the temporal coverage of the downscaled GWSA dataset was limited to the period from April 2002 to December 2020. This temporal restriction mainly reflects the availability of the high-resolution environmental covariates used for model training, particularly the meteorological variables.

Table 1. Summary of datasets and sources for downscaling model.

Variable	Spatial resolution	Temporal resolution	Temporal coverage	Data source
CSR GRACE/ GRACE-FO	0.25°	Monthly	Apr 2002– Nov 2025	https://www2.csr.utexas.edu/grace
<u>JPL GRACE/ GRACE-FO</u>	<u>0.5°</u>	<u>Monthly</u>	<u>Apr 2002– Jun 2026</u>	<u>https://grace.jpl.nasa.gov/data/get-data/jpl_global_mascons/</u>
GLDAS-2.1	0.25°	Monthly	Apr 2002– Dec 2020	https://disc.gsfc.nasa.gov/datasets
<u>H08 (ISIMIP3a)</u>	<u>0.5°</u>	<u>Monthly</u>	<u>Apr 2002– Dec 2021</u>	<u>https://data.isimip.org/search/query/h08_20crv3-era5_obsclim_histsoc_default_grou</u> <u>ndwstor_global_monthly/</u> <u>https://data.isimip.org/search/query/watergap2-2c_gswp3-ewembi_nobc_hist_nosoc_co2_grou</u> <u>ndwstor_global_monthly/</u> <u>https://www.WHYMAP.org/</u>
<u>WaterGAP 2.2c (ISIMIP2a)</u>	<u>0.5°</u>	<u>Monthly</u>	<u>Apr 2002– Dec 2016</u>	<u>https://data.isimip.org/search/query/watergap2-2c_gswp3-ewembi_nobc_hist_nosoc_co2_grou</u> <u>ndwstor_global_monthly/</u> <u>https://www.WHYMAP.org/</u>
<u>WHYMAP</u> Climate data (TAS, PRE)	<u>Vector</u> 1000m	<u>=</u> Monthly	<u>=</u> Apr 2002– Dec 2020	<u>https://www.WHYMAP.org/</u> (Karger et al., 2017)
ET	1000m	8-day	Apr 2002– Dec 2020	(Zhang et al., 2019)
SM	1000m	Monthly	Apr 2002– Dec 2020	(Zheng et al., 2023)
DEM	1000m	–	2010	https://www.usgs.gov/coastal-changes-and-impacts/gmted2010
Soil properties (Clay, Silt, Sand, CFVO, pH, CEC, SOC, TN)	250m	–	2020	(Poggio et al., 2021)
NDVI	500m	16-day	Apr 2002– Dec 2020	https://developers.google.com/earth-engine/datasets/catalog/MODIS_061_MOD13A1
LST	1000m	8-day	Apr 2002– Dec 2020	https://developers.google.com/earth-engine/datasets/catalog/MODIS_061_MOD11A2
<u>Global groundwater monitoring network</u>	<u>Point</u>	<u>Annual</u>	<u>1801-2022</u>	<u>Jasechko et al. (2024)</u>
<u>GROW</u>	<u>Point</u>	<u>Daily monthly annual</u>	<u>Dec 1838– Aug 2025</u>	<u>Bäthge et al. (2026)</u>

Variable	Spatial resolution	Temporal resolution	Temporal coverage	Data source
<u>China Groundwater Level Yearbook for Geo-environmental Monitoring</u>	<u>Point</u>	<u>Daily</u>	<u>Jan 2005–Dec 2017</u>	<u>China Geological Environment Monitoring Institute</u>

2.1.1 GRACE Terrestrial Water Storage

Our research employed the GRACE Mascon solutions provided by CSR, encompassing monthly Terrestrial Water Storage Anomaly (TWSA) sequences from April 2002 to November 2025. The reflected TWS anomalies were measured relative to the average baseline spanning from 2004 to 2009. Prior studies have indicated that Mascon solutions present substantial enhancements over spherical harmonics in terms of leakage errors and post-processing (Palazzoli et al., 2025; Scanlon et al., 2016). They do not necessitate additional smoothing, empirical de-striping, filtering, or scaling factors, and are more user-friendly (Save et al., 2016; Wang et al., 2024). The newly released ~~RL06~~RL0603 Mascon solutions adopt a novel grid structure founded on equidistant geodetic grids with a spatial resolution of 1° , which is resampled to a resolution of 0.25° for publication. This effectively mitigates the leakage between land and ocean signals. In this investigation, we utilized ~~Singular Spectrum Analysis (SSA)~~ to fill the 35-month data gaps. In combination with the ~~Global Land Data Assimilation System (GLDAS)~~, we constructed the ~~Groundwater Storage Anomaly (GWSA)~~ sequence. Moreover, we applied machine learning and deep learning methods to downscale the spatial resolution to 0.008333° (approximately 1 km at the equator), with the aim of attaining globally seamless high-resolution GWSA mapping with spatiotemporal resolution.

To assess the influence of different GRACE products on the GWSA downscaling model, we further introduced the JPL GRACE/GRACE-FO RL06.3Mv04 CRI-filtered mascon product for comparison with CSR. The JPL lwe₂ thickness variable represents TWSA in centimetres of equivalent water thickness and has a sampling resolution of 0.5° . We bilinearly resampled the monthly JPL grids to the CSR reference grid and used the common period from April 2002 to December 2020; missing months during the GRACE mission period were retained during model training (Scanlon et al., 2016).

2.1.2 GLDAS Noah V2.1

~~The National Aeronautics and Space Administration (NASA) Global Land Data Assimilation System (GLDAS)~~NASA GLDAS is designed to integrate satellite and ground-based observational data. By leveraging land surface modeling and data assimilation techniques, it aims to attain optimal

simulations of land surface conditions and flux fields (Rodell et al., 2004). It encompasses three land surface models: the Noah model, the Catchment (CLSM) model, and the Variable Infiltration Capacity (VIC) model.

The Noah land surface model includes four soil temperature and moisture layers, along with models for vegetation canopy, snow cover prediction, ~~evapotranspiration~~ET, and soil drainage runoff. Owing to its precise representation of surface energy and water budgets, it has been extensively employed in global hydrological and water resource change research (Campbell et al., 2019; Ek et al., 2003; Niu et al., 2011).

In this study, ~~the annual time series data of monthly~~ soil water storage (SWS), snow water equivalent (SWE), and vegetation canopy water storage (CWS) ~~provided by from GLDAS~~ Noah V2.1 were ~~utilized to calculate the groundwater used for the GWSA inversion. The soil-water-storage anomaly (GWSA). All input data~~ term was calculated as the vertically integrated storage of the four Noah soil layers, using 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). SWE and canopy water storage were ~~transformed into~~obtained from SWE_inst and CanopInt_inst, respectively. All of these variables are provided in kg m^{-2} and were converted to equivalent water thickness before component subtraction. Monthly anomalies were calculated relative to the ~~average baseline from corresponding~~ 2004–2009 (Wang et al., 2024; Zhang et al., 2021) ~~reference-period means~~ to ensure consistency with the ~~definitions of the Gravity Recovery and Climate Experiment (GRACE).~~ TWSA anomaly definition. The GLDAS soil-water-storage term used for GWSA inversion is distinct from the separate 1 km surface-soil-moisture dataset used as a high-resolution predictor in Sect. 2.1.4.

2.1.3 Global Hydrological Model Data

To benchmark the reconstructed groundwater-storage anomalies, we selected two ISIMIP global hydrological-model products: h08_20crv3-era5_obsclim_histsoc_default_groundwstor_global_monthly (H08, ISIMIP3a) and watgap2-2c_gswp3-ewembi_nobc_hist_nosoc_co2_groundwstor_global_monthly (WaterGAP 2.2c, ISIMIP2a). Both provide the monthly variable groundwstor in kg m^{-2} on a global 0.5° grid. H08 represents natural hydrological processes together with human water use (Hanasaki et al., 2008), whereas WaterGAP 2.2c combines global hydrology with sectoral water-use modules (Müller Schmied et al., 2016). To ensure grid consistency, we bilinearly interpolated both datasets to the CSR GRACE reference grid, retained

missing values in the time series, and standardized their analysis periods to April 2002 - December 2021 for H08 and April 2002 - December 2016 for WaterGAP.

2.1.4 Predictor Variables

The surface hydrological cycle demonstrates intricate dynamic variations governed by water, energy, and biogeochemical fluxes. For example, the equilibrium of ~~terrestrial water storage (TWS)~~TWS is affected by anomalous changes in constituent components across diverse climatic zones (Giroto & Rodell, 2019). Furthermore, terrestrial heat and TWSA themselves can influence the predictive precision of GLDAS (Campbell et al., 2023). To comprehensively consider various influencing factors, 19 variables closely associated with climate, hydrology, and vegetation dynamics were selected to construct a global environmental variable database at a 1 km scale as modeling predictors. This dataset encompasses 8 monthly-scale dynamic variables, 9 multi-year average variables, and two additional temporal coding variables (derived from sine and cosine transformations of monthly values to capture seasonal characteristics).

Monthly ~~average near surface air temperature (TAS)~~ and monthly ~~cumulative precipitation (PRE)~~ data are obtained from Climatologies at High resolution for the Earth's Land Surface Areas (CHELSA). This dataset, based on ERA climatic reanalysis and terrain correction algorithms, offers global meteorological forcing data with a spatial resolution of 1 km at the monthly scale, effectively supporting high-precision model simulations at local scales (Karger et al., 2017).

The ~~evapotranspiration (ET)~~ data are sourced from the PML_V2 high-resolution dataset (Zhang et al., 2019), which provides a spatial resolution of 500 m and a temporal resolution of 8 days for ET measurements along with three key independent components (i.e., vegetation transpiration, soil direct evaporation, and canopy interception evaporation). These data have undergone cross-validation in the Global Flux Network (FLUXNET) and exhibit robust representation of surface water consumption patterns. In this study, monthly synthetic ET was generated by rescaling the 8-day resolution data according to temporal scales, followed by cumulative ET calculations to obtain monthly cumulative evapotranspiration (ETC) values since April 2002.

~~Soil moisture (SM)~~ was derived from the Global Daily Surface Soil Moisture Dataset (Zheng et al., 2023), which was jointly inverted using ERA5 reanalysis data and ESA-CCI SSM products. This dataset

provides seamless global surface soil moisture data with a daily resolution of 1 km for the period 2000–2021. The monthly-scale product was selected.

The DEM is derived from the Global Multi-resolution Topographic Elevation Dataset (GMTED2010) provided by the United States Geological Survey (USGS). This study specifically extracts the spatial mean (Mean) product from this dataset, aiming to precisely characterize the average elevation variation of the underlying surface at the target grid scale.

The soil attributes are derived from the High Precision Global Soil Information System (SoilGrids) published by the International Soil Reference Information Centre (ISRIC). This dataset is generated through machine learning algorithms coupled with multi-source environmental covariate physical coupling inversion, providing global-scale products at a resolution of 250 m. The study extracts variables including Proportion of Clay ~~particles (Clay)~~, ~~Silt~~, Proportion of ~~silt particles (Silt)~~, ~~Proportion of sand particles (Sand)~~, ~~Coarse fragments ($\geq$ CFVO)~~, ~~Cation Exchange Capacity ($\geq$ CEC)~~, Soil pH, ~~Soil organic carbon content (SOC)~~, and ~~Total nitrogen (TN)~~ from depths of 0–30 cm to reflect surface soil (or topsoil) characteristics.

NDVI and LST were derived from MOD13A1 and MOD11A2 products, respectively. This study conducted quality validation based on the ~~GEE~~ Google Earth Engine platform and synthesized monthly average data for the period 2002–2020.

2.1.5 In-situ Groundwater-level Observations

To independently evaluate the spatiotemporal trends in the reconstructed GWSA, we used three types of in-situ groundwater observations: the global groundwater-monitoring network compiled by Jasechko et al. (2024), the Global-Scale Time Series Dataset for Groundwater Studies within the Earth System (GROW; B  thge et al., 2026), and the China Groundwater Level Yearbook for Geo-environmental Monitoring.

The global monitoring network of Jasechko et al. (2024) comprises 152,069 monitoring wells, with records spanning 1801–2022 and groundwater depth reported at annual intervals. We converted groundwater depth to groundwater-level change by reversing its sign and excluded stations with fewer than eight valid annual observations. Because groundwater-level change and groundwater-storage change have different physical dimensions, an effective specific-yield coefficient was introduced to calibrate the two quantities (2.5.1), and its sensitivity was further evaluated (3.1.2). In addition,

uncalibrated standardized comparisons were provided (3.2.1) to assess the consistency between the reconstructed GWSA and global in-situ observations.

The GROW dataset covers 55 countries and contains more than 200,000 groundwater time series, together with 36 Earth-system-related variables and 34 time-series feature flags. Daily and monthly records account for approximately 6% and 9% of the dataset, respectively (Bäthge et al., 2026). Following the GROW processing guidelines (Bäthge et al., 2026), we removed yearly records, calculated monthly means from the remaining observations, and retained stations with records throughout April 2002–December 2020. After spatial matching with the WHYMAP global aquifer database and retaining aquifers with more than five valid stations, 55 stations covering three aquifer classes were selected.

The China Groundwater Level Yearbook for Geo-environmental Monitoring compiles daily groundwater-level monitoring records across China. The data cover January 2005–December 2017 and include 1,074 monitoring stations with geographic coordinates across 31 provincial-level regions, of which 266 provide continuous observations over the 13-year period. Applying the same screening criteria yielded 87 stations covering aquifers in the North China Plain.

For both datasets, groundwater depth was converted to groundwater-level change, values exceeding three standard deviations were removed, and missing values resulting from the removal were linearly interpolated. Anomalies were then calculated relative to the 2004–2009 and 2005–2009 baseline periods, respectively. The processed station data were subsequently matched to WHYMAP aquifers and a common 1° grid for monthly validation.

2.2 Singular Spectrum Analysis for Data Gap Filling

~~Singular Spectrum Analysis (SSA)~~SSA is a technique for processing nonlinear time series (Vautard et al., 1992). Through the decomposition, sorting, and reconstruction of the trajectory matrix of the target time series, SSA can extract trends and noise from the data and has been extensively utilized for missing value imputation (Kondrashov and Ghil, 2006; Wang et al., 2021). In the present study, SSA was utilized to impute 22 discrete short gaps that occurred prior to June 2017 during the GRACE mission period, which were caused by periodic power system outages (defined as SSA-filling-a). Additionally, it was used to impute 11 consecutive months of missing data after July 2017 between the GRACE/GRACE-FO mission periods and two months of missing data within the GRACE-FO mission period (collectively defined as SSA-filling-b). This approach aims to minimize the influence of missing data on subsequent

334 downscaling modeling and temporal analysis. To ensure the accuracy of SSA reconstruction, the full
 335 available GRACE/GRACE-FO monthly record from April 2002 to November 2025 was used for the SSA
 336 calculation. The SSA calculation method is as follows:
 337 First, define the time series for uniform X sampling:

$$X = [x_1, x_2, \dots, x_N] \quad (1)$$

338 where x_i ($i = 1, 2, \dots, N$) represents the monthly GRACE observation values, where missing months
 339 were initialized with zero as starting values for the SSA iteration. N represents the sample size from
 340 April 2002 to November 2025. For the periods of October 2011 and April 2015, repeated observations
 341 were averaged, hence $N = 284$.

342 Define the trajectory matrix Y for time series X :

$$Y = \begin{bmatrix} x_1 & x_2 & \cdots & x_L \\ x_2 & x_3 & \cdots & x_{L+1} \\ \vdots & \vdots & \ddots & \vdots \\ x_M & x_{M+1} & \cdots & x_N \end{bmatrix} \quad (2)$$

$$L = N - M + 1 \quad (3)$$

$$K = \min(M, L) \quad (4)$$

343 where L represents the number of columns in the trajectory matrix; M represents the window width;
 344 and K represents the maximum modal number decomposable by the trajectory matrix (Eq. (4)). This
 345 research employed the optimization parameter search scheme proposed by Yi and Sneeuw (Yi & Sneeuw,
 346 2021), with $M = 24$ and $K = 12$ for SSA-filling-a, and $M = 48$ with K ranging from 1 to 12 for SSA-
 347 filling-b. These parameters were ascertained through iterative procedures and cross-validation. Conduct
 348 singular value decomposition on the trajectory matrix Y :

$$Y = U \Sigma V^T \quad (5)$$

$$e_i = U_{(:,i)} \sigma_i \quad (6)$$

$$p_i = V_{(:,i)} \quad (7)$$

$$Z_i = e_i \times p_i^T \quad (8)$$

$$\mathbf{Y} = \sum_{i=1}^K \mathbf{Z}_i \quad (9)$$

where Σ represents the singular value diagonal matrix; σ_i represents the i th singular value; $\mathbf{U}$ and $\mathbf{V}$ are orthogonal matrices; $\mathbf{e}_i$ represents the i th empirical orthogonal function; $\mathbf{p}_i$ represents the i th principal component; and $\mathbf{Z}_i$ represents the i th modal matrix. The trajectory matrix $\mathbf{Y}$ can be expressed as the sum of the K -modal $\mathbf{Z}_i$ matrices.

Calculate the reconstructed component (RC) sequence $\mathbf{g}_i$ by diagonal averaging of each modal matrix $\mathbf{Z}_i$, and the reconstructed time series $\mathbf{X}'$ was calculated as follows:

$$\mathbf{g}_i^p = \text{mean}(\mathbf{Z}_i^{j,k}), (j + k = p + 1) \quad (10)$$

$$\mathbf{X}' = \sum_{i=1}^{K'} \mathbf{g}_i \quad (11)$$

where $\mathbf{g}_i^p$ represents the value of the i th reconstructed component at the p th position of the time series; $\mathbf{g}_i$ represents the complete reconstructed component sequence composed of $\mathbf{g}_i^p$, calculated as the average of the skew-diagonal elements of each modal matrix $\mathbf{Z}_i$; p represents the position index of the time series ($p = 1, 2, \dots, N$); j and k represent the row and column numbers of $\mathbf{Z}_i$, respectively; K' represents the actual number of retained components; and $\mathbf{X}'$ represents the reconstructed time series.

Finally, a two-layer iterative approach is employed for $\mathbf{X}$ interpolation. In the outer iteration, K' is incrementally increased from 1 to determine the optimal number of modes. When $K' = K$, the sum of all reconstructed components equals $\mathbf{X}$. In the inner iteration, $\mathbf{X}'$ is continuously calculated for each fixed K' until the difference between consecutive reconstruction results ($\mathbf{X}'$) falls below a predefined threshold, indicating stabilization of the SSA results. This stabilized state is then adopted as the optimal reconstruction outcome for the current K' value. Additionally, further details on this methodology can be referred to the paper by Yi and Sneeuw (Yi & Sneeuw, 2021).

Ultimately, the system evaluates the uncertainty of the SSA reconstruction by integrating the error metrics, evaluation metric ε_1 from the SSA-filling-a phase and with the metric ε_2 from the SSA-filling-b phase:

$$\varepsilon_1 = \text{Std}(\mathbf{X}_{\text{obs}} - \hat{\mathbf{X}}) \quad (12)$$

$$\varepsilon_2 = \min \sqrt{\frac{1}{12} \sum \varepsilon^2} \quad (13)$$

$$\varepsilon = \sqrt{\frac{1}{N} \sum (X_{\text{obs}} - \hat{X})^2} \quad (14)$$

Both SSA ~~filling~~fitting and error evaluation were carried out on a pixel-by-pixel basis for the original ~~Gravity Recovery and Climate Experiment (GRACE)~~ data grid. In the SSA-filling-a phase, the reconstruction error was defined as the standard deviation of the discrepancy between the observed sequences from non-missing months and the SSA-reconstructed sequences. Here, X_{obs} and $\hat{X}$ represent the original observation sequences for all non-missing months within the GRACE mission period and the reconstructed sequences obtained by applying the SSA-filling-a algorithm to the corresponding months, respectively. ~~The standard deviation of the difference between the two sequences, denoted as ε_1 , represents the reconstruction uncertainty in this phase, with larger, implies that higher values indicating~~ lower reconstruction fitting accuracy during this phase.

In the SSA-filling-b phase, annual masking was applied to the original observation series using 2003–2016 as the cross-validation background, ~~and while~~ each year from 2004 to 2015 was ~~sequentially~~successively masked to form 12 annual gap-filling experiments. ~~For each masked year~~ First, the root mean square error (RMSE), ~~denoted as ε ,~~ for each masked year was ~~calculated between~~computed, followed by the ~~original observations and the reconstructed values using Eq. (14).~~ aggregation of the 12 annual RMSE values ~~were then aggregated using Eq. (according to Formula 13).~~ Subsequently, a sequential search was conducted within the range of the maximum modal number K (ranging from 1 to 12), and the minimum aggregated RMSE was defined as the fitting error for this phase. The corresponding combination of the window width M and the maximum modal number K was identified as the optimal fitting parameters for this phase. Here, X_{obs} and $\hat{X}$ represent the original observation series of a masked year and the reconstructed series obtained by applying the SSA-filling-b algorithm to that year, respectively. ~~N represents, N denotes~~ the number of non-missing months in the original GRACE data for that year, signifying the actual sequence length; ~~ε , ε represents the~~ RMSE of the fitting results for that year, while ε_2 denotes the minimum ~~aggregated RMSE from~~ value obtained by aggregating the 12 annual RMSE values from the masking experiments, ~~with larger values indicating.~~ Similarly, a higher ε_2 value indicates lower fitting accuracy for this phase.

2.3 GWSA Inversion and Masking Strategy

Terrestrial water storage (TWS) represents the aggregate vertically integrated water reserves both stored above and beneath below the land surface, encompassing surface water, including soil moisture water, groundwater, snow cover, ice, and, vegetation water storage canopy water, surface water, and land ice (Giroto ~~&~~ Rodell, 2019). ~~Through Following the utilization of a vertical water commonly used residual GRACE-land-surface-model in conjunction with GRACE and GLDAS data, framework, GWSA were estimated by subtracting modeled soil-water storage, snow-water equivalent, and canopy-water storage anomalies in groundwater storage (GWSA) can be segregated from TWSA GRACE-derived terrestrial water storage anomalies via the subsequent calculation approach: (TWSA):~~

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

The ~~data~~ monthly SWS, SWE, and CWS components were obtained from GLDAS Noah V2.1. Specifically, SWS was calculated as the sum of the four Noah soil-moisture layers, including 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), thereby representing vertically integrated soil-water storage (SMS), snow water equivalent (SWE), over the complete 0–200 cm Noah soil column. SWE and canopy water storage (CWS) were obtained from Noah V2.1, where Δ represents the were obtained from SWE_inst and CanopInt_inst, respectively. These GLDAS storage variables are provided in units of kg m⁻². Because 1 kg m⁻² of water is equivalent to 1 mm of water thickness, the GLDAS storage components were converted to centimetres of equivalent water height before being combined with the GRACE TWSA. All components were subsequently expressed as anomalies relative to the their respective mean values over the 2004–2009 baseline period (2004–2009). The anomalies in to ensure consistency with the GRACE anomaly definition.

Here, ΔTWS denotes the terrestrial water storage (ΔTWS) were anomaly derived from the Gravity Recovery and Climate Experiment (GRACE) data GRACE, while ΔGWS represents the anomalies in the isolated GRACE-derived groundwater storage (GWSA) anomaly estimated using the residual component-separation framework. We calculated the observed GWSA sequence (GWSAobs; ~~N=249=249~~) from the original GRACE data, and the reconstructed GWSA sequence (GWSASSA; ~~N=284=284~~) from the SSA-reconstructed GRACE data. The two sequences were fully spatially aligned and directly comparable at coincident time points. To validate the reliability of the reconstruction, the

observed GWSA sequence was kept intact, and only for the months with missing GRACE data, the GWSA_{SSA} data (N=35) were used to replace it. ~~This alternative~~ The resulting temporally continuous GWSA sequence was ~~then~~ subsequently used as the ~~benchmark~~ target dataset for ~~subsequent reconstruction~~ downscaling and ~~precision evaluation~~ related accuracy assessments.

