https://doi.org/10.5194/egusphere-2026-4563
Preprint. Discussion started: 4 August 2026

Pushing the Boundaries of NGGM and MAGIC Derived Terrestrial Water Storage Anomalies with Unsupervised Deep-Learning Downscaling

Gilberto Goracci[1], Ilias Daras[2], and Nicolas Longepe[1]

[1]European Space Agency ϕ-lab Explore Office (EOP-S8E) - Climate Action, Sustainability and Science Dept. – Earth Observation Programmes Directorate, Via Galileo Galilei 1, 00044 Frascati, Italy
[2]European Space Agency Earth and Mission Sciences Division - Climate Action, Sustainability and Science Dept. – Earth Observation Programmes Directorate, Keplerlaan 1, 2201 AZ Noordwijk, the Netherlands

Correspondence: Gilberto Goracci (gilberto.goracci@esa.int)

Abstract. This work presents an unsupervised deep-learning approach for the spatial downscaling of Terrestrial Water Storage Anomalies (TWSA) from their native coarse resolution to a 1° grid, using simulated Next-Generation Gravity Mission (NGGM) and ESA–NASA Mass-change And Geosciences International Constellation (MAGIC) products obtained from realistic end-to-end (E2E) closed-loop experiments carried out in the context of ESA studies, together with real Gravity Recovery and Climate Experiment (GRACE) and Follow-On (GRACE-FO) observations. A U-Net Convolutional Neural Network integrates hydro-climatic variables from ERA5-Land and topographic information to retrieve fine-scale TWSA patterns while keeping consistency with the original coarse gravimetric observations. Results over five regions of interest—Africa, the Amazon basin, Europe, the Ganges basin, and the Mississippi basin—have been evaluated through a composite score combining pixel-wise and structure-wise metrics. The analysis shows that the quality and native resolution of the input gravimetric products strongly influence the downscaled fields. In the simulated experiments, NGGM and MAGIC generally achieve higher high-resolution composite scores than the GRACE-C-like configuration, while monthly products outperform the corresponding 5-daily solutions. At the native satellite resolution, the aggregated downscaled fields remain highly consistent with the input products, with composite scores generally above 0.90 in the simulated and real scenarios. Water balance equation (WBE) analyses over the Congo and Danube basins show that the downscaled products generally preserve the basin-scale closure behavior of their corresponding satellite inputs, particularly for the 5-daily simulations, where the errors of the aggregated and 1° downscaled products remain close to the satellite baseline. The improved quality of the simulated observing scenarios is also reflected in lower WBE errors: from the GRACE-C-like to the MAGIC configuration, the 1° downscaled RMSE decreases from 4.82 to 3.32 mm over Congo and from 14.47 to 7.10 mm over Danube. For real GRACE(-FO) observations, the aggregated downscaled products reproduce the satellite-scale WBE residuals with NRMSE values of 0.014 over Congo and 0.007 over Danube. Although the fine-scale redistribution remains underconstrained in the absence of independent high-resolution observations, these findings demonstrate that unsupervised deep-learning downscaling can improve the spatial detail of satellite-derived TWSA while preserving the large-scale gravimetric and hydrological consistency of both simulated future-mission products and real GRACE(-FO) observations.

https://doi.org/10.5194/egusphere-2026-4563
Preprint. Discussion started: 4 August 2026

1 Introduction

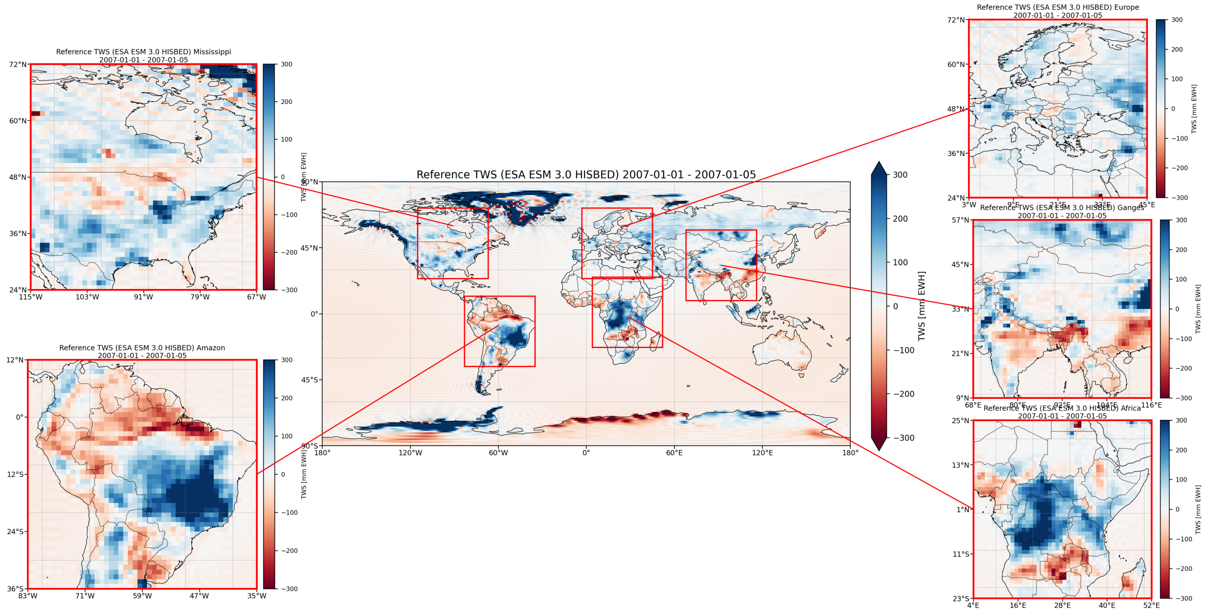
The Gravity Recovery and Climate Experiment (GRACE) and its follow-on mission (GRACE-FO) have transformed the observation of terrestrial water storage (TWS) by providing global, vertically integrated measurements of mass redistribution within the Earth system. These missions have enabled the monitoring of large-scale hydrological variability, including droughts, floods, groundwater depletion, and long-term storage trends. However, the coarse spatial resolution of GRACE and GRACE-FO products, typically on the order of several hundred kilometers, limits their direct applicability to regional hydrological monitoring and water-resource management. Many hydrological processes and management decisions occur at spatial scales ranging from a few kilometers to a few tens of kilometers, creating a mismatch between the native resolution of satellite gravimetry and the spatial detail required for basin-scale assessment, groundwater regulation, and drought monitoring. Through the years, this limitation has been highlighted by the international scientific community (Pail, 2015; International Union of Geodesy and Geophysics, 2023) and led to decisions by space agencies to collaborate for the future of sustained enhanced satellite gravimetry measurements. NASA and DLR are planning the continuation of the GRACE and GRACE-FO programme of record with GRACE-Continuity (GRACE-C), which is expected to be launched in 2028 in a near-polar orbit around 500 km orbital altitude (Flechtner et al., 2024). ESA is currently preparing the Next Generation Gravity Mission (NGGM), which is expected to be launched in 2032 in an inclined controlled repeat orbit at 70 degrees inclination and around 400 km orbital altitude (Daras et al., 2023; Pail et al., 2024). ESA and NASA will collaborate on the MAss-change and Geosciences International Collaboration (MAGIC) which consists of a staggered deployment of GRACE-C and NGGM with the goal to address the needs of the international user community (Daras et al., 2024). While GRACE-C will ensure continuity and global coverage, NGGM will provide the enhancements in the measurements and the consistency in product quality required by operational applications. During the expected overlap between the two pairs, MAGIC is expected to significantly improve the accuracy, temporal and spatial resolution of temporal gravity field and derived terrestrial water storage anomaly (TWSA) or groundwater storage anomaly (GWSA) products, while also reducing temporal aliasing errors and dependence on de-aliasing models and post-processing. These improvements will enable more accurate TWSA products, including high-temporal-resolution solutions, such as 5-daily fields, as well as monthly products with improved quality compared to current GRACE-like observations. Future satellite gravity missions are therefore expected to substantially improve the observation of mass redistribution in the Earth system, however, even for NGGM and the full MAGIC constellation, the native spatial resolution of the products (~ 150 km for the MAGIC monthly products (Daras et al., 2024)) will remain coarser than the scale required by many hydrological applications, making downscaling a necessary step to enhance the applicability of current and future TWSA observations for regional hydrological analyses (i.e. spatial scales from 10 to 100 km). In literature, the coarse resolution of satellite gravimetry products has motivated a wide range of studies aimed at downscaling GRACE-derived TWSA, namely estimating higher-resolution TWSA or GWSA fields that preserve consistency with the original coarse gravimetric signal. Existing approaches can be broadly divided into model-based and statistical methods. Model-based strategies assimilate GRACE observations into land surface or hydrological models, redistributing the satellite-derived storage signal within a physically consistent modelling framework (Zhong et al., 2021). These methods can provide spatially detailed estimates of individual water-storage compartments, but they are com-

https://doi.org/10.5194/egusphere-2026-4563
Preprint. Discussion started: 4 August 2026

