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 Green and High-end Utilization of Salt Lake Resources, Qinghai Institute of Salt Lakes, Chinese Academy of Sciences, Xining, 810008, 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 Qinghai Provincial Key Laboratory of Geology and Environment of Salt Lakes, Qinghai Institute of Salt Lakes, Chinese Academy of Sciences, Xining, 810008, China

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: Groundwater depletion poses a challenge to water security, but the coarse spatial resolution of Gravity Recovery and Climate Experiment (GRACE) observations limits the characterization of groundwater variability at regional and local scales. Here, we present a 1 km monthly global groundwater storage anomaly (GWSA) dataset covering April 2002–December 2020. The dataset was generated using Singular Spectrum Analysis (SSA) for temporal gap filling, 19 high-resolution hydroclimatic and geophysical predictors, and aquifer-stratified machine-learning downscaling. Extreme Gradient Boosting (XGBoost) and deep neural networks (DNN) were evaluated as candidate models, and XGBoost was selected for dataset production based on internal, cross-scale, and in-situ evaluations. When aggregated to the original GRACE scale, the 1 km product agreed closely with the original GWSA (R^2 = 0.972, *RMSE* = 2.10 cm). Independent evaluation used a global archive of 152,069 groundwater-monitoring wells and a 1° spatial-block calibration–validation design. Effective specific yield (S_{yeff}) was estimated 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}. The association between modeled and observation-derived trends remained similar under ±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 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-level, abstraction, and hydrogeological observations.

Keywords: Groundwater storage anomaly; GRACE downscaling; Machine learning; High spatiotemporal resolution; Product uncertainty

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 (Bierkens and Wada, 2019; Rodell et al., 2018; Wada et al., 2010). 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 an effective spatial resolution 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.

Statistical downscaling has 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), scaling-factor methods, random forests (RF), and Extreme Gradient Boosting (XGBoost). However, most applications have been conducted at regional or basin scales (Miro and Famiglietti, 2018; Seyoum et al., 2019), which limits their use for global groundwater assessment. In addition, many downscaled products rely on coarse-resolution global hydrological models for model development and lack constraints from independent groundwater observations (Scanlon et al., 2018). 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 global 1 km monthly GWSA dataset covering April 2002–December 2020. The framework combines SSA for temporal gap filling (Yi and Sneeuw, 2021) with 19 high-

resolution environmental predictors. XGBoost and DNN were evaluated as candidate downscaling models (Reichstein et al., 2019; Shen, 2018), and the production model was selected using 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) was divided into 1° spatial blocks following spatial cross-validation principles (Roberts et al., 2017). Effective specific yield was estimated from calibration blocks and transferred to held-out blocks before groundwater-level and 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 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 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 using 152,069 in-situ groundwater monitoring stations 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.

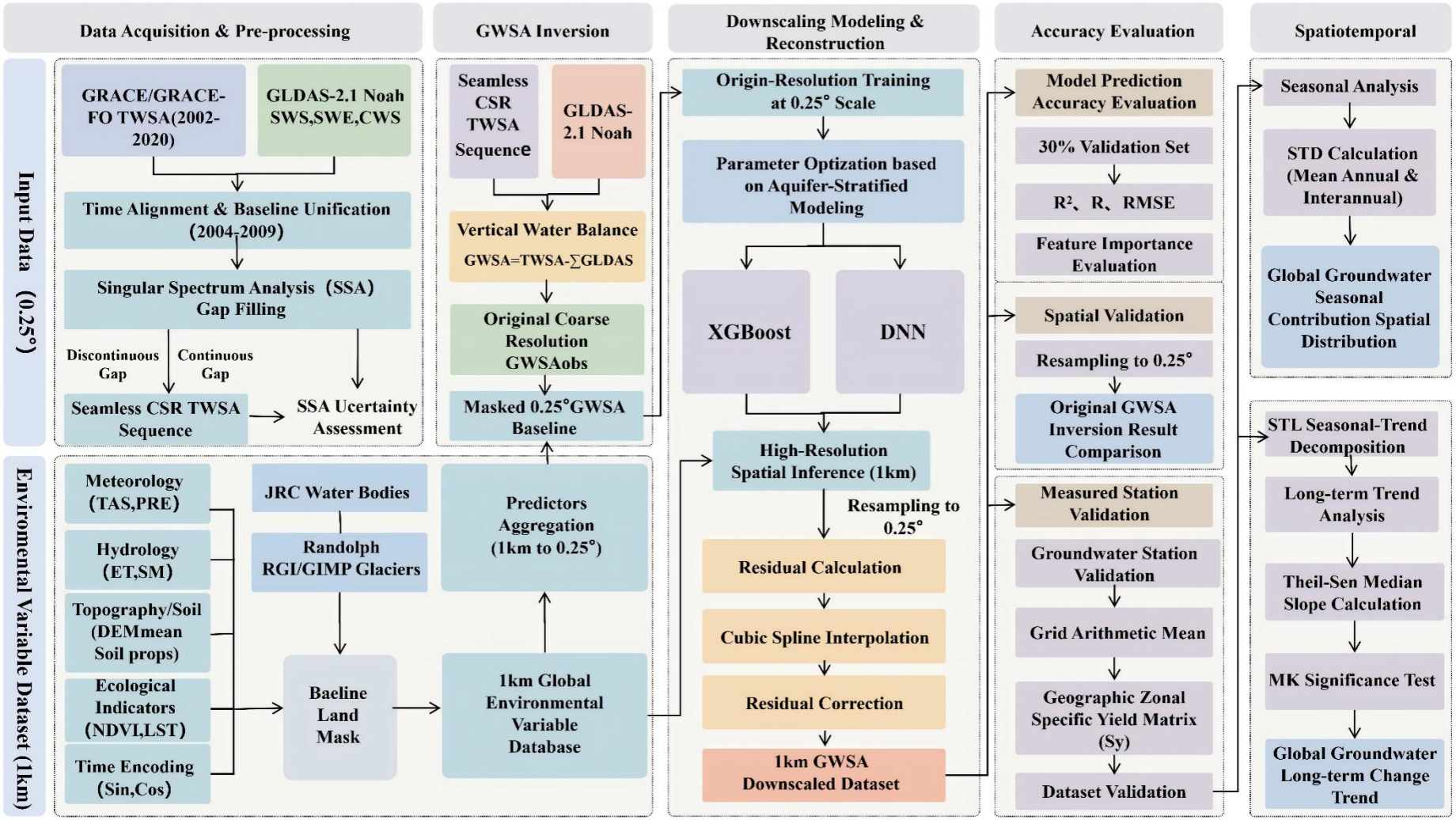
Data Acquisition & Pre-processing
GWSA Inversion
Downscaling Modeling & Reconstruction
Accuracy Evaluation
Spatiotemporal
Input Data (0.25°)
GRACE/GRACE-FO TWSA(2002-2020)
GLDAS-2.1 Noah SWS,SWE,CWS
Time Alignment & Baseline Unification (2004-2009)
Singular Spectrum Analysis (SSA) Gap Filling
Discontinuous Gap
Continuous Gap
Seamless CSR TWSA Sequence
SSA Ucertainty Assessment
Seamless CSR TWSA Sequence
GLDAS-2.1 Noah
Vertical Water Balance
GWSA=TWSA-∑GLDAS
Original Coarse Resolution GWSAobs
Masked 0.25°GWSA Baseline
Origin-Resolution Training at 0.25° Scale
Parameter Optization based on Aquifer-Stratified Modeling
XGBoost
DNN
High-Resolution Spatial Inference (1km)
Resampling to 0.25°
Residual Calculation
Cubic Spline Interpolation
Residual Correction
1km GWSA Downscaled Dataset
Model Prediction Accuracy Evaluation
30% Validation Set
R², R、RMSE
Feature Importance Evaluation
Spatial Validation
Resampling to 0.25°
Original GWSA Inversion Result Comparison
Measured Station Validation
Groundwater Station Validation
Grid Arithmetic Mean
Geographic Zonal Specific Yield Matrix (Sy)
Dataset Validation
Seasonal Analysis
STD Calculation (Mean Annual & Interannual)
Global Groundwater Seasonal Contribution Spatial Distribution
STL Seasonal-Trend Decomposition
Long-term Trend Analysis
Theil-Sen Median Slope Calculation
MK Significance Test
Global Groundwater Long-term Change Trend
Environmental Variable Dataset (1km)
Meteorology (TAS,PRE)
Hydrology (ET,SM)
Topography/Soil (DEMmean Soil props)
Ecological Indicators (NDVI,LST)
Time Encoding (Sin,Cos)
JRC Water Bodies
Randolph RGI/GIMP Glaciers
Baeline Land Mask
Predictors Aggregation (1km to 0.25°)
1km Global Environmental Variable Database

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 difference 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
JPL GRACE/ GRACE-FO	0.5°	Monthly	Apr 2002–Jun 2026	https://grace.jpl.nasa.gov/data/get-data/jpl_global_mascons/
GLDAS-2.1	0.25°	Monthly	Apr 2002–Dec 2020	https://disc.gsfc.nasa.gov/datasets
H08 (ISIMIP3a)	0.5°	Monthly	Apr 2002–Dec 2021	https://data.isimip.org/search/query/h08_20crv3-era5_obsclim_histsoc_default_groundwstor_global_monthly/
WaterGAP 2.2c (ISIMIP2a)	0.5°	Monthly	Apr 2002–Dec 2016	https://data.isimip.org/search/query/watergap2-2c_gswp3-ewembi_nobc_hist_nosoc_co2_groundwstor_global_monthly/
WHYMAP	Vector	–	–	https://www.WHYMAP.org/
Climate data (TAS, PRE)	1000m	Monthly	Apr 2002–Dec 2020	(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
Global groundwater monitoring network	Point	Annual	1801-2022	Jasechko et al. (2024)
GROW	Point	Daily monthly annual	Dec 1838–Aug 2025	Bäthge et al. (2026)

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

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 TWSA 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 (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; Scanlon et al., 2016). The newly released 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 SSA to fill the 35-month data gaps. In combination with the GLDAS, we constructed the 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 lwe_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; Wiese et al., 2016).

2.1.2 GLDAS Noah V2.1

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, 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 (Ek et al., 2003; Niu et al., 2011; Rodell et al., 2004).

