To: The Editor and Reviewer 2, *Earth System Science Data (ESSD)*

Subject: Point-by-point response to Reviewer 2 Comments

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

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.

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

Major Comments

1. 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 surface-water-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. 9, 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.

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 land-water-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 GLDAS-Noah 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, the Noah soil-moisture profile was treated as the vertically integrated storage of its four soil layers (0–10, 10–40, 40–100, and 100–200 cm), rather than as surface soil moisture alone. Snow water equivalent and vegetation canopy water storage were taken from the corresponding Noah storage outputs. These storage components were converted to a common water-equivalent thickness representation in centimetres before the component subtraction. The GLDAS variables used here are therefore distinct from the separate 1 km surface-soil-moisture dataset described in Sect. 2.1.3, which serves only as a high-resolution predictor in the downscaling model and does not enter Eq. (15).

The GLDAS components were processed at the monthly scale, and anomalies were calculated relative to the same 2004–2009 reference period used for GRACE. Specifically, the corresponding 2004–2009 mean was removed from each monthly storage component before applying Eq. (15). The resulting GWSA calculation was therefore performed consistently at the monthly scale as

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

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 non-groundwater 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.

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.

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, Syeff was estimated only from calibration blocks, and final statistics were calculated only from held-out 0.25° grids.

This block design provides the spatial-generalization evidence used to assess final product reliability; 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. 2.4.1, 2.5, and 4.3 now make this limitation and scope explicit.

5. Line 425: I agree with the comment from the first referee regarding the specific-yield data. Given that Sy 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 Sy 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 scale-dependent effective specific yield, S_{yeff}, used as a level-to-storage conversion coefficient for harmonizing groundwater-level trends with equivalent-water-height groundwater-storage trends. S_{yeff} is treated as an effective observational conversion parameter rather than as a direct estimate of local aquifer-specific yield, and it is used only in the in-situ evaluation framework.