putationally demanding and depend on model structure, forcing data, and uncertain parameterizations. Statistical approaches, instead, establish empirical relationships between GRACE anomalies and higher-resolution predictors such as precipitation, evapotranspiration, soil moisture, temperature, or vegetation-related variables, and then apply these relationships at finer spatial scales (Yin et al., 2018; Vishwakarma et al., 2021; Tourian et al., 2023; Seyoum et al., 2019; Sahour, 2020). Although these methods are generally easier to implement, they often rely on assumptions of linearity or scale invariance and may require dense in-situ observations for calibration and validation. Consequently, traditional downscaling can suffer from biased trends, discontinuous spatial patterns, and limited water-balance consistency (Tao et al., 2023; Fatolazadeh et al., 2022; Chen et al., 2025). In recent years, machine learning (ML) and deep learning (DL) have increasingly been adopted to overcome some of these limitations. Data-driven models such as random forests, gradient boosting methods, and neural networks have been used to infer nonlinear relationships between GRACE observations and multi-source hydro-climatic predictors, producing downscaled TWSA or GWSA estimates at resolutions ranging from approximately 1 to $25\ \mathrm{km}$ (Ahmed et al., 2019; Ali et al., 2023; Uz et al., 2024). Other studies have embedded ML methods within gap-filling or data-fusion schemes that combine GRACE(-FO) data with land surface model outputs (Tao et al., 2023; Chen et al., 2025). More recent developments include self-supervised and unsupervised deep-learning strategies, such as global model-assisted learning (Gou and Soja, 2024), generative super-resolution approaches (Wang et al., 2025), and representation learning applied to gravity and climate datasets (Ram, 2022). These methods show the potential of AI-based techniques to extract complex, nonlinear, and multivariate relationships from heterogeneous datasets. Nevertheless, AI-based downscaling remains challenging because models can be data-demanding, prone to overfitting, weakly interpretable, and often insufficiently constrained by physical consistency or uncertainty information (Fatolazadeh et al., 2022; Sahour, 2020; Ram, 2022). In this study, an unsupervised DL-based framework for downscaling TWSA fields from their native coarse resolution to a 1° target grid is presented. The method relies on a Convolutional Neural Network architecture driven by multi-source hydro-climatic and topographic predictors. The proposed framework is applied in two different settings. First, it is tested on simulated next-generation gravity mission products generated through realistic end-to-end (E2E) closed-loop simulations carried out in the context of ESA studies. This simulated dataset includes both 5-daily products at 3° native resolution and monthly products at 2° native resolution, allowing the downscaling performance to be quantitatively assessed against an available reference hydrological state. This controlled setting therefore provides a direct evaluation of the ability of the method to reconstruct fine-scale TWSA patterns while remaining consistent with the original coarse-resolution satellite-like observations. Second, the framework is applied to real monthly GRACE(-FO) TWSA products at 3° resolution to assess its practical applicability beyond the simulated environment. Because no independent high-resolution TWSA ground truth is available, the real-data experiment cannot provide a complete validation of the downscaled 1° fields. Instead, it evaluates whether the original gravimetric signal is preserved after aggregation to the native satellite resolution. In addition, GRACE-C-like simulations and GRACE(-FO) observations are compared over their common period in terms of water balance equation closure and spatial long-term trends.

https://doi.org/10.5194/egusphere-2026-4563
Preprint. Discussion started: 4 August 2026

Figure 1. Regions of interest: Africa, Amazon, Europe, Ganges, Mississippi.



2 Methodology

This work presents an unsupervised deep-learning approach for the spatial downscaling of terrestrial water storage anomaly (TWSA) grids from their native coarse resolution to a 1° target grid. The method relies on Convolutional Neural Networks (CNNs) to infer spatial links between coarse TWSA observations and higher-resolution auxiliary predictors derived from meteorological and topographic datasets. The framework is based on an unsupervised high-to-high (H2H) strategy, in which the network takes a high-resolution input and directly generates a high-resolution TWSA estimate that is subsequently aggregated back to the original satellite-like resolution and compared with the satellite product through the loss function. In this way, the model can be trained without requiring a direct high-resolution TWSA reference, making the approach suitable for present and future satellite gravimetry missions, characterized by the absence of a high-resolution ground truth.

The analysis is carried out over five regions of interest, namely Africa, the Amazon basin, Europe, the Ganges basin, and the Mississippi basin. These areas have been selected to cover different hydro-climatic regimes, geographical settings, and gravity-driven TWSA variability. For each region, an independent network is trained in order to account for region-dependent relationships between TWSA variability and the auxiliary climatic predictors. Each study area is represented by a $48° \times 48°$ spatial window, whose latitude and longitude limits are reported in Table 1.

https://doi.org/10.5194/egusphere-2026-4563
Preprint. Discussion started: 4 August 2026

Table 1. Latitude and longitude limits of the five study regions used for model training and validation

Region	lat_min	lat_max	lon_min	lon_max
Africa	-23	25	4	52
Amazon	-36	12	-83	-35
Europe	24	72	-3	45
Ganges	9	57	68	116
Mississippi	24	72	-115	-67

2.1 Model architecture

The adopted neural network is based on a U-Net-like architecture, characterized by a symmetric encoder–decoder structure with skip connections between layers at matching spatial scales (Fig. 2). The encoder branch progressively compresses the spatial information through convolutional and pooling operations, allowing the model to extract increasingly abstract spatial features. The decoder branch then reconstructs the target high-resolution field by means of upsampling and convolutional layers, while the skip connections help retain fine-scale spatial information from the encoder.

The input to the network is a multi-channel tensor of size (C_{in}, H, W), where C_{in} denotes the number of input channels, including the auxiliary meteorological and topographic variables described in Sect. 3.3, all resampled onto the 1° target grid. Each convolutional block is composed of two consecutive 3×3 convolutional layers followed by ReLU activation functions. A separate U-Net model is trained for each of the five selected regions. Since each regional patch has a spatial extent of $48° \times 48°$, the two pooling operations reduce the feature maps to a 12×12 bottleneck resolution. This provides a compromise between capturing large-scale hydrological patterns and preserving sufficient spatial detail for the reconstruction. The architecture is summarized in Table 2.

https://doi.org/10.5194/egusphere-2026-4563
Preprint. Discussion started: 4 August 2026

Figure 2. U-Net-like architecture used for TWSA downscaling.

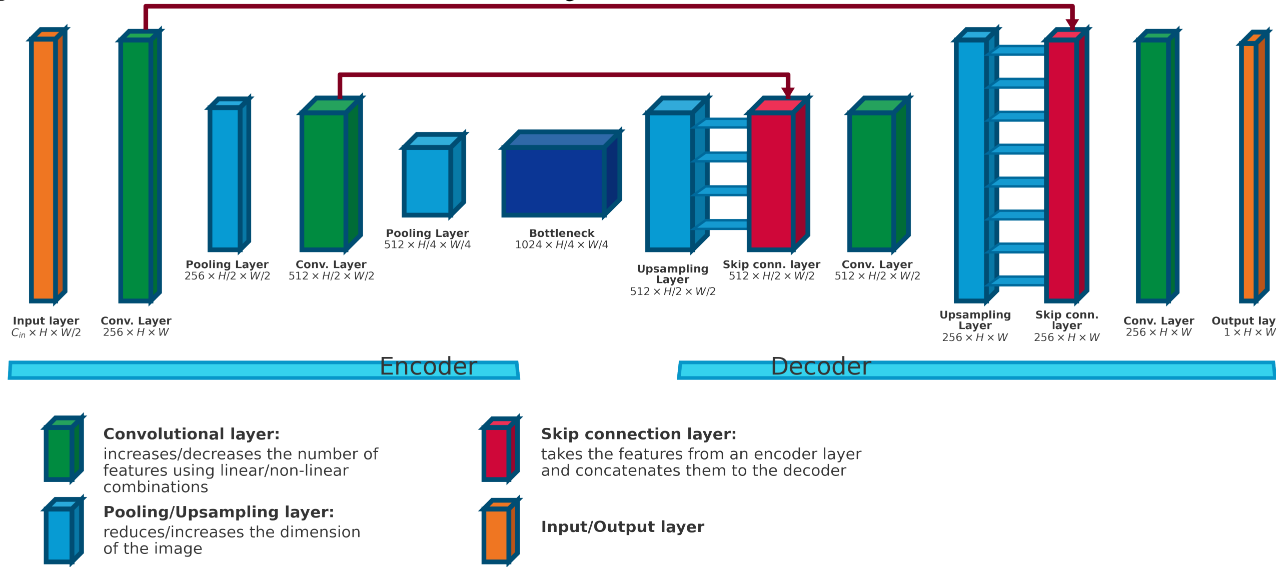


Table 2. Summary of the U-Net-like architecture used for TWSA downscaling. H and W denote the spatial dimensions of the input patch.