In this study, monthly soil water storage (SWS), snow water equivalent (SWE), and vegetation canopy water storage (CWS) from GLDAS Noah V2.1 were used for the GWSA inversion. The soil-water-storage 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 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 2004–2009 means to match the GRACE TWSA anomaly baseline. 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 watergap2-2c_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

Surface hydrological states and fluxes vary across climatic zones, and GLDAS integrates satellite- and ground-based observations with land-surface modeling to simulate these processes (Rodell et al., 2004). 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 TAS and monthly 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 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.

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, Silt, Proportion of Sand, CFVO, CEC, Soil pH, SOC, and 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 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; Bäthge 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 (Sect. 2.5.1), and its sensitivity was further evaluated (Sect. 3.1.2). In addition, uncalibrated standardized comparisons were provided (Sect. 3.2.1) to assess the consistency between the reconstructed GWSA and global in-situ observations.

The GROW dataset covers 53 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, 53 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

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; Yi and Sneeuw, 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 downscaling modeling and temporal analysis. To ensure the accuracy of SSA reconstruction, the full available GRACE/GRACE-FO monthly record from April 2002 to November 2025 was used for the SSA calculation. The SSA calculation method is as follows:

First, define the time series for uniform $\boldsymbol{X}$ sampling:

$$\boldsymbol{X} = [x_1, x_2, \ldots, x_N] \tag{1}$$

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

Define the trajectory matrix $\mathbf{Y}$ for time series $\boldsymbol{X}$:

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

$$L = N - M + 1 \tag{3}$$

$$K = \min(M, L) \tag{4}$$

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

$$\mathbf{Y} = \mathbf{U}\,\mathbf{\Sigma}\,\mathbf{V}^{\mathrm{T}} \tag{5}$$

$$\boldsymbol{e}_i = \mathbf{U}_{(:,i)}\sigma_i \tag{6}$$

$$\boldsymbol{p}_i = \mathbf{V}_{(:,i)} \tag{7}$$

$$\mathbf{Z}_i = \boldsymbol{e}_i \times \boldsymbol{p}_i^{\mathrm{T}} \tag{8}$$

$$\mathbf{Y} = \sum_{i=1}^{K} \mathbf{Z}_i \tag{9}$$

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

The reconstructed component sequence $\boldsymbol{\vartheta}_i$ was obtained by diagonal averaging of each modal matrix $\mathbf{Z}_i$, and the reconstructed time series $\boldsymbol{X}'$ was calculated as follows:

$$\vartheta_i^p = \mathrm{mean}(Z_i^{j,k}),\ (j + k = p + 1) \tag{10}$$

$$\boldsymbol{X}' = \sum_{i=1}^{K'} \boldsymbol{\vartheta_i} \tag{11}$$

where ϑ_i^p represents the value of the ith reconstructed component at the pth position of the time series; $\boldsymbol{\vartheta_i}$ represents the complete reconstructed component sequence composed of ϑ_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, …, N); j and k represent the row and column numbers of $\mathbf{Z}_i$, respectively; K' represents the actual number of retained components; and $\boldsymbol{X}'$ represents the reconstructed time series.

Finally, a two-layer iterative approach is employed for $\boldsymbol{X}$ 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 $\boldsymbol{X}$. In the inner iteration, $\boldsymbol{X}'$ is continuously calculated for each fixed K' until the difference between consecutive reconstruction results ($\boldsymbol{X}'$) 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 refered to the paper by Yi and Sneeuw (Yi & Sneeuw, 2021).

The uncertainty of the SSA reconstruction was evaluated using two error metrics, ε_1 from the SSA-filling-a phase and ε_2 from the SSA-filling-b phase:

$$\varepsilon_1 = \mathrm{Std}(\boldsymbol{X}_{\mathrm{obs}} - \widehat{\boldsymbol{X}}) \tag{12}$$

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

$$\varepsilon = \sqrt{\frac{1}{N}\sum (\boldsymbol{X}_{\mathrm{obs}} - \widehat{\boldsymbol{X}})^2} \tag{14}$$

Both SSA filling and error evaluation were carried out on a pixel-by-pixel basis for the original 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, $\boldsymbol{X}_{\mathrm{obs}}$ and $\widehat{\boldsymbol{X}}$ 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; ε_1 represents the reconstruction uncertainty in this phase, with larger values indicating lower reconstruction accuracy.

In the SSA-filling-b phase, annual masking was applied to the original observation series using 2003–2016 as the cross-validation background, and each year from 2004 to 2015 was sequentially masked to form 12 annual gap-filling experiments. For each masked year, the root mean square error (*RMSE*), denoted as ε, was calculated between the original observations and the reconstructed values using Eq. (14). The 12 annual *RMSE* values were then aggregated using Eq. (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, $\boldsymbol{X}_{\text{obs}}$ and $\widehat{\boldsymbol{X}}$ 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 the number of non-missing months in the original GRACE data for that year, signifying the actual sequence length; ε_2 represents the minimum aggregated *RMSE* from the 12 annual masking experiments, with larger values indicating lower fitting accuracy.

2.3 GWSA Inversion and Masking Strategy

Terrestrial water storage (TWS) represents the vertically integrated water stored above and below the land surface, including soil water, groundwater, snow, vegetation canopy water, surface water, and land ice (Rodell and Famiglietti, 2001). Through the utilization of a vertical water model in conjunction with GRACE and GLDAS data, GWSA can be segregated from TWSA via the subsequent calculation approach:

$$\Delta\text{GWS} = \Delta\text{TWS} - (\Delta\text{SWS} + \Delta\text{SWE} + \Delta\text{CWS}) \tag{15}$$

The 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 over the complete 0–200 cm Noah soil column. SWE and canopy water storage 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 their respective mean values over the 2004–2009 baseline period to ensure consistency with the GRACE anomaly definition.

Here, ΔTWS denotes the terrestrial water storage anomaly derived from GRACE, while ΔGWS represents the GRACE-derived groundwater storage anomaly estimated using the residual component-separation framework. We calculated the observed GWSA sequence ($GWSA_{obs}$; N = 249) from the original GRACE data, and the reconstructed GWSA sequence ($GWSA_{SSA}$; N = 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. The resulting temporally continuous GWSA sequence was subsequently used as the target dataset for downscaling and related accuracy assessments.

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. GRACE signal attenuation can affect seasonal basin-scale water-storage estimates (Chen et al., 2007), while glacier and ice-sheet mass variations introduce additional uncertainty in polar and alpine regions (Mankoff et al., 2020).

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 (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 GWSA at the global scale, separate models were developed for each aquifer type after rasterizing the aquifer map and assigning each pixel to an aquifer category. For each aquifer category, 200,000 grid-month records were randomly sampled across both space and time for hyperparameter optimization. Subsequently, all samples within each aquifer category were used to develop the XGBoost and DNN models, with 70% used for training and 30% for 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. We therefore recommend further reference to the in situ validation results (Sect. 2.5).

(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) Accuracy assessment: This 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 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 data at the original spatial resolution (0.25°) to quantify the deviations from the baseline data.

(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 GWSA and multi-source hydro-climatic variables, two representative machine-learning architectures were employed: XGBoost and DNN. XGBoost is a scalable gradient-boosted tree ensemble that can represent nonlinear statistical mappings (Chen and Guestrin, 2016); machine-learning approaches have also been used to downscale GRACE-derived groundwater signals in regional studies (Miro and Famiglietti, 2018; Seyoum et al., 2019). Its regularization, shrinkage, and row/column subsampling help control model complexity and overfitting (Chen and Guestrin, 2016; Hastie et al., 2009). Complementarily, DNN 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 DNN 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 for hyperparameter optimization. The model was trained using 70% of all samples from each aquifer, while the remaining 30% were used for 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.

2.4.2 Statistical Evaluation Indicators

We evaluated model performance using the coefficient of determination (R^2), Pearson's correlation coefficient (*CC*), bias (*Bias*), mean absolute error (*MAE*), and root mean square error (*RMSE*). R^2 and CC quantify the explained variance and linear agreement between predicted and observed values, respectively, whereas $Bias$, MAE, and $RMSE$ characterize the systematic deviation, mean absolute deviation, and overall magnitude of prediction errors, respectively.

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

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

$$Bias = \frac{1}{n}\sum_{i=1}^{n} (y_i - o_i) \tag{18}$$

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |y_i - o_i| \tag{19}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (o_i - y_i)^2}{n}} \tag{20}$$

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 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:

$$\Delta GWS_{obs} = S_{yeff} \times \Delta GWL_{obs} \tag{21}$$

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 Comparison with Published Estimates

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.

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 strongly affected by periodic seasonal variability, and seasonal extremes can offset or amplify longer-term changes (Cooley et al., 2021). Because this behavior is often underrepresented in groundwater assessments, we used the reconstructed high-resolution GWSA series to quantify the seasonal contribution to global groundwater dynamics at the pixel scale.

$$C_{seasonality} = \frac{\bar{\sigma}}{\bar{\sigma} + \delta} \tag{22}$$

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. $C_{\text{seasonality}}$ 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 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 $\boldsymbol{S_t}$ from the original time series $\boldsymbol{Y_t}$, the de-seasonalized sequence $\boldsymbol{X_t}$ 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).

$$\boldsymbol{X_t} = \boldsymbol{Y_t} - \boldsymbol{S_t} \tag{23}$$

where the seasonally adjusted $\boldsymbol{X_t}$ series comprehensively preserves both the trend component signal of the 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:

$$\beta = \text{Median}\left(\frac{X_j - X_i}{j - i}\right), \quad \forall j > i \tag{24}$$

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, 1970; Mann, 1945).

The test statistic was calculated using the conventional Mann–Kendall formulation. This study conducted trend tests for GWSA changes at significance levels of $p < 0.05$ and $p < 0.01$.

3 Results

3.1 Model Comparison and Validation

3.1.1 Temporal Continuity and SSA-based Reconstruction Quality

To ensure the spatiotemporal continuity of the downscaling baseline without resorting to artificial annual averaging, SSA was employed to reconstruct missing GRACE observations (Fig. 2). Global evaluation 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.

To empirically represent the reconstruction capability under conditions of long-term continuous missing data, 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 1st and 99th percentiles at 0.732 cm and 10.204 cm, respectively. Focusing specifically on terrestrial regions (excluding Antarctica), the median error was 2.708 cm, with 1st and 99th percentiles of 0.862 cm and 15.295 cm. While the errors in land areas are slightly higher than the global average, they remain consistently low and 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 associated with large natural signal amplitudes, significant seasonal oscillations, and complex local processes rather than algorithmic instability. Consequently, the SSA reconstruction provides a temporally continuous baseline with quantified reconstruction uncertainty for the subsequent high-resolution downscaling.