~~It is worth noting that this formula still has certain limitations. For example, in polar and alpine regions, TWSA is primarily governed by mass changes from intense iceberg discharge, submarine melting, and melt water runoff. In these areas, TWSA dynamics are not solely driven by snowfall, precipitation, or evapotranspiration (Chen et al., 2007; Mankoff et al., 2020; Rodell & Famiglietti, 2001). In tropical regions, the anomalies in surface water reserves, including those in rivers, inland water bodies, reservoirs, wetlands, and floodplains, play a dominant role in the changes of TWSA, which may introduce uncertainties into the inversion of the Groundwater Storage Anomaly (GWSA). Therefore, this study employed rigorous glacier and water body masking strategies to minimize the errors in the global GWSA estimation. Equation (15), however, represents a residual component-separation framework and does not imply complete physical isolation of groundwater storage from all other terrestrial water-storage components. In particular, GLDAS Noah V2.1 does not provide within this calculation a globally homogeneous dynamic surface-water-storage term representing the combined temporal variations of rivers, lakes, reservoirs, wetlands, and seasonal floodplains. Consequently, residual surface-water variability may remain in the GRACE-derived GWSA estimate, especially in tropical and floodplain-dominated regions where surface-water storage can contribute substantially to TWSA. Similarly, glacier and ice-sheet mass variations can introduce additional uncertainty in polar and alpine regions (Chen et al., 2007; Mankoff et al., 2020; Rodell and Famiglietti, 2001).~~

The glacier and permanent-waterbody masks applied in this study were therefore used as error-reduction measures rather than as substitutes for explicit dynamic component subtraction. These masks reduce contamination over the most evident permanent-water and land-ice domains but cannot completely remove time-varying surface-water storage or GRACE leakage into adjacent land pixels. Accordingly, the reconstructed GWSA should be interpreted as a GRACE-constrained residual groundwater-storage estimate that inherits uncertainty from component separation and spatial leakage. Incorporating globally consistent dynamic surface-water-storage observations represents an important direction for further improvement of the dataset.

2.3.1 Permanent Water Body Mask

To minimize signal leakage errors induced by permanent water bodies and maximize the preservation of coastal terrestrial information, this study employed the Global Surface Water Explorer (GSWE) dataset furnished by the European Commission Joint Research Centre (JRC) (Pekel et al., 2016). This dataset documents the spatiotemporal dynamics of global surface water bodies at a spatial resolution of 30 m. Regions with pixel values of 3 (Permanent water) were extracted from its Yearly Classification product as reliable water body benchmarks. Pixel-wise set union processing was conducted on the temporal classification results spanning from 2002 to 2020, and resampling to a 1 km resolution was carried out, ultimately generating a unified global permanent water body range mask.

2.3.2 Glacier and Ice Sheet Mask

The global mountain glacier mask was established by utilizing data from The Randolph Glacier Inventory (RGI) supplied by the National Snow and Ice Data Center (NSIDC) (<https://nsidc.org/data/nsidc-0770/versions/7#anchor-documentation>). This dataset encompasses single-contour vector maps of each glacier dating back to approximately 2000. These maps were integrated within ArcGIS Pro 3.5.4 to demarcate the global mountain glacier extent.

The Greenland land ice extent was obtained from the MEaSURES Greenland Ice Mapping Project (GIMP) Land Ice and Ocean Classification Mask version 1 data provided by NSIDC (<https://nsidc.org/data/nsidc-0714/versions/1>), making use of the 30m spatial resolution classification data from the GimpIceMask_30m_2015 product.

Ultimately, the two masks were combined into a unified global glacier mask, which was then merged with the water body mask to generate the baseline land ice extent.

2.4 Machine Learning Downscaling Framework

The fundamental premise of the downscaling model posits that the statistical correlation between the target variable (~~Groundwater Storage Anomaly~~, GWSA) and the independent variables remains invariant across diverse spatial scales. Building upon this premise, we constructed the model utilizing lower-resolution (0.25°) data spanning from 2002 to 2020, while leveraging higher-resolution (1 km) data for prediction purposes. The specific methodology is delineated as follows:

(1) Data pre-processing: Initially, all 19 input variables with a 1 km resolution were subjected to masking processing to ensure complete congruence with the reference mask. Scattered missing values were imputed using inverse distance weighting, ultimately generating a spatiotemporally continuous 1 km environmental variable database. Subsequently, the 19 variables were resampled to a 0.25° resolution and aligned with the GWSA grid, a process accomplished through mean aggregation via the gdal.Warp function.

(2) Model development: Given the substantial spatial heterogeneity of hydrogeological conditions and Groundwater Storage Anomalies (GWSA) at the global scale, ~~it is challenging to formulate a unified model that reflects the overall scenario. Therefore,~~ separate models were developed for each aquifer type after rasterizing the aquifer ~~data, map and assigning~~ each pixel ~~was assigned a specific~~ to an aquifer category. ~~First, 200,000 samples were randomly selected from each aquifer to optimize the model parameters, and then the models were trained using all samples from~~ For each aquifer category, ~~Random Forest (RF), 200,000 grid-month records were randomly sampled. Seventy percent were used for model fitting and Deep Neural Network (DNN) models were constructed, yielding lower resolution training outcomes to assess model accuracy~~ hyperparameter optimization and 30% for internal testing. ~~This random split was used only for internal model development and does not constitute a spatially or temporally independent generalization test because adjacent spatial and temporal samples may occur in both subsets.~~

(3) Prediction and correction: The 1 km environmental variable database was inputted to obtain high-resolution global GWSA reconstruction results. These results were resampled to a lower resolution, and the residuals between them and the $GWSA_{obs}$ or $GWSA_{SSA}$ sequences were calculated. The low-resolution residuals were then interpolated to a 1 km resolution using cubic spline interpolation and added to the reconstructed results to implement residual correction.

(4) ~~Validation~~ Accuracy assessment: This ~~mainly encompasses two aspects. First,~~ assessment focuses on the internal accuracy of the reconstruction framework without introducing external observations and ~~comprises three components. Firstly, SSA reconstruction errors were quantified by comparing the reconstructed results were compared with the values from the SSA-filling-a and SSA-filling-b phases with the available GRACE/GRACE-FO observations. Secondly, by utilizing the 30% test samples set aside during the training phase, we measured the prediction accuracy of the model across diverse aquifer zones via evaluation metrics and carried out an analysis of the relative significance of feature variables.~~

Finally, cross-validation was carried out between the reconstructed outcomes and the original GWSA inversion results data at the original spatial resolution (0.25°). Second) to quantify the deviations from the baseline data. Finally, the high-resolution reconstructed-reconstruction results (0.008333°) were compared with the extracted in situ observational data verified using global long-term groundwater monitoring well data (covering the period from 2002 to 2020) to systematically evaluate the overall accuracy of the downscaling products.

(5) in-situ observational validation: Independent validation was conducted using annual observations from 152,069 monitoring sites and monthly observations from 140 monitoring sites, as described in detail in Sect. 2.5.

2.4.1 Machine Learning Architectures and Model Configuration

To capture the intricate nonlinear relationships between groundwater storage anomaly (GWSA) and multi-source hydro-climatic variables, two representative machine-learning architectures were employed: Extreme Gradient Boosting (XGBoost) and Deep Neural Networks (DNN). XGBoost, a scalable ensemble grounded in the gradient-boosting decision tree algorithm (Chen & Guestrin, 2016), demonstrates superiority in establishing statistical mappings for GWSA downscaling (Ali et al., 2023; Zhang et al., 2021). It effectively mitigates multicollinearity and overfitting through the randomization of samples and variables (Chen et al., 2020; Khan et al., 2024). Complementarily, DNNs leverage interconnected hierarchical layers to automatically extract deep-level variable interactions from extensive datasets (LeCun et al., 2015; Shen, 2018). The high model capacity of DNNs effectively compensates for the potential limitations of tree-based decision spaces in smoothing nonlinear surfaces, especially under extreme hydrological forcing conditions (Reichstein et al., 2019; Shen, 2018).

To account for global hydrogeological heterogeneity, both models were independently trained for each aquifer category. For each aquifer, 200,000 pixels were randomly sampled. Seventy percent of the data was allocated for training and hyperparameter optimization. The model was trained using 70% of all samples from each aquifer, while the remaining 30% was reserved were used for validation (e.g., calculating the root mean square error (RMSE)). model accuracy assessment. For XGBoost, the optimal hyperparameters, including *n_estimators*, *learning_rate*, *max_depth*, and *min_child_weight*, were identified via the *RandomizedSearchCV* method. For the DNN, a Multi-Layer Perceptron (MLP) architecture was constructed, and a Robust Scaler was applied to both input and target variables to

eliminate interference from extreme climate outliers (Shen, 2018). The optimal network configuration (3–6 layers and 128–768 units) was autonomously determined using the Optuna optimization framework (Akiba et al., 2019). The DNN was trained using the Adam optimizer with a mean squared error (MSE) loss function, incorporating an Early Stopping mechanism to prevent over-fitting on the large-scale spatiotemporal dataset.

2.5 Model Accuracy Evaluation

The model evaluation procedure consists of three crucial elements. Firstly, by utilizing the 30% test samples set aside during the training phase, we measured the prediction accuracy of the model across diverse aquifer zones via evaluation metrics and carried out an analysis of the relative significance of feature variables. Secondly, cross-validation was carried out between the reconstructed outcomes and the original Groundwater Storage Anomaly (GWSA) data at the original spatial resolution (0.25°) to quantify the deviations from the baseline data. Finally, the high-resolution reconstruction results (0.008333°) were verified using global long-term groundwater monitoring well data (covering the period from 2006 to 2020) to systematically evaluate the overall accuracy of the downscaling products.

4.2.5.1 Statistical Evaluation Indicators

We assess the downscaledevaluated model through three metrics: performance using the coefficient of determination (R^2), the R^2 , Pearson correlation coefficient (CC), $RMSE$, and the root-mean squareabsolute error ($RMSE$). The coefficient of determination (R^2) denotes the model's explanatory capacity for the MAE . R^2 measures the fraction of variance of the dependent variable, signifying the proportion of the total variance that can be ascribed to the model. The correlation coefficient (CC) gauges the intensity of the linear correlation between the explained by the model. CC quantifies the strength of linear association between predicted and observed values, whereas the root $RMSE$ characterizes the overall magnitude of prediction error, and MAE represents the mean square error ($RMSE$) quantifies the overallabsolute deviation between the predicted and observed results. The calculation methods are as follows: values.

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - o_i)^2}{\sum_{i=1}^n (o_i - \bar{o})^2} \quad (16)$$

$$CC = \frac{\sum_{i=1}^n (y_i - \bar{y})(o_i - \bar{o})}{\sqrt{\sum_{i=1}^n (y_i - \bar{y})^2 \sum_{i=1}^n (o_i - \bar{o})^2}} \quad (17)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (o_i - y_i)^2}{n}} \quad (18)$$

where o represents the observed value; y represents the predicted value; n represents the sample size; and $\bar{o}$ and $\bar{y}$ are the means of o and y respectively.

~~2.5.2 In-Situ Cross Validation and Dimensional Conversion~~

~~To objectively evaluate the physical reliability of the 1 km downscaled global groundwater storage anomaly (GWSA) product, we collected and integrated in-situ groundwater records from 1,518 monitoring wells with long-term continuous observations worldwide for the period 2006–2020. This in-situ dataset integrates multiple high-quality observation networks, including national and regional monitoring stations reported in the China Groundwater Yearbook and the latest global groundwater-level database compiled by Jasechko et al. (2024). To suppress random high-frequency noise at the individual-well scale caused by local geological heterogeneity or irregular pumping, all valid well time series within the same 0.25° grid cell were arithmetically averaged to represent the regional groundwater-level anomaly dynamics at that grid scale (groundwater level anomaly, GWLA). However, the equivalent water height (EWH) anomalies derived from satellite gravimetry and the in-situ groundwater level anomalies (GWLA) differ in physical dimension, and their relationship can be expressed as:~~

2.5 in-situ observational validation

To evaluate the reconstructed GWSA product against groundwater-level observations, three complementary validation procedures were implemented. First, long-term groundwater-level trends were converted to equivalent-water-height groundwater-storage trends using a calibration-derived effective specific yield (S_{yeff}) within a spatially blocked calibration–validation framework. The monitoring network was partitioned into 1° spatial blocks, with approximately 70% of the blocks used for coefficient calibration and the remaining 30% retained for validation. All wells within the same block were assigned to the same subset, thereby reducing spatial dependence between coefficient estimation and validation. Second, annual and monthly groundwater-level observations were compared with reconstructed GWSA

without level-to-storage conversion using independently standardized time series to assess temporal co-variation. Third, the reconstructed product was compared with published groundwater-storage estimates for representative groundwater regions to evaluate temporal variability, magnitude, and long-term trends.

2.5.1 in-situ Validation with S_{yeff} Correction and Sensitivity Analysis

Monitoring sites with at least eight valid annual observations were retained for the long-term trend evaluation. Groundwater depth was converted to groundwater-level (GWL) change by reversing its sign, such that positive trends indicate groundwater-level recovery and negative trends indicate groundwater-level decline. For each monitoring site, the long-term GWL trend was estimated using the Theil-Sen slope estimator and expressed in cm yr^{-1} . The corresponding long-term GWSA trend was extracted from the 1 km reconstructed product at the same monitoring location.

Because groundwater-level change and groundwater-storage change have different physical dimensions, a scale-dependent effective specific yield, S_{yeff} , was introduced as a level-to-storage conversion coefficient. The purpose of S_{yeff} is to harmonize observation-derived GWL trends with GWSA trends expressed as equivalent water height; it is not intended to represent a direct measurement of local aquifer-specific yield.

To reduce information leakage between coefficient estimation and validation, the monitoring network was partitioned into 1° spatial blocks. Approximately 70% of the blocks were assigned to the calibration subset and the remaining 30% to the held-out validation subset. All wells belonging to the same block were retained in the same subset. S_{yeff} was estimated exclusively from observations in the calibration blocks and was subsequently transferred to monitoring sites in the held-out blocks without refitting or using observations from the validation subset.

Within the calibration subset, S_{yeff} was estimated from the multiplicative relationship between long-term GWL trends and the corresponding reconstructed GWSA trends:

$$\Delta GWS = \Delta GWL \times S_{yeff} \quad \Delta GWS_{obs} = S_{yeff} \times \Delta GWL_{obs} \quad (19)$$

Where ΔGWS and ΔGWL represent the anomaly change in groundwater storage and the anomaly change in observed groundwater level, respectively, both expressed in cm; S_{yeff} is the specific yield, which is used to convert between the two quantities and ensure unit consistency. Notably, given

the strong heterogeneity of deep aquifer systems at the global macroscale, directly introducing high-resolution spatial grids of S_x with substantial uncertainty would likely introduce spurious local variation noise during the validation process. Therefore, following widely adopted standards and robust evaluation practices in large-scale GRACE-based groundwater studies (Famiglietti et al., 2011; Rodell et al., 2009; Scanlon et al., 2012), we constructed a geographically zoned dynamic specific yield parameter matrix based on the hydrogeological settings of the major continental regions and representative aquifer systems worldwide (Table 2). Representative regional empirical mean S_x values ranging from 0.08 to 0.18 were then assigned to different latitude-longitude domains to linearly scale the observed groundwater level trends, thereby enabling an independent validation of the multi-year trends in the reconstructed GWSA. **Table 2**, where ΔGWS_{obs} is the GWS trend derived from the in-situ observations, ΔGWL_{obs} is the long term GWL trend, and S_{yeff} is the effective conversion coefficient used to convert GWL changes into equivalent GWS changes.

The converted in-situ GWS trends and reconstructed GWSA trends were then aggregated separately onto the original 0.25° GWSA grid. Only grid cells containing valid values from both datasets were retained. Anomalous values were removed using a three standard deviation criterion. CC , R^2 , $RMSE$, $Bias$, regression slope, intercept, and trend sign consistency were then calculated. The S_{yeff} coefficient was used only to harmonize the units of the in-situ validation data and did not contribute to the production of the GWSA dataset.

To evaluate uncertainty in the conversion coefficient, the calibration derived S_{yeff} values were perturbed by -30%, -15%, +15, and +30%. The complete calculation procedure was repeated for each perturbation scenario. Changes in correlation, error, regression amplitude, and trend sign consistency were used to quantify the influence of S_{yeff} uncertainty on the validation results.

2.5.2 Conversion Free in-situ Validation

Conversion free in-situ validation was conducted at annual and monthly scales. For the annual scale analysis, the GWL and reconstructed GWSA series for each aquifer were converted to anomalies relative to the 2004 to 2009 reference period, smoothed using a three year moving average to reduce short term interannual fluctuations, and independently standardized. Pearson correlation coefficients were then

compared among different aquifer categories and representative aquifer systems to evaluate the consistency of interannual trajectories and long term covariation without introducing uncertainty from the S_{yeff} conversion.

The monthly scale analysis was designed to further evaluate the ability of the reconstructed GWSA to reproduce long term trends, seasonal variations, and short term fluctuations. Monitoring sites with continuous observations throughout the analysis period were first selected. Aquifers containing more than five valid monitoring sites were then retained. Anomalous values in the continuous monthly series were removed using a three standard deviation criterion. The remaining station series were aggregated to the aquifer scale using the median, and the corresponding reconstructed GWSA series were extracted and aggregated over the same aquifer domains.

The aquifer scale observation and reconstructed GWSA series were independently standardized and decomposed using STL into long term trend, seasonal, and residual components. Pearson correlation coefficients were calculated for the original monthly series, trend components, and seasonal components. Seasonal contributions were calculated following the method described in Sect. 2.6. This procedure evaluates the consistency of monthly variability, long term evolution, and seasonal phase, and provides a basis for assessing the ability of the reconstruction to represent regional signals and local seasonal processes across different aquifer categories.

2.5.3 Regional Specific Yield (S_y) parameter matrix-Comparison with Published Estimates

Region (Inferred)	Latitude Range	Longitude Range	Specific Yield (S_y)
Asia	40°N to 80°N	60°E to 150°E	0.11
Europe & Middle East	35°N to 75°N	25°W to 60°E	0.14
North America	45°N to 75°N	170°W to 50°W	0.16
South America	60°S to 15°N	90°W to 30°W	0.18
Africa	35°S to 35°N	20°W to 55°E	0.08
Australia / Oceania	50°S to 0°	110°E to 180°E	0.09
Global Default (Other regions)	-	-	0.15

The 1 km GWSA product was compared with published estimates for the North China Plain, northwestern India, California, and the High Plains Aquifer. The spatial extent of each region was reconstructed as closely as possible to the corresponding reference study. Public numerical data were used preferentially, and published curves were digitized only when numerical data were unavailable. The

reconstructed and reference series were matched in time and adjusted to a common baseline according to the temporal coverage and data structure of each study.

Pearson *CC*, *RMSE*, *Bias*, standard deviation ratio, seasonal amplitude, and trend slope were used to evaluate regional temporal variability, amplitude agreement, and long term trend consistency. For studies reporting a specific analysis period, ordinary least squares trends were additionally calculated over the same time window. This regional comparison links the global 1 km reconstruction with well studied groundwater systems, including the North China Plain, northwestern India, California, and the High Plains Aquifer. It provides a regional assessment of the product's ability to represent persistent groundwater depletion, seasonal variability, and trend magnitude, and establishes a common basis for comparison with existing regional knowledge.

2.6 Seasonal Analysis

Hydrological time-series are generally affected by robust periodic seasonal rhythms, and seasonality molds the dynamics of the hydrosphere on a global scale. Seasonal-induced extreme hydrological anomalies can even notably counterbalance or intensify the effects of long-term trends (Cooley et al., 2021; Li et al., 2025). Nevertheless, research in this domain frequently neglects these aspects. Consequently, we employed the methodology put forward by Li et al. (2025), leveraging reconstructed ~~groundwater storage anomaly (GWSA)~~ sequences with high spatiotemporal resolution to quantify the seasonal contributions to global groundwater dynamics at the pixel scale.

$$C_{seasonality} = \frac{\bar{\sigma}}{\bar{\sigma} + \delta} \frac{\bar{\sigma}}{\bar{\sigma} + \delta} \quad (20)$$

where $\bar{\sigma}$ and δ represent the annual standard deviation of GWSA and the standard deviation of the multi-year average of GWSA for the period 2002–2020, respectively. $C_{seasonality}$ represents the relative seasonal contribution, with a value range of 0% to 100%. A value above 50% signifies that seasonality dominates the groundwater dynamics at that pixel.

2.7 Long-term Trend Analysis

To precisely extract the long-term evolution characteristics of ~~Groundwater Storage Anomaly (GWSA)~~, this study utilized a seasonal-trend decomposition approach based on local weighted regression (STL, Seasonal-Trend decomposition using LOESS) to conduct de-seasonalization of the

original time series. STL is a commonly employed non-parametric estimation method. Its fundamental premise is that time series are composed of seasonal components, long-term trends, and residuals (Cleveland et al., 1990). By separating the seasonal term S_t from the original time series Y_t , the deseasonalized sequence X_t is acquired. In comparison with traditional linear regression based on the least-squares method, its non-parametric slope calculation exhibits greater robustness towards missing data and outliers in the time series (Sen, 1968; Theil, 1950).

$$X_t = Y_t - S_t \quad (21)$$

where the seasonally adjusted X_t series comprehensively preserves both the trend component signal of the groundwater storage anomaly (GWSA) and the natural residual component signal, which encompasses extreme hydrological events. On this basis, the Theil-Sen Median Trend method is utilized to ascertain the long-term variation trend. The calculation method is as follows:

$$\beta = \text{Median}\left(\frac{X_j - X_i}{j - i}\right), \quad \forall j > i \quad (22)$$

where X_j and X_i represent the deseasonalized GWSA values at months j and i , respectively; β is the median slope of all pairwise combinations in the time series; $\beta > 0$ indicates an increasing trend in groundwater storage, whereas $\beta < 0$ indicates a decreasing trend.

The Mann-Kendall Significance Test serves as a crucial complement to Theil-Sen trend estimation and is a non-parametric test method recommended by the World Meteorological Organization (WMO). Its advantage lies in not requiring sample data to follow a specific distribution, making it widely applicable for trend significance testing in hydro-meteorological series (Kendall, 1975; Mann, 1945). The calculation methodology can be referenced in (Wang et al., 2022). This study conducted trend tests for GWSA changes at significance levels of $p < 0.05$ and $p < 0.01$.

3 Results

3.1 Model ~~Performance Comparison~~ and ~~Multi-scale~~ Validation

3.1.1 Temporal Continuity and SSA-based Reconstruction Quality

To ensure the spatiotemporal continuity of the downscaling baseline without ~~relying on~~ ~~resorting to~~ artificial annual averaging, ~~Singular Spectrum Analysis (SSA)~~ was employed to reconstruct missing

GRACE observations. ~~The reconstruction quality was evaluated using two error metrics corresponding to the SSA-filling-a and SSA-filling-b phases~~ (Fig. 2). Global evaluation ~~showed~~confirms that the SSA framework successfully maintained low reconstruction errors while preserving the integrity of the original temporal dynamics across two distinct gap-filling phases.

During the SSA-filling-a phase, the reconstruction error was quantified as the standard deviation of residuals between observed and reconstructed values over all non-missing months. Globally, the error ranged from 0.232 to 62.458 cm, with a tightly constrained median of 1.156 cm. The 1% and 99% percentiles were 0.475 and 7.550 cm, respectively, indicating robust reconstruction precision during this stage.

~~During~~To empirically represent the ~~SSA-filling-b phase, which was designed to evaluate reconstruction performance capability~~ under conditions of long-term continuous missing data gaps, a rigorous cross-validation approach was implemented in the second phase. Specifically, entire years of data within known observation periods were sequentially masked and reconstructed. At the global scale, the median error for this phase was 1.679 cm, with the 1%1st and 99%99th percentiles at 0.732 cm and 10.204 cm, respectively. OverFocusing specifically on terrestrial regions ~~(excluding Antarctica)~~, the median error ~~increased to was~~ 2.708 cm, with ~~the 1%1st and 99%99th~~ percentiles of 0.862 cm and 15.295 cm, ~~respectively. Although reconstruction. While the errors over in land were areas are~~ slightly higher than the global average, they ~~remained~~remain consistently low and ~~within an acceptable range bounded~~.

Spatially, the high-error centers across both phases exhibited high consistency, primarily clustering in the high-latitude cryosphere of the Northern Hemisphere, the Amazon Basin, southern South America, East Africa, and Southeast Asia. Although uncertainties in these specific regions are relatively higher, these variations are primarily ~~driven by~~associated with large natural signal amplitudes, significant seasonal oscillations, and complex local processes rather than algorithmic instability. Consequently, the SSA reconstruction provides a ~~physically reliable, temporally~~ continuous ~~temporal~~-baseline with quantified reconstruction uncertainty for the subsequent high-resolution downscaling.