Layer (type)	In $\rightarrow$ Out Channels	Output size
Input	C_{in}	$[C_{\text{in}}, H, W]$
`enc1` (Conv2D $\times 2$ + ReLU)	$C_{\text{in}} \rightarrow 256$	$[256, H, W]$
`pool1` (MaxPool2D)	–	$[256, H/2, W/2]$
`enc2` (Conv2D $\times 2$ + ReLU)	$256 \rightarrow 512$	$[512, H/2, W/2]$
`pool2` (MaxPool2D)	–	$[512, H/4, W/4]$
`bottleneck` (Conv2D $\times 2$ + ReLU)	$512 \rightarrow 1024$	$[1024, H/4, W/4]$
`up2` (ConvTranspose2D)	$1024 \rightarrow 512$	$[512, H/2, W/2]$
`dec2` (Conv2D $\times 2$ + ReLU)	$1024 \rightarrow 512$	$[512, H/2, W/2]$
`up1` (ConvTranspose2D)	$512 \rightarrow 256$	$[256, H, W]$
`dec1` (Conv2D $\times 2$ + ReLU)	$512 \rightarrow 256$	$[256, H, W]$
`out_conv` (Conv2D, 1×1)	$256 \rightarrow 1$	$[1, H, W]$

2.2 Unsupervised downscaling scheme

The network is trained using an unsupervised high-to-high (H2H) downscaling scheme. The objective is to estimate 1° TWSA fields whose coarse-scale representation remains consistent with the original satellite product at its native resolution, namely 3° for the simulated 5-daily and GRACE(-FO) products, and 2° for the simulated monthly products. At each training step, the model receives the auxiliary variables onto the 1° grid, then it produces a high-resolution prediction, denoted as TWSA_{HR}.

https://doi.org/10.5194/egusphere-2026-4563
Preprint. Discussion started: 4 August 2026

To enforce consistency with the original gravimetric observation, the predicted high-resolution field is aggregated back to the native coarse resolution through an area-weighted average, yielding TWSA_{LR}. This low-resolution representation is then compared with the original satellite product, TWSA_{SAT}, using a mean squared error (MSE) loss:

$$\mathcal{L}_{\mathrm{MSE}} = \frac{1}{N}\sum_{i=1}^{N}\left(\mathrm{TWSA}_{LR}^{(i)} - \mathrm{TWSA}_{SAT}^{(i)}\right)^2, \tag{1}$$

where N is the number of valid pixels in the image, corresponding to the number of land pixels. This formulation constrains the predicted high-resolution map to reproduce the large-scale signal observed by the satellite while allowing the network to infer sub-grid spatial variability from the auxiliary predictors.

2.3 Training strategy

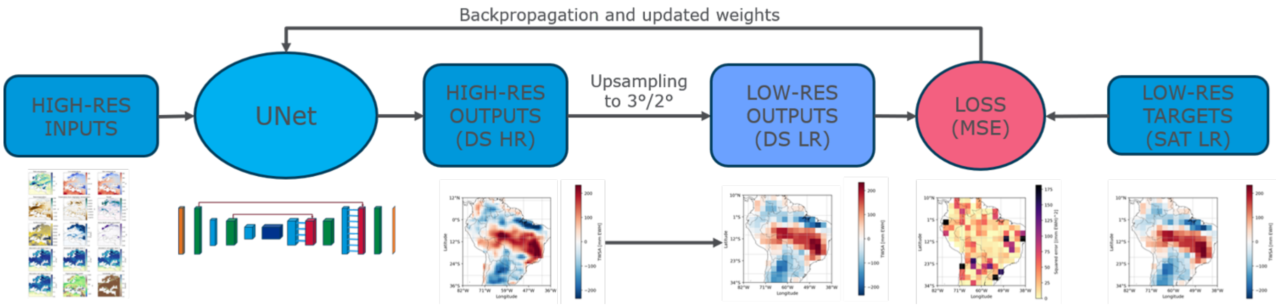
The training procedure is applied to filtered real and simulated Level-3 TWSA grids. For each gravimetric time step, the corresponding 1° auxiliary fields are provided as input to the network. The TWSA products, auxiliary variables, and preprocessing steps are described in detail in Sect. 3. Because TWSA exhibits strong temporal persistence, the dynamic input variables are not limited to the current time step t, but also include information from previous time steps. Depending on the variable, temporal lags of up to three months are considered. In the 5-daily case, each month corresponds to $n = 6$ samples, while in the monthly case it corresponds to $n = 1$ sample. For a time-lag of three months, for example, this results in a total of $N_{\mathrm{lag}} = 19$ samples in the 5-daily case and $N_{\mathrm{lag}} = 4$ in the monthly case, including also the current timestep t. The lag length adopted for each input variable is reported in Table 3.

Seasonal information is included through two additional 48 × 48 grids, whose values are given by $\sin(2\pi\,\mathrm{doy}/N_y)$ and $\cos(2\pi\,\mathrm{doy}/N_y)$, respectively, where doy indicates the day of the year and $N_y = 365.25$. This temporal encoding provides the network with information on the seasonal phase and helps represent the memory effects associated with hydrological processes controlling storage variations.

To capture region-specific hydrological behavior, each of the five study regions is processed using an independently trained network. In this way, each model is tailored to the region on which it is trained, improving the performance over a single global model. Moreover, processing 48×48 images is computationally less demanding than processing global-domain images, as they can be fed to the network as a whole without requiring prior partitioning into patches. Training is performed with the Adam optimizer, using a learning rate of 10^{-4} and a total of 50 epochs. Each epoch loops over the full time series, with a batch size equal to 1, so that the model parameters are updated after each individual sample. This configuration has been found to perform better than larger batch sizes. Although the batch size is small, the optimization is carried out over the complete dataset, allowing the model to learn general relationships between the coarse TWSA fields and the auxiliary predictors. In principle, the same procedure could also be applied to a single image; however, in that case the optimization would be image-specific and would not directly imply generalization to unseen samples. The training workflow is illustrated in Fig. 3.

https://doi.org/10.5194/egusphere-2026-4563
Preprint. Discussion started: 4 August 2026

Figure 3. U-Net training scheme.



2.4 Ensemble learning

An ensemble of 10 independently trained models has been generated using different random seeds to assess the sensitivity of the model to the random initialization of the network parameters. The final downscaled TWSA estimate has been obtained as the pixel-wise mean of the ensemble predictions, while the corresponding standard deviation has been used to quantify the spread among the individual model outputs. Deep ensembles are commonly employed in ML applications (Gou and Soja, 2024) to characterize model uncertainty, particularly the variability associated with parameter initialization and the optimization process. However, the resulting standard deviation should not be interpreted as the physical uncertainty of the final downscaled TWSA product. It represents only a part of the epistemic uncertainty associated with the trained model and does not account for other uncertainty sources, such as errors associated to the satellite observations or input predictors.

3 Data preparation

3.1 Simulated terrestrial water storage anomaly grids

Simulated Terrestrial Water Storage fields are provided by end-to-end (E2E) closed-loop experiments carried out in the frame of the ESA-funded SING (Studying the Impact of NGGM and MAGIC Missions) project (Pfeffer and the SING consortium, 2026; European Space Agency (ESA), 2025b). This work uses 5-daily and monthly Level-3 filtered (L3B) data provided by the project with a temporal coverage from 2007 to 2020. These simulations integrate instrument noise, background model errors, temporal aliasing and realistic satellite orbits in order to take into account all error contributions (European Space Agency (ESA), 2025a). The reference TWSA grids are obtained from the hydrology (H), cryosphere (I), solid Earth (S), sea-level (B), earthquakes (E) and regional deep ocean transport and sedimentation (D) combined layers (HISBED) of the ESA Earth System Model (ESM) 3.0 (Dobslaw et al., 2015; Shihora et al., 2026), provided as spherical harmonics coefficients up to d/o 180, corresponding to approximately 1° resolution using the conversion $\theta = 180/l_{max}$ where l_{max} is the maximum degree and θ is the native angular resolution. The coefficients have been downloaded from the GFZ Information System and Data Center (GFZ German Research Centre for Geosciences, 2026). The simulated L3B data is produced from the L2B data provided in terms of spherical harmonics coefficients up to d/o 70 for the 5-daily products (≈2.57°) and d/o 120 (≈1.5°) for the monthly

https://doi.org/10.5194/egusphere-2026-4563
Preprint. Discussion started: 4 August 2026

products, replacing the C_{20} term by more accurate Satellite Laser Ranging (SLR) data (L2PB) and correcting for earthquake and solid Earth signals removing the E and S layers of the ESA ESM 3.0. In particular, the E layer includes co-seismic and post-seismic signals while the S layer includes Glacial Isostatic Adjustment (GIA), Little Ice Age (LIA) and core signals. The coefficients are then synthesized on a latitude/longitude grid, whose spatial sampling is defined by the native angular resolution of the products approximated to 3° for the 5-daily and 2° grid for the monthly products. To evaluate the performance of the downscaling method, the full reference (d/o 180) is used for the comparison at 1°, while for the comparison at the satellite resolution the reference is truncated to the resolution of the simulations to avoid treating unresolved high-degree signals as omission error (Kusche et al., 2025).

3.2 GRACE(-FO) terrestrial water storage anomaly grids

For real data analysis, the GRACE and GRACE-FO COST-G TWSA fields (RL02, v. 0001) (Boergens et al., 2025) have been downloaded from the GFZ GravIS data portal (Dahle et al., 2025). The products are provided as gridded monthly TWSA L3 on a 1° grid, therefore they have been re-gridded to 3° to match the effective spatial resolution of the satellite using an area-weighted average to preserve the total water mass within each coarse grid cell. The time series from April 2002 to March 2025 has been used.