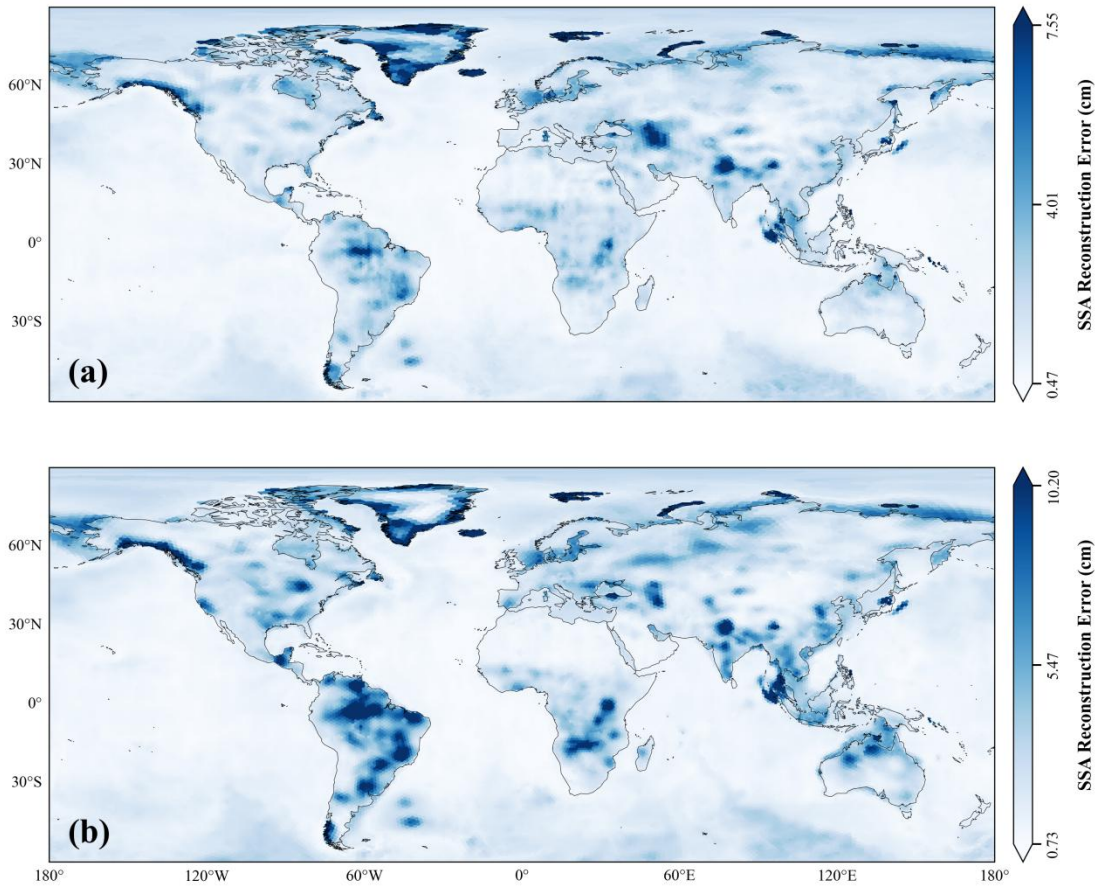

Figure 2. Evaluation of the reconstruction accuracy of Singular Spectrum Analysis (SSA). (a) Reconstruction error in the SSA-filling-a phase (cm), quantified as the standard deviation of residuals between observed and reconstructed values. (b) Reconstruction error in the SSA-filling-b phase (cm), quantified as the minimum aggregated *RMSE* from annual masking experiments.

3.1.2 Aquifer-stratified Model Comparison and Multi-scale Validation

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.

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

Cross-scale aggregation to the original 0.25° grid confirmed the same overall ranking (Fig. 3). At the global scale, XGBoost achieved $R^2 = 0.972$, $CC = 0.986$, and $RMSE = 2.097$ cm, compared with $R^2 = 0.961$, $CC = 0.981$, and $RMSE = 2.661$ cm for DNN. Across the 11 evaluated aquifer categories, XGBoost achieved higher R^2 and CC and lower $RMSE$ in 10 categories. Major groundwater basins (AQU categories 11–15) and regions with complex hydrogeological structures (AQU categories 22–25) maintained an average R^2 of 0.961 and $RMSE$ of 0.730 cm, whereas local and shallow aquifers (AQU categories 33–34) had a higher average $RMSE$ of 2.091 cm. The cross-scale assessment therefore shows that the relative performance advantage of XGBoost is retained when the 1 km reconstructions are evaluated against the original-resolution GWSA.

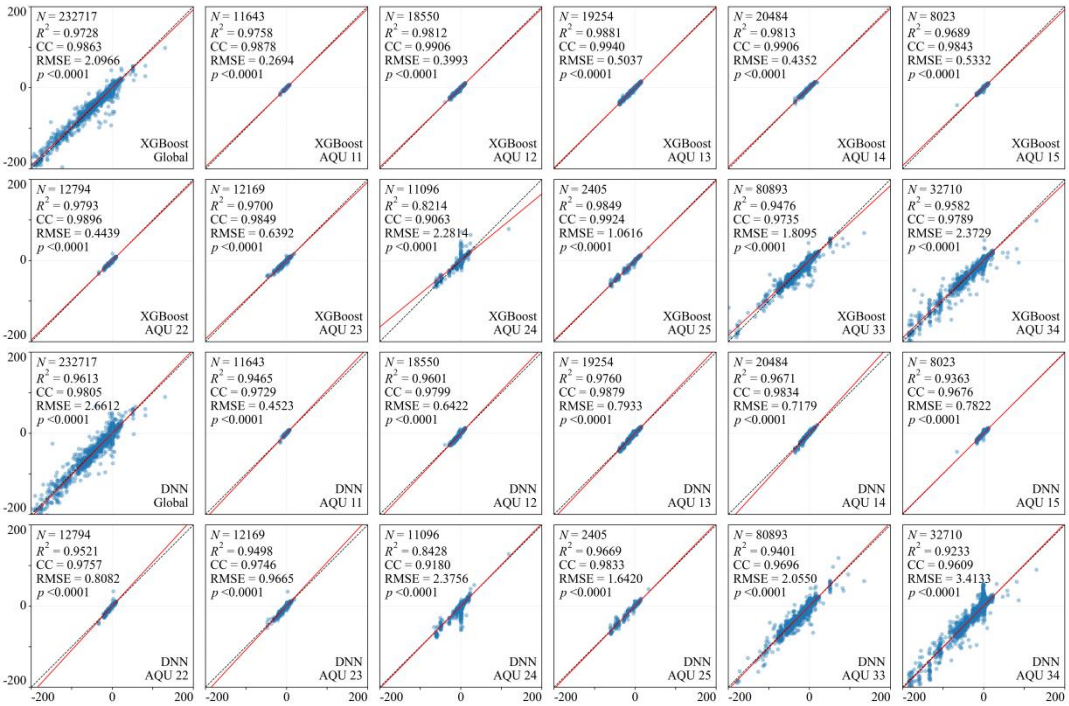

Figure 3. cross-scale consistency comparison of the downscaled GWSA and the original-resolution GWSA multi-year mean for 2002–2020.

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 (AQU categories 11–15), Complex hydrogeological structures (AQU categories 22–25), and Local and shallow aquifers (AQU categories 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), R^2, CC, $RMSE$, and significance level (p) are shown in the upper-left corner.

A 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 $RMSE = 0.34$ cm yr^{-1} and $Bias = 0.09$ 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 $CC = 0.73$, $R^2 = 0.54$, $RMSE = 0.38$ cm yr^{-1}, $Bias = 0.11$ 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.

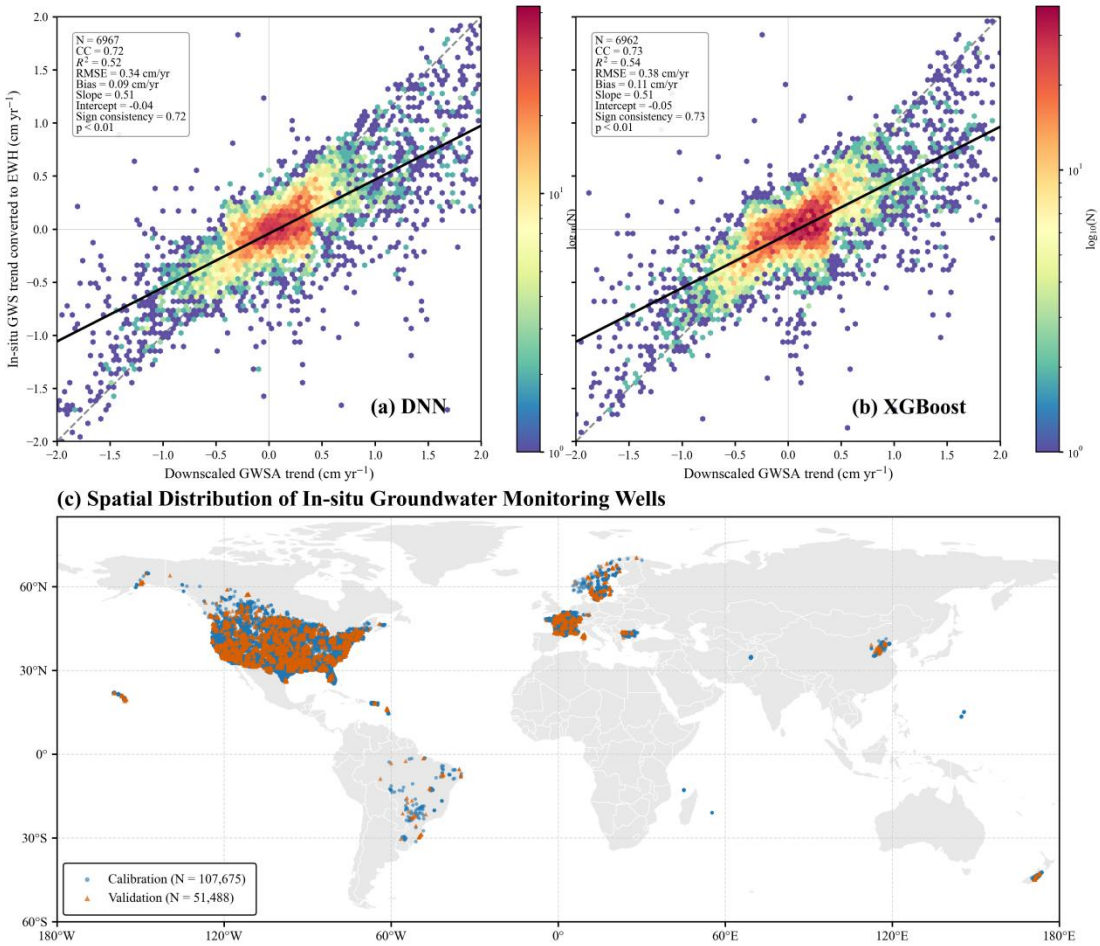

Figure 4. Spatially independent, quality-controlled in-situ evaluation of long-term downscaled groundwater-storage trends and spatial distribution of the monitoring network.