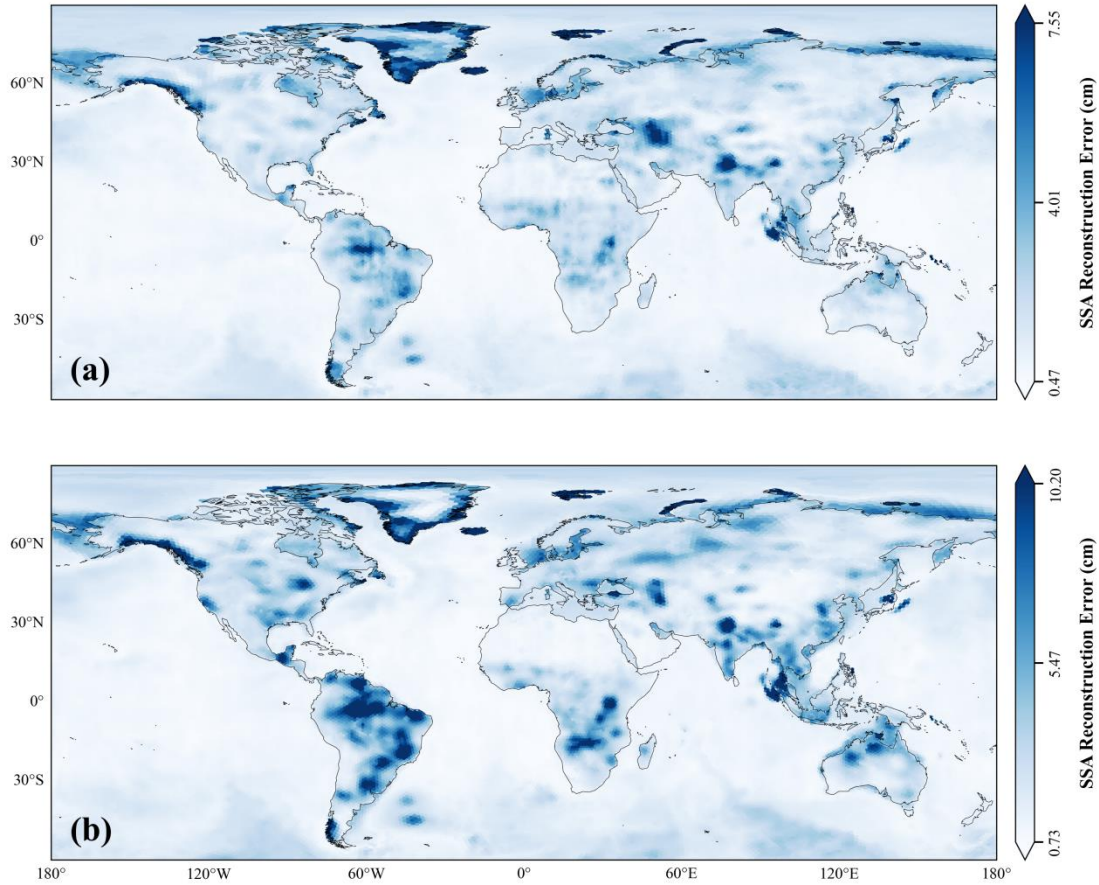


Figure 2. Evaluation of the reconstruction accuracy of Singular Spectrum Analysis (SSA).

(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.

3.1.2 ~~Internal Consistency Across Global~~ Aquifer Systems

To guarantee the robustness ~~stratified Model Comparison~~ and spatial consistency of the final dataset across various hydro-climatic zones, an assessment based on the WHYMAP Global Mega Aquifer Systems was carried out. In structurally stable or naturally climate-driven porous aquifer systems (e.g., AQU 11 and AQU 12), the downscaled outcomes exhibited high accuracy with negligible absolute errors. For example, in AQU 11, the selected XGBoost architecture attained a Root Mean Square Error (RMSE) of 1.15 mm. Nevertheless, in regions exposed to intensive agricultural irrigation and over-exploitation of fossil groundwater (e.g., AQU 22 and AQU 23), the necessity of our architectural selection became evident. The adopted tree-based framework (XGBoost) achieved highly compact error convergence (e.g., $R^2 = 0.90$, $RMSE = 2.56$ mm in the Indus Basin), demonstrating significantly superior performance compared to the continuous DNN baseline ($R^2 = 0.85$, $RMSE = 3.23$ mm). This quality control procedure validates that the tree-based ensemble has an inherent structural advantage in capturing the sharp

boundaries caused by localized human pumping, thus justifying its function as the core engine for our dataset production. **Multi-scale Validation**

Table 3. Cross-validation performance by aquifer system and mask.

Table 2 first compares the two candidate architectures within the random 70%/30% development split for each WHYMAP mega-aquifer system or specific mask. XGBoost yielded lower RMSE and MAE and higher R^2 than DNN in all 11 listed aquifer systems; for AQU 11, for example, XGBoost achieved RMSE = 1.15 cm and $R^2 = 0.921$, compared with 1.61 cm and 0.845 for DNN. These values are within-distribution diagnostics used for model comparison, while product reliability is evaluated below through cross-scale consistency and spatially held-out in-situ observations.

Table 2. Internal-test performance of the XGBoost and DNN downscaling models across major global mega-aquifer systems and specific masks.

Aquifer ID	Mega-Aquifer System / Specific Mask	XGBoost			DNN		
		RMSE	R^2	MAE	RMSE	R^2	MAE
11	Cuvelai-Upper Zambezi Basin	1.15	0.921	0.76	1.61	0.845	1.11
12	Stampriet-Kalahari Basin	1.72	0.929	1.16	2.34	0.868	1.66
13	Karoo Basin	2.37	0.94	1.62	2.93	0.908	2.07
14	Northern Great Plains Aquifer	2.8	0.902	1.91	3.25	0.867	2.3
15	Cambrian-Ordovician Aquifer	3.88	0.942	2.51	4.34	0.928	2.93
22	Arabian Aquifer System	1.95	0.9	1.31	2.42	0.846	1.74
23	Indus Basin	2.56	0.903	1.78	3.23	0.846	2.32
24	Ganges-Brahmaputra Basin	3.17	0.916	1.99	3.96	0.87	2.64
25	West Siberian Basin	4.42	0.926	2.42	5.91	0.867	3.42
33	Russian Platform Basins	4.12	0.916	2.32	4.9	0.881	2.93
34	North Caucasus Basin	5.4	0.943	3.4	5.97	0.93	3.88

3.1.3 Cross-scale Consistency and Independent In-situ Validation

To precisely evaluate the mass conservation and spatial fidelity of the resolution enhancement process, the 1 km downscaled GWSA results were spatially aggregated to the original 0.25° resolution and cross-validated against the original 2002–2020 mean observations (Fig. 3). The results demonstrate high consistency across both modeling architectures ($R^2 > 0.8$), effectively reproducing the long-term spatial patterns of macroscopic groundwater storage anomalies.

grid confirmed the same overall ranking (Fig. 3). At the global scale, the Extreme Gradient Boosting (XGBoost) model outperformed the Deep Neural Network (DNN) model. As illustrated in the global panels of Fig. 3, XGBoost achieved an R^2 of 0.972, a correlation coefficient (CC) of 0.986, and a root-

mean-square error ($RMSE$) of ≈ 2.097 cm. In comparison, the DNN model achieved an R^2 of, compared with $R^2 = 0.961$, a CC of ≈ 0.981 , and an $RMSE$ of ≈ 2.661 cm for DNN. Across the 11 evaluated aquifer categories, XGBoost exhibited achieved higher R^2 and CC values, alongside and lower $RMSE$, in 10 categories, indicating a more accurate characterization of spatial distributions with reduced systemic bias. Further spatial breakdown reveals that, Major groundwater basins (AQUs 11–15) and regions with complex hydrogeological structures (AQUs 22–25) showed superior fitting accuracy, maintaining maintained an average R^2 of 0.961 and an $RMSE$ of 0.730 cm. Although localized, whereas local and shallow aquifers (AQUs 33–34) exhibited slightly greater uncertainty with an had a higher average $RMSE$ increasing to of 2.091 cm, this variance is physically plausible. Such regions are heavily influenced by localized micro topography, intensive land use changes, and complex near surface recharge processes, which inherently introduce high frequency spatial variations. Consequently, the highly robust XGBoost architecture was selected to generate the final dataset for subsequent in-depth analysis.

Beyond internal, The cross-scale consistency, rigorous independent validation was conducted utilizing long term physical evolutionary trends (2006–2020) from 1,518 global groundwater monitoring stations (Fig. 4). To systematically resolve the dimensional discrepancy between satellite derived Equivalent Water Height (EWH) and in situ Groundwater Level (GWL), a geographically zoned Specific Yield (S_y) matrix was introduced. Furthermore, quantile binning was applied to the grid trends to mitigate high frequency local noise caused by heterogeneous geological conditions or isolated pumping events, thereby extracting representative macroscopic evolution patterns assessment therefore shows that the relative performance advantage of XGBoost is retained when the 1 km reconstructions are evaluated against the original-resolution GWSA.

As shown in Fig. 4, after eliminating high frequency spatial noise, both downscaling architectures demonstrated robust linear relationships with the ground measured trends ($p < 0.01$). However, they exhibited physically significant differences in overall fitting accuracy and error magnitude. The XGBoost model (Fig. 4a) demonstrated a superior capacity to explain the variance of observed trends, yielding an R^2 of 0.44 and a lower $RMSE$ of 0.51 cm yr^{-1} . Notably, its scattered data points were densely and symmetrically distributed around the 1:1 reference line. In the context of dynamic range analysis, XGBoost effectively captured deep groundwater depletion signals driven by intensive anthropogenic activities ($< 1.0 \text{ cm yr}^{-1}$) and successfully reproduced reservoir recovery characteristics in the positive ranges. In contrast, although the DNN model (Fig. 4b) maintained a comparable linear correlation (CC

=0.67), it yielded a lower explained variance ($R^2=0.39$) and a higher prediction error ($RMSE=0.57$ cm yr^{-1}). Moreover, the DNN predictions exhibited greater deviations from the reference line at extreme depletion and recharge values.

When interpreting these validation metrics, it is imperative to acknowledge the inherent point-to-pixel scale mismatch. In situ wells record micro scale groundwater fluctuations influenced by highly localized fractured media or individual pumping events, whereas the downscaled GWSA represents spatially averaged storage dynamics over a 1 km grid cell. Given these significant dimensional and physical scaling challenges, an R^2 of 0.44 confirms the strong efficacy of the tree-based architecture in reconstructing core macroscopic evolutionary trends, rather than attempting precise point scale replication.

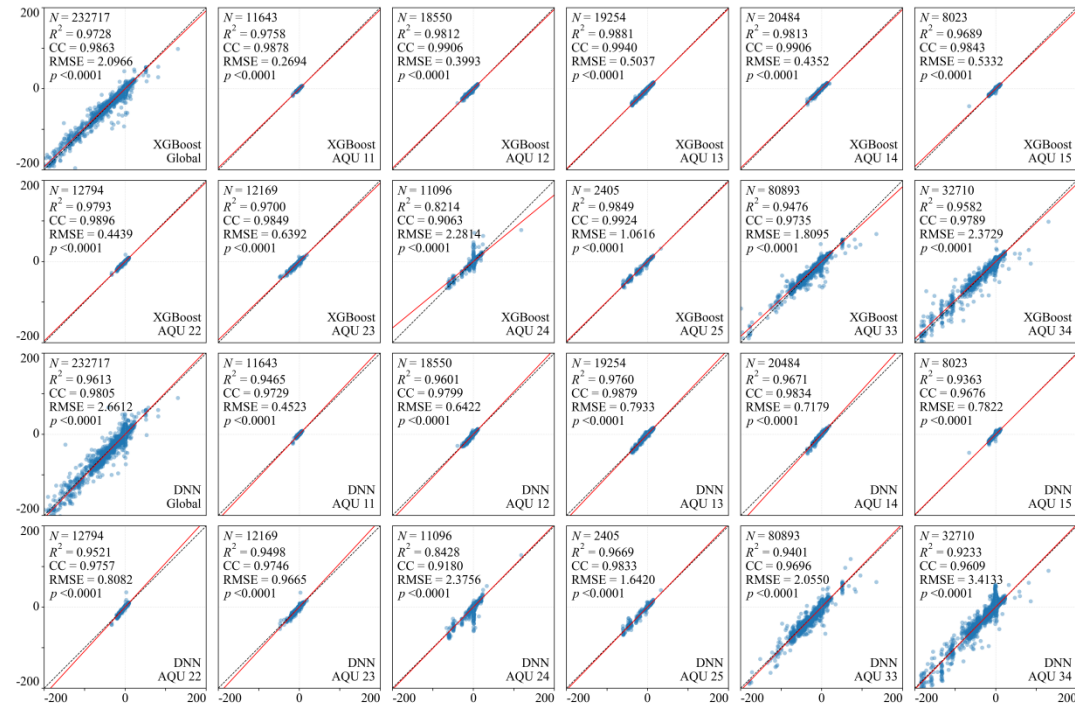


Figure 3. Cross validation of the downscaled groundwater storage anomalies (GWSA) and the original-resolution GWSA multi-year mean for 2002–2020.

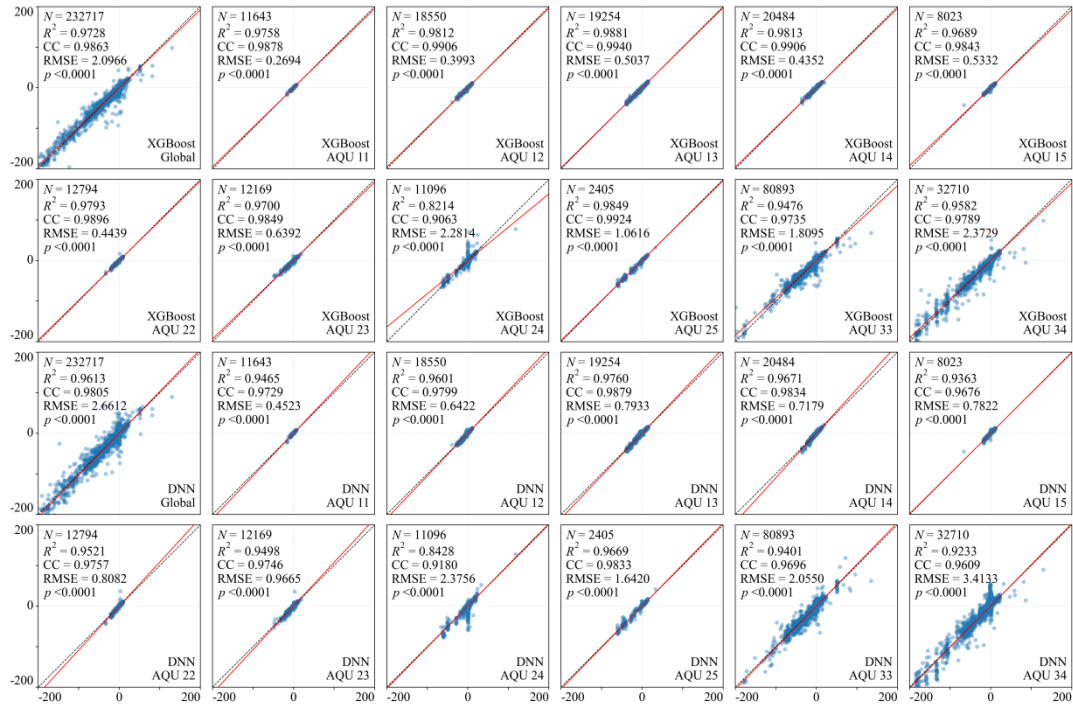


Figure 3. cross-scale consistency comparison of the downscaled GWSA and the original-resolution GWSA multi-year mean for 2002–2020.

The first and second rows show the XGBoost results, and the third and fourth rows show the DNN results. Panels are arranged in the following order: global, Major groundwater basins (AQUs 11–15), Complex hydrogeological structures (AQUs 22–25), and Local and shallow aquifers (AQUs 33–34). In each panel, the x axis represents the original GWSA, and the y axis represents the downscaled GWSA aggregated to the same resolution. The red line indicates the fitted line. The corresponding sample size (N), coefficient of determination (R^2), correlation coefficient (CC), root mean square error ($RMSE$), and significance level (p) are shown in the upper-left corner.

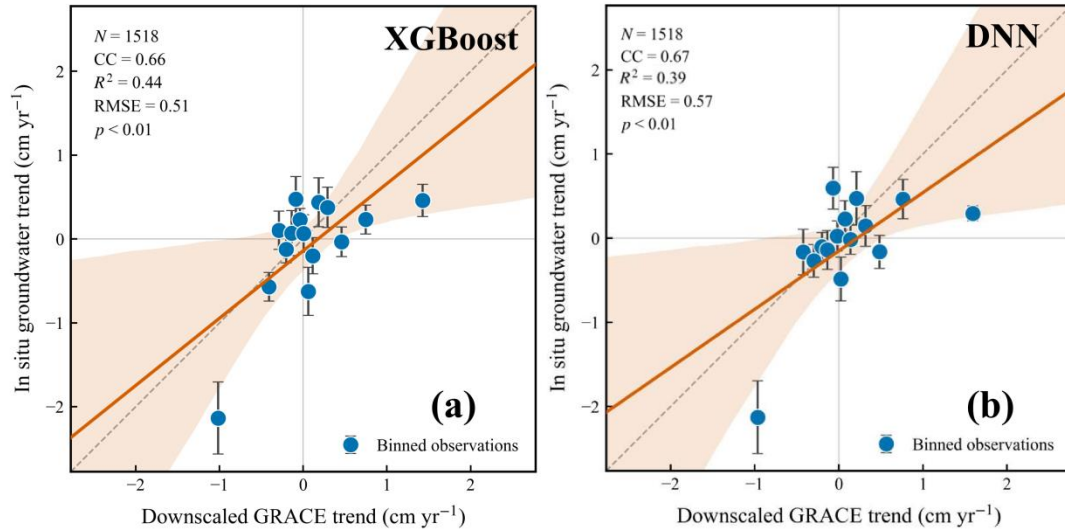


Figure 4. Cross-validation of downscaled groundwater storage anomalies (GWSA) and long-term evolutionary trends of ground-measured groundwater levels (2006–2020).

Figures (a) and (b) respectively present the independent validation results for the XGBoost and DNN models. The x axis represents the long term trends of GWSA predicted by the downscaling models, while the y axis displays the trends of measured groundwater equivalent water height (EWH) converted through geographic Specific Yield (S_y) matrices, with the unit being cm yr^{-1} . To mitigate the heterogeneity resulting from individual well observations and high frequency noise, 1,518 validation grids with continuous observational records were globally aggregated into 15 bins (Binned observations, blue dots). The blue dots indicate the arithmetic means of grid trends within each bin, and the vertical error bars represent the standard errors of the mean (SEM) for the measured EWH trends. The gray dashed line serves as the 1:1 reference line, and the dark orange solid line and its light translucent shaded area respectively represent the best linear fit and the 95% confidence interval. Comprehensive statistical evaluations reveal that, despite having comparable linear correlations, XGBoost ($CC = 0.66$, $R^2 = 0.44$, $RMSE = 0.51 \text{ cm yr}^{-1}$, $p < 0.01$) significantly outperforms DNN ($CC = 0.67$, $R^2 = 0.39$, $RMSE = 0.57 \text{ cm yr}^{-1}$, $p < 0.01$) in minimizing prediction errors and explaining trend variances. This validates its superior robustness and accuracy in reproducing the true physical evolutionary dynamics of groundwater storage at the global macro scale.

3.2A global observation-constrained in-situ evaluation was conducted using long-term groundwater-monitoring records. Following the 1° spatial-block design described in Sect. 2.5.1, the comparison was performed on retained 0.25° validation grids after dimensional harmonization, common-grid aggregation, and uniform quality control. Figure 4 summarizes the resulting grid-level agreement for the DNN and XGBoost products.

For DNN, CC was 0.72, corresponding to $R^2 = 0.52$, with $RMSE = 0.34 \text{ cm yr}^{-1}$ and $Bias = 0.09 \text{ cm yr}^{-1}$. The fitted regression slope and intercept were 0.51 and -0.04 cm yr^{-1} , respectively, and trend-sign consistency was 0.72. XGBoost exhibited slightly stronger spatial covariance and directional consistency, with $CC = 0.73$, $R^2 = 0.54$, $RMSE = 0.38 \text{ cm yr}^{-1}$, $Bias = 0.11 \text{ cm yr}^{-1}$, a fitted slope of 0.51, an intercept of -0.05 cm yr^{-1} , and trend-sign consistency of 0.73. Although slopes below unity indicate residual differences in the absolute magnitude of long-term groundwater-storage change, both models reproduced the dominant spatial covariance and direction of groundwater depletion and recovery across the monitoring network. XGBoost showed slightly stronger correlation and trend-direction agreement, whereas DNN exhibited marginally lower absolute-error statistics. Together with the cross-scale consistency assessment, these results support the use of XGBoost as the production architecture without implying uniform superiority across every metric.

The in-situ evaluation complements the cross-scale consistency assessment by providing an external comparison with groundwater-monitoring records. The 0.25° grid-based implementation reduces the mismatch between point-scale well observations and spatially integrated GWSA estimates, while the Z-score comparisons without S_{yeff} in Figs. 6–8 provide an additional assessment of temporal co-variation.

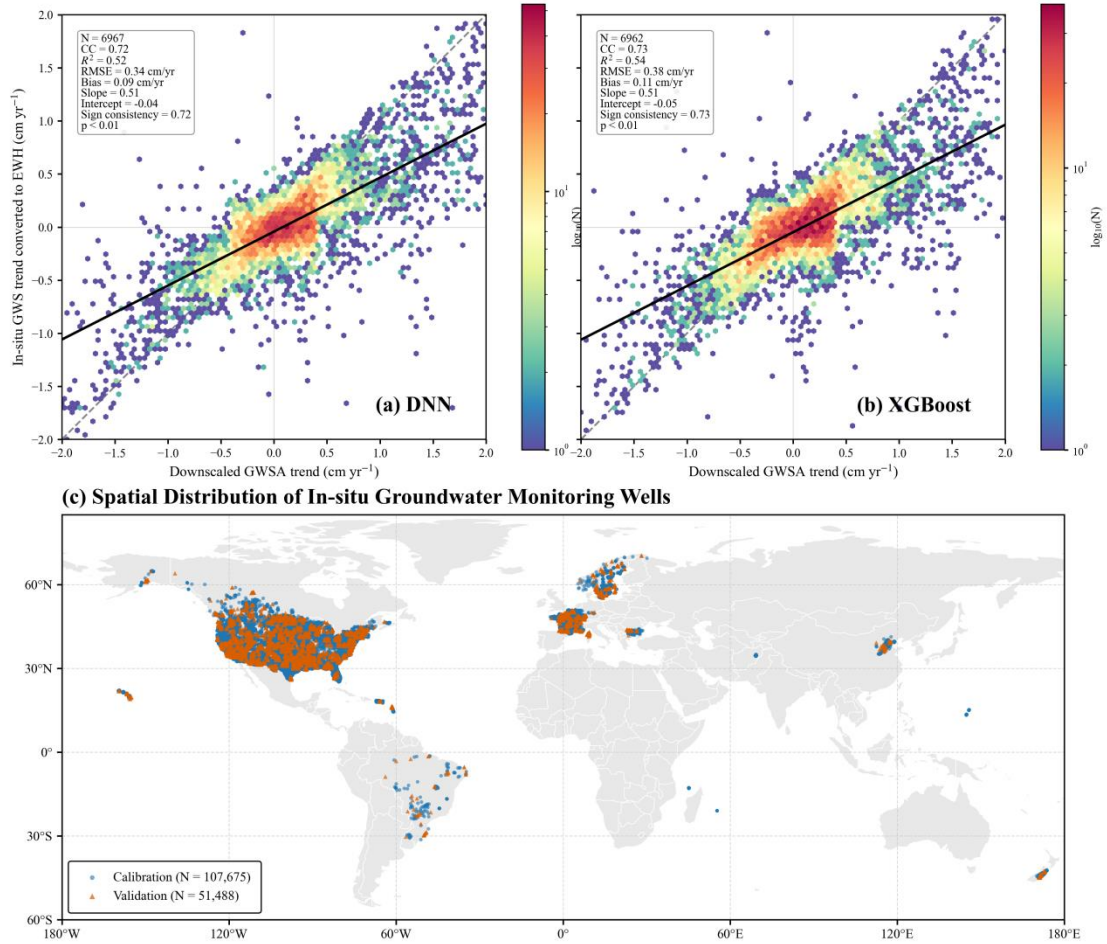


Figure 4. Spatially independent, quality-controlled in-situ evaluation of long-term downscaled groundwater-storage trends and spatial distribution of the monitoring network.

(a) Held-out 0.25° grid comparison between DNN-derived GWSA trends and groundwater-level trends converted to equivalent water height using S_{eff} 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_{eff} 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.

The sensitivity of the in-situ evaluation to S_{veff} was further quantified by perturbing the baseline conversion coefficients by $\pm 15\%$ and $\pm 30\%$ (Table 3). For the XGBoost product, the baseline evaluation yielded $CC = 0.734$, $R^2 = 0.538$, $RMSE = 0.378 \text{ cm yr}^{-1}$, $Bias = 0.106 \text{ cm yr}^{-1}$, and a regression slope of 0.506. Across the full $\pm 30\%$ perturbation range, CC varied from 0.691 to 0.734 and R^2 from 0.477 to 0.538, whereas $RMSE$ remained within 0.359–0.408 cm yr^{-1} and $Bias$ within 0.104–0.107 cm yr^{-1} . The regression slope showed a comparatively stronger response, ranging from 0.379 to 0.593, while trend-sign consistency remained between 0.696 and 0.743. DNN exhibited a similar response, with R^2 ranging from 0.476 to 0.522 and $RMSE$ from 0.343 to 0.393 cm yr^{-1} . These results indicate that uncertainty in S_{veff} has a stronger influence on amplitude-related quantities, particularly the regression slope, than on the overall association and trend-direction agreement between the modeled and in-situ-derived groundwater-storage trends.