3.3 Auxiliary fields

The inputs for the neural network are a set of auxiliary climate variables from the ERA5-Land dataset (Muñoz-Sabater et al., 2021; Copernicus Climate Change Service, 2019), the ETOPO2022 Digital Elevation Model (DEM) (MacFerrin et al., 2025; NOAA National Centers for Environmental Information, 2022) and the ESA WorldCover 2021 v200 (Zanaga et al., 2022) for land cover mapping. Data from ERA5-Land and the DEM have been downloaded at their original spatial resolution (0.1° for ERA5-Land, 60 arcsec for the DEM) and then aggregated to the target angular resolution of 1°. The WorldCover processing required two extra steps. First: the downloaded map at 10 m resolution has been binarized setting to 1 the pixels belonging to the *Permanent Water Bodies* class and the rest to 0 to act as a proxy for surface water. Second: the map has been spatially aggregated to 1° resolution using an area weighted average. The result is a non-binary map with values between 0 and 1 representing the percentage of water in each pixel.

The different inputs are listed in Tab. 3.

https://doi.org/10.5194/egusphere-2026-4563
Preprint. Discussion started: 4 August 2026

Table 3. Neural Network Inputs

Input	Unit	Native Resolution	Time-lag
Total precipitation	m	0.1°	3 months
2 m temperature	K	0.1°	3 months
Skin temperature	K	0.1°	2 months
Total evaporation	m	0.1°	3 months
Evaporation from vegetation transpiration	m	0.1°	3 months
Runoff	m	0.1°	3 months
Surface pressure	Pa	0.1°	1 month
Snow cover	None	0.1°	1 month
Snow depth water equivalent	m	0.1°	3 months
Volumetric soil water layer 1	$\mathrm{m}^3/\mathrm{m}^3$	0.1°	1 month
Volumetric soil water layer 2	$\mathrm{m}^3/\mathrm{m}^3$	0.1°	2 months
Volumetric soil water layer 3	$\mathrm{m}^3/\mathrm{m}^3$	0.1°	3 months
Volumetric soil water layer 4	$\mathrm{m}^3/\mathrm{m}^3$	0.1°	3 months
Elevation	m	60 arcsec	None (static)
Land/Water mask	None	10 m	None (static)

Considering the time-lag configuration adopted for each dynamic variable described in Sec. 2.3 the total number of input channels C_{in} is given by:

$$C_{\mathrm{in}} = \sum_{j=1}^{C_{\mathrm{dyn}}} N_{\mathrm{lag},j} + C_{\mathrm{static}} + C_{\mathrm{time}}, \tag{2}$$

where:

- $C_{\mathrm{dyn}} = 13$;
- $C_{\mathrm{static}} = 2$;
- $C_{\mathrm{time}} = 2$;
- $N_{\mathrm{lag},j}$ is the total number of temporal samples included for the j-th dynamic variable, including the current time step.

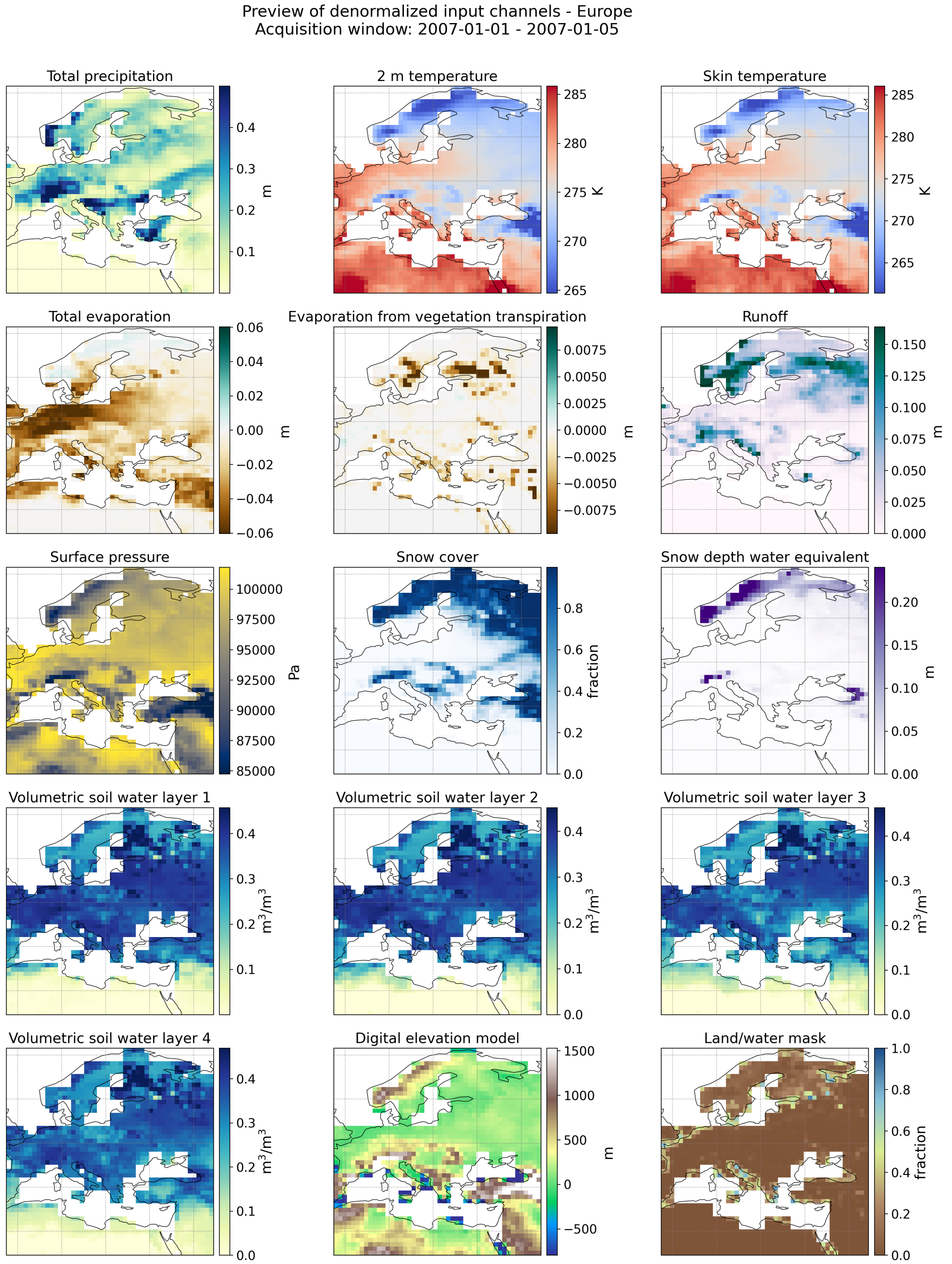
The number of input channels therefore corresponds to 203 channels for the 5-daily and 48 channels for the monthly analysis. This, together with the shorter time series, makes the training significantly faster in the case of monthly acquisitions. The input fields have been temporally aggregated to match the temporal resolution of the TWSA products taking into account the extensive or intensive nature of each variable. For the GRACE(-FO) case, dates not explicitly contributing to the final gravity

https://doi.org/10.5194/egusphere-2026-4563
Preprint. Discussion started: 4 August 2026

field solutions are nevertheless included in the aggregation of the input fields, to avoid introducing an artificial temporal sub-sampling of the climatological forcing that can potentially lead to mass imbalance and aliasing effects. Finally, the target spatial resolution is obtained via area-weighted averaging onto the target grid cells. Figure 4 shows a timestep of the 5-daily input channels (excluding the time-encoding channels) for Europe.

https://doi.org/10.5194/egusphere-2026-4563
Preprint. Discussion started: 4 August 2026

Figure 4. Snapshot of the denormalized input channels (excluding time) considering the first time sample of each variable (Europe).



https://doi.org/10.5194/egusphere-2026-4563
Preprint. Discussion started: 4 August 2026

All the time-series (TWSA and auxiliary) and the DEM are normalized using the z-score (i.e. centered to zero mean and scaled by their standard deviation) in order to harmonize the dynamic ranges of the otherwise very heterogeneous variables and stabilize the network optimization. The mean and standard deviation of the coarse TWSA are used to descale the predicted TWSA maps.

4 Evaluation metrics

The final downscaled products are organized as a $5 \times T \times 48 \times 48$ array, where 5 is the number of regions, T is the number of timesteps, and 48×48 is the spatial grid at 1° resolution. For each region and metric, a single representative value is derived through two complementary aggregation strategies to provide a balanced evaluation of model performance:

1. Temporo-spatial (TS): the metric is first computed temporally at pixel level and then reduced spatially over the valid domain.

2. Spatio-temporal (ST): the metric is first computed spatially for each map and then reduced over time, yielding one value per region.