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

The sensitivity of the in-situ evaluation to S_{yeff} was further quantified by perturbing the baseline conversion coefficients by ±15% and ±30% (Table 3). For the XGBoost product, the baseline evaluation yielded *CC* = 0.734, R^2 = 0.538, *RMSE* = 0.378 cm yr^{-1}, *Bias* = 0.106 cm yr^{-1}, and a regression slope of 0.506. Across the full ±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_{yeff} 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_{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

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 WaterGAP 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. These differences indicate that the relative importance assigned to static heterogeneity and temporal variation depends on the target source and represents a source-related structural uncertainty.

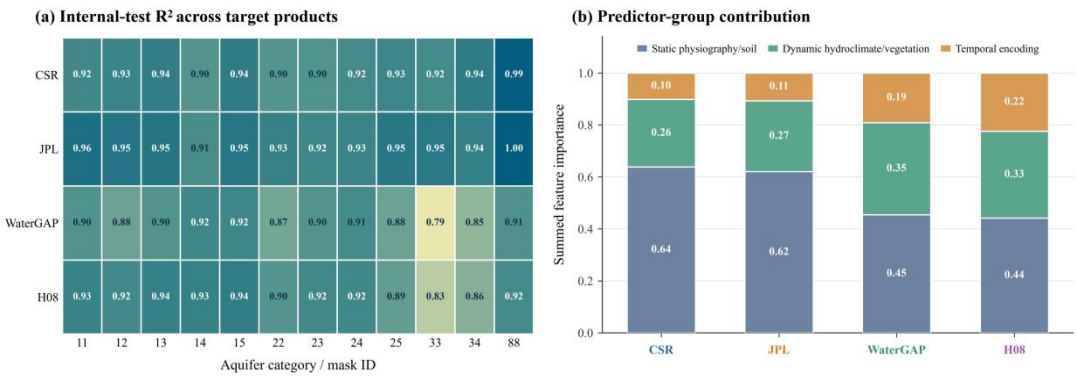

Figure 5. Controlled sensitivity of the aquifer-stratified XGBoost downscaling workflow to alternative target products.

(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_{yeff}-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 (AQU 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 (AQU 11, N = 1713, CC = 0.72, Fig. 6c), the Russian Platform Basins (AQU 33, N = 32340, CC = 0.67, Fig. 6f), and the Ganges-Brahmaputra Basin (AQU 24, N = 2208, CC = 0.54, Fig. 6d). Although the Arabian Aquifer System (AQU 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 (AQU 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 (AQU 14, N = 10052, CC = -0.23, Fig. 7b) and the North Caucasus Basin (AQU 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 (AQU 15, N = 634, CC = 0.24, Fig. 8b) and the West Siberian Basin (AQU 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.

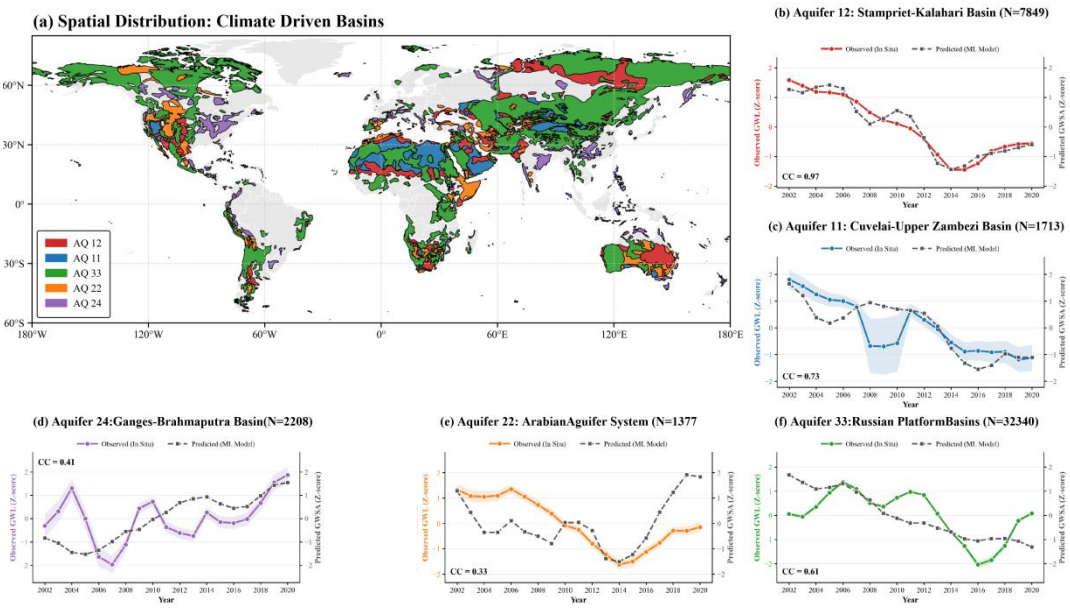

Figure 6. Spatial distribution and interannual variations of GWL and downscaled GWSA for climate-driven/natural-dominated major aquifer basins.

(a) Spatial locations of selected aquifers by AQU category (AQU 12: Stampriet-Kalahari Basin, AQU 11: Cuvelai-Upper Zambezi Basin, AQU 33: Russian Platform Basins, AQU 22: Arabian Aquifer System, and AQU 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.

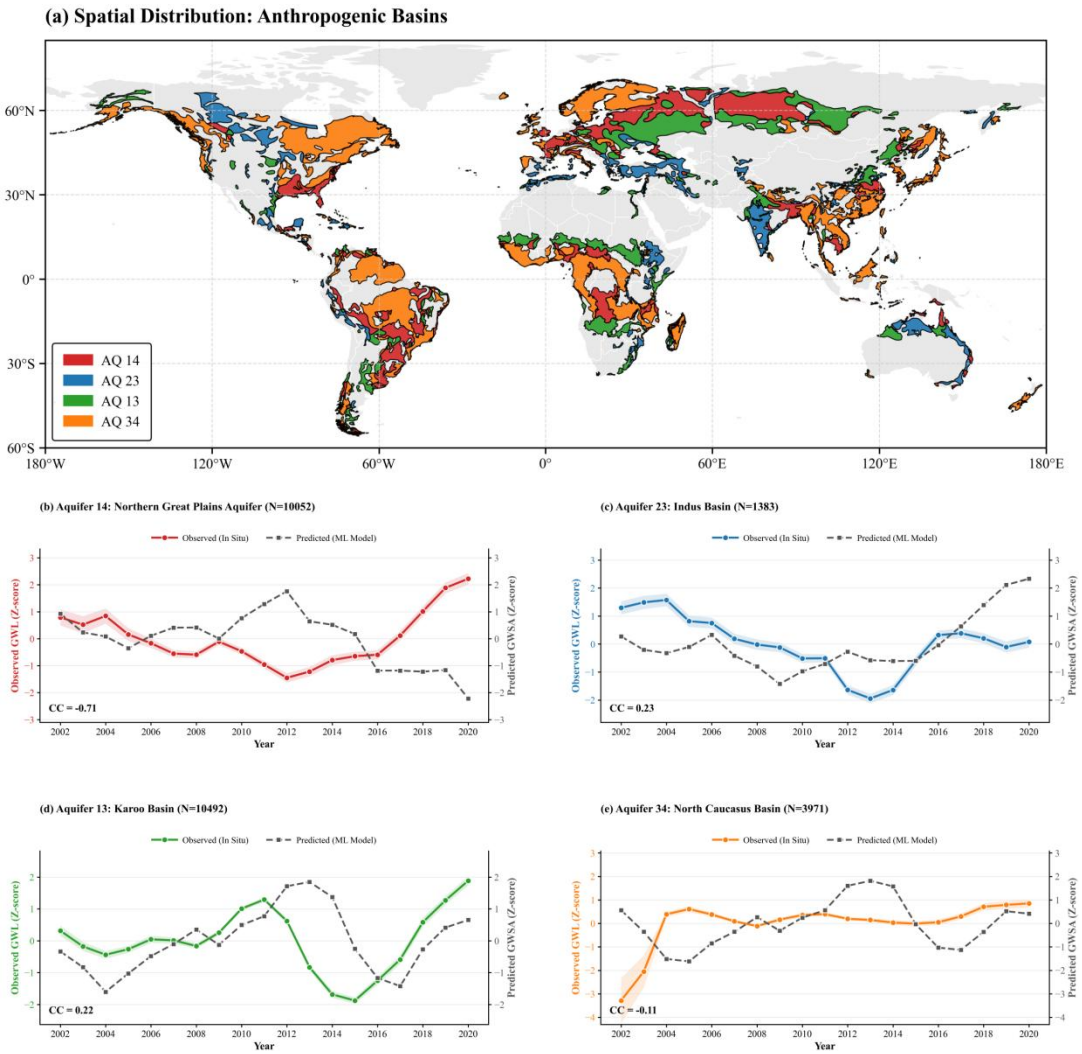

Figure 7. Spatial distribution and interannual variations of groundwater dynamics for anthropogenic-dominated major aquifer basins.

(a) Spatial locations of selected human-impacted aquifers by AQU category (AQU 14: Northern Great Plains Aquifer, AQU 23: Indus Basin, AQU 13: Karoo Basin, and AQU 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.