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

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

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

Table 3

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

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

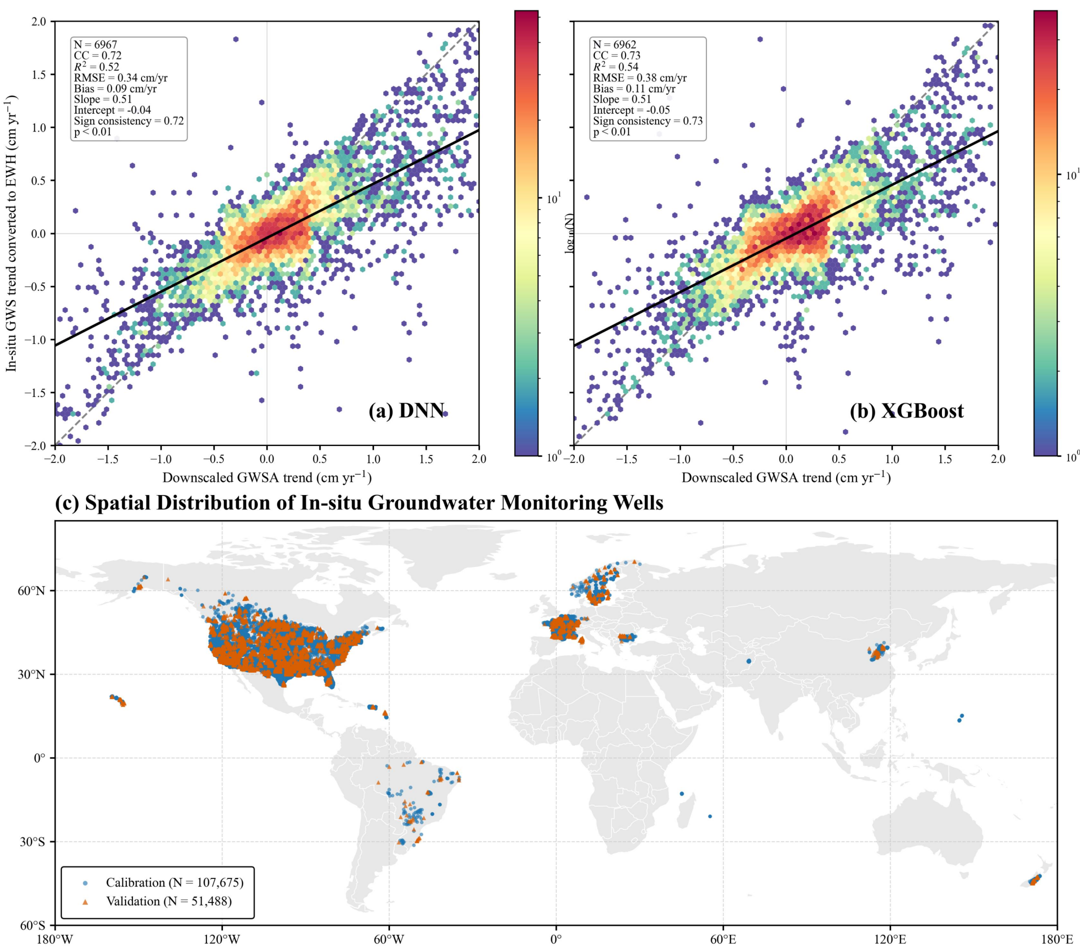

Figure 5. Spatially independent, quality-controlled in-situ evaluation of long-term downscaled groundwater-storage trends and spatial distribution of the monitoring network. (a) Held-out 0.25° grid comparison between DNN-derived GWSA trends and groundwater-level trends converted to equivalent water height using S_{yeff} estimated only from calibration spatial blocks and blindly transferred to the held-out blocks. (b) Corresponding held-out comparison for XGBoost. The network was partitioned into 1° spatial blocks; approximately 70% were calibration blocks and 30% were held-out validation blocks, with all wells from a block assigned to one subset. Neither held-out wells nor their blocks contributed to 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, R2, RMSE, Bias, slope, intercept, trend-sign consistency, and significance) are calculated exclusively from held-out 0.25° grids. (c) Monitoring-well distribution 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 land-management 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.

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. 5, 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” in the revised manuscript refers to the removal of the previous 15-bin trend aggregation, rather than to the elimination of the physically necessary point-to-grid harmonization.

Revised Fig. 5 reports all retained held-out grid-level comparisons, with N = 6,967 for DNN and N = 6,962 for XGBoost. The figure and accompanying text report CC, R^2, RMSE, Bias, regression slope and intercept, trend-sign consistency, and p-value. Panel (c) additionally shows the spatial distribution of the groundwater monitoring network and the calibration/validation allocation, allowing the geographical representativeness of the evaluation dataset to be assessed directly.

The revised methods now also state explicitly that N denotes the number of valid held-out 0.25° grid comparisons after common-grid aggregation and quality control, rather than the number of individual monitoring wells. This distinction makes the remaining spatial harmonization transparent while avoiding the smoothing effect introduced by the previous 15-bin presentation.