In the TS view, the metric is first evaluated over the temporal dimension to obtain a spatial map, and this map is then reduced over space to a scalar summary. In the ST view, the metric is first evaluated over space at each time step to obtain a time series, which is then aggregated over time into a scalar. Thus, TS emphasizes temporal consistency at the pixel level, whereas ST emphasizes spatial consistency at the snapshot level. Using both views makes it possible to evaluate whether a model reproduces both the temporal evolution of local signals and the spatial organization of individual fields. The final reduction is median-based to reduce the sensitivity to outliers.

The evaluation includes pixel-wise metrics, namely Pearson correlation coefficient (r), Nash–Sutcliffe efficiency (NSE), normalized root mean square error (nRMSE), normalized median absolute error (nMdAE) and variance ratio (β), together with structure-wise metrics, namely the Structural Similarity Index Measure (SSIM), the Gradient Magnitude Similarity Deviation (GMSD), and the Fractions Skill Score (FSS). Low-pass versions of the structural metrics are also considered to explicitly assess the reproduction of large-scale spatial patterns. The complete mathematical formulation of the individual metrics, the adopted aggregation procedure, and the window sizes used for multiscale structural metrics are reported in Appendix A.

4.1 Composite score

Since each individual metric captures only one aspect of model performance, the final assessment is based on a composite score that combines the most relevant pixel-wise and structure-wise indicators into a single ranking criterion across regions, observing configurations, and temporal resolutions. The metric families included in the composite score are Pearson correlation, variance ratio, NSE, SSIM, GMSD, FSS, nMdAE, nRMSE and bias. The considered metrics are listed in Table 4, with metrics in the same family sharing the same empirical weight reflecting the relative importance attributed to the corresponding performance dimension. Higher weights are assigned to correlation, variance consistency, and NSE, while structural metrics

https://doi.org/10.5194/egusphere-2026-4563
Preprint. Discussion started: 4 August 2026

are used to complement pixel-wise scores by accounting for the spatial realism of the downscaled fields. A more detailed formulation of the structure-wise metrics, including the window sizes adopted for SSIM and FSS can be found in Appendix A. In Table 4 the *Evaluated field* column refers to the product(s) being evaluated for the specific metric. For instance, DS_{HR} and DS_{LR} denote the downscaled product at high and native resolution, respectively. For the latter, given the absence of a ground truth signal, the scores are evaluated against the native satellite product, therefore only a low-resolution composite score is computed. For the downscaled simulated products, on the other hand, the high-resolution evaluation is performed against the corresponding ESA ESM 3.0 HISBED TWSA (Sec. 3.1). Before aggregation, all metrics are converted into positively oriented quantities, so that larger values always indicate better performance, and are normalized to the interval $[0, 1]$. When multiple realizations of the same metric family are available, for example SSIM or FSS evaluated at different window sizes, they are first averaged within the corresponding family. If both TS and ST aggregations are available, they are then combined with equal weight. The detailed implementation of the composite score is provided in Appendix B.

Table 4. Summary of the evaluation metrics considered in this study and raw weights assigned to each metric family in the composite score. TS and ST denote temporo-spatial and spatio-temporal aggregation, respectively. Evaluated field indicates the TWSA field being assessed for each metric. Variants belonging to the same metric family share the same raw weight w_k. The effective weights $\widehat{w}_k$ used in a given experiment are obtained by renormalizing the raw weights over the subset of available metric families in the specific comparison scenario as described in Appendix B.

Metric	Aggregation	Evaluated field	Raw weight w_k
Pearson correlation r	TS, ST	DS_{HR}, DS_{LR}	0.18
β variance ratio	TS, ST	DS_{HR}, DS_{LR}	0.17
NSE	TS, ST	DS_{HR}, DS_{LR}	0.15
SSIM (global)	ST	DS_{HR}, DS_{LR}	
SSIM (windowed)	ST	DS_{HR}, DS_{LR}	0.11
LP-SSIM (global)	ST	DS_{HR}	
GMSD	ST	DS_{HR}	0.11
LP-GMSD	ST	DS_{HR}	
FSS (windowed)	ST	DS_{HR}	0.11
LP-FSS (windowed)	ST	DS_{HR}	
nMdAE	TS, ST	DS_{HR}, DS_{LR}	0.07
nRMSE	TS, ST	DS_{HR}, DS_{LR}	0.06
Bias	TS, ST	DS_{HR}, DS_{LR}	0.04

Since not all metric families are necessarily available for every experiment, comparison setup, or aggregation mode, the raw weights w_k are renormalized over the subset of available metric families $\mathcal{A}$:

$$\widehat{w}_k = \frac{w_k}{\sum_{\ell \in \mathcal{A}} w_\ell}, \qquad k \in \mathcal{A}. \tag{3}$$

https://doi.org/10.5194/egusphere-2026-4563
Preprint. Discussion started: 4 August 2026

The final composite score is then defined as

$$\mathrm{CS} = \sum_{k \in \mathcal{A}} \widehat{w}_k F_k, \tag{4}$$

where F_k is the normalized score of metric family k. This formulation ensures that all metrics contribute on a common scale, that multiple realizations within the same family are equally weighted, and that missing metric families do not introduce artificial penalties. Further implementation details are presented in Appendix B.

4.2 Water balance equation closure

To assess the physical consistency of the provided results, downscaled TWSA grids have been evaluated in terms of water balance equation (WBE) closure considering precipitation (P), evapotranspiration (ET) and runoff (RO) from ERA5-Land. Variations in TWSA have been computed as

$$\Delta TWSA(t_i) := TWSA(t_i) - TWSA(t_{i-1}), \tag{5}$$

and the WBE is defined as

$$\Delta TWSA(t_i) = P(t_i) - ET(t_i) - RO(t_i). \tag{6}$$

The right-hand side of Eq. (6) indicates accumulated quantities from t_{i-1} to t_i. In the ERA5-Land convention, upward latent water fluxes from the surface to the atmosphere are stored as negative values in `total_evaporation`. To ensure consistency with the WBE, the components are therefore defined as

$$P = \texttt{total\_precipitation};$$

$$ET = -\texttt{total\_evaporation};$$

$$RO = \texttt{runoff},$$

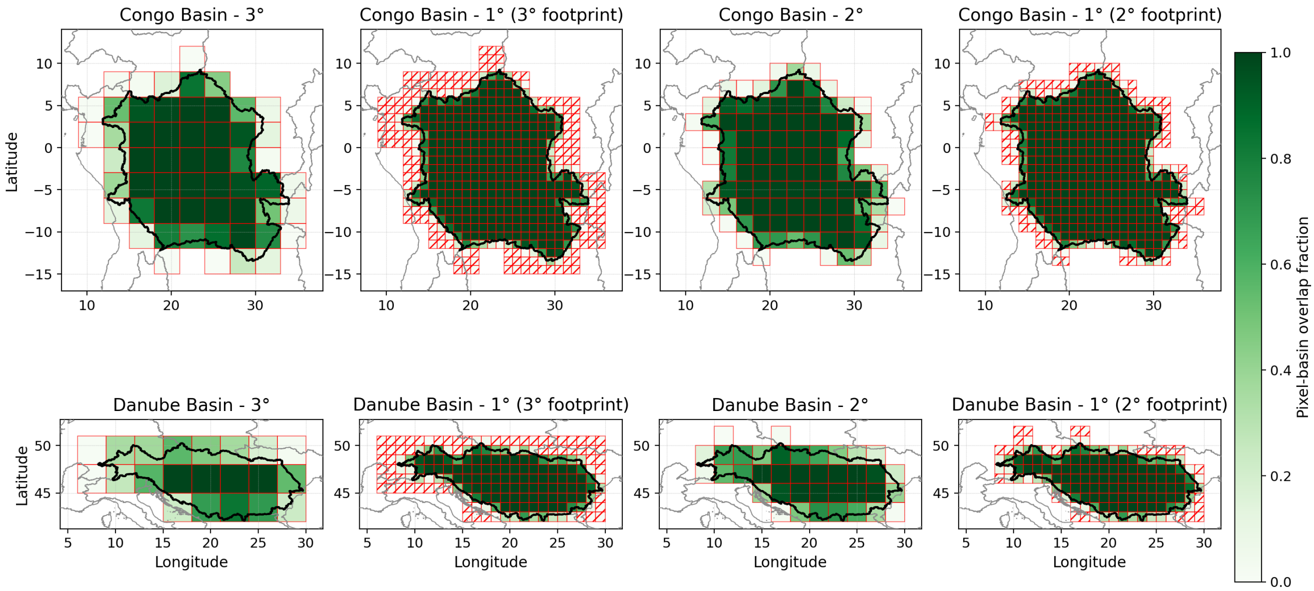
such that positive values of ET represent water loss from the surface. The residual is defined as the difference between the left-hand and right-hand sides of Eq. (6) and is evaluated at basin scale over the Danube and Congo basins by applying an area-weighted average in which each pixel is weighted according to the fraction of its area overlapping the basin:

$$\mathrm{Res}_{WBE} := \Delta TWSA(t_i) - (P(t_i) - ET(t_i) - RO(t_i)). \tag{7}$$

Fig. 5 shows the 3° and 2° footprints for the Congo and Danube basins, together with the corresponding 1° representations. In Fig. 5, the red hatched cells denote 1° pixels included in the coarse-resolution footprint but assigned zero weight because they do not overlap the basin, while the color shading represents the fractional overlap between each pixel and the basin. Basin boundaries have been derived from HydroBASINS (Lehner and Grill, 2013) Level-3 polygons. The residuals from the different products are compared against the reference ESA ESM 3.0 HISBED residuals.

https://doi.org/10.5194/egusphere-2026-4563
Preprint. Discussion started: 4 August 2026

Figure 5. Coarse-resolution (3° and 2°) footprints of the Congo and Danube basins and corresponding 1° representations. In the 1° panels, shading denotes the fractional pixel overlap with the basin and red hatched cells indicate pixels within the coarse-resolution footprint that receive zero weight because they do not intersect the basin. Basin boundaries are derived from HydroBASINS Level-3 polygons (Lehner and Grill, 2013).