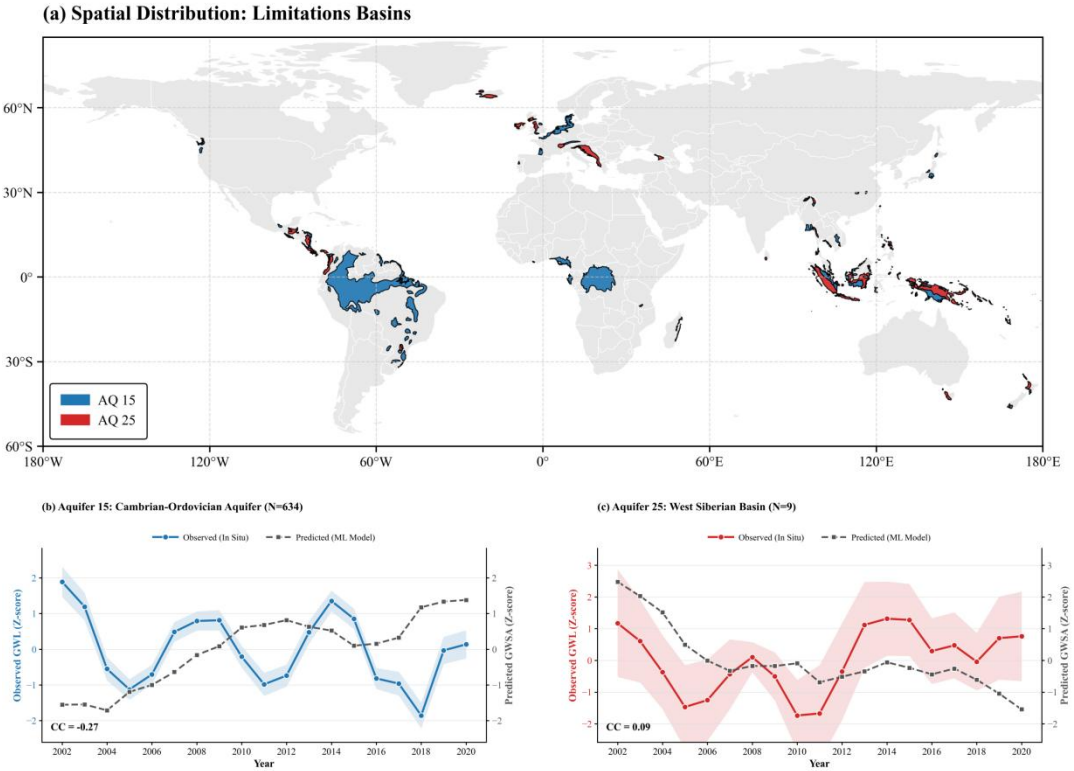

Figure 8. Spatial distribution and interannual variations of groundwater anomalies for aquifers under physical limitations or specific climatic-hydrogeological constraints.

(a) Geographic locations of the selected aquifers by AQU category (AQU 15: Cambrian-Ordovician Aquifer and AQU 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.

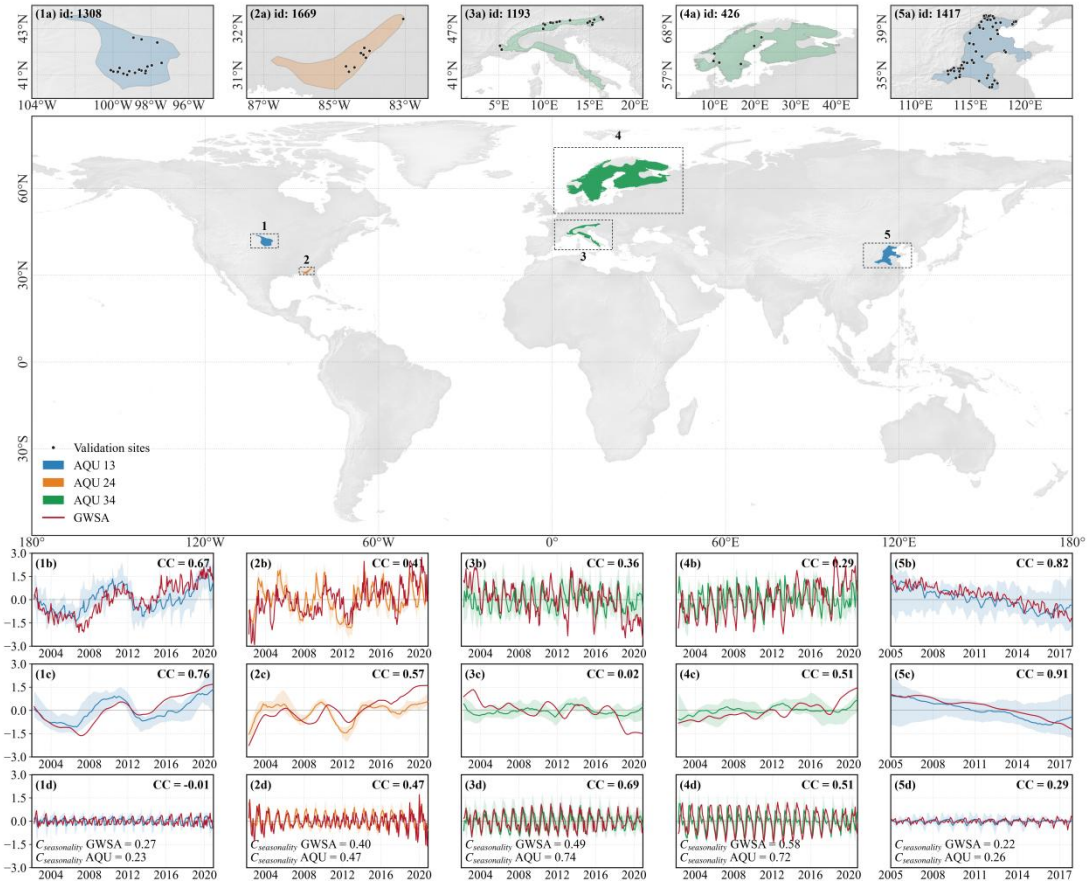

Figure 9. Spatial distributions and monthly STL decompositions of in-situ groundwater-level observations and reconstructed GWSA across five representative validation regions.

(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 ± 3.9 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 mm. The product Sen slope was nearly neutral (−0.10 mm yr^{-1}; 95% confidence interval −5.10 to +5.41 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 mm yr^{-1}, opposite to the published −27.6 ± 3.3 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.

Metric	North China Plain	Northwestern India	California	High Plains Aquifer
Reference	Zhang et al. (2023)	Rodell et al. (2009)	Liu et al. (2022)	Rateb et al. (2020)
Temporal resolution	Monthly	Monthly	Monthly	Annual
Temporal coverage	2003–2020	Aug 2002–Oct 2008	Oct 2003–Dec 2020	2003–2017
N	216	73	175	15
CC	0.963	0.731	0.789	0.526
RMSE (mm)	32.0	66.4	77.3	48.2
Bias (mm)	+10.2	0.0	+48.3	+19.4
SD ratio	0.912	0.593	0.673	0.625
Seasonal amplitude: product / reference (mm)	23.2 / 8.9	63.9 / 23.5	82.0 / 82.7	–
Sen slope: Product / reference (mm yr^{-1})	−17.71 / −20.27	−22.96 / −39.47	−6.96 / −15.80	−0.10 / −9.28
Sen-slope difference (mm yr−1)	+2.56	+16.51	+8.84	+9.19
Sen-trend sign agreement	Yes	Yes	Yes	Yes

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.

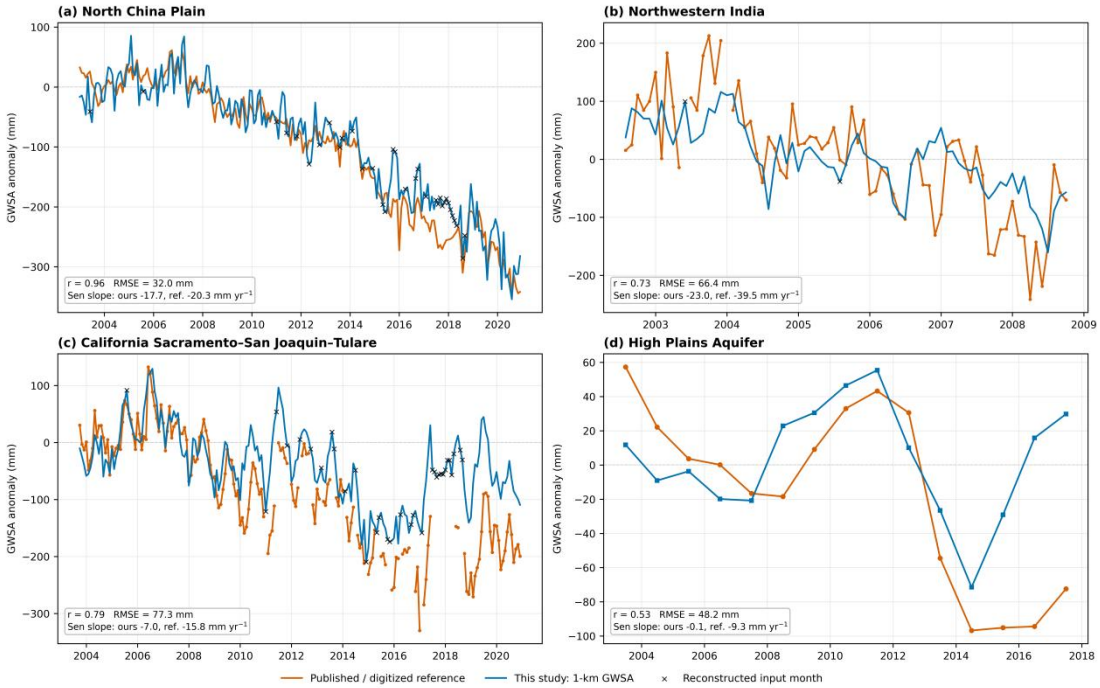

Figure 10. Regional evaluation of the 1 km GWSA reconstruction against published groundwater-storage estimates.

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 *CC*, *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 representing localized hydrological features, a detailed comparative analysis was conducted across four globally representative regions (Fig. 11). 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. 11, Column a) that obscure fine-scale groundwater dynamics. In contrast, the 1 km downscaled products provide predictor-informed spatial detail while retaining coarse-scale GRACE constraints.

In the Amazon Basin (Area 1), the downscaled fields contain complex dendritic textures that are not resolved in the coarse GRACE field. These patterns cannot be interpreted as groundwater-only structures: residual dynamic storage in rivers, wetlands, and floodplains, together with GRACE leakage, are plausible contributors. In southern Tibet (Area 2), the finalized XGBoost product (Fig. 11, panel 2c) resolves predictor-informed textures aligned with canyon topography, whereas the DNN baseline (Fig. 11, panel 2b) produces broader regional patches. This contrast indicates model-dependent use of static geophysical covariates rather than independent kilometre-scale satellite information.