Table 3

Sensitivity of the global in-situ evaluation to perturbations in S_{veff} . 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_{veff} 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
	-30%	0.691	0.477	0.408	0.105	0.379	0.696
XGBoost	-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

Taken together, the aquifer-stratified internal diagnostics, cross-scale consistency assessment, and spatially independent in-situ evaluation support XGBoost as the production architecture. XGBoost provided more consistent performance across aquifer categories and stronger global cross-scale agreement, while DNN retained slightly lower absolute errors in the in-situ trend comparison. The selection therefore reflects the combined evidence rather than uniform superiority on every metric.

3.1.3 Sensitivity to Target Product

The target-product experiments show that model fit was reproducible but depended on the source product (Fig. 5). 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 produced the highest R^2 in 10 categories,

while CSR and H08 each ranked first in one category. The full ranges were 0.900–0.994 for CSR, 0.915–0.996 for JPL, 0.794–0.924 for WaterGAP, and 0.828–0.941 for H08. All four products produced high within-product fit, with systematic differences among target sources.

Predictor importance also depended on the target source. Static physiography and soil variables accounted for 0.638 and 0.620 of total importance in the CSR and JPL experiments, compared with 0.454 and 0.442 for WaterGAP and H08. Conversely, temporal encoding contributed 0.102 and 0.107 for CSR and JPL but increased to 0.191 and 0.225 for WaterGAP and H08. Feature-importance profiles were similar within the two GRACE-derived targets (Spearman $\rho = 0.954$) and within the two GHM targets ($\rho = 0.909$), but cross-family correlations were lower ($\rho = 0.382$ – 0.425). Reconstruction performance was therefore consistent across products, but the importance assigned to static heterogeneity and temporal variation depended on the target and represents a source-related structural uncertainty.

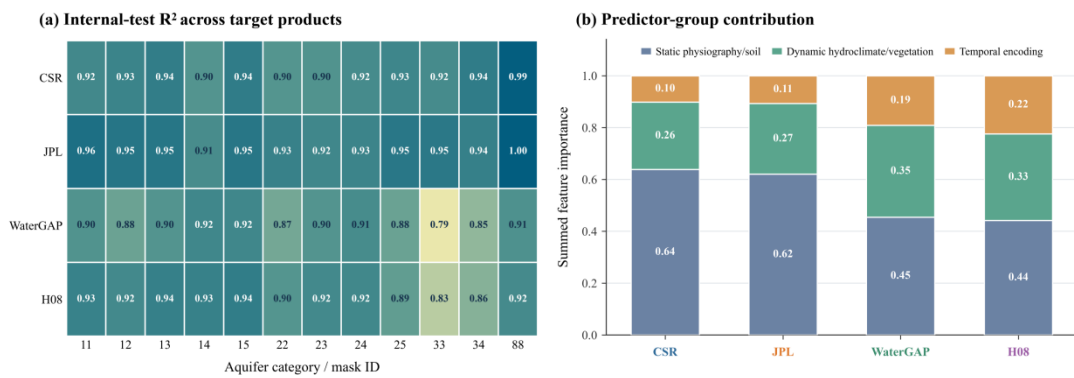


Figure 5. Controlled sensitivity of the aquifer-stratified XGBoost downscaling workflow to alternative target products.

(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.

3.2 Validation Across Aquifer Basins

The in-situ assessment contains two conceptually distinct components. The S_{veff} -based analysis presented above converts groundwater-level trends to equivalent-water-height storage trends, allowing

their magnitudes to be compared in common physical units. This conversion is used for dimensional harmonization and for evaluating magnitude consistency, but it is not a prerequisite for assessing whether the downscaled product reproduces the observed temporal evolution of groundwater. The basin-scale analyses below therefore use independently standardized GWL and GWSA series without applying S_{yeff} . These conversion-free comparisons constitute the main temporal validation in this section and directly evaluate interannual variability and long-term co-variation between the downscaled product and in-situ observations. The S_{yeff} -based and standardized analyses should thus be viewed as complementary rather than sequential validation steps.

3.2.1 Conversion-free In-situ Temporal Comparisons

To assess temporal fidelity under diverse hydrogeological and anthropogenic settings, complementary in-situ comparisons were categorized into climate-driven, anthropogenic-dominated, and data-limited basins (Figs. 6–8). These analyses compare independently standardized GWL and downscaled GWSA time series, and thus provide a conversion-free assessment of interannual variability and long-term temporal co-variation.

In natural, climate-dominated regions (Fig. 6), the downscaled dataset demonstrates exceptional temporal alignment with ground observations. For instance, within the Stampriet-Kalahari Basin (Aquifer 12, $N = 7849$), the localized time-series comparison yielded a high $CC = 0.82$, accurately reproducing the multi-year depletion trajectory from 2004 to 2014 and subsequent recovery (Fig. 6b). Similarly, robust agreement was observed in the Cuvelai-Upper Zambezi Basin (Aquifer 11, $N = 1713$, $CC = 0.72$, Fig. 6c), the Russian Platform Basins (Aquifer 33, $N = 32340$, $CC = 0.67$, Fig. 6f), and the Ganges-Brahmaputra Basin (Aquifer 24, $N = 2208$, $CC = 0.54$, Fig. 6d). Although the Arabian Aquifer System (Aquifer 22, $N = 1377$, Fig. 6e) exhibited a weaker linear correlation ($CC = 0.07$) due to complex local fossil groundwater dynamics, the overall multi-year deficit trend remains observable. These findings confirm that the machine learning framework effectively translates macro-scale climatic forcing into high-resolution, localized groundwater storage variations.

Conversely, evaluations in intensive anthropogenic basins (Fig. 7) reveal distinct dynamic behaviors. Notably, in the Indus Basin (Aquifer 23, $N = 1383$, Fig. 7c), the application of the localized mask and variance-normalized Z-scores demonstrates a highly consistent temporal trajectory ($CC = 0.58$), successfully capturing the severe depletion trough around 2013 and the subsequent recovery phase driven

by regional water management. Meanwhile, in the Northern Great Plains Aquifer (Aquifer 14, $N = 10052$, $CC = -0.23$, Fig. 7b) and the North Caucasus Basin (Aquifer 34, $N = 3971$, $CC = -0.01$, Fig. 7e), predicted climate-driven baselines indicate storage recovery or dynamic stability responding to favorable meteorological conditions, whereas in-situ observations exhibit sustained depletion. This decoupling indicates that the downscaled GWSA and local monitoring wells may respond differently in intensively managed aquifers. The resulting discrepancy is therefore interpreted as a signal potentially consistent with anthropogenic groundwater depletion rather than as a direct estimate of groundwater extraction.

In data-limited or complex regions (Fig. 8), such as the Cambrian-Ordovician Aquifer (Aquifer 15, $N = 634$, $CC = 0.24$, Fig. 8b) and the West Siberian Basin (Aquifer 25, $N = 9$, $CC = 0.48$, Fig. 8c), the validation metrics reflect the inherent challenges of sparse observation networks and complex hydrogeological constraints, yet the downscaled product still manages to capture the broad multi-year wave patterns corresponding to the limited in-situ records.

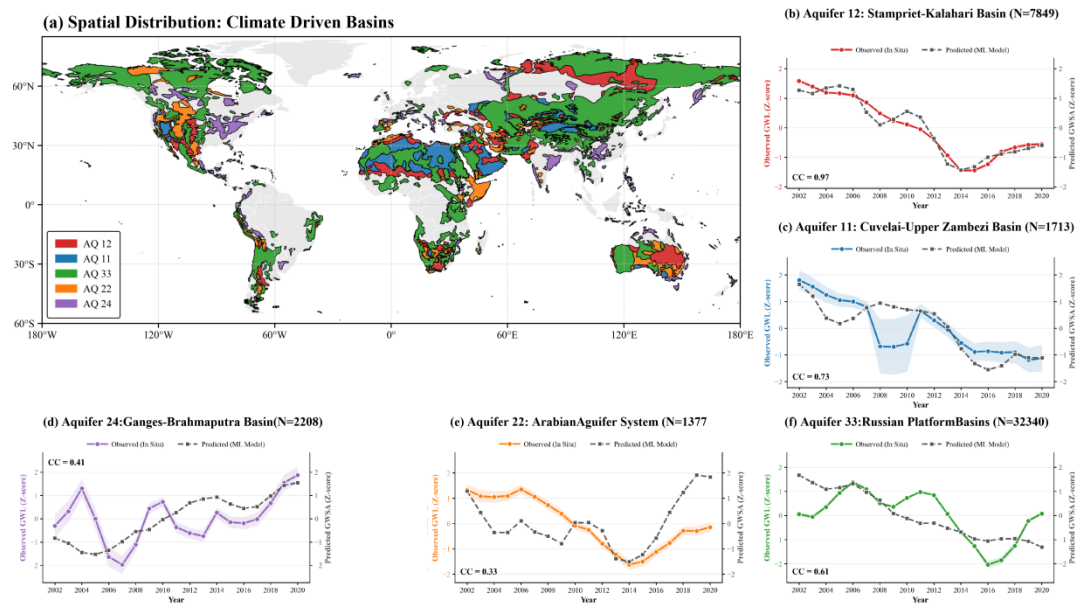


Figure 6. Spatial distribution and interannual variations of GWL and downscaled GWSA for climate-driven/natural-dominated major aquifer basins.

(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 CC denotes the Pearson correlation coefficient between the standardized observations and predictions.

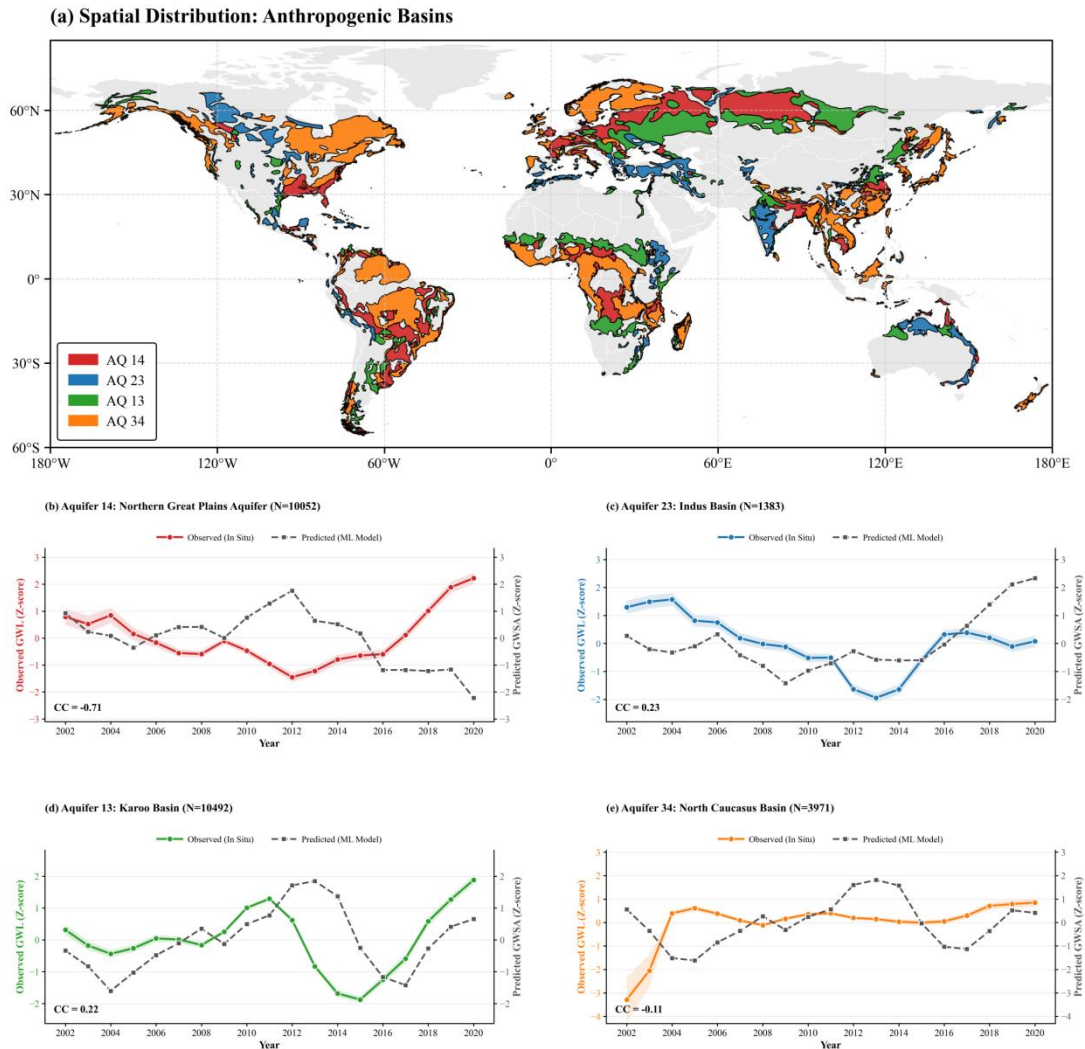


Figure 7. Spatial distribution and interannual variations of groundwater dynamics for anthropogenic-dominated major aquifer basins.

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

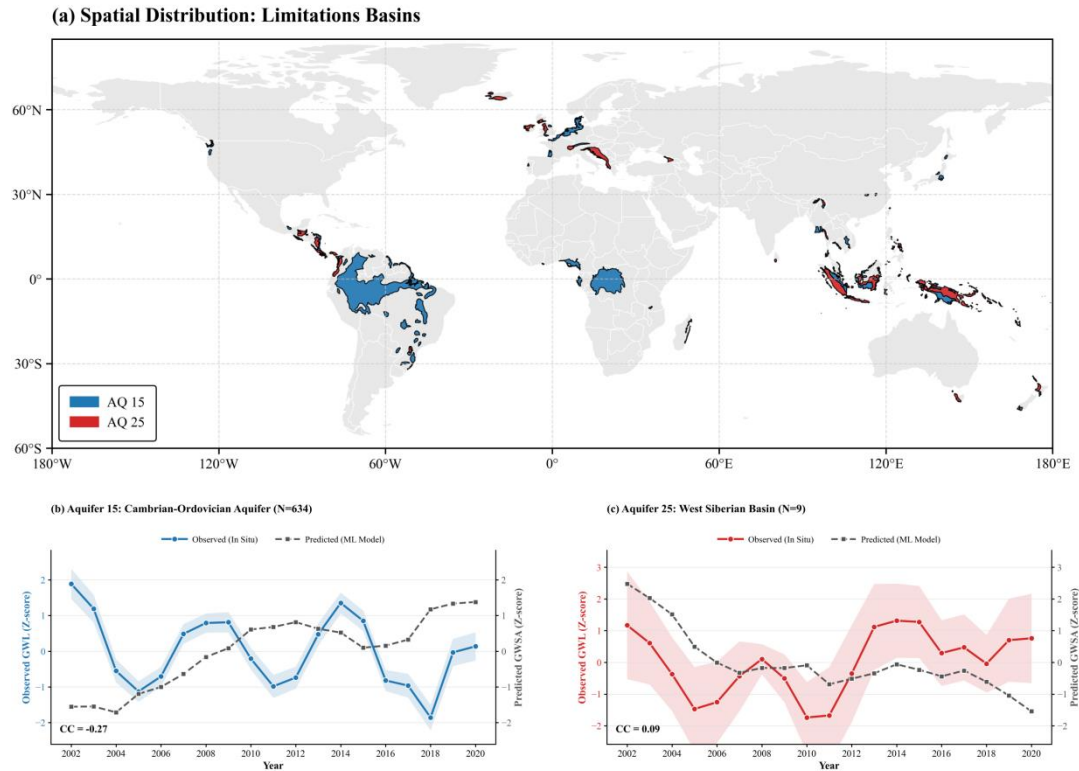


Figure 8. Spatial distribution and interannual variations of groundwater anomalies for aquifers under physical limitations or specific climatic-hydrogeological constraints.

(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.

3.2.2 Monthly Validation of Representative Aquifers

Figure 9 provides a monthly-scale validation of the reconstructed GWSA using five representative aquifer-site groups distributed across North America, Europe, and East Asia. The maps show the aquifer masks and the available monitoring sites, while the time-series panels decompose the standardized station

and reconstructed records into STL trend, seasonal, and residual components. Four records span April 2002–December 2020; the ID 1417 record spans 2005–2017 because of its available observations.

To further quantify the monthly-scale consistency between the reconstructed GWSA and in-situ observations, we first aggregated the monitoring-site records to the aquifer scale using the median method. The GWSA and aquifer series were then standardized to remove dimensional differences. Finally, STL decomposition was applied to separate the original series into long-term trend and seasonal components, and correlation coefficients for the original, trend, and seasonal series, together with the seasonal contribution, were calculated.

Overall, the mean correlation coefficients across the five groups were 0.51, 0.55, and 0.39 for the original, trend, and seasonal series, respectively. This indicates that the reconstruction reproduces long-term variability slightly better than monthly and seasonal fluctuations. The highest mean correlations were obtained for major groundwater basins (AQU 13; 0.75, 0.85, and 0.13), followed by aquifers with complex hydrogeological structures (AQU 24; 0.41, 0.57, and 0.46). Local and shallow aquifers (AQU 34) showed the lowest correlations for the original and trend series (0.33, 0.27, and 0.61). These differences indicate that coherent basin-scale storage variations favor agreement with in-situ observations, whereas local hydrogeological heterogeneity, strong seasonality, and anthropogenic perturbations reduce temporal consistency.

Major groundwater basins (AQU 13) include the High Plains Aquifer (ID 1308) and the North China Plain (ID 1417), with correlation coefficients of (0.67, 0.76, and 0.00) and (0.82, 0.91, and 0.26), respectively. In the High Plains Aquifer, the reconstruction captures the transition from early depletion to later recovery, and its seasonal contribution is close to that of the in-situ observations (0.27 and 0.23). However, the near-zero seasonal correlation suggests a phase offset or lag in the seasonal cycle. In the North China Plain, the reconstruction reproduces the persistent groundwater deficit and shows a similar seasonal contribution to the in-situ observations (0.22 and 0.26), probably because the large alluvial plain produces a relatively coherent regional response despite intensive irrigation and groundwater abstraction. Aquifers with complex hydrogeological structures (AQU 24) are represented by the Floridan Aquifer (ID 1669), for which the correlations are (0.41, 0.57, and 0.46). The reconstructed and in-situ seasonal contributions are relatively close (0.40 and 0.47), whereas complex carbonate-aquifer architecture and spatially variable recharge may weaken the correspondence between local observations and the regional reconstruction.

Local and shallow aquifers (AQU 34) include the Apennine Peninsula aquifer systems (ID 1193) and the Scandinavian Peninsula aquifer systems (ID 426), with correlation coefficients of (0.36, 0.02, and 0.70) and (0.29, 0.51, and 0.51), respectively. In the Apennine region, the reconstruction retains substantial seasonal co-variation but fails to reproduce the long-term trend. This may reflect strong local recharge contrasts associated with mountainous terrain, carbonate and fractured aquifers, together with local pumping effects. In the Scandinavian region, the reconstruction partly captures the overall trend, but its monthly and seasonal agreement is weaker, probably because of pronounced snowmelt and freeze–thaw cycles, shallow groundwater responses, sparse monitoring, and local hydrogeological heterogeneity. Thus, the low original-series correlations in both AQU 34 regions mainly reflect the influence of local processes and high-frequency seasonal variability on monthly matching.

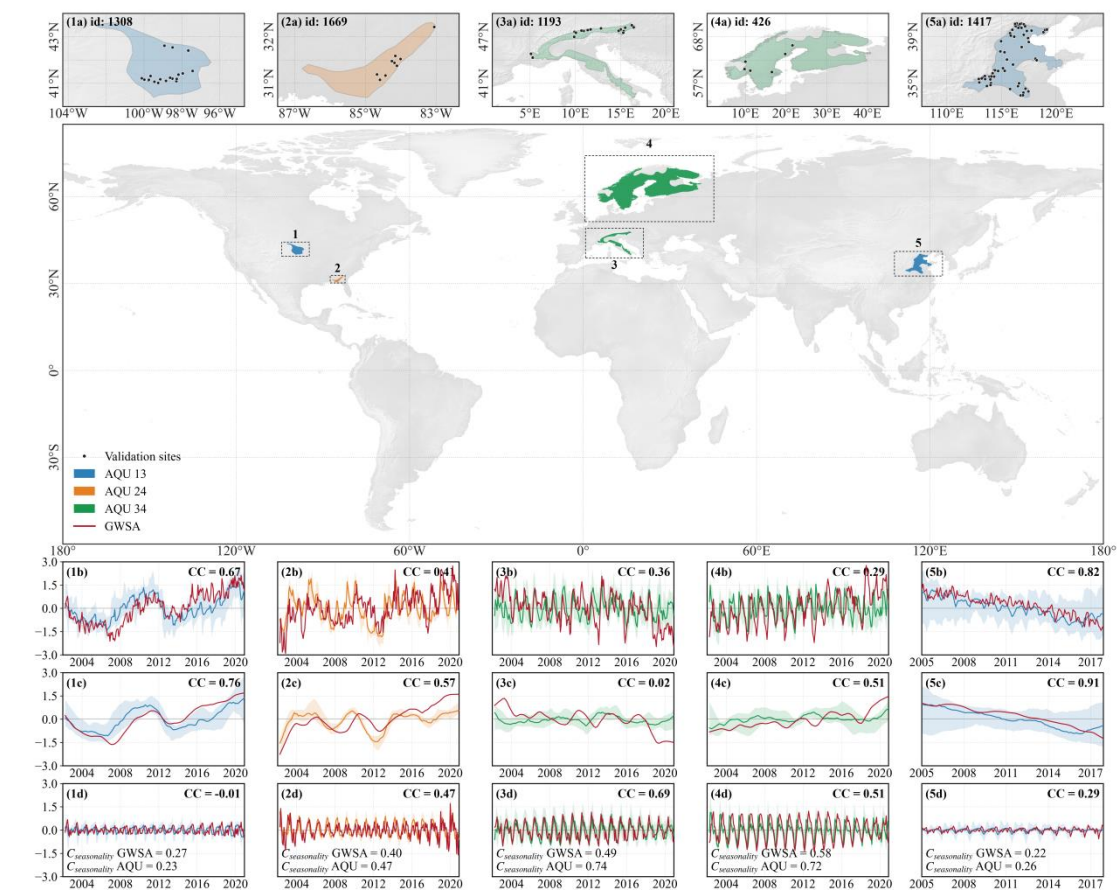


Figure 9. Spatial distributions and monthly STL decompositions of in-situ groundwater-level observations and reconstructed GWSA across five representative validation regions.

(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 $C_{\text{seasonality}}$ denotes the seasonal contribution.

The seasonal-contribution results show that the mean contribution is approximately 0.39 for the reconstruction, lower than approximately 0.48 for the in-situ observations. In AQU 13, the reconstructed and in-situ seasonal contributions are relatively close for both the High Plains Aquifer and the North China Plain. The difference is also small in AQU 24. By contrast, in AQU 34, the in-situ contributions reach 0.74 and 0.72 in the Apennine and Scandinavian regions, respectively, compared with 0.49 and 0.58 for the reconstructions. The product therefore captures the regional seasonal rhythm of large basins more effectively but tends to underestimate seasonal amplitudes in local and shallow aquifers, particularly in high-latitude environments.

3.2.3 Regional Literature Benchmarks

Figure 10 and Table 4 compare the unsmoothed regional series on matched dates. Agreement was strongest in the NCP ($r = 0.963$; $RMSE = 32.0$ mm), followed by California ($r = 0.789$; $RMSE = 77.3$ mm) and northwestern India ($r = 0.731$; $RMSE = 66.4$ mm). The NCP had the closest total variability magnitude (SD ratio = 0.912), whereas variability was attenuated in northwestern India (0.593) and California (0.673). Product and reference seasonal amplitudes were similar in California (82.0 and 82.7 mm), but the product had stronger seasonality in the NCP (23.2 versus 8.9 mm) and northwestern India (63.9 versus 23.5 mm). Thus, strong correlation did not guarantee agreement in seasonal or total amplitude.