However, due to the difficulty of reproducing the actual hydrological variability over the Congo basin in the LISFLOOD hydrological model, the simulations yield unrealistic outputs over the region that are not directly comparable with observations.

5 Results

This section presents the results of the AI-based downscaling of TWSA grids for the simulated 5-daily, simulated monthly, and GRACE(-FO) scenarios. In the following, the high-resolution downscaled products at 1°, obtained as output of the U-Net, are referred to as Downscaled HR (DS HR). The same products aggregated back to the native satellite resolution (3° for the 5-daily and GRACE(-FO) cases, and 2° for the monthly case) are referred to as Downscaled LR (DS LR). The original satellite products, available only at coarse resolution, are denoted as NATIVE. The ESA ESM 3.0 HISBED reference TWSA fields are indicated as ESM 3.0 HISBED d/o 70 (3°), ESM 3.0 HISBED d/o 120 (2°), and ESM 3.0 HISBED d/o 180 (1°), according to the truncation procedure described in Sec. 3.1. Model performance is assessed through the composite score described in Sec. 4.1 and WBE closure (Sec. 4.2).

5.1 Composite score results

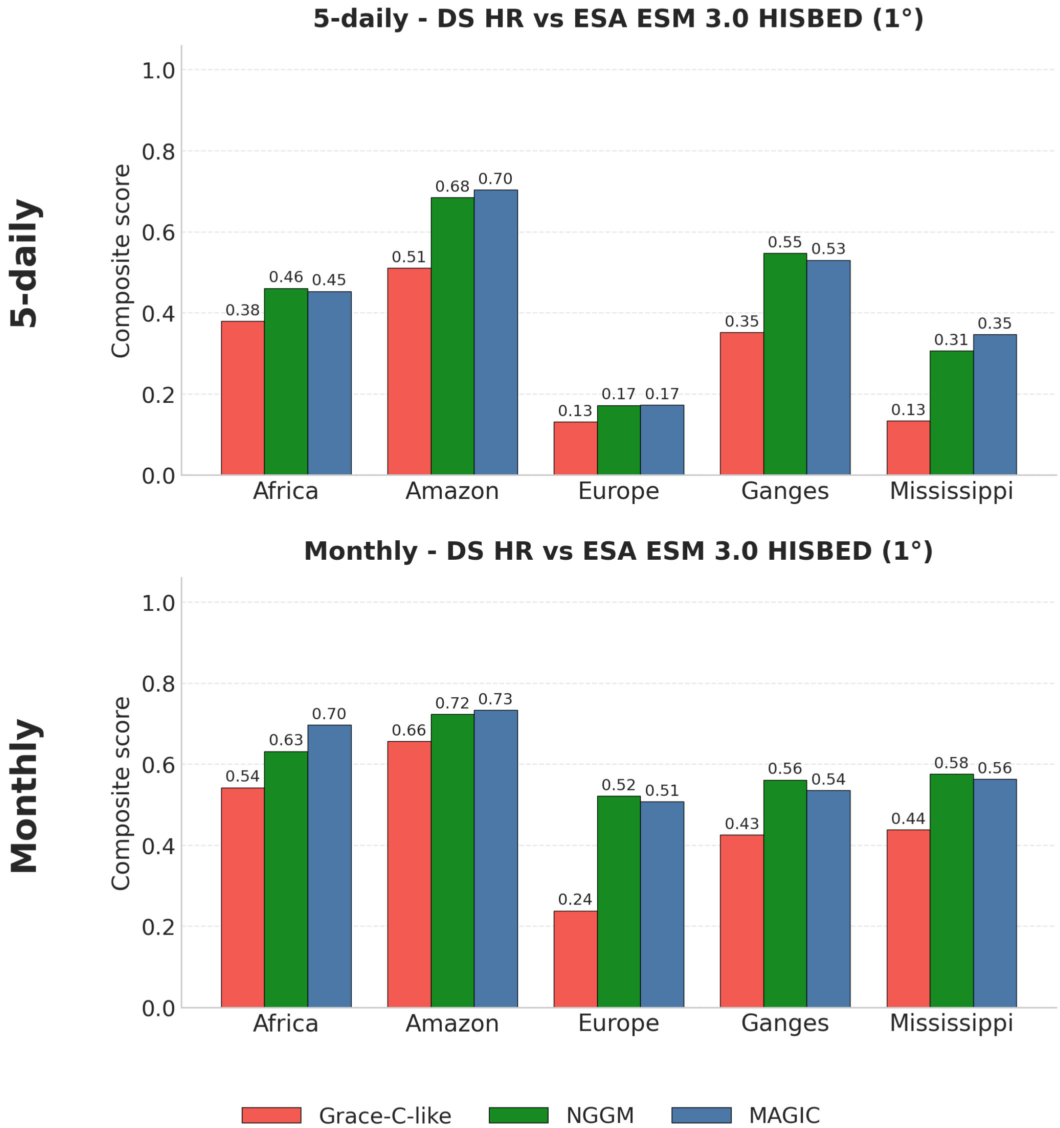
For the simulated 5-daily and monthly configurations, two composite scores have been computed for each configuration–region combination using the ensemble downscaled products. The first composite score (Fig. 6) evaluates the Downscaled HR TWSA

https://doi.org/10.5194/egusphere-2026-4563
Preprint. Discussion started: 4 August 2026

fields against ESA ESM 3.0 HISBED at 1°, whereas the second (Fig. 7) compares the Downscaled LR TWSA fields with the NATIVE products. Three main aspects emerge. First, the Downscaled HR NGGM and MAGIC products generally achieve higher scores than the Downscaled HR GRACE-C-like product at both temporal resolutions. Secondly, composite scores are generally higher for the monthly scenarios, as expected from the higher quality and native resolution of the monthly satellite products. Thirdly, the LR composite scores in Fig. 7 are consistently close to unity, indicating strong structural coherence between the Downscaled LR and NATIVE fields. These LR scores should not be interpreted primarily as an inter-configuration quality ranking, but rather as a measure of how well each downscaled product preserves its corresponding mission-specific satellite observation when aggregated back to the native coarse resolution. Values close to 1 indicate that the downscaling procedure does not introduce a substantial loss of information at the satellite scale, and that possible U-Net-generated structures at 1° do not significantly affect the consistency of the product once evaluated at LR. In the monthly scenarios, the Downscaled LR products from GRACE(-FO) over the shared 2007–2020 period are also included for visual comparison. GRACE(-FO) scores over the full 2002–2025 period are similar: Africa: 0.90, Amazon: 0.96, Europe: 0.92, Ganges: 0.92, and Mississippi: 0.96.