The additional spatial detail is also apparent in regions associated with intensive 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 reveals localized groundwater-storage gradients spatially aligned with agricultural boundaries and 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 differences in spatial representation in these intensively managed areas. The DNN baseline tends to produce spatially smoothed anomaly zones, such as broad depletion cones (Fig. 11, panel 3b), whereas the adopted XGBoost architecture preserves sharper nonlinear gradients associated with local environmental heterogeneity. The integration of multi-source 1 km covariates therefore provides a predictor-informed reconstruction of spatial heterogeneity under coarse-scale GRACE constraints.

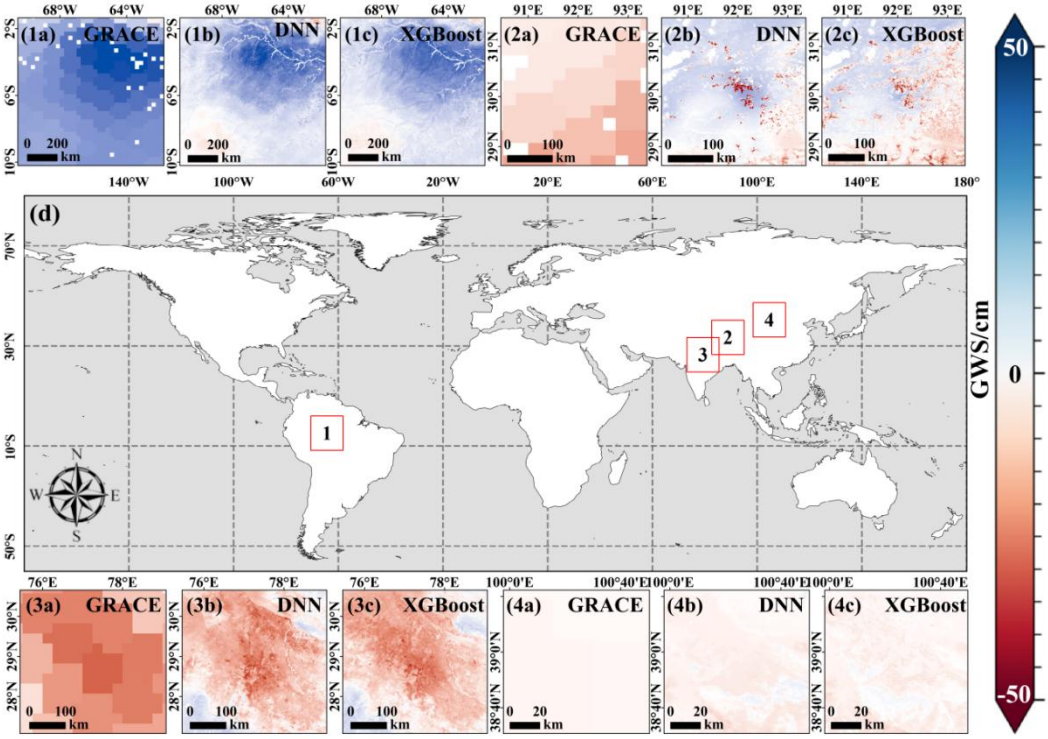

Figure 11. Spatial enhancement and feature reconstruction of 1 km resolution GWSA across four globally representative regions.

Columns present (a) the original 0.25° GRACE observations, alongside downscaled outputs generated by (b) DNN and (c) XGBoost models at the 1 km scale. The four selected regions (indicated by red boxes) encompass: (1) the Amazon Basin, associated with natural hydro-climatic variability; (2) the mountainous region of southern Tibet, characterized by extreme topographical complexity; (3) the agricultural zones surrounding New Delhi, India; and (4) the Zhangye agricultural oasis within the hyper-arid Heihe River Basin. Original GRACE observations (Column a) display coarse macroscopic anomalies due to spatial leakage. The XGBoost model (Column c) provides predictor-informed fine-scale textures aligned with canyon micro-topography in southern Tibet (2c) and localized groundwater-storage gradients spatially associated with agricultural boundaries in New Delhi (3c) and Zhangye (4c). These fine-scale patterns are predictor-informed under coarse-scale GRACE constraints, and their correspondence with agricultural boundaries is interpreted descriptively.

3.4 Global Climatology and Spatiotemporal Variability Patterns

3.4.1 Baseline Imbalances and Meridional Profiles

The spatial pattern of the global Mean GWSA was reconstructed at a 1 km resolution for the period from 2002 to 2020, along with the corresponding meridional mean profiles (Fig. 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. 12b), the 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. 12a) 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 groundwater depletion hotspots located in regions characterized by intensive human water use on the map. From the perspective of the dynamic mechanism, this pattern is spatially associated with the combined influence of climatic drought stress and intensive human water use. This latitude belt not only lies within the long-term recharge-deficient region 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, 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 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.

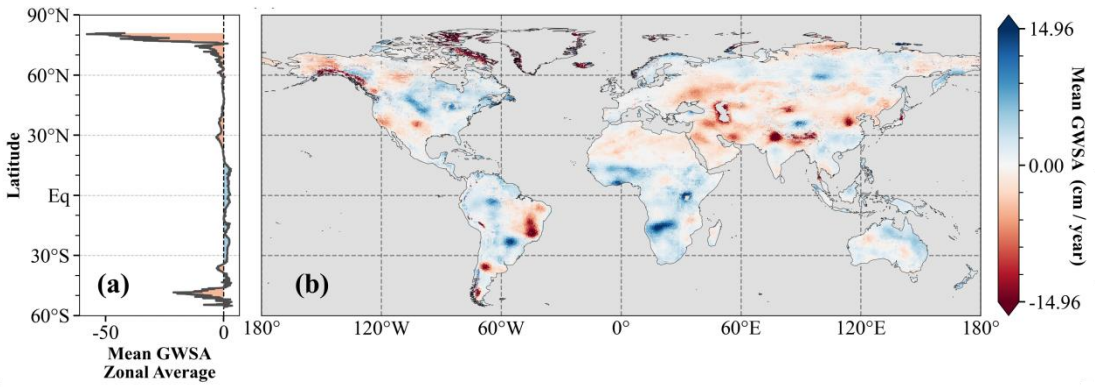

Figure 12. Spatial pattern and latitudinal profile of global Mean GWSA during the period from 2002 to 2020. (a) Latitudinal mean profile, presenting area-weighted averages of inland land pixels across diverse latitudes. The black solid line denotes the latitudinal mean, and the red and blue filled areas signify groundwater depletion and surplus states, respectively. (b) Global 1 km spatial pattern map depicting the high-resolution mean 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.

3.4.2 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. 13). 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 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 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.

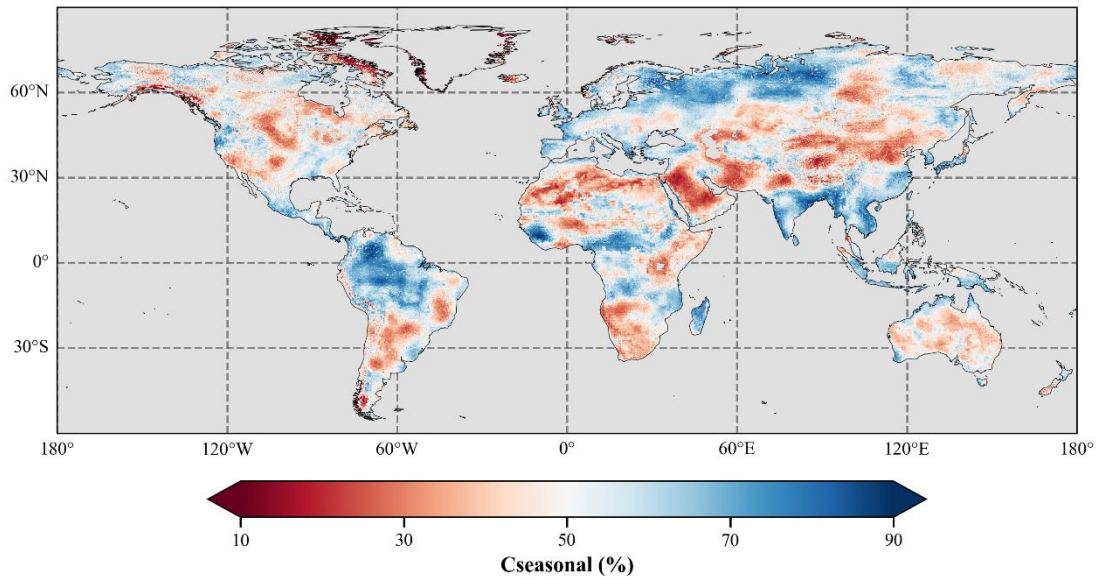

Figure 13. Seasonal contribution to global 1 km GWSA.

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.

3.5 Spatiotemporal Evolution and Trend Polarization

3.5.1 Spatial Heterogeneity of Global GWSA Trends

As illustrated in 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 reduce the visual effects of spatial smoothing inherent in traditional coarse-resolution GRACE observations and reveal predictor-informed local hydrological 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 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 GWSA demonstrates extensive and spatially continuous significant surplus. Additionally, high-latitude regions in North America, such as northern Canada with a polar long-winter climate, and localized areas in northern

Eurasia, such as the southeastern Siberian Plain in Russia where permafrost is prevalent, show groundwater accumulation signals in trend maps. These signals potentially reflect deep water infiltration associated with accelerated hydrological cycles or cryosphere degradation at high latitudes. These surplus zones, dominated by large-scale climatic or natural hydrological processes, exhibit relatively smooth spatial transitions at 1 km scales, forming distinct morphological contrasts with the sharper deficit boundaries spatially coincident with intensively managed agricultural regions.