The paired-series Sen slopes had the same negative sign in the three monthly comparisons, but depletion was weaker in the product: -17.71 versus -20.27 mm yr⁻¹ in the NCP, -22.96 versus -39.47 mm yr⁻¹ in northwestern India, and -6.96 versus -15.80 mm yr⁻¹ in California. Exact-window OLS comparisons likewise gave -13.74 mm yr⁻¹ for NCP during 2003–2016, versus -16.1 mm yr⁻¹ reported by Zhang et al. (2023); -24.47 mm yr⁻¹ for northwestern India during August 2002–October 2008, versus -40.0 ± 10.0 mm yr⁻¹ in Rodell et al. (2009); and -6.93 mm yr⁻¹ for California during October 2003–

March 2010, versus $-20.4 \pm 3.9 \text{ mm yr}^{-1}$ in Famiglietti et al. (2011). The full 2003–2021 trend of Liu et al. (2022) was not treated as an exact-window test because the product ends in December 2020.

The HPA comparison was less favorable. The annual well-based composite from Rateb et al. (2020) yielded $r = 0.526$ and $RMSE = 48.2 \text{ mm}$. The product Sen slope was nearly neutral (-0.10 mm yr^{-1} ; 95% confidence interval -5.10 to $+5.41 \text{ mm yr}^{-1}$), compared with -9.28 mm yr^{-1} in the reference. More importantly, over the exact March 2003–February 2013 window of Breña-Naranjo et al. (2014), the product OLS trend was $+4.50 \text{ mm yr}^{-1}$, opposite to the published $-27.6 \pm 3.3 \text{ mm yr}^{-1}$. The reconstruction therefore did not reproduce the reported HPA-wide depletion during that interval. These contrasts identify regional limitations in depletion magnitude and show why literature benchmarking is informative beyond correlation-only evaluation.

Table 4. Regional agreement with published groundwater-storage estimates.

<u>Metric</u>	<u>North China Plain</u>	<u>Northwestern India</u>	<u>California</u>	<u>High Plains Aquifer</u>
<u>Reference</u>	<u>Zhang et al. (2023)</u>	<u>Rodell et al. (2009)</u>	<u>Liu et al. (2022)</u>	<u>Rateb et al. (2020)</u>
<u>Paired period</u>	<u>2003–2020</u>	<u>Aug 2002–Oct 2008</u>	<u>Oct 2003–Dec 2020</u>	<u>2003–2017</u>
<u>Frequency / N</u>	<u>Monthly / 216</u>	<u>Monthly / 73</u>	<u>Monthly / 175</u>	<u>Annual / 15</u>
<u>Pearson r</u>	<u>0.963</u>	<u>0.731</u>	<u>0.789</u>	<u>0.526</u>
<u>RMSE (mm)</u>	<u>32.0</u>	<u>66.4</u>	<u>77.3</u>	<u>48.2</u>
<u>Bias (mm)</u>	<u>+10.2</u>	<u>0.0</u>	<u>+48.3</u>	<u>+19.4</u>
<u>SD ratio</u>	<u>0.912</u>	<u>0.593</u>	<u>0.673</u>	<u>0.625</u>
<u>Seasonal amplitude: product / reference (mm)</u>	<u>23.2 / 8.9</u>	<u>63.9 / 23.5</u>	<u>82.0 / 82.7</u>	<u>Not evaluated</u>
<u>Sen slope: product / reference (mm yr⁻¹)</u>	<u>-17.71 / -20.27</u>	<u>-22.96 / -39.47</u>	<u>-6.96 / -15.80</u>	<u>-0.10 / -9.28</u>
<u>Sen-slope difference (mm yr⁻¹)</u>	<u>+2.56</u>	<u>+16.51</u>	<u>+8.84</u>	<u>+9.19</u>
<u>Sen-trend sign agreement</u>	<u>Yes</u>	<u>Yes</u>	<u>Yes</u>	<u>Yes*</u>

N is the number of paired observations. *Bias* and slope difference are product minus reference; *SD* ratio is product/reference. Seasonal amplitude is the peak-to-trough amplitude of a detrended annual harmonic. All metrics use the available paired dates after baseline alignment. The HPA product Sen slope is effectively neutral and its 95% confidence interval includes zero (-5.10 to $+5.41$ mm yr⁻¹). Sign agreement over 2003–2017 does not imply agreement with the separate March 2003–February 2013 benchmark, for which the product and published OLS trends have opposite signs.

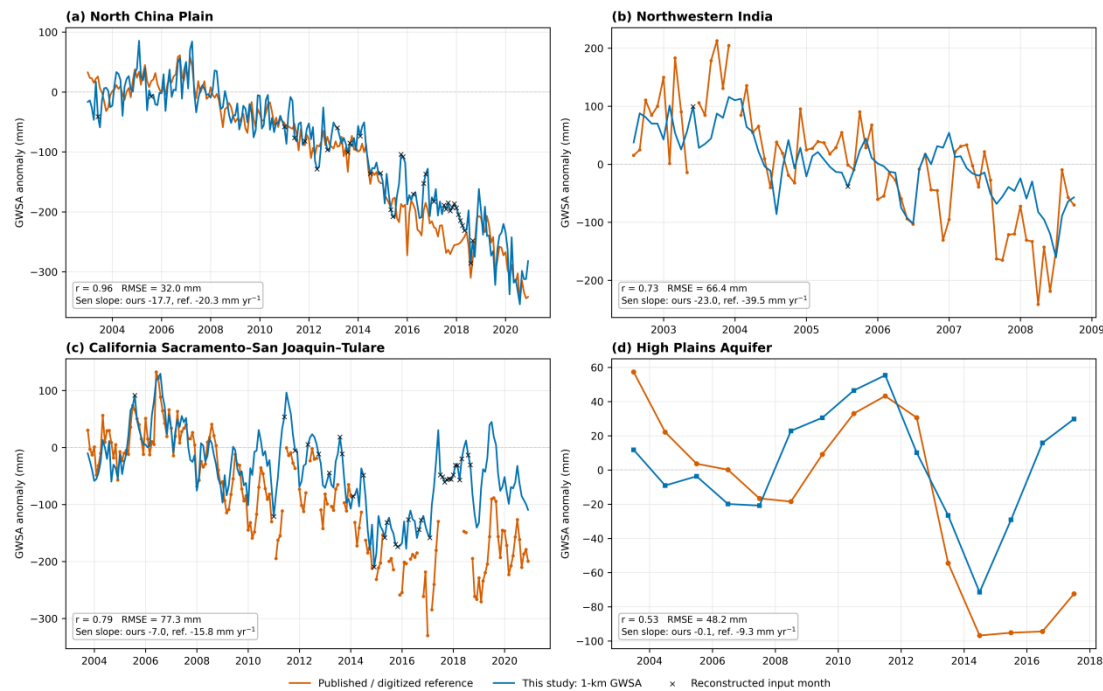


Figure 10. Regional evaluation of the 1 km GWSA reconstruction against published groundwater-storage estimates.

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.

3.3 Spatial Fidelity and Resolution Enhancement of Downscaled GWSA

To systematically evaluate the capability of the downscaling framework in ~~characterizing~~representing localized hydrological features, a detailed comparative analysis was conducted across four globally representative regions (Fig. 511). Original GRACE observations, restricted by the truncation of spherical harmonic coefficients, exhibit significant spatial smoothing and signal leakage effects. This manifests as coarse macroscopic anomalies (Fig. 511, Column a) that obscure fine-scale groundwater dynamics. In contrast, the 1 km downscaled products ~~effectively mitigate these observational artifacts while demonstrating high~~provide predictor-informed spatial fidelity~~detail while retaining coarse-scale GRACE constraints.~~

In the Amazon Basin (Area 1), ~~driven by natural hydro-climatic forcing, the downscaling framework successfully reconstructs~~downscaled fields contain complex dendritic ~~drainage~~ textures that are ~~entirely absent~~not resolved in the coarse-resolution GRACE data. ~~Furthermore, field. These patterns cannot be interpreted as groundwater-only structures: residual dynamic storage in the mountainous region of rivers, wetlands, and floodplains, together with GRACE leakage, are plausible contributors. In~~ southern Tibet (Area 2), ~~characterized by extreme topographical complexity, the finalized dataset (XGBoost, product (Fig. 511, panel 2c) captures fine-scale geomorphological signals, resolving spatial~~resolves predictor-informed textures ~~that align closely~~aligned with the ~~micro-canyon~~ topography of mountain gorges. ~~To demonstrate the necessity of our architectural selection, whereas the DNN baseline (Fig. 511, panel 2b) is shown to exhibit a tendency to fit dominant forcing signals, producing more generalized~~produces broader regional patches ~~that fail to delineate the sharp gradients of the deep valley systems.~~ This disparity confirms that the selected tree-based ensemble possesses superior capacity in processing high-frequency spatial information derived from contrast indicates model-dependent use of static geophysical covariates ~~such as digital elevation models~~ rather than independent kilometre-scale satellite information.

The ~~advantages of this high-resolution reconstruction are particularly pronounced~~additional spatial detail is also apparent in regions ~~dominated by~~associated with intensive ~~anthropogenic intervention.~~water management. In the agricultural zones surrounding New Delhi, India (Area 3)) and the Zhangye agricultural oasis within the hyper-arid Heihe River Basin (Area 4), the finalized dataset ~~accurately resolves~~reveals localized ~~depletion hotspots. It clearly delineates fine-scale~~ groundwater ~~depletion patterns (Fig. 5, panels 3e and 4e) that correspond highly~~storage gradients spatially aligned with ~~actual~~ agricultural boundaries and ~~exploitation~~managed-land extents. These spatial correspondences are

descriptive and reflect predictor-informed heterogeneity under coarse-scale GRACE constraints; attribution to groundwater pumping requires independent high-resolution observations.

The contrast with the continuous baseline architecture further highlights the robustness of the finalized product differences in spatial representation in these heavily pumped intensively managed areas. The DNN baseline tends to produce spatially smoothed anomaly zones, such as broad depletion cones (Fig. 511, panel 3b), which obscure abrupt transitions between irrigated oases and the surrounding arid background. In contrast, whereas the adopted XGBoost architecture exhibits an inherent edge-preserving property, successfully capturing the sharp, preserves sharper nonlinear boundaries induced by localized groundwater pumping. These findings demonstrate that gradients associated with local environmental heterogeneity. The integration of multi-source 1 km covariates is not merely therefore provides a mathematical downscaling procedure but a necessary predictor-informed reconstruction of the underlying hydrological heterogeneity, providing a robust digital benchmark for localized water resource management spatial heterogeneity under coarse-scale GRACE constraints.

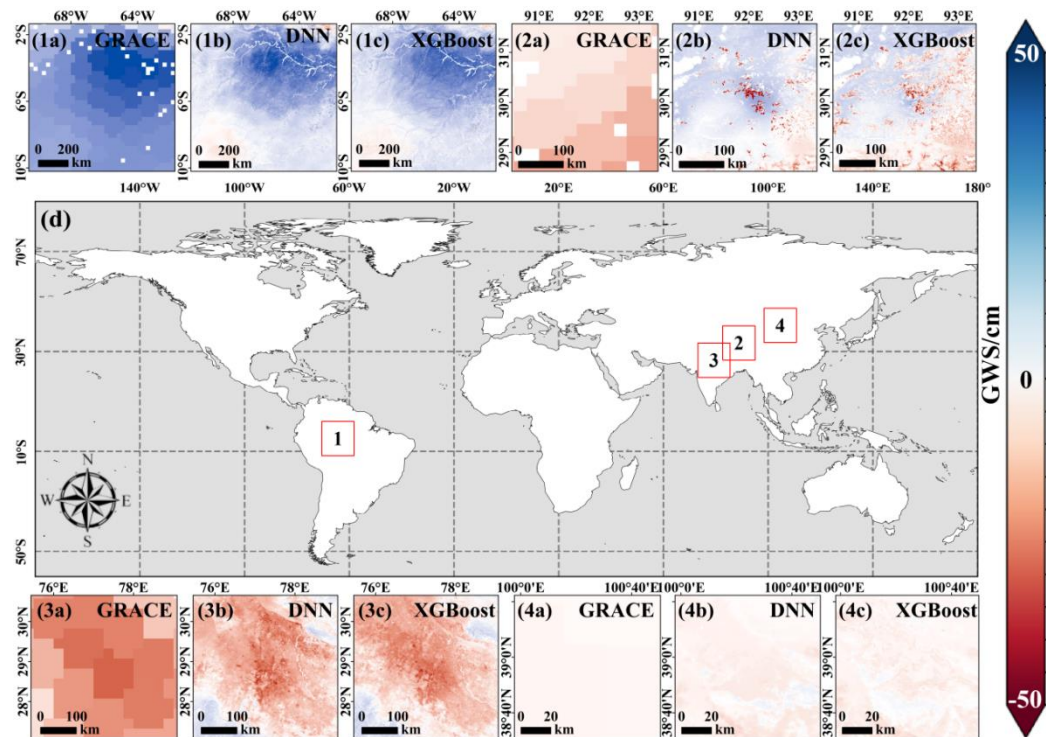


Figure 511. Spatial enhancement and feature reconstruction of 1 km resolution groundwater storage anomalies (GWSA) across four globally representative regions.

Columns present (a) the original 0.25° GRACE observations, alongside downscaled outputs generated by (b) Deep Neural Network (DNN) and (c) Extreme Gradient Boosting (XGBoost) models at the 1 km scale. The four selected regions (indicated by red boxes) encompass: (1) the Amazon Basin, driven

by associated with natural hydro-climatic forcing variability; (2) the mountainous region of southern Tibet, characterized by extreme topographical complexity; (3) the agricultural zones surrounding New Delhi, India, driven by intensive anthropogenic groundwater extraction; and (4) the Zhangye agricultural oasis within the hyper-arid Heihe River Basin. A comparative assessment highlights that Original GRACE observations (Column a) display coarse macroscopic anomalies due to spatial leakage. Following downscaling, The XGBoost model (Column c) captures provides predictor-informed fine-scale geomorphological and anthropogenic signals, resolving spatial textures aligned with canyon micro-topography in southern Tibet (2c) and clearly delineating localized groundwater depletion patterns corresponding to storage gradients spatially associated with agricultural boundaries in New Delhi (3c) and Zhangye (4c). Conversely, the DNN model (Column b) exhibits a tendency to fit dominant forcing signals, resulting in more generalized, spatially smoothed anomaly zones (e.g., broad depletion cones) that obscure These fine-scale hydrological details in both topographically complex (2b) and heavily pumped regions (3b). patterns are predictor-informed under coarse-scale GRACE constraints, and their correspondence with agricultural boundaries is interpreted descriptively.

3.3.4 Global Climatology and Spatiotemporal Variability Patterns

3.3.4.1 Baseline Imbalances and Meridional Profiles

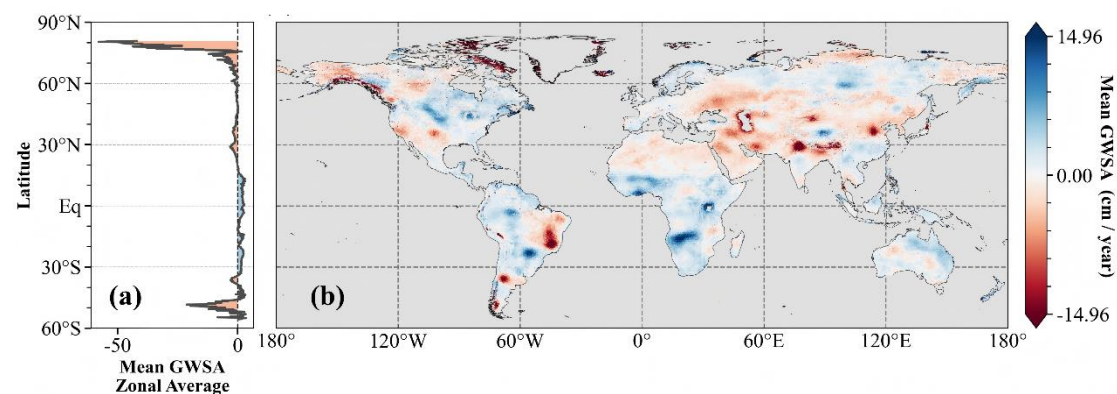
The spatial pattern of the global Mean groundwater storage anomalies (Mean-GWSA) was reconstructed at a 1 km resolution for the period from 2002 to 2020, along with the corresponding meridional mean profiles (Fig. 6.12). This reconstruction clarifies the baseline state of mass balance in the global groundwater system during the study period and offers deeper insights into the current imbalanced distribution patterns of the terrestrial water cycle on Earth. From the perspective of the global spatial pattern (Fig. 6.12b), the terrestrial water storage TWS exhibits systematic mass imbalances, manifested primarily as a net surplus in tropical and mid-to-high-latitude humid zones, whereas cumulative depletion occurs in areas of intensive human intervention and arid regions.

Significant negative anomalies are not randomly distributed but are highly concentrated in several typical large alluvial plain aquifers and arid regions globally. These localized depletion centers include the Indus-Ganges Basin, the North China Plain, the Arabian Peninsula, the Caspian Sea region, and the High Plains of the United States. In sharp contrast, significant positive anomalies, functioning as global groundwater accumulation zones, are stably distributed in tropical rainforest areas influenced by the

Intertropical Convergence Zone (ITCZ) and certain mid-to-high-latitude regions affected by persistent moist air currents.

The meridional mean profile (Fig. 6a12a) further condenses the two-dimensional heterogeneity into distinct zonal distribution characteristics along the latitudinal gradient. A core feature of this profile is the global groundwater depletion zone situated between 25°N and 45°N, with its trough spatially corresponding to ~~the most densely distributed anthropogenic groundwater~~ depletion hotspots located in regions characterized by intensive human water use on the map. From the perspective of the dynamic mechanism, this ~~phenomenon results from the high spatial coupling between pattern is spatially associated with the combined influence of~~ climatic drought stress ~~under natural geographic conditions~~ and intensive human activities water use. This latitude belt not only lies within the long-term recharge-deficient region ~~caused by associated with~~ the descending branch of the Hadley Cell but also represents the most densely populated and agriculturally irrigated area globally.

The reduction in the depth of the mean profile indicates that the groundwater system in this zone has fundamentally deviated from its climate-driven natural equilibrium state, leading to persistent mass loss. The profile exhibits stable positive peaks near the equator (Eq), consistent with the high concentration of reserves in tropical rainforest regions, while a pronounced negative shift occurs at extremely high latitudes (around 80°N). This shift effectively captures land mass loss signals ~~driven by associated with~~ cryosphere retreat or related dynamic processes at the polar margins. Ultimately, this spatial distribution of multi-year averages serves as a physical framework for understanding the evolution of the global groundwater system, providing essential contextual support for the subsequent analysis of dynamic evolution rates.



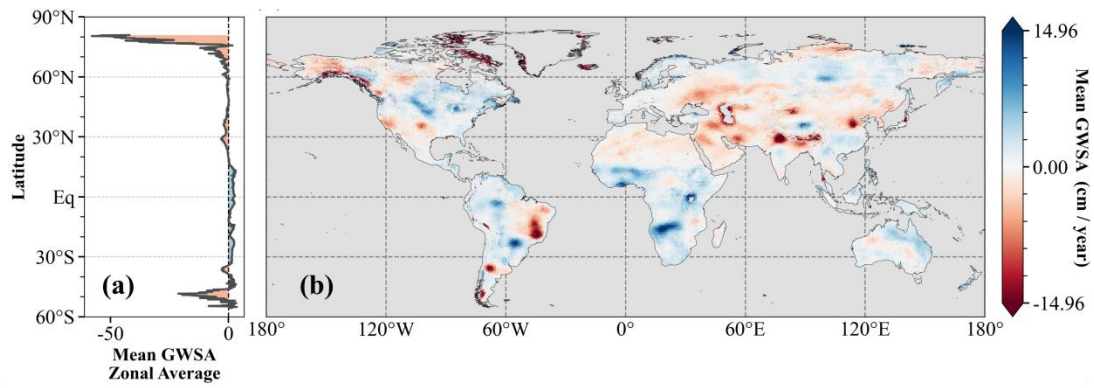


Figure 612. Spatial pattern and latitudinal profile of global Mean ~~groundwater storage anomalies (Mean GWSA)~~ during the period from 2002 to 2020.

(a) Latitudinal mean profile, presenting area-weighted averages of inland land pixels across diverse latitudes. The black solid line denotes the latitudinal mean, and the red and blue filled areas signify groundwater depletion and surplus states, respectively. (b) Global 1 km spatial pattern map depicting the high-resolution mean ~~groundwater storage anomaly~~GWSA distribution reconstructed via downscaling techniques. The color bands exhibit a smooth transition from dark red (depletion) to dark blue (surplus), and the white areas represent regions in a near-equilibrium state.

3.34.2 Seasonal Contributions and Global Hydrological Rhythms

To further evaluate the influence of cyclic factors on the dynamics of global groundwater storage, this study quantified the relative contributions of seasonal and non-seasonal components using the downscaled monthly GWSA time series (Fig. 713). In this framework, seasonality is characterized by the average intra-annual standard deviation for each year from 2002 to 2020, whereas non-seasonality is measured by the standard deviation of the annual mean values over the same 19-year period. The relative seasonal contribution is determined by the ratio of the seasonal component to the sum of both components, with a threshold of 50% used to identify regions predominantly governed by seasonal factors.

The results indicate that global GWSA dynamics generally exhibit distinct seasonally-regulated characteristics. Based on equal-area projection statistics, the seasonal contribution exceeds 50% across 52.30% of the global land area. This suggests that over half of the terrestrial domains are predominantly governed by stable and recurrent intra-annual cycles, rather than non-seasonal anomalies at interannual scales. Furthermore, in 26.09% of the land area, the seasonal contribution is greater than 66%, meaning the seasonal component is more than 1.5 times the magnitude of non-seasonal fluctuations, indicating a pronounced dominance of seasonality.

Spatially, regions with high seasonal contributions (60% to 90%) are primarily concentrated in areas with regular hydrological rhythms, including the Amazon and Congo Basins, the South and East Asian monsoon regions, northern Europe to the southern margin of Siberia, and high-latitude zones in North America. These patterns reflect a robust response of groundwater storage to periodic variations in precipitation, evapotranspiration, surface water-groundwater interactions, and freeze-thaw processes.

In contrast, regions with low seasonal contributions are predominantly distributed in inland arid zones, such as North Africa, the Arabian Peninsula, Central Asia, inland Australia, and western North America. In these areas, non-seasonal fluctuations dominate, manifesting primarily as persistent GWSA deficits ~~driven by~~ associated with long-term trends. The frequent occurrence of extreme heat, compound droughts, and heatwave events in these regions exacerbates long-term soil moisture imbalances. These groundwater systems are heavily ~~driven by~~ associated with persistent drought, intensive overexploitation, and complex topography-climate coupling. Consequently, they have transitioned from structured, seasonal recharge-depletion cycles to a state dominated by non-periodic disturbances.

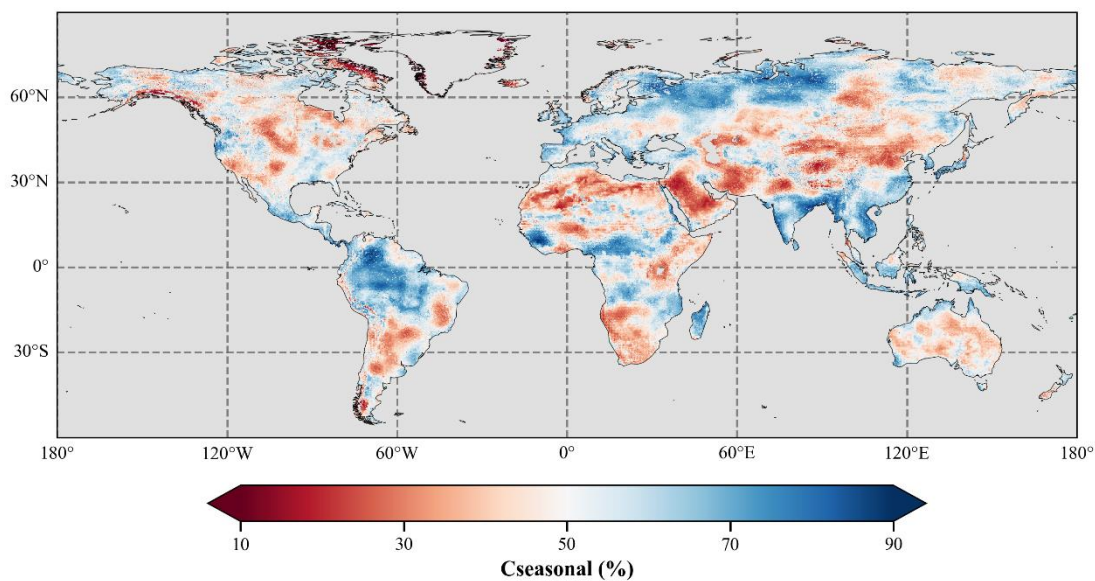


Figure 7.13. Seasonal contribution to global 1 km ~~groundwater storage anomalies (GWSA)~~.

Color bar shows the contribution of seasonality to GWSA variability during 2002–2020 (%). Red denotes a higher seasonal contribution, whereas blue denotes a lower seasonal contribution. Regions with a seasonal contribution greater than 50% were defined as seasonality-dominated areas.