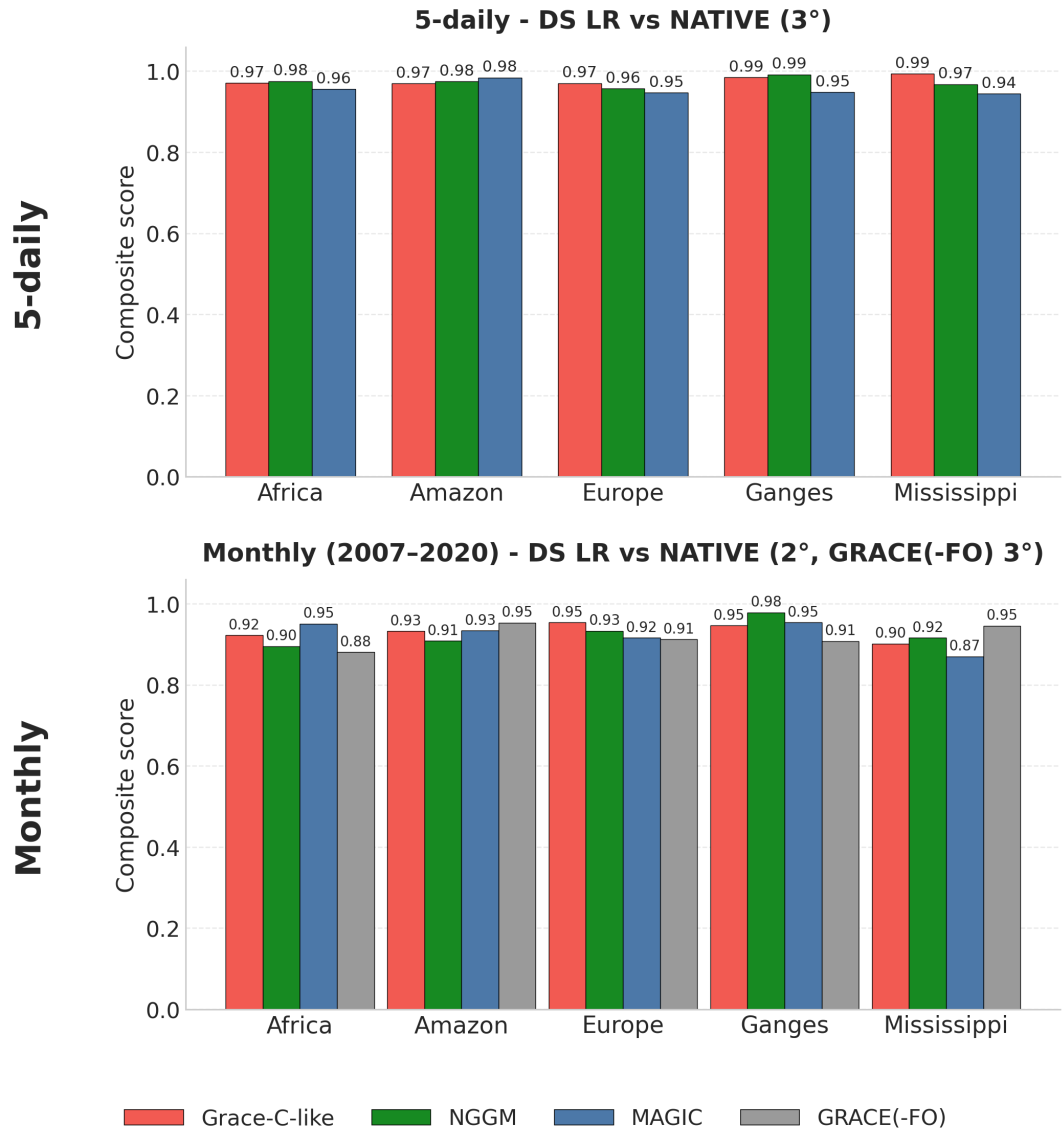
https://doi.org/10.5194/egusphere-2026-4563
Preprint. Discussion started: 4 August 2026

Figure 6. High-resolution composite scores for each region and configuration in the 5-daily (top) and monthly (bottom) configurations. The comparison of the Downscaled HR products is made against the ESA ESM 3.0 HISBED reference.



https://doi.org/10.5194/egusphere-2026-4563
Preprint. Discussion started: 4 August 2026

Figure 7. Low-resolution composite scores for each region and configuration in the 5-daily (top) and monthly (bottom) configurations. The comparison of the Downscaled LR products is made against the native satellite product. In the monthly scenarios, the Downscaled LR products from GRACE(-FO) from the shared timespan (2007-2020) are also included for visual comparison.



5.2 WBE closure results

Tables 6 and 5 report the WBE closure summaries for the Danube and Congo basins, respectively, including the simulated 5-daily, simulated monthly, and GRACE(-FO) products. For each product, the reported statistics are computed from the difference between its WBE residual time series and the corresponding ESA ESM 3.0 HISBED residual time series at the same spatial resolution. Therefore, the mean and standard deviation reported in the tables refer to the residual mismatch with respect to

https://doi.org/10.5194/egusphere-2026-4563
Preprint. Discussion started: 4 August 2026

the reference, not to the absolute WBE residual of the product itself. For GRACE(-FO), the satellite is used in place of the HISBED reference. For the simulated products, the mean and standard deviation of the WBE residuals from Eq. (7) are very similar between the coarse and 1° HISBED references, indicating that the high- and low-resolution reference fields provide nearly identical basin-scale closure. This confirms that the change in spatial resolution does not substantially affect the large-scale WBE residual when evaluated over the basin. The downscaled products generally preserve the closure behavior of their corresponding low-resolution inputs, although the level of agreement depends on the basin, temporal resolution, and evaluation scale. This preservation is particularly evident for the 5-daily products, for which the DS LR and DS HR errors remain close to the NATIVE baseline in both basins. Within each 5-daily configuration, the similarity among NATIVE, DS LR, and DS HR therefore indicates that the downscaling procedure does not substantially alter the basin-scale WBE consistency of the original satellite product. A separate effect is observed when comparing the different simulated observing scenarios. Over the Danube, the improvement from the GRACE-C-like configuration to MAGIC is strong and systematic across temporal resolutions and spatial scales. For example, the 5-daily DS HR RMSE decreases from 14.47 to 7.10 mm, while the monthly DS HR RMSE decreases from 6.05 to 3.09 mm. Similar reductions are obtained for the corresponding NATIVE and DS LR products. This indicates that the higher quality of the simulated observing scenarios translates into improved basin-scale WBE consistency that is retained after downscaling. Over the Congo, the behavior is more scale- and resolution-dependent. In the 5-daily case, the DS LR and DS HR products remain close to the corresponding NATIVE baselines, and the RMSE decreases consistently from the GRACE-C-like to the MAGIC scenario, from 4.64 to 3.06 mm for DS LR and from 4.82 to 3.32 mm for DS HR. In the monthly case, however, the downscaled products preserve the satellite-scale closure only after reaggregation to 2°. At 1°, the DS HR RMSE values range from 7.02 to 7.53 mm, substantially exceeding both the NATIVE and DS LR errors, which remain between approximately 2 and 3 mm. Moreover, the configuration-dependent improvement is weak and non-monotonic at high resolution, with NGGM providing the lowest DS HR RMSE and MAGIC remaining close to the GRACE-C-like result. This aspect, observed only in the monthly-Congo scenario, may be related to the difficulty of reproducing the actual hydrological variability of the Congo basin with the OS LISFLOOD hydrological model (Knijff et al., 2010; Jensen et al., 2025), which is used as the land-hydrology component of ESA ESM 3.0 (Shihora et al., 2026). These limitations may affect the consistency between the simulated storage variations and the other water-budget components, making the resulting WBE closure more difficult to interpret. This point is currently under investigation. For the real GRACE(-FO) products, the downscaled fields aggregated back to 3° remain close to the original satellite residuals: the RMSE is 4.35 mm for Congo and 3.03 mm for Danube, corresponding to NRMSE values of 0.014 and 0.007, respectively. Overall, these results show that the downscaling procedure robustly preserves the satellite-scale WBE consistency, while preservation at 1° is generally strong in the 5-daily experiments and over the Danube, but less effective for the monthly Congo case. Compared with the individual-seed products, the ensemble does not produce a substantial additional change in the WBE diagnostics, indicating that the composite-score improvement associated with ensemble averaging is not obtained at the expense of basin-scale water-balance consistency.

https://doi.org/10.5194/egusphere-2026-4563
Preprint. Discussion started: 4 August 2026

Table 5. Danube basin WBE closure summary for the ensemble 5-daily, monthly, and GRACE(-FO) products. Mean, standard deviation, RMSE, and NRMSE are computed from the difference between the product WBE residual and the corresponding reference WBE residual, namely the HISBED for the simulated products and original satellite product for GRACE(-FO). NRMSE is normalized with respect to the standard deviation of the reference residual.

Scenario	Product	Mean [mm]	Std [mm]	RMSE [mm]	NRMSE [-]
Simulated 5-daily					
HISBED d/o 70 (3°): mean WBE residual = 9.865 mm, std = 111.29 mm					
HISBED d/o 180 (1°): mean WBE residual = 10.203 mm, std = 118.34 mm					
GRACE-C-like	SAT (3°)	0.007	14.17	14.17	0.127
NGGM	SAT (3°)	0.006	8.06	8.06	0.072
MAGIC	SAT (3°)	0.003	6.98	6.98	0.063
GRACE-C-like	DS (3°)	0.004	13.96	13.96	0.125
NGGM	DS (3°)	0.007	7.90	7.90	0.071
MAGIC	DS (3°)	0.005	6.86	6.86	0.062
GRACE-C-like	DS (1°)	0.010	14.47	14.47	0.122
NGGM	DS (1°)	0.012	8.11	8.11	0.069
MAGIC	DS (1°)	0.009	7.10	7.10	0.060
Simulated Monthly					
HISBED d/o 120 (2°): mean WBE residual = 62.926 mm, std = 479.04 mm					
HISBED d/o 180 (1°): mean WBE residual = 63.239 mm, std = 485.89 mm					
Scenario	Product	Mean [mm]	Std [mm]	RMSE [mm]	NRMSE [-]
GRACE-C-like	SAT (2°)	0.044	6.09	6.09	0.013
NGGM	SAT (2°)	0.025	3.04	3.04	0.006
MAGIC	SAT (2°)	0.027	2.85	2.85	0.006
GRACE-C-like	DS (2°)	0.052	6.20	6.20	0.013
NGGM	DS (2°)	0.024	2.82	2.82	0.006
MAGIC	DS (2°)	0.023	2.50	2.50	0.005
GRACE-C-like	DS (1°)	0.007	6.05	6.05	0.012
NGGM	DS (1°)	0.002	3.43	3.43	0.007
MAGIC	DS (1°)	0.001	3.09	3.09	0.006
GRACE(-FO)					
GRACE(-FO) SAT 3°: mean WBE residual = 65.3 mm, std = 458.9 mm					
Scenario	Product	Mean [mm]	Std [mm]	RMSE [mm]	NRMSE [-]
GRACE(-FO)	DS (3°)	-0.025	3.03	3.03	0.007

https://doi.org/10.5194/egusphere-2026-4563
Preprint. Discussion started: 4 August 2026

Table 6. Congo basin WBE closure summary for the ensemble 5-daily, monthly, and GRACE(-FO) products. Mean, standard deviation, RMSE, and NRMSE are computed from the difference between the product WBE residual and the corresponding reference WBE residual, namely the HISBED for the simulated products and original satellite product for GRACE(-FO). NRMSE is normalized with respect to the standard deviation of the reference residual.