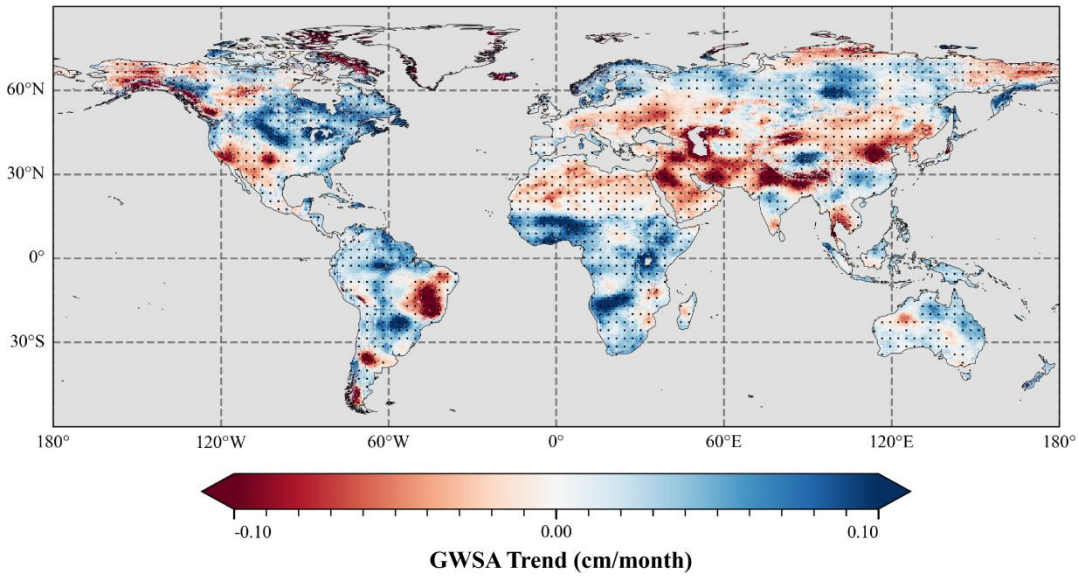

Figure 14. Long-term evolutionary trends and significant spatial distributions of global 1 km resolution GWSA.

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 MK significance test ($p < 0.01$). Light gray regions correspond to oceans, polar permafrost areas, or extremely arid deserts where valid data are lacking.

3.5.2 Statistical Classification and Graded Evaluation of Groundwater Evolution

To summarize the magnitude and statistical persistence of global groundwater trends, Theil–Sen trend rates were combined with MK significance thresholds to define seven descriptive classes (Fig. 15). This classification identifies the direction, rate, and significance category of each mapped trend; it does not by itself diagnose the external forcing 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.

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.

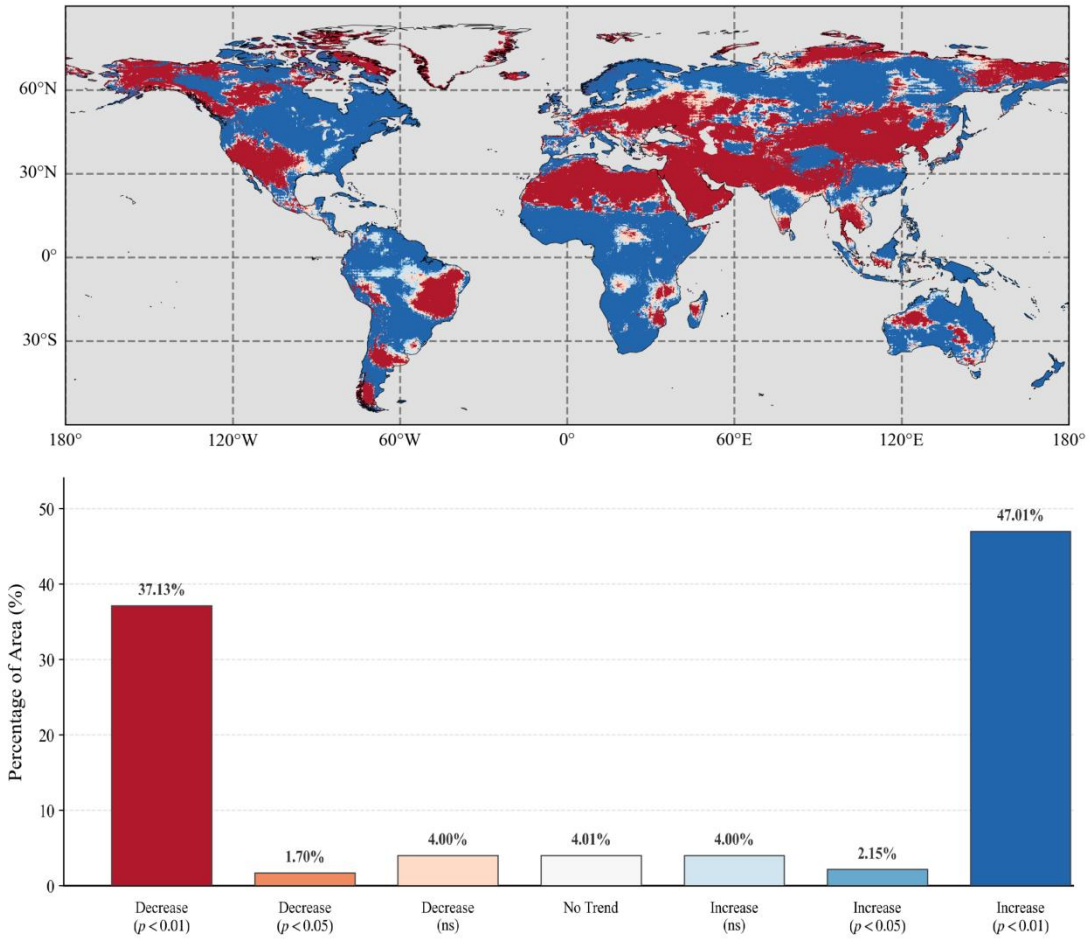

Figure 15. Spatial classification and area proportions of global 1 km GWSA trends.

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.

3.6 Residual Signals Not Explained by the GHM-simulated Background

Regional depletion and accumulation hotspots show substantial spatial heterogeneity. To describe departures from a GHM-simulated background, this study uses a residual diagnostic framework that compares high-resolution reconstructed sequences (with missing values filled through SSA) with GHM simulation benchmarks. The 0.5° GHM-simulated background signals are bilinearly interpolated to 1 km for spatial matching with the downscaled GWSA before subtraction. 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 above, residual sequences after STL decomposition to remove periodic components exhibit marked spatial heterogeneity (Figure 16). In intensively irrigated regions, including the Indus Basin (AQU 23) and the North China Plain, residual trend maps 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 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.

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.

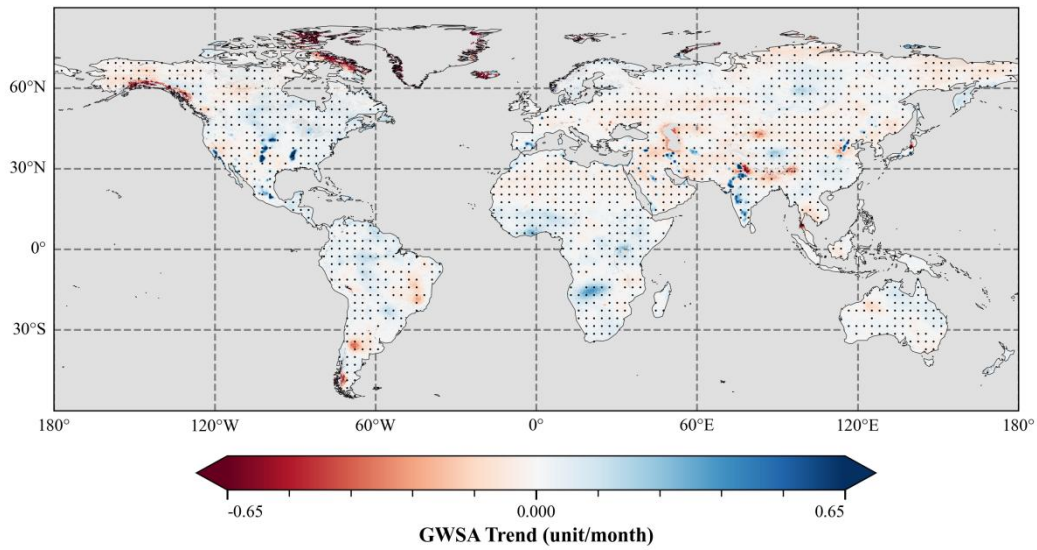

Figure 16. Residual signals not explained by the GHM-simulated background (p < 0.01).

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 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 GHMs, such as GLDAS, PCR-GLOBWB, and WaterGAP, this dataset 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), their coarse grids and differing representations of groundwater withdrawals can yield departures from the downscaled product in intensive agricultural regions. As shown in Section 3.6, residual depletion signals in the Indus Basin and the North China Plain are not explained by the GHM-simulated background, but they are not quantification of unsustainable 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 statistical downscaling studies have focused on individual regions, including California's Central Valley and the North China Plain (Miro and Famiglietti, 2018; Yin et al., 2018). The current 1 km global reconstruction leverages the XGBoost architecture and high-resolution auxiliary environmental covariates to represent localized spatial heterogeneity under coarse-scale GRACE constraints. This approach can reveal predictor-informed hydro-geomorphological textures, such as canyon-aligned patterns in southern Tibet, and groundwater-storage gradients spatially associated with managed agricultural areas. Such information provides a useful basis for spatial screening of hydrogeological heterogeneity and managed-land associations.

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, 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 elucidation of these complementary mechanisms can not only clarify the logical basis of the dataset selection strategy adopted in this study, but also provide crucial theoretical references for the construction of hybrid integration frameworks with strengthened physical constraints in future Earth system science research (Karniadakis et al., 2021). The high-resolution dataset constructed based on the XGBoost algorithm exhibits improved capability in characterizing complex topographic features and localized groundwater storage gradients, which renders it applicable to spatial screening and water resource management practices on the premise that the coarse-scale GRACE constraint and attribution limitations of the dataset are explicitly acknowledged.

4.3 Sources of Uncertainty in Downscaling Models and Future Research Perspectives

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 ±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 global hydrological models (GHMs) introduce an additional structural uncertainty component. The residuals between the downscaled groundwater storage anomaly (GWSA) and the background signals simulated by GHMs integrate discrepancies originating from multiple sources, including the forcing of GHMs, parameterization of recharge and evapotranspiration, characterization of groundwater processes, component separation based on GRACE observations, and the statistical downscaling procedure itself. Therefore, these residuals are most appropriately interpreted as groundwater storage variations that cannot be explained by the background simulated by GHMs. In regions with independently recorded intensive groundwater exploitation, sustained negative residuals provide robust 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 enables the residual analysis framework to possess practical application value for spatial screening and hypothesis generation, particularly when integrated with independent datasets of 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 scientific knowledge and process-based models with machine learning (Willard et al., 2022) 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

This study presents a global 1 km monthly GWSA dataset covering April 2002–December 2020. The dataset was generated using SSA for temporal gap filling, 19 high-resolution environmental predictors, and aquifer-stratified machine-learning downscaling. XGBoost was selected as the production model after comparison with DNN using internal evaluation, cross-scale consistency, and independent groundwater observations.

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 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 multiband Tagged Image File Format (TIFF) 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). 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.

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, 401, 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.