3.4.5 Spatiotemporal Evolution and Trend Polarization

3.4.5.1 Spatial Heterogeneity of Global GWSA Trends

As illustrated in ~~Fig. 8~~Figure 14, the evolution of global groundwater storage over the past two decades has exhibited notable regional clustering characteristics and spatial differentiation trends. With a spatial resolution of 1 km, the downscaled trend maps ~~mitigate~~reduce the visual effects of spatial smoothing ~~effects~~ inherent in traditional coarse-resolution GRACE observations, ~~thereby precisely revealing complex~~ and ~~diverse~~reveal predictor-informed local hydrological ~~evolution~~ gradients under coarse-scale GRACE constraints.

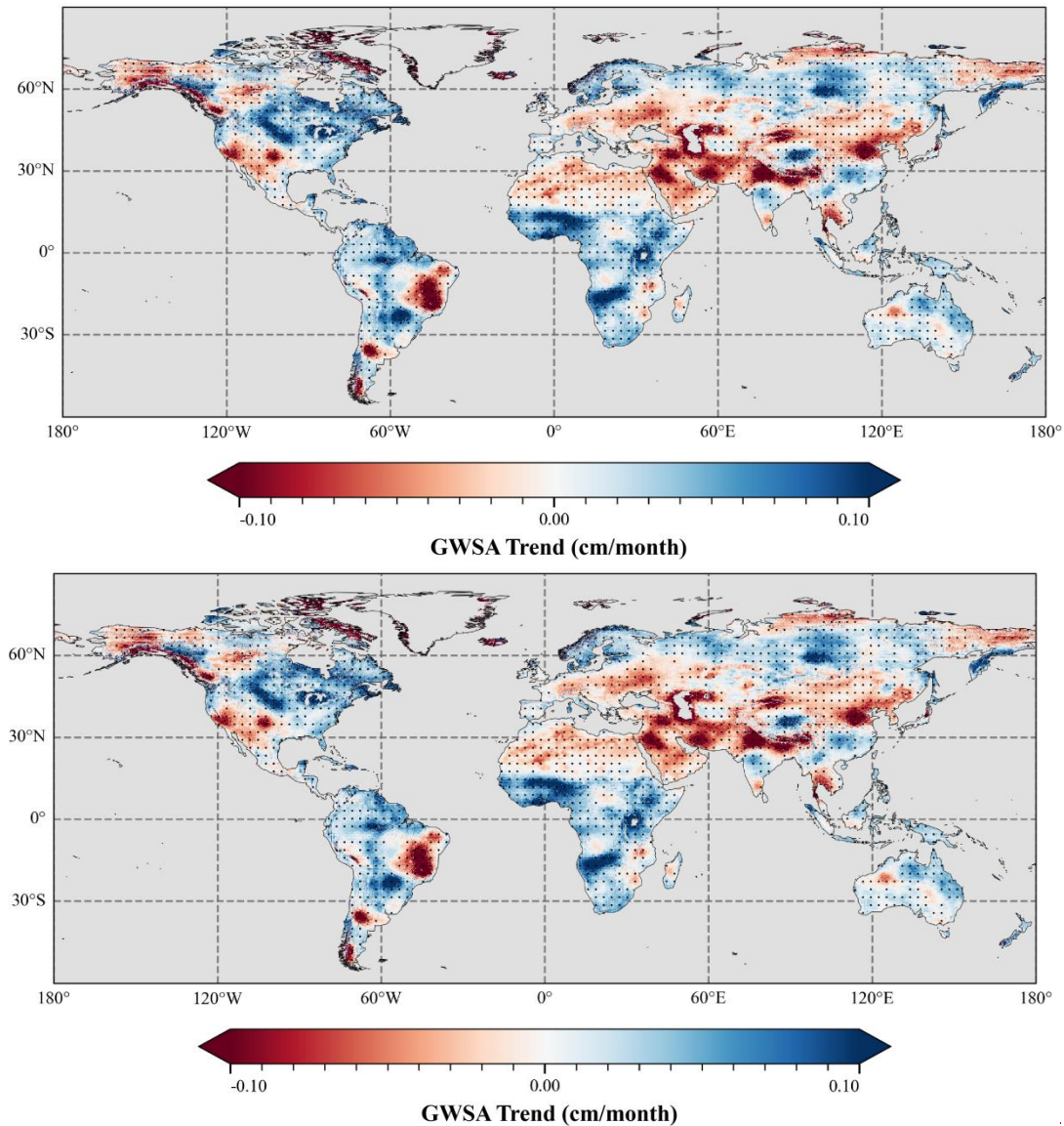
Globally, significant groundwater depletion signals ($p < 0.01$ and trend rate < -0.10 cm/month) are mainly manifested in two typical spatial patterns.

The first pattern consists of relatively clustered, patchy, and linear depletion centers, the spatial distribution of which shows a high degree of overlap with global high-intensity agricultural irrigation basins. For example, the 1 km dataset effectively captures steep depletion gradients along specific pre-mountain alluvial fans or typical farmland boundaries in regions such as the Central Valley of California, the High Plains Aquifer (Ogallala), northwestern India, and the North China Plain of China. This spatial characterization does not establish accurate basin-wide depletion magnitude: the separate regional benchmark in Sect. 3.2.3 identifies an HPA-wide trend mismatch over 2003–2013.

The second pattern involves large-scale, contiguous regional depletion, primarily concentrated in arid and semi-arid climate zones and major fossil groundwater extraction areas, such as the Middle East, the Arabian Peninsula, the hinterland around the Caspian Sea, and the La Plata Plain at the border of Argentina and Brazil in South America. Most of the Middle East region is characterized by a tropical desert climate, while the hinterland around the Caspian Sea is located deep within the continent, featuring arid conditions and intense evaporation, with inherently scarce water resources. The dense distribution of significant depletion points in these areas (indicated by black stippling in ~~Fig. 8~~Figure 14) indicates that their groundwater systems remained in a state of sustained and significant depletion during the study period.

Spatially corresponding to the aforementioned deficit patterns, significant groundwater accumulation regions worldwide ($p < 0.01$ with trend rates > 0.10 cm/month) predominantly display diffuse spatial distribution characteristics in accordance with major surface water systems and natural topography. In equatorial and tropical monsoon climate zones, including the transverse extension areas of Africa's Sahel region, the Victoria Lake Basin, and the core Amazon Basin, the ~~Groundwater Storage Anomalies (GWSA)~~GWSA demonstrate extensive and spatially continuous significant surplus.

1280 Additionally, high-latitude regions in North America, such as northern Canada with a polar long-winter
1281 climate, and localized areas in northern Eurasia, such as the southeastern Siberian Plain in Russia where
1282 permafrost is prevalent, show groundwater accumulation signals in trend maps. These signals potentially
1283 reflect deep water infiltration ~~driven by~~associated with accelerated hydrological cycles or cryosphere
1284 degradation at high latitudes. These surplus zones, dominated by large-scale climatic or natural
1285 hydrological processes, exhibit relatively smooth spatial transitions at 1 km scales, forming distinct
1286 morphological contrasts with the ~~sharply defined~~sharper deficit boundaries ~~caused by localized high-~~
1287 ~~intensity human activities~~spatially coincident with intensively managed agricultural regions.
1288



1289 **Figure 814.** Long-term evolutionary trends and significant spatial distributions of global 1 km resolution
1290 **groundwater storage anomalies (GWSA).**
1291

The color scale of the base map denotes the long-term change rate of GWSA (unit: cm/month) calculated on a pixel-by-pixel basis using the Theil-Sen median method. The red color series indicates continuous groundwater depletion, whereas the blue color series represents cumulative groundwater accumulation. Areas covered by dense black stippling points signify that the evolutionary trend of the pixel has passed the non-parametric ~~Mann-Kendall~~MK significance test ($p < 0.01$). Light gray regions correspond to oceans, polar permafrost areas, or extremely arid deserts where valid data are lacking.

3.4.5.2 Statistical Classification and Graded Evaluation of Groundwater Evolution

To ~~systematically quantify~~summarize the magnitude and statistical ~~robustness and physical significance-persistence~~ of global groundwater ~~evolution~~ trends, ~~this study integrated the absolute magnitude of~~ Theil-Sen trend rates were combined with ~~non-parametric Mann-Kendall (MK)~~ significance ~~test~~ thresholds to ~~establish a spatially stratified evaluation system composed of~~define seven ~~evolutionary stages~~descriptive classes (Fig. 9.15). This ~~dual-indicator~~ classification ~~framework not only~~ identifies the direction, rate of reservoir variations but also determines whether groundwater systems exhibit statistically significant responses to, and significance category of each mapped trend; it does not by itself diagnose the external forcing ~~factors, such as climate oscillations and high intensity extraction, through significance thresholds~~ responsible for that trend.

Pixel-level statistical results indicate that groundwater reserves in most global regions have significantly deviated from long-term dynamic equilibrium states from 2002 to 2020. Data reveal that approximately 38.84% of the effective land area is experiencing significant or extremely significant groundwater depletion ($p < 0.05$ or $p < 0.01$), whereas about 49.16% of regions show significant or extremely significant recharge. In contrast, only 12.00% of land areas maintained a relatively stable dynamic equilibrium with non-significant trends during the study period.

~~Based on the quantitative description provided by the MK test, the study further reveals a representative spatial polarization pattern in global groundwater evolution (Fig. 9). Analysis shows that in all regions displaying significant change signals, the evolutionary gradients are extremely steep, lacking smooth transitional zones across space. Specifically, among the total deficit areas accounting for 38.84%, the proportion of areas showing extremely significant reduction ($p < 0.01$) reaches 37.13%, while only 1.70% exhibit significant reduction ($p < 0.05$). Accumulation areas present similar~~

characteristics, with the proportion of extremely significantly increased areas (47.01%) far exceeding that of significantly increased areas (2.15%).

This distribution pattern, characterized by the absence of a broad buffer zone between non-significant and extremely significant changes, is manifested as clearly defined spatial-physical boundaries on the 1-km high-resolution classification base map. From the perspective of physical mechanisms, this phenomenon implies that the global groundwater system responds to intense external pressures, such as structural overextraction or extreme precipitation deficits, with abrupt, unidirectional physical evolution rather than gradual processes.

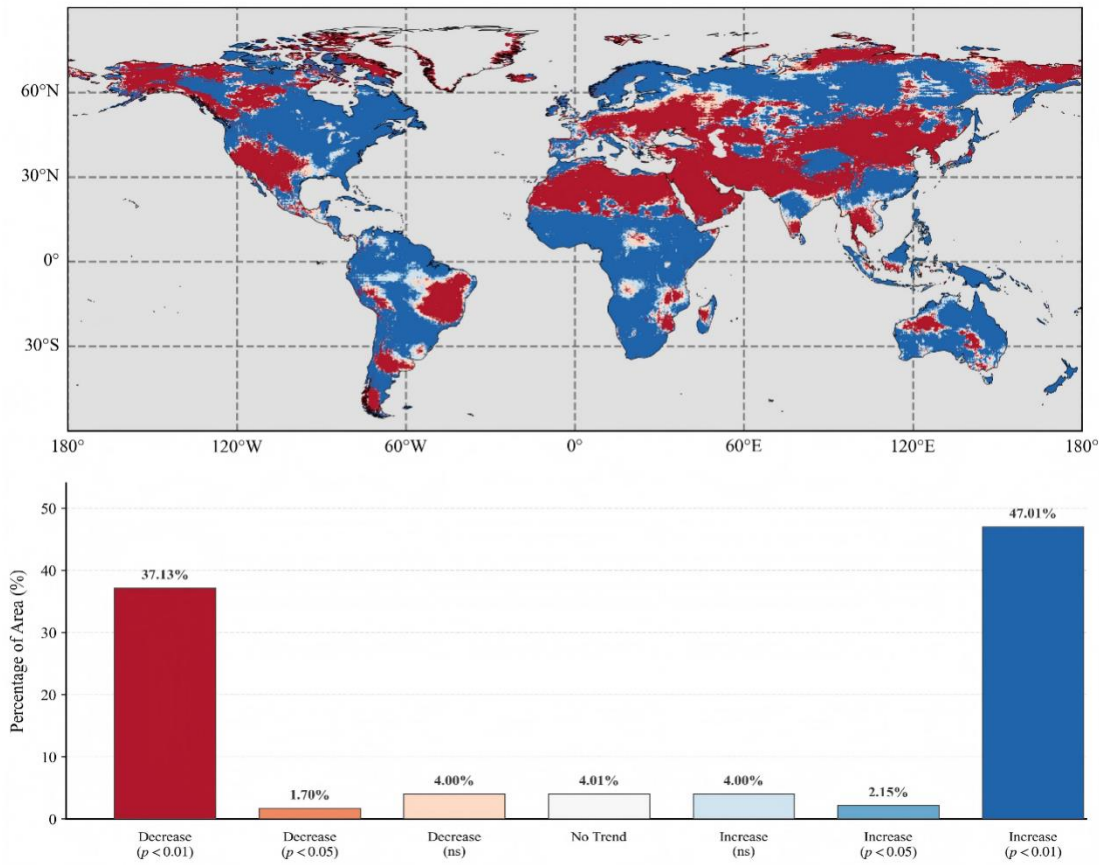


Figure 9. Spatial Classification and Statistical Analysis of the Evolutionary Status and Area Proportion of Global 1 km Groundwater Reserve Anomalies (GWSA). The upper figure depicts the global spatial distribution of groundwater evolutionary status (from 2002 to 2020, which is constructed by integrating Theil-Sen trend rates and Mann-Kendall significance tests. Different evolutionary stages are represented by RdBu divergence color bands. Dark red and dark blue respectively signify highly significant reduction and highly significant increase ($p < 0.01$), while lighter colors correspond to dynamically balanced states with insignificant changes (ns) or no trend. The lower figure presents a bar chart that shows the global

proportion of effective land area, maintaining a strict color correspondence with the spatial distribution map. The statistical results disclose a prominent "spatial polarization" in global groundwater evolution: regions with highly significant reduction (37.13%) and highly significant increase (47.01%) account for the majority of the global land area, whereas intermediate transitional zones with significant changes ($p < 0.05$) constitute a minimal proportion (less than 2.5%). This distribution pattern implies that the global groundwater system demonstrates abrupt, unidirectional physical evolutionary responses under intense external pressures rather than gradual processes.

The MK-based classification indicates that statistically significant positive and negative trends occupy large fractions of the mapped land area (Fig. 15). Among areas classified as depleted, 37.13% of the total effective land area falls in the $p < 0.01$ category and 1.70% in the $p < 0.05$ category; the corresponding accumulation fractions are 47.01% and 2.15%. These proportions describe the adopted significance classes and should not be interpreted as proof of abrupt physical thresholds or a unique forcing mechanism.

The mapped contrast between non-significant and highly significant classes may reflect a combination of persistent regional trends, the discrete MK thresholds used for classification, predictor-informed spatial structure, and residual-correction effects. Accordingly, sharp class boundaries in the 1 km product are treated as statistical features requiring independent verification, not as direct evidence of abrupt aquifer boundaries, structural overextraction, or any single causal process.

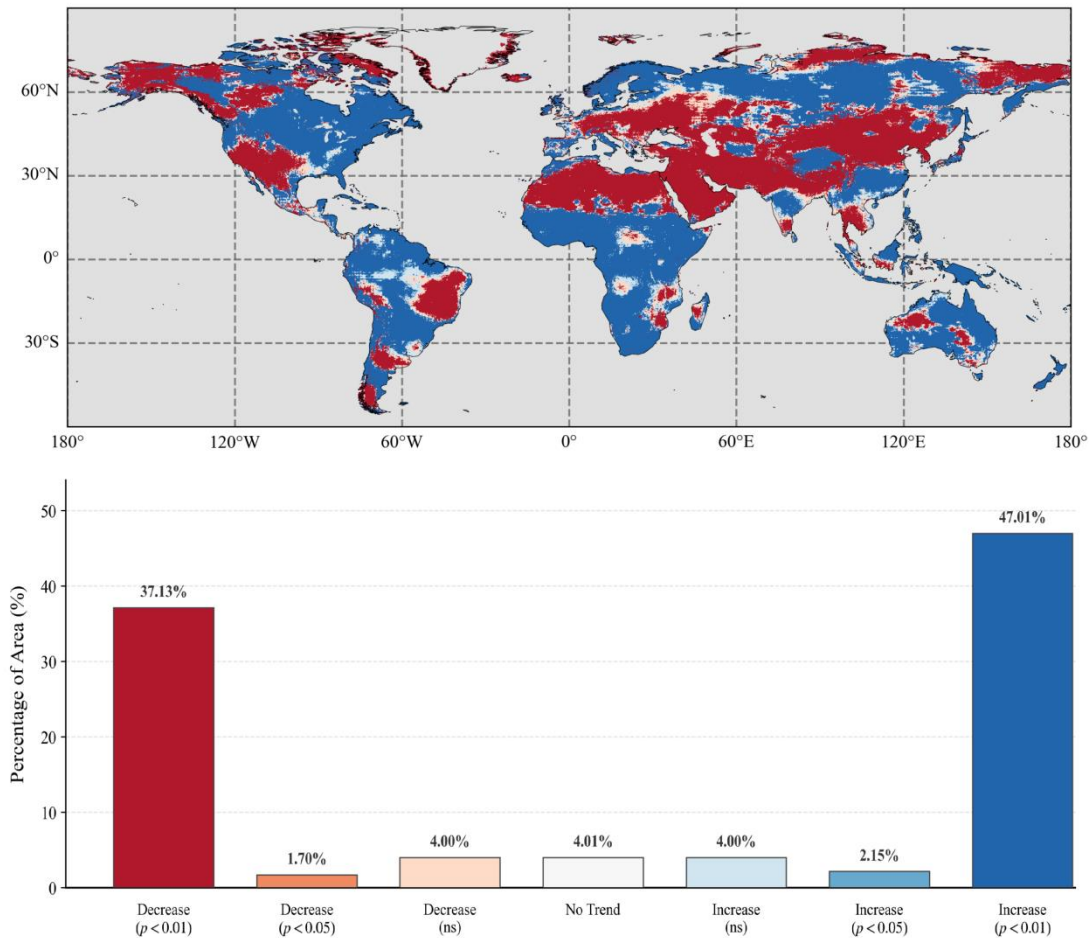


Figure 15. Spatial classification and area proportions of global 1 km GWSA trends.

The map integrates Theil–Sen trend direction with MK significance classes for 2002–2020. Dark red and dark blue denote significant decrease and increase at $p < 0.01$; lighter colors denote $p < 0.05$ classes, and the neutral class denotes non-significant trends. The lower bar chart reports the area proportion of each class. The prevalence of the $p < 0.01$ classes describes statistical persistence under the adopted classification and does not by itself establish abrupt physical thresholds or causal attribution.

3.5 Anthropogenic Fingerprint Identification via Residual Attribution Signals Not Explained by the GHM-simulated Background

In the long-term evolutionary panorama of global 1-km-resolution groundwater storage anomalies (GWSA), regional depletion and accumulation hotspots not only manifest significant spatial polarization but also exhibit distinct spatial dynamics at the morphological level. To effectively isolate the influence of human intervention from composite signals, this study establishes a substantial spatial heterogeneity. To describe departures from a GHM-simulated background, this study uses a

residual diagnostic framework ~~based on residual attribution. By comparing that compares~~ high-resolution reconstructed sequences (~~obtained by filling with~~ missing values ~~via Singular Spectrum Analysis~~filled through SSA) with GHM simulation benchmarks ~~from global hydrological models (GHMs) founded on physical mechanisms, we quantitatively determined the estimated magnitude of intensive anthropogenic extraction. Specifically, the 0.5° natural background signals simulated by the GHMs were. The 0.5° GHM-simulated background signals are~~ bilinearly interpolated to ~~a~~ 1 km ~~resolution to achieve for~~ spatial matching with the downscaled GWSA before subtraction. ~~Maintaining the 1 km resolution provides precise spatial morphological criteria for distinguishing anthropogenic anomalies from natural gradients~~The resulting residuals are not direct estimates of pumping or anthropogenic extraction; they are depletion signals not explained by the GHM-simulated background and may reflect a combination of human influences, GHM structural error, GRACE-derived GWSA separation uncertainty, and downscaling uncertainty.

Among the global deficit hotspots identified ~~in the previous sections, the above,~~ residual sequences ~~following after~~ STL decomposition ~~for the elimination of to remove~~ periodic components exhibit ~~remarkable~~marked spatial heterogeneity (~~Fig. 10~~Figure 16). In ~~high intensity irrigation intensively irrigated~~ regions, ~~such as including~~ the Indus Basin (AQU 23) and the North China Plain, ~~the~~ residual trend maps ~~feature distinct, steep gradient deep red patches. In the North China Plain, long term intensive agricultural practices have led to severe and continuous groundwater deficits, such as those persisting from 2019 to 2020. Significantly, these abrupt changes in spatial morphology demonstrate high physical consistency with actual administrative irrigation boundaries or farmland exploitation extents. Conversely, regions mainly influenced by natural factors, such as the core Amazon Basin, display show localized negative patches. Their spatial association with irrigation boundaries or managed-land extents is consistent with, but does not establish, an influence of intensive groundwater use. Conversely, residual values approaching zero after subtracting the model background, suggesting a stable distribution. This morphological contrast verifies that the significant deficits observed at the 1 km scale ($p < 0.01$) are predominantly driven by human exploitation surpassing natural recharge thresholds in areas such as the core Amazon Basin indicate closer agreement between the downscaled product and the modeled background. Significance at $p < 0.01$ characterizes the residual trend, not the cause of that trend.~~

~~The residual attribution framework further reveals the structural limitations of physical hydrological models in simulating complex human interventions. In the Arab Aquifer (AQU 22), where the over-~~

extraction of fossil groundwater is severe, traditional GHMs typically generate simulation trajectories characterized by gradual natural attenuation, failing to accurately replicate the observed sharp nonlinear depletion processes. Using the reconstructed time series data with missing values filled through SSA and temporal coherence preserved, our study identifies significant negative deviations between observed trends and model simulated trajectories. The continuous accumulation of such time series residuals offers a direct quantification of groundwater imbalance driven by human pumping activities, effectively overcoming the limitations of physical models in characterizing anthropogenic extraction processes. By maintaining the 1 km resolution throughout the attribution process, this residual distribution map precisely identifies groundwater depletion hotspots driven by localized high intensity irrigation activities, effectively eliminating the spatial smoothing effects from the original GRACE observations.

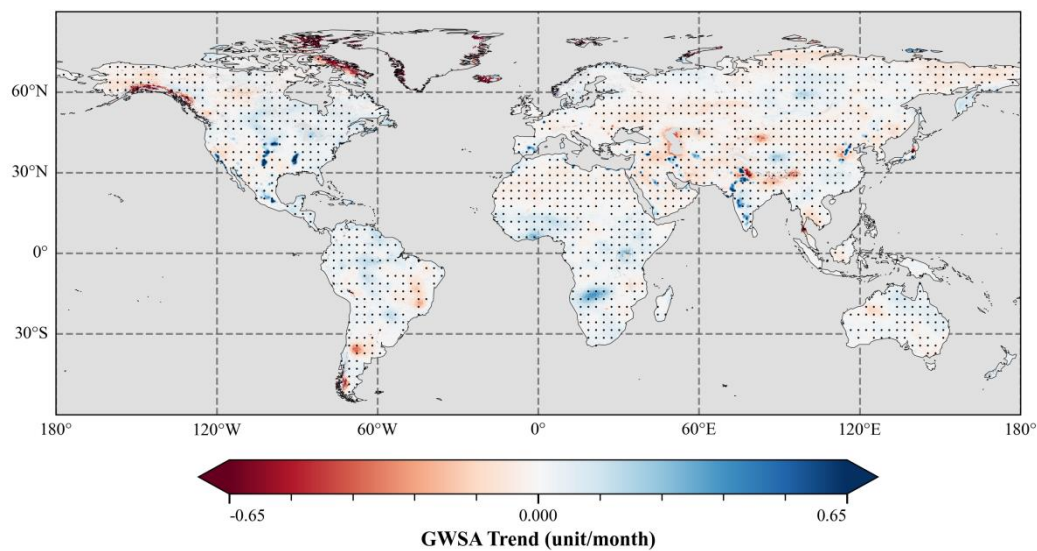


Figure 10. Human Pumping Fingerprint Identification Based on Observation and Model Residual Attribution ($p < 0.01$).