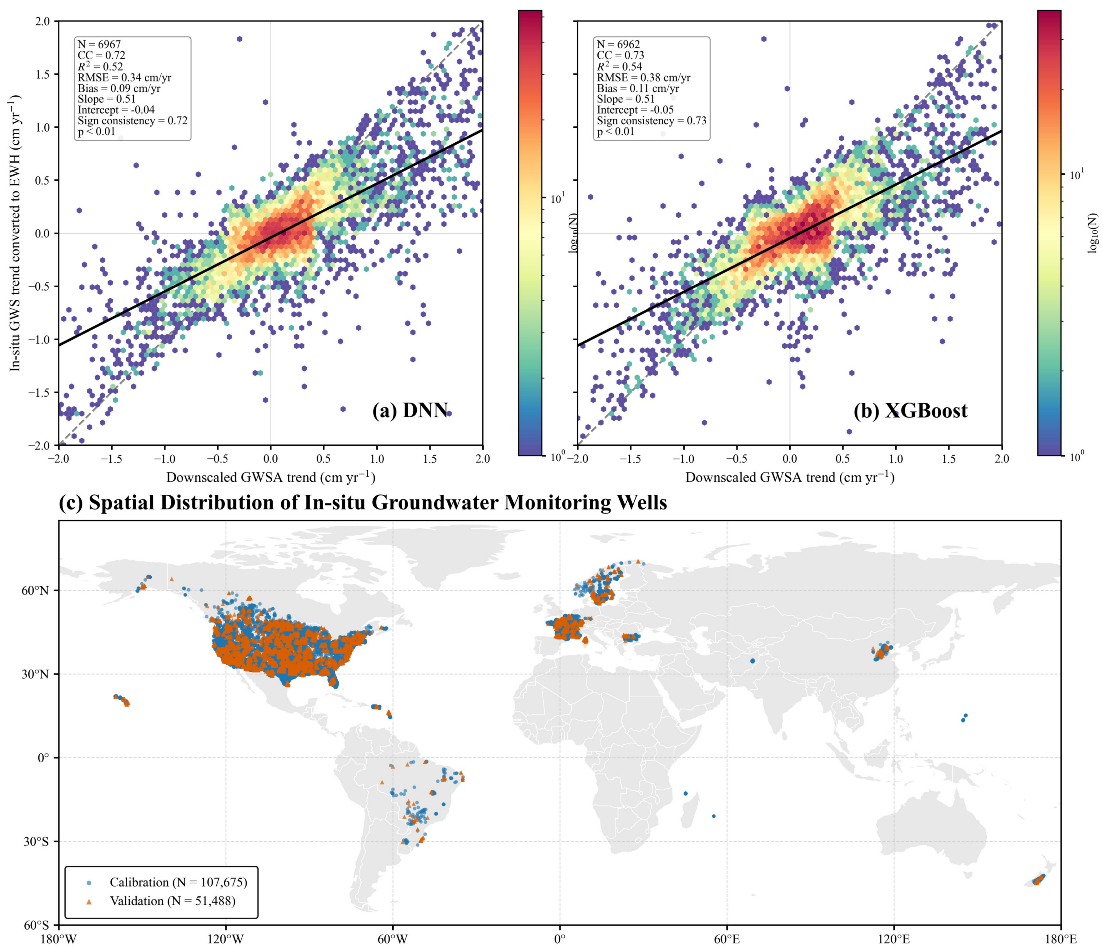

Figure 5. Spatially independent, quality-controlled in-situ evaluation of long-term downscaled groundwater-storage trends and spatial distribution of the monitoring network. (a) Held-out 0.25° grid comparison between DNN-derived GWSA trends and groundwater-level trends converted to equivalent water height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

reported statistics (CC, R2, RMSE, Bias, slope, intercept, trend-sign consistency, and significance) are calculated exclusively from held-out 0.25° grids. (c) Monitoring-well distribution 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 modeled natural background, rather than treating them as direct estimates of groundwater pumping.

Response:

We agree and adopted this interpretation throughout. The section has been retitled 'Residual Signals Not Explained by the Modeled Natural Background.' Residuals are now described as diagnostic departures that may contain human influence, GHM structural error, 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.

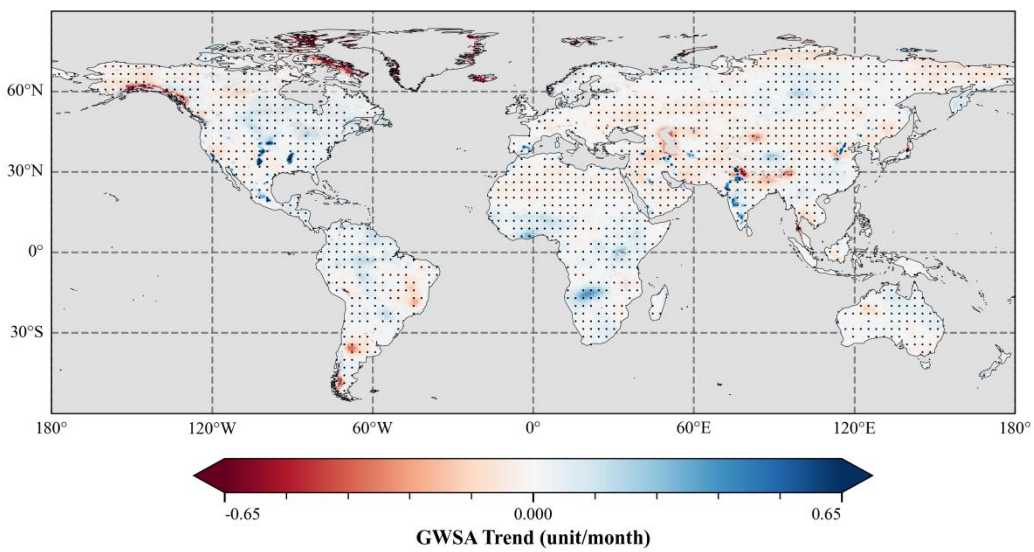

Figure 14. Residual signals not explained by the modeled natural background ($p < 0.01$). The analysis bilinearly interpolates GHM-simulated natural-background signals to 1 km and subtracts them from the 1 km downscaled GWSA reconstruction sequence. Trend computations use the Theil-Sen median method, followed by STL de-seasonalization. Red regions indicate negative residual depletion signals not explained by the modeled natural 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