Chen, J. L., Wilson, C. R., Famiglietti, J. S., and Rodell, M.: Attenuation effect on seasonal basin-scale water storage changes 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.

Cleveland, R. B., Cleveland, W. S., McRae, J. E., and Terpenning, I.: STL: A seasonal-trend decomposition procedure based on loess, Journal of Official Statistics, 6, 3–73, 1990.

Cooley, S. W., Ryan, J. C., and Smith, L. C.: Human alteration of global surface water storage variability, Nature, 591, 78–81, https://doi.org/10.1038/s41586-021-03262-3, 2021.

Ek, M. B., Mitchell, K. E., Lin, Y., Rogers, E., Grunmann, P., Koren, V., Gayno, G., and Tarpley, J. D.: Implementation of Noah land surface model advances in the National Centers for Environmental Prediction operational mesoscale Eta model, J. Geophys. Res. Atmos., 108(D22), 8851, https://doi.org/10.1029/2002JD003296, 2003.

Famiglietti, J. S.: The global groundwater crisis, Nat. Clim. Change, 4, 945–948, https://doi.org/10.1038/nclimate2425, 2014.

Famiglietti, J. S., Lo, M., Ho, S. L., Bethune, J., Anderson, K. J., Syed, T. H., Swenson, S. C., de Linage, C. R., and Rodell, M.: Satellites measure recent rates of groundwater depletion in California's Central Valley, Geophys. Res. Lett., 38, L03403, https://doi.org/10.1029/2010GL046442, 2011.

Fan, Y., Zhang, C., Han, W., Che, T., and Ge, Y.: 1 km high-resolution monthly-scale GWSA dataset, National Cryosphere Desert Data Center [data set], https://doi.org/10.12072/ncdc.hydrology.db7323.2026.

Hanasaki, N., Kanae, S., Oki, T., Masuda, K., Motoya, K., Shirakawa, N., Shen, Y., and Tanaka, K.: An integrated model for the assessment of global water resources – Part 2: Applications and assessments, Hydrol. Earth Syst. Sci., 12, 1027–1037, https://doi.org/10.5194/hess-12-1027-2008, 2008.

Hastie, T., Tibshirani, R., and Friedman, J.: The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd Edn., Springer, New York, NY, USA, 745 pp., https://doi.org/10.1007/978-0-387-84858-7, 2009.

Jasechko, S., Seybold, H., Perrone, D., Fan, Y., Shamsudduha, M., Taylor, R. G., Fallatah, O., and Kirchner, J. W.: Rapid groundwater decline and some cases of recovery in aquifers globally,

Nature, 625, 715–721, https://doi.org/10.1038/s41586-023-06879-8, 2024.

Karger, D. N., Conrad, O., Böhner, J., Kawohl, T., Kreft, H., Soria-Auza, R. W., Zimmermann, N. E., Linder, H. P., and Kessler, M.: Climatologies at high resolution for the earth's land surface areas, Scientific Data, 4, 170122, https://doi.org/10.1038/sdata.2017.122, 2017.

Karniadakis, G. E., Kevrekidis, I. G., Lu, L., Perdikaris, P., Wang, S., and Yang, L.: Physics-informed machine learning, Nat. Rev. Phys., 3, 422–440, https://doi.org/10.1038/s42254-021-00314-5, 2021.

Kendall, M. G.: Rank Correlation Methods, 4th Edn., Charles Griffin, London, 202 pp., 1970.

Kondrashov, D. and Ghil, M.: Spatio-temporal filling of missing points in geophysical data sets, Nonlinear Processes in Geophysics, 13, 151–159, https://doi.org/10.5194/npg-13-151-2006, 2006.

LeCun, Y., Bengio, Y., and Hinton, G.: Deep learning, Nature, 521, 436–444, https://doi.org/10.1038/nature14539, 2015.

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

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

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

Miro, M. E. and Famiglietti, J. S.: Downscaling GRACE remote sensing datasets to high-resolution groundwater storage change maps of California's Central Valley, Remote Sens., 10, 143, https://doi.org/10.3390/rs10010143, 2018.

Müller Schmied, H., Adam, L., Eisner, S., Fink, G., Flörke, M., Kim, H., Oki, T., Portmann, F. T., Reinecke, R., Riedel, C., Song, Q., Zhang, J., and Döll, P.: Variations of global and continental water balance components as impacted by climate forcing uncertainty and human water use, Hydrol. Earth Syst. Sci., 20, 2877–2898, https://doi.org/10.5194/hess-20-2877-2016, 2016.

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

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

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

Poggio, L., de Sousa, L. M., Batjes, N. H., Heuvelink, G. B. M., Kempen, B., Ribeiro, E., and Rossiter, D.: SoilGrids 2.0: producing soil information for the globe with quantified spatial uncertainty, SOIL, 7, 217–240, https://doi.org/10.5194/soil-7-217-2021, 2021.

Rateb, A., Scanlon, B. R., Pool, D. R., Sun, A., Zhang, Z., Chen, J., Clark, B., Faunt, C. C., Haugh, C. J., Hill, M., Hobza, C., McGuire, V. L., Reitz, M., Müller Schmied, H., Sutanudjaja, E. H., Swenson, S., Wiese, D., Xia, Y., and Zell, W.: 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.

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

Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J. J., Schröder, B., Thuiller, W., Warton, D. I., Wintle, B. A., Hartig, F., and Dormann, C. F.: Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure, Ecography, 40, 913–929, https://doi.org/10.1111/ecog.02881, 2017.

Rodell, M. and Famiglietti, J. S.: An analysis of terrestrial water storage variations in Illinois with implications for the Gravity Recovery and Climate Experiment (GRACE), Water Resour. Res., 37, 1327–1339, https://doi.org/10.1029/2000WR900306, 2001.

Rodell, M., Houser, P. R., Jambor, U., Gottschalck, J., Mitchell, K., Meng, C. J., Arsenault, K., Cosgrove, B., Radakovich, J., Bosilovich, M., Entin, J. K., Walker, J. P., Lohmann, D., and Toll, D.: The Global Land Data Assimilation System, Bull. Amer. Meteorol. Soc., 85, 381–394, https://doi.org/10.1175/BAMS-85-3-381, 2004.

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.

Rodell, M., Famiglietti, J. S., Wiese, D. N., Reager, J. T., Beaudoing, H. K., Landerer, F. W., and Lo, M. H.: Emerging trends in global freshwater availability, Nature, 557, 651–659, https://doi.org/10.1038/s41586-018-0123-1, 2018.

Save, H., Bettadpur, S., and Tapley, B. D.: High-resolution CSR GRACE RL05 mascons, J. Geophys. Res. Solid Earth, 121, 7547–7569, https://doi.org/10.1002/2016JB013007, 2016.

Scanlon, B. R., Faunt, C. C., Longuevergne, L., Reedy, R. C., Alley, W. M., McGuire, V. L., and McMahon, P. B.: Groundwater depletion and sustainability of irrigation in the US High Plains and Central Valley, Proc. Natl. Acad. Sci. U.S.A., 109, 9320–9325, https://doi.org/10.1073/pnas.1200311109, 2012.

Scanlon, B. R., Zhang, Z., Save, H., Wiese, D. N., Landerer, F. W., Long, D., Longuevergne, L., and Chen, J.: Global evaluation of new GRACE mascon products for hydrologic applications, Water Resour. Res., 52, 9412–9429, https://doi.org/10.1002/2016WR019494, 2016.

Scanlon, B. R., Zhang, Z., Save, H., Sun, A. Y., Müller Schmied, H., van Beek, L. P. H., Wiese, D. N., Wada, Y., Long, D., Reedy, R. C., Longuevergne, L., Döll, P., and Bierkens, M. F. P.: Global models underestimate large decadal declining and rising water storage trends relative to GRACE satellite data, Proc. Natl. Acad. Sci. U.S.A., 115, E1080–E1089, https://doi.org/10.1073/pnas.1704665115, 2018.

Sen, P. K.: Estimates of the regression coefficient based on Kendall's tau, Journal of the American Statistical Association, 63, 1379–1389, https://doi.org/10.1080/01621459.1968.10480934, 1968.

Seyoum, W. M., Kwon, D., and Milewski, A. M.: Downscaling GRACE TWSA data into high-resolution groundwater level anomaly using machine learning-based models in a glacial aquifer system, Remote Sens., 11, 824, https://doi.org/10.3390/rs11070824, 2019.

Shen, C.: A transdisciplinary review of deep learning research and its relevance for water resources scientists, Water Resour. Res., 54, 8558–8593, https://doi.org/10.1029/2018WR022643, 2018.

Taylor, R. G., Scanlon, B., Döll, P., Rodell, M., van Beek, R., Wada, Y., Longuevergne, L., Leblanc, M., Famiglietti, J. S., Edmunds, M., Konikow, L., Green, T. R., Chen, J., Taniguchi, M., MacDonald, A., Fan, Y., Maxwell, R. M., Yechieli, Y., Gurdak, J. J., Allen, D. M., Shamsudduha, M., Hiscock, K., Yeh, P. J.-F., Holman, I., and Treidel, H.: Ground water and climate change, Nat. Clim. Change, 3, 322–329, https://doi.org/10.1038/nclimate1744, 2013.

Theil, H.: A rank-invariant method of linear and polynomial regression analysis, Parts I–III, Proc. K. Ned. Akad. Wet., 53, 386–392, 521–525, and 1397–1412, 1950.

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

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

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

Willard, J., Jia, X., Xu, S., Steinbach, M., and Kumar, V.: Integrating Scientific Knowledge with Machine Learning for Engineering and Environmental Systems, ACM Comput. Surv., 55(4), 1–37, https://doi.org/10.1145/3514228, 2022.

Yi, S. and Sneeuw, N.: Filling the Data Gaps Within GRACE Missions Using Singular Spectrum Analysis, J. Geophys. Res. Solid Earth, 126, e2020JB021227, https://doi.org/10.1029/2020JB021227, 2021.

Yin, W., Hu, L., Zhang, M., Wang, J., and Han, S.-C.: Statistical Downscaling of GRACE-Derived Groundwater Storage Using ET Data in the North China Plain, J. Geophys. Res. Atmos., 123, 5973–5987, https://doi.org/10.1029/2017JD027468, 2018.

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.

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

Zheng, C., Jia, L., and Zhao, T.: A 21-year dataset (2000–2020) of gap-free global daily surface soil moisture at 1-km grid resolution, Scientific Data, 10, 139, https://doi.org/10.1038/s41597-023-01991-w, 2023.