This visualization presents the distribution characteristics of physical signals subsequent to the attribution of groundwater trends. The computational approach entails bilinearly interpolating the natural background signals simulated by global hydrological models (GHMs) to a 1 km resolution, and subsequently subtracting these natural baselines from the generated 1 km downscaled groundwater storage anomaly (GWSA) reconstruction sequence. Trend computations utilize the Theil Sen median method, followed by seasonal time series decomposition (STL) for de-seasonalization. The red regions signify observed depletion rates that significantly surpass the explanations provided by natural climate variability, representing an approximation of localized depletion anomalies likely driven by intensive

anthropogenic extraction, following the removal of modeled natural fluctuations. The blue areas may indicate artificial recharge or model overestimation of natural deficits. Dense black dots denote residual trends that pass the Mann-Kendall significance test ($p < 0.01$). By maintaining the 1 km resolution throughout the attribution process, this residual distribution map precisely identifies groundwater depletion hotspots driven by localized high intensity irrigation activities, effectively eliminating the spatial smoothing effects from the original Gravity Recovery and Climate Experiment (GRACE) observations.

This residual diagnostic also illustrates limitations in comparing a predictor-informed downscaled product with coarse GHM simulations. In the Arab Aquifer (AQU 22), negative deviations between reconstructed trends and modeled trajectories may be potentially influenced by groundwater use, but they can also incorporate errors in GHM forcing, process parameterization, and representation of withdrawals, as well as uncertainty in separating GRACE-derived GWSA components and in the downscaling procedure. Maintaining a 1 km grid after interpolation does not create independent 1 km observations from either GRACE or the GHM. Accordingly, residual patterns are best treated as locations for hypothesis generation and follow-up with independent high-resolution observations, rather than as causal attribution or quantification of pumping.

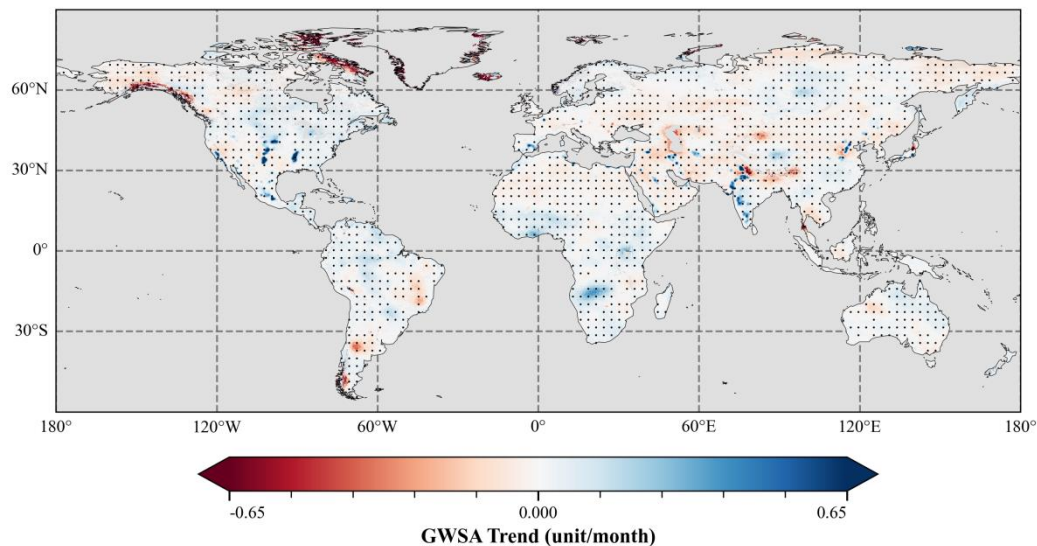


Figure 16. Residual signals not explained by the GHM-simulated background ($p < 0.01$).

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.

4. Discussion

4.1 Comparison with Existing Macro-scale Groundwater Datasets

This dataset serves as a macro-scale alternative for regions where computationally intensive three-dimensional numerical models, such as MODFLOW, are unfeasible. High-fidelity regional models require dense borehole data and precise aquifer geometry, which are often unavailable in Africa and Central Asia. Although our data-driven framework does not explicitly simulate vertical flow fields, the spatially independent in-situ evaluation provides observation-constrained evidence of long-term grid-scale agreement in held-out areas. The product can therefore serve as a useful prior hydrological proxy for ungauged basins and support future development of regional groundwater digital twins.

To objectively evaluate the scientific positioning of the 1 km global ~~groundwater storage anomaly (GWSA)~~ dataset, it is essential to systematically compare it with existing mainstream monitoring products. Native GRACE satellite observations are fundamentally constrained by their coarse spatial resolution, which frequently precludes the capture of fine-scale groundwater dynamics and restricts their utility in sub-basin scale management. In contrast, the 1 km dataset presented in this study functions as a high-fidelity bridge across this cross-scale observation gap, providing a refined spatial benchmark that is difficult to achieve through macro-scale gravity inversion alone.

Compared with traditional ~~Global Hydrological Models (GHMs)~~, such as GLDAS, PCR-GLOBWB, and WaterGAP, this dataset ~~offers significant improvements in characterizing localized anthropogenic interventions~~. provides additional predictor-informed spatial detail for examining localized groundwater-storage gradients. Although the latest generation of GHMs exhibits high physical rigor in simulating natural climatic cycles under natural forcings (Scanlon et al., 2018; Müller Schmied

et al., 2021), ~~these models often exhibit limitations in simulating the highly nonlinear trajectories~~their
~~coarse grids and differing representations of water resource depletion~~groundwater withdrawals can yield
~~departures from the downscaled product~~in intensive agricultural regions. This discrepancy primarily
~~stems from their reliance on coarse spatial grids and the widespread absence of localized groundwater~~
~~extraction modules.~~As demonstrated~~shown~~in Section 3.5, while traditional GHMs predominantly reflect
~~gradual natural attenuation backgrounds,~~residual depletion signalsin the Indus Basin and the North
~~China Plain,~~our 1 km downscaled dataset successfully captures sharp depletion troughs that exceed
~~natural resilience thresholds. Consequently, in the absence of large scale pumping records, this dataset~~
~~functions as a robust observational reconstruction and provides an objective benchmark for quantifying~~
~~the localized impacts are not explained by the GHM-simulated background, but they are not~~
~~quantification~~of unsustainable anthropogenic extraction.These differences can guide targeted
evaluation where independent high-resolution hydrological and pumping observations are available.

Furthermore, our framework provides a distinct methodological alternative to previous statistical
downscaling efforts. Most existing studies remain confined to specific regional basins (Miro and
Famiglietti, 2018) or exhibit a spatial resolution plateau at the 10-km scale (Yin et al., 2022). The current
1 km global reconstruction ~~is achieved not through simple mathematical interpolation, but by~~
~~leveraging~~leverages the ~~Extreme Gradient Boosting (XGBoost) architecture to extract localized spatial~~
~~heterogeneity from~~and high-resolution auxiliary environmental covariates; to represent localized spatial
heterogeneity under coarse-scale GRACE constraints. This approach ~~accurately reconstructs natural~~can
reveal predictor-informed hydro-geomorphological ~~pattern~~textures, such as canyon-aligned
~~textures~~patterns in southern Tibet, ~~while simultaneously isolating~~and groundwater-storage gradients
spatially associated with managed agricultural ~~water use hotspots in hyper arid regions like~~
~~Zhangye areas.~~ Such ~~fine-scale resolution is critical~~information provides a useful basis for ~~resolving the~~
spatial ~~polarization~~screening of ~~groundwater evolution identified in this study~~hydrogeological
heterogeneity and managed-land associations.

Finally, this dataset serves as a valuable macro-scale alternative for regions where computationally
~~intensive three-dimensional numerical models, such as MODFLOW, are unfeasible. The development~~
~~of high-fidelity regional numerical models relies on dense in-situ borehole data and precise aquifer~~
~~geometries, presenting substantial challenges in data-scarce regions across Africa and Central Asia.~~
~~Although our data-driven paradigm does not explicitly model vertical flow fields, it exhibits robust~~

~~statistical consistency with storage trends validated against a comprehensive global network of 1,518 monitoring wells (Jasechko et al., 2024). Therefore, this 1 km GWSA product can function as a crucial prior hydrological proxy for ungauged basins, providing essential initial conditions and validation fields for the future development of regional groundwater digital twins.~~

4.2 Mechanistic Differences Between Deep Learning and Tree-Based Models in Downscaling

Identifying the distinct physical representation disparities among underlying algorithms under extreme hydrological forcing conditions is critical for understanding the applicability and limitations of the 1 km dataset presented in this study. Although the cross-validation results (Section 3.1) indicate that both DNN and XGBoost achieve high prediction accuracies globally, their mathematical architectures reveal systematic physical representation discrepancies when processing highly nonlinear hydrological forcing and complex spatial heterogeneity.

~~In terms of spatial texture representation, as illustrated in Section 3.2 (Fig. 5), XGBoost exhibits relatively continuous transitions at aquifer boundaries while preserving rich spatial micro-heterogeneity textures at the pixel scale. This spatial representation characteristic stems from its underlying random subspace mechanism and ensemble averaging effect (Breiman, 2001; Tyrakis et al., 2019). During training, XGBoost enforces node splitting of each decision tree solely within randomly selected feature subsets. This design effectively reduces the dominance of macro-scale forcing factors (such as precipitation), enabling the model to capture local heterogeneity from secondary environmental variables like digital elevation models (DEM) and soil physical properties. Simultaneously, the independent prediction expectations generated by hundreds of decision trees partially mitigate the hard boundary segmentation effects of individual trees, allowing the model to represent spatial textures shaped by micro-terrain and local capillary hydrodynamics consistent with natural gradient patterns (e.g., Fig. 5-2c).~~

~~In contrast, the spatial patterns generated by DNNs display sharply defined boundary features with relatively limited local details (e.g., Fig. 5-2b and Fig. 5-3b). This phenomenon is attributable to the priority capture characteristics of gradient descent-based multi-layer perceptrons for macroscopic dominant signals. During high-dimensional nonlinear mapping, DNN hierarchical networks tend to prioritize fitting climate forcing signals with the highest variance contribution rates. Consequently, localized weak topographic or soil texture variations may be processed as high-frequency noise by the networks, leading to implicit filtering or excessive spatial smoothing (Reichstein et al., 2019; Shen, In~~

terms of spatial texture representation, XGBoost preserves relatively detailed predictor-defined heterogeneity while maintaining coherent transitions at broader scales (Sect. 3.3; Fig. 11). This behavior follows from recursive partitioning and additive gradient boosting, with tree depth, child-weight constraints, learning rate, and column or row subsampling controlling how interactions among terrain, soil, and hydroclimatic predictors are represented (Chen and Guestrin, 2016). The resulting fine-scale textures remain model outputs constrained by the predictor database and should not be interpreted as direct observations of local capillary processes or aquifer boundaries.

In contrast, the DNN generally produces smoother, broader regional anomaly fields, with sharper transitions appearing only in some high-contrast zones (Fig. 11). Its hierarchical continuous mapping tends to prioritize dominant high-variance relationships, while weaker local terrain or soil contrasts may be attenuated through regularization and optimization. This interpretation is descriptive of the observed model outputs and does not imply that one architecture is universally more physically realistic.

In intensively managed regions such as New Delhi, the two architectures generate different spatial textures: DNN emphasizes broad depletion patches, whereas XGBoost retains finer predictor-defined gradients. Apparent edges or patches in either result can arise from model structure and predictor contrasts; their spatial association with managed land is therefore used for hypothesis generation rather than as proof of pumping.

The two architectures also differ mathematically outside the range represented in the training targets. Tree ensembles typically cannot extrapolate beyond the range of their terminal-node responses, whereas continuous neural networks can in principle produce values outside the observed target range (Hastie et al., 2009). However, the present study did not perform a temporally blocked extrapolation experiment. Consequently, no empirical claim is made that DNN provides superior prediction of future droughts or that XGBoost quantifies post-2020 extremes; both products are evaluated as retrospective reconstructions over 2002–2020.

A systematic elucidation2018).

~~Additionally, when localized areas experience extreme human-induced pumping activities (e.g., agricultural zones in New Delhi, Fig. 5-3) causing sudden water storage fluctuations, nonlinear activation functions (e.g., ReLU) in DNN hidden layers may induce directional high-frequency activation. This can result in numerical jumps in predicted values at critical hydrological boundaries, forming large-scale deficit patches with sharp visual polarization on maps. Consequently, DNNs tend to emphasize~~

macroscopic trends, whereas XGBoost demonstrates greater sensitivity to heterogeneity at the micro-scale.

Beyond spatial texture representation, the temporal responses of groundwater systems to extreme climate events highlight disparities in the extrapolation capabilities between the two models. The mathematical nature of tree-based models inherently confines their prediction outputs within the bounding envelope of the training set label data (Hastie et al., 2009). This implies that during extreme drought events, XGBoost prediction curves may exhibit truncation at the lower end, potentially underestimating the absolute physical magnitude of extreme groundwater deficits. In contrast, DNNs, with continuous mapping capabilities, demonstrate extrapolation potential beyond the training set distribution boundaries, enabling the simulation of deeper trough amplitudes. However, without explicit constraints from hydrological physical laws, such mathematical extrapolation risks non-physical oscillations.

A comprehensive understanding of these complementary mechanisms can not only explain/clarify the rationale behind our logical basis of the dataset selection strategy adopted in this study, but also provides critical insights/provide crucial theoretical references for deploying the construction of hybrid integration frameworks with enhanced/strengthened physical constraints in future Earth system science research (Karniadakis et al., 2021). The XGBoost-based high-resolution dataset presented in this study demonstrates an enhanced/constructed based on the XGBoost algorithm exhibits improved capability to resolve/in characterizing complex terrain/topographic features and localized human pumping boundaries, making groundwater storage gradients, which renders it particularly suitable for applicable to spatial screening and water resource management practices on the premise that the coarse-scale GRACE constraint and attribution research requiring refined spatial inputs/limitations of the dataset are explicitly acknowledged.

4.3 Sources of Uncertainty in Downscaling Models and Future Research Perspectives

Although the 1 km high-resolution dataset constructed in this study exhibits favorable accuracy and spatiotemporal continuity across global aquifer systems, it is imperative to objectively acknowledge the sources of model uncertainty and structural limitations stemming from the complexity of Earth system observations and the inherent constraints of purely data-driven paradigms.

1586 Firstly, the cascading propagation of prior errors in the input feature fields represents a crucial
1587 external factor influencing local downscaling precision. Meteorological forcing data (e.g., precipitation
1588 and temperature) and surface parameters, which serve as baseline covariates for model predictions,
1589 inherently display observational biases in regions characterized by extreme topographical complexity or
1590 a scarcity of ground monitoring stations (such as alpine mountain ranges and hyper arid desert
1591 hinterlands). Additionally, the native GRACE satellite data, constrained by the truncation of spherical
1592 harmonic coefficients and spatial filtering processes, may introduce signal leakage effects that
1593 incorporate mass anomalies from adjacent areas into target pixels (Landerer and Swenson, 2012; Long
1594 et al., 2015). These multi-source heterogeneous data uncertainties, existing at both physical and
1595 observational levels, can propagate through deep nonlinear mapping networks during high-resolution
1596 reconstruction processes, potentially amplifying localized noise or creating spatial artifact biases.

1597 Secondly, purely data-driven downscaling models persistently demonstrate structural limitations in
1598 dealing with physical mechanism constraints. Current machine learning architectures are highly reliant
1599 on the Scale Invariance Assumption, which postulates that the statistical mapping relationships between
1600 coarse 0.25° groundwater storage anomalies (GWSA) and surface environmental covariates can be
1601 linearly extrapolated to 1-km scales. Nevertheless, actual groundwater flow dynamics display highly
1602 nonlinear characteristics at local scales, propelled by micro topography and the complex three-
1603 dimensional geometry of aquifers. Moreover, purely data-driven models lack explicit constraints from
1604 mass conservation laws or groundwater dynamic equations. Consequently, aggregated 1-km water
1605 storage values within specific closed basins may encounter difficulties in attaining absolute physical
1606 mass closure with the original 0.25° observations. Additionally, for sudden localized water resource
1607 management interventions or extreme land use changes that occur outside the temporal range of the
1608 training dataset, purely statistical models lacking explicit anthropogenic proxy variables frequently fail
1609 to offer accurate real-time dynamic feedback (Kratzert et al., 2019).

1610 Thirdly, the residual attribution framework proposed in Section 3.4 entails its own inherent
1611 structural uncertainties. By subtracting GHM-simulated natural backgrounds from the downscaled
1612 GWSA, the resulting residuals are assumed to represent human-induced depletion. However, these
1613 residuals inevitably incorporate the structural errors of the GHMs themselves—such as biases in
1614 precipitation routing, evapotranspiration parameterization, and soil moisture accounting under extreme

climatic stress. Therefore, these residuals should be interpreted as observation-constrained approximations of anthropogenic impact, rather than isolated, absolute pumping signals.

It is also crucial to clarify the role of the Specific Yield (S_y) parameter used during the independent validation phase. The published 1 km GWSA dataset is provided strictly in units of Equivalent Water Height (EWH). The regional effective average S_y was introduced solely as an intermediate dimensionality conversion parameter to translate satellite-derived EWH into Groundwater Level (GWL) variations for validation against in situ well data. While this approach inevitably introduces uncertainties related to the 3D heterogeneity of local aquifers (Rodell et al., 2009; Scanlon et al., 2012), preliminary sensitivity tests demonstrated that even when introducing $\pm 20\%$ random perturbations to the empirical S_y averages across subregions, the significant linear correlation between model predictions and observed in situ water level trends ($p < 0.01$) remained robust. This confirms that the macro-scale evolutionary trends revealed in this dataset are fundamentally driven by actual hydro-climatic forcings, rather than artifacts of the dimensionality conversion process.

To address the aforementioned uncertainties and limitations, future high-resolution hydrological reconstruction endeavors can be improved via two dimensions: multi-source data fusion and algorithmic architecture evolution. On one hand, incorporating explicit, high spatiotemporal resolution proxy variables of human activities, such as high-frequency land subsidence data derived from Synthetic Aperture Radar Interferometry (InSAR) into model forcing inputs can substantially enhance the algorithms' capability to directly represent intensive anthropogenic extraction (Castellazzi et al., 2016). On the other hand, integrating hydrological physical prior knowledge into data-driven black-box models has become a critical research frontier for overcoming statistical generalization bottlenecks. Developing Physically Informed Neural Networks (PINNs) that strictly enforce regional water balance closure (Willard et al., 2022), or establishing loose coupling frameworks with high-resolution 3D groundwater flow models (e.g., MODFLOW), represent promising pathways. This integrated research paradigm, bridging massive Earth observation data with classical hydrodynamic laws, will provide the essential scientific foundation for constructing global groundwater digital twin systems that combine high spatial fidelity with robust physical interpretability.

The global in-situ validation reveals multiple uncertainty sources associated with discrepancies in observation scales, subsurface hydrogeological characteristics, and monitoring network densities. Groundwater monitoring wells provide point-scale measurements of hydraulic-head variations, while

GRACE-derived GWSA characterizes vertically integrated water mass variations over a substantially larger spatial support. Such scale mismatch is particularly pronounced in deep confined aquifer systems, where variations in hydraulic head do not necessarily correspond proportionally to variations in total-column groundwater storage. Additionally, sparse monitoring networks further reduce the representativeness of regional comparative analyses, which is especially prominent in high-latitude regions and aquifer systems with inadequate data availability. Therefore, the in-situ validation implemented in this study mainly provides effective evaluation for long-term grid-scale correlation, magnitude consistency and trend direction, rather than verifying strict point-to-grid equivalence. The above limitations are clearly manifested in typical aquifer systems: for the Cambrian-Ordovician Aquifer, confined hydrogeological conditions complicate the quantitative relationship between hydraulic head and storage variation; for the West Siberian Basin, the limited number of available monitoring wells constrains the statistical representativeness of validation results.

A second source of uncertainty arises from the conversion of groundwater-level trends to equivalent-water-height storage trends through the scale-dependent effective specific yield, S_{yeff} . In this study, S_{yeff} is used as an observation-constrained level-to-storage conversion coefficient that accounts for the dimensional difference between point-scale groundwater-level observations and grid-scale GWSA estimates. Its uncertainty therefore propagates primarily into the magnitude of the converted in-situ storage trends. The sensitivity experiments provide a quantitative assessment of this effect. For XGBoost, perturbing S_{yeff} by up to $\pm 30\%$ changed R^2 from the baseline value of 0.538 to 0.477–0.520 under the perturbed scenarios, while CC remained within 0.691–0.721 and trend-sign consistency within 0.696–0.743. $Bias$ remained comparatively stable at 0.104–0.107 cm yr⁻¹, whereas the fitted regression slope varied more substantially from 0.379 to 0.593. DNN showed a similar response. These results indicate that uncertainty in S_{yeff} influences amplitude scaling more strongly than the overall association structure or trend-direction agreement. The independently standardized GWL and GWSA comparisons provide a complementary assessment of interannual variability and long-term temporal co-variation without requiring level-to-storage conversion.

Uncertainty originates from both groundwater-component separation and statistical downscaling. The 1 km product inherits errors related to GRACE leakage, component subtraction, SSA reconstruction, and coarse observational support, while its fine-scale variability is supplied by the environmental predictors. The kilometre-scale patterns are therefore observation-constrained, predictor-informed

reconstructions rather than independent satellite observations. Model architecture adds further structural uncertainty because XGBoost and DNN produce different spatial textures and responses. In addition, the random 70%/30% development split does not remove spatial or temporal autocorrelation and may yield optimistic within-distribution performance estimates. This random split was used for internal model development and does not constitute a spatially or temporally independent generalization test because adjacent spatial and temporal samples may occur in both subsets. Spatial generalization is assessed separately through the blocked in-situ evaluation, but temporal extrapolation was not tested with a chronological holdout; the dataset should therefore not be interpreted as a forecasting product.

The regional literature benchmarks add a further limitation (Sect. 3.2.3): temporal agreement is not uniform, and depletion magnitude is attenuated in northwestern India and California. The HPA is a clearer counterexample, with a trend-sign mismatch during the published 2003–2013 window. Opposing behavior in the Northern and Central–Southern HPA can produce spatial cancellation (Rateb et al., 2020), but this does not by itself explain away the discrepancy with Breña-Naranjo et al. (2014). Shared GRACE inputs also mean that high regional correlation is not independent evidence at 1 km. Consequently, the dataset is suitable for regional screening and comparative analysis under the stated constraints, while quantitative depletion estimates in heterogeneous, intensively managed aquifers require corroboration with regional observations.

Comparisons with GHMs introduce an additional component of structural uncertainty. The residuals between the downscaled GWSA and the GHM-simulated background signals simulated by GHMs integrate discrepancies derived from multiple sources, including the forcing of GHMs, parameterization of recharge and evapotranspiration, representation of groundwater processes, component separation based on GRACE observations, and the statistical downscaling procedure itself. Accordingly, these residuals are most appropriately interpreted as groundwater storage variations that cannot be explained by the GHM-simulated background. In regions with independently recorded intensive groundwater exploitation, sustained negative residuals provide effective evidence for identifying areas where anthropogenic disturbances may contribute to observed groundwater depletion, while spatially consistent residuals in other regions may reflect differences in model structure or unrepresented hydrological processes. This interpretation endows the residual analysis framework with practical value for spatial screening and hypothesis generation, especially when integrated with independent data on irrigation, groundwater abstraction, land use, or regional hydrogeology.

Accordingly, future research improvements should focus on reducing uncertainty at both the observational and modeling levels. The construction of higher-density, temporally consistent groundwater monitoring networks will facilitate more effective independent evaluation in currently under-sampled regions, while higher-resolution hydrogeological data can better constrain the conversion relationship between groundwater level and storage. Additional observations of surface water storage, groundwater abstraction, recharge, and aquifer properties will further enhance the physically-based interpretation of the reconstructed signals. In terms of modeling, spatially explicit physical constraints, multi-model ensemble frameworks, and hybrid approaches that integrate machine learning with process-based groundwater models can improve extrapolation performance under extreme hydrological conditions and reduce dependence on individual predictor sets or model structures. These advances will facilitate the transformation of the current observation-constrained 1 km resolution reconstruction from a high-resolution groundwater monitoring resource into an increasingly physically constrained framework for regional water resource assessment and groundwater system analysis.

5. Conclusion

~~To overcome the spatial resolution limitations of current satellite gravimetry, This study presents a globally seamless, global 1 km resolution monthly groundwater storage anomaly (GWSA) dataset spanning covering April 2002 to December 2020. The dataset was generated using an aquifer stratified Extreme Gradient Boosting (XGBoost) downscaling framework, integrated with Singular Spectrum Analysis (SSA) and for temporal gap filling, 19 high-resolution environmental covariates.~~

~~The scientific reliability of this dataset is firmly established through rigorous multi-scale quality control. This includes predictors, and aquifer-stratified machine-learning downscaling. XGBoost was selected as the production model after comparison with DNN using internal evaluation, cross-scale consistency evaluations against native GRACE observations, and independent validation utilizing 1,518 global in situ monitoring wells combined with a geographically zoned specific yield matrix. groundwater observations.~~

Unlike traditional coarse-resolution global hydrological models, this 1 km product uniquely resolves the fine-scale spatial heterogeneity of deep-aquifer systems. It successfully captures sharp, nonlinear boundaries driven by localized intensive agricultural pumping, revealing critical depletion hotspots that were previously obscured by spatial smoothing. Ultimately, this publicly available dataset provides a robust, observation-constrained physical benchmark for regional water-resource management, Earth-system modeling, and the development of groundwater digital twins.

The downscaled product retained the large-scale information of the original GRACE-derived GWSA. After aggregation to the original resolution, the global comparison yielded $R^2 = 0.972$ and $RMSE = 2.10$ cm. The spatially independent in-situ evaluation used S_{yeff} estimated only from calibration blocks and transferred to held-out blocks before comparison at a common 0.25° grid. For XGBoost, the held-out evaluation yielded ($R^2 = 0.538$, $RMSE = 0.378$ cm yr⁻¹, and $Bias = 0.106$ cm yr⁻¹). Sensitivity experiments showed that uncertainty in S_{yeff} affected amplitude-related metrics more strongly than correlation and trend-sign agreement. Standardized and monthly time-series comparisons further showed that the product generally reproduced long-term groundwater variability better than seasonal variability.

Regional comparisons with published groundwater-storage estimates provided additional information on product performance. Agreement was strongest in the North China Plain, whereas depletion magnitudes were weaker than published estimates in northwestern India and California. In the High Plains Aquifer, the reconstructed trend during 2003–2013 had the opposite sign to the published estimate. These comparisons show that temporal correlation alone is not sufficient to evaluate agreement in storage magnitude and long-term trend. Experiments using CSR, JPL, H08, and WaterGAP also showed that reconstruction performance and predictor importance depend on the target product used for downscaling.

The 1 km dataset provides finer spatial information than the original GRACE-scale GWSA, but the kilometre-scale patterns are inferred from environmental predictors under coarse-scale GRACE constraints. They should therefore not be interpreted as direct observations of local groundwater pumping or as evidence of a single causal process. Residual differences from global hydrological models can be used to identify areas where groundwater changes are not explained by the GHM-simulated background, but local attribution requires additional observations of groundwater level, groundwater abstraction, recharge, and aquifer properties. Within these limitations, the dataset can support regional groundwater-

change analysis, hydrological-model evaluation, and studies of groundwater responses to climatic and human influences.

6. Data available

The datasets used to develop the High-resolution Global ~~Groundwater Storage Anomalies~~GWSA Dataset are publicly available and are described in detail in the Data Sources section. The monthly downscaled GWSA data and trend products can be previewed through the Google Earth Engine application (<https://yifeifan.projects.earthengine.app/view/high-resolution-gwsa>). In addition, we provide the original high-precision dataset in ~~multi-band~~multiband Tagged Image File Format (TIFF format) files, covering the period from April 2002 to December 2020. The dataset is openly available at <https://doi.org/10.12072/ncdc.hydrology.db7323.2026> (Fan et al., ~~2026~~2026). Public regional reference data used in the additional benchmarking are available from Figshare (<https://doi.org/10.6084/m9.figshare.22819976>) and Zenodo (<https://doi.org/10.5281/zenodo.7392554>); the remaining reference curves were digitized from Rodell et al. (2009) and Rateb et al. (2020).

7. Author contributions

Chen Zhang: Writing – review & editing, Writing – original draft, Data curation, Visualization, Formal analysis, Investigation. Yifei Fan: Writing – review & editing, Writing – original draft, Data curation, Visualization, Formal analysis, Investigation. Wenxia Han: Writing – review & editing, Writing – original draft, Funding acquisition, Supervision. Tao Che: Writing – review & editing, Writing – original draft, Funding acquisition, Supervision. Jiabei Zhu: Formal analysis, Supervision. Yingchun Ge: Conceptualization, Methodology, Supervision, Investigation, Formal analysis.

8. Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

9. Acknowledgments

This work was supported by the National Natural Science Foundation of China (grant no. 42595534), the Gansu Provincial Science and Technology Program (grant no. 23ZDKA0004), the Qinghai Province Major Science and Technology Project (grant no. 2025-SF-A2), and financial support from Qinghai Province through the Kunlun Talent Program for Wenxia Han. We sincerely acknowledge the National Earth System Science Data Center and the China Centre for Resources Satellite Data and Application for providing the data used in this study. We also acknowledge the Google Earth Engine platform for providing cloud-based geospatial data access and processing capabilities.

Reference

- Akiba, T., Sano, S., Yanase, T., Ohta, T., and Koyama, M.: Optuna: A Next-generation Hyperparameter Optimization Framework, Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2623–2631, <https://doi.org/10.1145/3292500.3330701>, 2019.
- Ali, S., Pham, Q. B., Sajjad, A., et al.: A comparative study of machine learning models for groundwater storage anomaly downscaling, *Journal of Hydrology*, 620, 129424, <https://doi.org/10.1016/j.jhydrol.2023.129424>, 2023.
- [Bäthge, A., Ruz Vargas, C., Lischeid, G., Collenteur, R., Cuthbert, M., Fleckenstein, J., Flörke, M., de Graaf, I., Gnann, S., Hartmann, A., Huggins, X., Moosdorf, N., Wada, Y., Wagener, T., and Reinecke, R.: A Global-Scale Time Series Dataset for Groundwater Studies within the Earth System, *Scientific Data*, 13, 225, <https://doi.org/10.1038/s41597-026-06966-1>, 2026.](#)
- Bierkens, M. F. P. and Wada, Y.: Non-renewable groundwater use and groundwater depletion: a review, *Environ. Res. Lett.*, 14, 063002, <https://doi.org/10.1088/1748-9326/ab1a5f>, 2019.
- [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.](#)
- Breiman, L.: Random forests, *Mach. Learn.*, 45, 5–32, <https://doi.org/10.1023/A:1010933404324>, 2001.
- Campbell, P. C., et al.: Global-scale evaluation of the GLDAS-2.1 Noah Land Surface Model, *Journal of Hydrometeorology*, 20, 819–839, <https://doi.org/10.1175/JHM-D-18-0133.1>, 2019.
- Castellazzi, P., Martel, R., Rivera, A., Huang, J., Pavlic, G., Calderhead, A. I., Chaussard, E., Garfias, J., and Salas, J.: Groundwater depletion in Central Mexico: Use of GRACE and InSAR to support water resources management, *Water Resour. Res.*, 52, 5985–6003, <https://doi.org/10.1002/2015WR018211>, 2016.
- Chen, J., Wilson, C. R., Famiglietti, J. S., and Rodell, M.: Attenuation effects on seasonal signals from GRACE time-variable gravity, *Journal of Geodesy*, 81, 237–245, <https://doi.org/10.1007/s00190-006-0104-2>, 2007.
- Chen, T. and Guestrin, C.: XGBoost: A Scalable Tree Boosting System, Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794, <https://doi.org/10.1145/2939672.2939785>, 2016.
- Chen, Y., Zhang, Z., and Khan, S.: Understanding the impact of irrigation on groundwater storage using

1825 XGBoost, *Agricultural Water Management*, 240, 106297,
 1826 <https://doi.org/10.1016/j.agwat.2020.106297>, 2020.
 1827 Cleveland, R. B., Cleveland, W. S., McRae, J. E., and Terpenning, I.: STL: A seasonal-trend
 1828 decomposition procedure based on loess, *Journal of Official Statistics*, 6, 3–73, 1990.
 1829 Cooley, S. W., Ryan, J. C., and Smith, L. C.: Human alteration of global surface water storage variability,
 1830 *Nature*, 591, 78–81, <https://doi.org/10.1038/s41586-021-03262-3>, 2021.
 1831 Ek, M. B., Mitchell, K. E., Lin, Y., Rogers, E., Grunmann, P., Koren, V., Gayno, G., and Tarpley, J. D.:
 1832 Implementation of Noah land surface model changes in the National Centers for Environmental
 1833 Prediction operational mesoscale Eta model, *J. Geophys. Res.*, 108, 8851,
 1834 <https://doi.org/10.1029/2002JD003296>, 2003.
 1835 Famiglietti, J. S.: The global groundwater crisis, *Nat. Clim. Change*, 4, 945–948,
 1836 <https://doi.org/10.1038/nclimate2425>, 2014.
 1837 Famiglietti, J. S., Lo, M., Ho, S. L., Bethune, J., Anderson, K. J., Syed, T. H., Swenson, S. C., de Linage,
 1838 C. R., and Rodell, M.: Satellites measure recent rates of groundwater depletion in California's
 1839 Central Valley, *Geophys. Res. Lett.*, 38, L03403, <https://doi.org/10.1029/2010GL046442>, 2011.
 1840 Fan Yifei, Zhang Chen, Han Wenxia, Che Tao, Ge Yingchun. 1km high-resolution monthly-scale GWSA
 1841 dataset. National Cryosphere Desert Data Center(<http://www.ncdc.ac.cn>), 2026.
 1842 <https://str.ei/CSTR-11738-11-doi.org/10.12072/ncdc.hydrology.db7323.2026>.
 1843 Giroto, M. and Rodell, M.: Terrestrial water storage anomalies and the Earth's energy imbalance, *Current*
 1844 *Climate Change Reports*, 5, 290–301, <https://doi.org/10.1007/s40641-019-00140-5>, 2019.
 1845 Hastie, T., Tibshirani, R., and Friedman, J.: *The Elements of Statistical Learning: Data Mining, Inference,*
 1846 *and Prediction*, 2nd Edn., Springer, New York, NY, USA, 745 pp., 2009.
 1847 Jasechko, S., Seybold, H., Perrone, D., Fan, Y., Kirchner, J. W., Taylor, R. G., and Tindimugaya, C.:
 1848 Rapid groundwater decline and some cases of recovery in aquifers globally, *Nature*, 625, 715–
 1849 721, <https://doi.org/10.1038/s41586-023-06879-8>, 2024.
 1850 Karger, D. N., Conrad, O., Böhrer, J., Kawege, T. W., Kreft, H., RW, M., Zimmermann, N. E., and Kessler,
 1851 M.: Climatologies at high resolution for the earth's land surface areas, *Scientific Data*, 4, 170122,
 1852 <https://doi.org/10.1038/sdata.2017.122>, 2017.
 1853 Karniadakis, G. E., Kevrekidis, I. G., Lu, L., Perdikaris, P., Wang, S., and Yang, L.: Physics-informed
 1854 machine learning, *Nat. Rev. Phys.*, 3, 422–440, <https://doi.org/10.1038/s42254-021-00314-5>,
 1855 2021.
 1856 Kendall, M. G.: *Rank Correlation Methods*, 4th Edn., Charles Griffin, London, 1975.
 1857 Khan, S., et al.: Global groundwater storage anomaly downscaling using tree-based ensembles, *Journal*
 1858 *of Hydrology*, 630, 130635, <https://doi.org/10.1016/j.jhydrol.2024.130635>, 2024.
 1859 Kondrashov, D. and Ghil, M.: Spatio-temporal filling of missing points in geophysical data sets,
 1860 *Nonlinear Processes in Geophysics*, 13, 151–159, <https://doi.org/10.5194/npg-13-151-2006>,
 1861 2006.
 1862 Kratzert, F., Klotz, D., Shalev, G., Klambauer, G., Hochreiter, S., and Nearing, G.: Towards learning
 1863 universal, regional, and local hydrological behaviors via machine learning applied to large-
 1864 sample datasets, *Hydrol. Earth Syst. Sci.*, 23, 5089–5110, [https://doi.org/10.5194/hess-23-5089-](https://doi.org/10.5194/hess-23-5089-2019)
 1865 2019, 2019.
 1866 Landerer, F. W. and Swenson, S. C.: Accuracy of scaled GRACE terrestrial water storage estimates,
 1867 *Water Resour. Res.*, 48, W04531, <https://doi.org/10.1029/2011WR011453>, 2012.
 1868 LeCun, Y., Bengio, Y., and Hinton, G.: Deep learning, *Nature*, 521, 436–444,

1869 <https://doi.org/10.1038/nature14539>, 2015.

1870 Li, J., et al.: Global quantification of groundwater seasonality and its contribution to total storage
1871 variability, *Earth System Science Data*, 17, 2025.

1872 Liu, P.-W., Famiglietti, J. S., Purdy, A. J., Adams, K. H., McEvoy, A. L., Reager, J. T., Bindlish, R., Wiese,
1873 D. N., David, C. H., and Rodell, M.: Groundwater depletion in California's Central Valley
1874 accelerates during megadrought, *Nat. Commun.*, 13, 7825, [https://doi.org/10.1038/s41467-022-](https://doi.org/10.1038/s41467-022-35582-x)
1875 [35582-x](https://doi.org/10.1038/s41467-022-35582-x), 2022.

1876 Long, D., Longuevergne, L., and Scanlon, B. R.: Global analysis of spatiotemporal variability in merged
1877 total water storage changes using multiple GRACE products and global hydrological models,
1878 *Remote Sens. Environ.*, 170, 117–133, <https://doi.org/10.1016/j.rse.2015.09.016>, 2015.

1879 Mankoff, K. D., Solgaard, A., Colgan, W., Ahlstrøm, A. P., Khan, S. A., and Fausto, R. S.: Greenland Ice
1880 Sheet solid ice discharge from 1986 through 2019, *Earth Syst. Sci. Data*, 12, 1367–1383,
1881 <https://doi.org/10.5194/essd-12-1367-2020>, 2020.

1882 Mann, H. B.: Nonparametric tests against trend, *Econometrica*, 13, 245–259,
1883 <https://doi.org/10.2307/1907187>, 1945.

1884 Miro, M. E. and Famiglietti, J. S.: Downscaling GRACE remote sensing datasets to high-resolution
1885 groundwater storage changes in California's Central Valley, *Remote Sens.*, 10, 143,
1886 <https://doi.org/10.3390/rs10010143>, 2018.

1887 Müller Schmied, H., Cáceres, D., Eisner, S., Flörke, M., Herbert, C., Niemann, C., Peiris, T. A., Popat,
1888 E., Portmann, F. T., Reinecke, R., Schumacher, M., Shadkam, S., Telteu, C.-E., Trautmann, T.,
1889 and Döll, P.: The global water resources and use model WaterGAP v2.2d: model description and
1890 evaluation, *Geosci. Model Dev.*, 14, 1037–1079, <https://doi.org/10.5194/gmd-14-1037-2021>,
1891 2021.

1892 Niu, G. Y., Yang, Z. L., Mitchell, K. E., Chen, F., Ek, M. B., Barlage, M., Kumar, A., Manning, K., Niyogi,
1893 D., Rosero, E., Tewari, M., and Xia, Y.: The community Noah land surface model with
1894 multiparameterization options (Noah-MP): 1. Model description and evaluation with local-scale
1895 measurements, *J. Geophys. Res.*, 116, D12109, <https://doi.org/10.1029/2010JD015139>, 2011.

1896 Pekel, J.-F., Cottam, A., Gorelick, N., and Belward, A. S.: High-resolution mapping of global surface
1897 water and its long-term changes, *Nature*, 540, 418–422, <https://doi.org/10.1038/nature20584>,
1898 2016.

1899 Poggio, L., de Sousa, L. M., Batjes, N. H., Heuvelink, G. B. M., Kempen, B., Ribeiro, E., and Rossiter,
1900 D.: SoilGrids 2.0: producing soil information for the world with quantified uncertainty, *SOIL*,
1901 7, 217–244, <https://doi.org/10.5194/soil-7-217-2021>, 2021.

1902 Rateb, A., Scanlon, B. R., Pool, D. R., Sun, A., Zhang, Z., Chen, J., Clark, B., Faunt, C. C., Haugh, C. J.,
1903 Hill, M., Hobza, C., McGuire, V. L., Reitz, M., Müller Schmied, H., Sutanudjaja, E. H.,
1904 Swenson, S., Wiese, D., Xia, Y., and Zell, W.: Comparison of groundwater storage changes from
1905 GRACE satellites with monitoring and modeling of major U.S. aquifers, *Water Resour. Res.*,
1906 56, e2020WR027556, <https://doi.org/10.1029/2020WR027556>, 2020.

1907 Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., and Prabhat: Deep
1908 learning and process understanding for data-driven Earth system science, *Nature*, 566, 195–204,
1909 <https://doi.org/10.1038/s41586-019-0912-1>, 2019.

1910 Roberts, D. R., Bahn, V., Ciuti, S., et al.: Cross-validation strategies for data with temporal, spatial,
1911 hierarchical, or phylogenetic structure, *Ecography*, 40, 913–929,
1912 <https://doi.org/10.1111/ecog.02881>, 2017.

- 1913 Rodell, M. and Famiglietti, J. S.: An analysis of terrestrial water storage variations in Illinois with
1914 implications for the Gravity Recovery and Climate Experiment (GRACE), *Water Resour. Res.*,
1915 37, 2327–2339, <https://doi.org/10.1029/2000WR900306>, 2001.
- 1916 Rodell, M., Houser, P. R., Jambor, U., Gottschalk, J., Mitchell, K., Meng, C. J., Arsenault, K., Cosgrove,
1917 B., Radakovich, J., Bosilovich, M., Entin, J. K., Walker, J. P., Lohmann, D., and Toll, D.: The
1918 Global Land Data Assimilation System, *Bull. Amer. Meteorol. Soc.*, 85, 381–394,
1919 <https://doi.org/10.1175/BAMS-85-3-381>, 2004.
- 1920 Rodell, M., Velicogna, I., and Famiglietti, J. S.: Satellite-based estimates of groundwater depletion in
1921 India, *Nature*, 460, 999–1002, <https://doi.org/10.1038/nature08238>, 2009.
- 1922 Rodell, M., Famiglietti, J. S., Wiese, D. N., Reager, J. T., Beaudoin, H. K., Landerer, F. W., and Lo, M.
1923 H.: Emerging trends in global freshwater availability, *Nature*, 557, 651–659,
1924 <https://doi.org/10.1038/s41586-018-0123-1>, 2018.
- 1925 Save, H., Bettadpur, S., and Tapley, B. D.: High-resolution CSR GRACE RL05 mascons, *J. Geophys.*
1926 *Res. Solid Earth*, 121, 3544–3564, <https://doi.org/10.1002/2016JB013007>, 2016.
- 1927 Scanlon, B. R., Faunt, C. C., Longuevergne, L., Reedy, R. C., Alley, W. M., McGuire, V. L., and Anyamba,
1928 A.: Groundwater depletion and sustainability of irrigation in the US High Plains and Central
1929 Valley, *Proc. Natl. Acad. Sci. U.S.A.*, 109, 9320–9325,
1930 <https://doi.org/10.1073/pnas.1200311109>, 2012.
- 1931 Scanlon, B. R., Zhang, Z., Save, H., Wiese, D. N., Landerer, F. W., Royetti, D., Longuevergne, L., and
1932 Chen, J.: Global evaluation of new GRACE mascon products for hydrologic applications, *Water*
1933 *Resour. Res.*, 52, 9412–9429, <https://doi.org/10.1002/2016WR019494>, 2016.
- 1934 Scanlon, B. R., Zhang, Z., Save, H., Sun, A. Y., Müller Schmied, H., van Beek, L. P. H., Wiese, D. N.,
1935 Wada, Y., Long, D., Reedy, R. C., Longuevergne, L., Döll, P., and Bierkens, M. F. P.: Global
1936 models underestimate large decadal declining and rising water storage trends relative to
1937 GRACE satellite data, *Proc. Natl. Acad. Sci. U.S.A.*, 115, E1080–E1089,
1938 <https://doi.org/10.1073/pnas.1704665115>, 2018.
- 1939 Sen, P. K.: Estimates of the regression coefficient based on Kendall's tau, *Journal of the American*
1940 *Statistical Association*, 63, 1379–1389, <https://doi.org/10.1080/01621459.1968.10480934>,
1941 1968.
- 1942 Seyoum, W. M., Kwon, D., and Milewski, A. M.: Downscaling GRACE TWSA data into high-resolution
1943 groundwater level anomaly using machine learning-based models in a glacial aquifer system,
1944 *Remote Sens.*, 11, 824, <https://doi.org/10.3390/rs11070824>, 2019.
- 1945 Shen, C.: A transdisciplinary review of deep learning research and its relevance for water resources
1946 scientists, *Water Resour. Res.*, 54, 8558–8593, <https://doi.org/10.1029/2018WR022643>, 2018.
- 1947 Sun, A. Y., Scanlon, B. R., AghaKouchak, A., and Zhang, Z.: Using Big Data to Advance Our
1948 Understanding of the Earth's Hydrologic Cycle: A Case Study in Space-Based Monitoring of
1949 Water Resources, *Front. Earth Sci.*, 7, 214, <https://doi.org/10.3389/feart.2019.00214>, 2019.
- 1950 Theil, H.: A rank-invariant method of linear and polynomial regression analysis, *Nederl. Akad. Wetensch.*
1951 *Proc.*, 53, 1397–1412, 1950.
- 1952 Tripathi, V., Groh, A., Horwath, M., and Ramsankaran, R.: Scaling methods of leakage correction in
1953 GRACE mass change estimates revisited for the complex hydro-climatic setting of the Indus
1954 Basin, *Hydrol. Earth Syst. Sci.*, 26, 4515–4535, <https://doi.org/10.5194/hess-26-4515-2022>,
1955 2022.
- 1956 Tyralis, H., Papacharalampous, G., and Langousis, A.: A brief review of random forests for water

1957 scientists and practitioners and their recent history in water resources, *Water*, 11, 910,
1958 <https://doi.org/10.3390/w11050910>, 2019.

1959 Vautard, R., Yiou, P., and Ghil, M.: Singular-spectrum analysis: A toolkit for short, noisy chaotic signals,
1960 *Physica D: Nonlinear Phenomena*, 58, 95–126, [https://doi.org/10.1016/0167-2789\(92\)90103-T](https://doi.org/10.1016/0167-2789(92)90103-T),
1961 1992.

1962 Wada, Y., van Beek, L. P. H., van Kempen, C. M., Reckman, J. W. T. M., and Bierkens, M. F. P.: Global
1963 depletion of groundwater resources, *Geophys. Res. Lett.*, 37, L20402,
1964 <https://doi.org/10.1029/2010GL044571>, 2010.

1965 Wang, J., Yin, W., Hu, L., Zhang, M., and Han, S. C.: Spatiotemporal evaluation of GRACE-derived
1966 groundwater storage anomalies using machine learning, *Journal of Hydrology*, 612, 128219,
1967 <https://doi.org/10.1016/j.jhydrol.2022.128219>, 2022.

1968 Weber, T., et al.: Global depletion of groundwater is a critical threat to water security, *Environmental*
1969 *Research Letters*, 11, 064003, <https://doi.org/10.1088/1748-9326/11/6/064003>, 2016.

1970 Wiese, D. N., Landerer, F. W., and Watkins, M. M.: Quantifying and reducing leakage errors in the JPL
1971 RL05M GRACE mascon solution, *Water Resources Research*, 52, 7490–7502,
1972 <https://doi.org/10.1002/2016WR019344>, 2016.

1973 Willard, J., Jia, X., Xu, S., Steinbach, M., and Kumar, V.: Integrating physics-based modeling with
1974 machine learning: A survey, *ACM Comput. Surv.*, 55, 1–34, <https://doi.org/10.1145/3514228>,
1975 2022.

1976 Yi, S. and Sneeuw, N.: Filling the data gaps within and between GRACE and GRACE-FO missions using
1977 singular spectrum analysis, *Geophysical Journal International*, 224, 1393–1411,
1978 <https://doi.org/10.1093/gji/ggaa535>, 2021.

1979 Yin, W., Hu, L., Zhang, M., Wang, J., and Han, S. C.: Statistical downscaling of GRACE-derived
1980 groundwater storage using evapotranspiration and precipitation, *J. Hydrol.*, 612, 128219,
1981 <https://doi.org/10.1016/j.jhydrol.2022.128219>, 2022.

1982 Zhang, J., Hu, L., Sun, J., and Wang, D.: Reconstructing groundwater storage changes in the North China
1983 Plain using a numerical model and GRACE data, *Remote Sens.*, 15, 3264,
1984 <https://doi.org/10.3390/rs15133264>, 2023.

1985 Zhang, Y., Kong, D., Gan, R., Chiew, F. H. S., McVicar, T. R., Zhang, Q., and Yang, Y.: Coupled
1986 estimation of 500 m and 8-day resolution global actual evapotranspiration and gross primary
1987 production in 2002–2017, *Remote Sensing of Environment*, 222, 165–182,
1988 <https://doi.org/10.1016/j.rse.2018.12.031>, 2019.

1989 Zhang, Z., et al.: Global groundwater storage anomaly downscaling using tree-based ensembles, *Journal*
1990 *of Hydrology*, 601, 126620, <https://doi.org/10.1016/j.jhydrol.2021.126620>, 2021.

1991 Zheng, H., et al.: A global daily surface soil moisture dataset at 1 km resolution, *Scientific Data*, 10, 143,
1992 <https://doi.org/10.1038/s41597-023-01991-w>, 2023.