Simulated 5-daily

HISBED d/o 70 (3°): mean WBE residual = 20.531 mm, std = 67.13 mm

HISBED d/o 180 (1°): mean WBE residual = 20.308 mm, std = 68.19 mm

Scenario	Product	Mean [mm]	Std [mm]	RMSE [mm]	NRMSE [-]
GRACE-C-like	SAT (3°)	-0.008	5.86	5.86	0.087
NGGM	SAT (3°)	0.001	4.96	4.96	0.074
MAGIC	SAT (3°)	-0.000	4.09	4.09	0.061
GRACE-C-like	DS (3°)	-0.012	4.64	4.64	0.069
NGGM	DS (3°)	-0.003	3.70	3.70	0.055
MAGIC	DS (3°)	-0.002	3.06	3.06	0.046
GRACE-C-like	DS (1°)	-0.025	4.82	4.82	0.071
NGGM	DS (1°)	-0.015	3.89	3.89	0.057
MAGIC	DS (1°)	-0.013	3.32	3.32	0.049

Simulated Monthly

HISBED d/o 120 (2°): mean WBE residual = 122.948 mm, std = 292.91 mm

HISBED d/o 180 (1°): mean WBE residual = 122.552 mm, std = 293.97 mm

Scenario	Product	Mean [mm]	Std [mm]	RMSE [mm]	NRMSE [-]
GRACE-C-like	SAT (2°)	0.004	3.13	3.14	0.011
NGGM	SAT (2°)	-0.012	2.48	2.48	0.008
MAGIC	SAT (2°)	-0.014	1.94	1.94	0.007
GRACE-C-like	DS (2°)	-0.041	3.20	3.21	0.011
NGGM	DS (2°)	-0.031	2.51	2.51	0.009
MAGIC	DS (2°)	-0.029	2.65	2.65	0.009
GRACE-C-like	DS (1°)	-0.046	7.53	7.53	0.026
NGGM	DS (1°)	-0.032	7.02	7.02	0.024
MAGIC	DS (1°)	-0.030	7.48	7.48	0.025

GRACE(-FO)

GRACE(-FO) SAT 3°: mean WBE residual = 130.3 mm, std = 309.6 mm

Scenario	Product	Mean [mm]	Std [mm]	RMSE [mm]	NRMSE [-]
GRACE(-FO)	DS (3°)	0.010	4.35	4.35	0.014

https://doi.org/10.5194/egusphere-2026-4563
Preprint. Discussion started: 4 August 2026

5.3 GRACE-C-like/GRACE(-FO) comparison

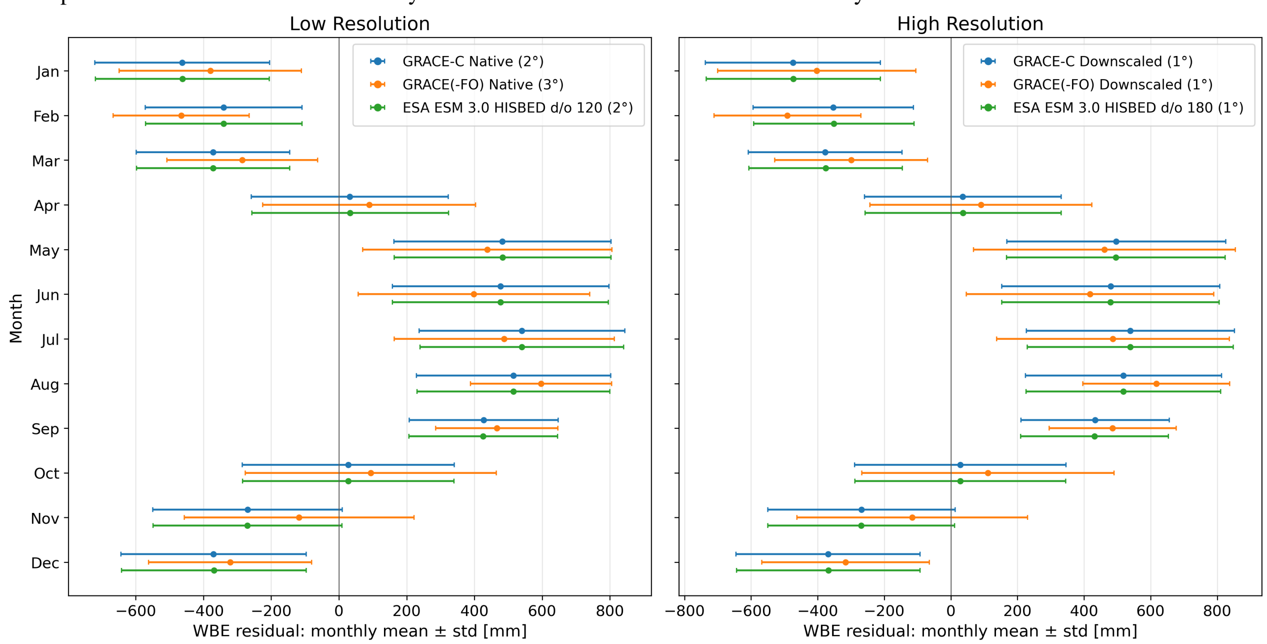
This framework enables a direct comparison between the simulated GRACE-C-like products and GRACE(-FO) observations over the common 2007–2020 period, both at their native satellite resolution and after the downscaling. The products have been evaluated in terms of WBE closure and compared through their long-term trends. However, due to the difficulty of reproducing the actual hydrological variability over the Congo basin, mentioned in Sect. 5.2, the simulations yield unrealistic outputs over the region that are not directly comparable with observations. Consequently, this area has been excluded from the present analysis, and the WBE closure assessment has been performed only over the Danube basin. Table 7 shows the mean WBE residual and the temporal standard deviation for the different products over the Danube basin. The mean residuals reported in the table are similar for GRACE-C-like, GRACE(-FO), and the ESA ESM 3.0 HISBED reference. However, the large variability of the residuals does not allow for a robust compatibility assessment. To reduce the influence of the pronounced seasonal cycle on the variability of the WBE residuals, the comparison between GRACE-C-like and GRACE(-FO) has been performed separately for each calendar month (Fig. 8). For each month, the difference between the mission-specific mean residuals has been quantified using Cohen's standardized mean difference (Cohen, 1988) and Welch's unequal-variance t-test (Welch, 1947). Cohen's d quantifies the difference between the mission-specific mean WBE residuals relative to their pooled interannual variability. Small absolute values of d therefore indicate that the difference between the two missions is limited compared with the natural variability of the residuals. Welch's test evaluates whether the difference between the two means is statistically distinguishable from zero while allowing for unequal variances and sample sizes. In this context, large p-values ($p > 0.05$) indicate insufficient evidence to reject the null hypothesis of equal mean residuals. The mean absolute Cohen's d has been approximately 0.28 for the Satellite LR, Downscaled LR, and Downscaled HR products, indicating generally small inter-mission differences. The maximum monthly $|d|$ ranged from 0.54 to 0.57, suggesting that moderate deviations occurred in individual months. However, Welch's test detected no statistically significant difference between the GRACE-C-like and GRACE(-FO) mean WBE residuals for any calendar month ($p > 0.05$ in all cases). Taken together, the small average effect sizes and the absence of statistically significant differences indicate that GRACE-C-like and GRACE(-FO) can be considered consistent over the Danube basin within the adopted statistical criteria. The compatibility between the two missions can also be assessed by examining the spatial correlation between their TWSA trend (Hartmann et al., 2013) fields. At $1°$ resolution, the downscaled GRACE-C-like and GRACE(-FO) products exhibit positive spatial correlations across the four studied areas, with correlation coefficients of 0.58 for the Amazon, 0.48 for Europe, 0.51 for the Ganges, and 0.72 for the Mississippi. These results indicate that the two missions generally identify consistent spatial patterns of TWSA change over the common 2007–2020 period, although differences in the local magnitude of the trends remain. The overall positive spatial correspondence therefore provides additional evidence of compatibility between the downscaled GRACE-C-like and GRACE(-FO) products.

https://doi.org/10.5194/egusphere-2026-4563
Preprint. Discussion started: 4 August 2026

Table 7. Mean signed WBE residual ± temporal standard deviation over the Danube basin during the common 2007–2020 period for GRACE-C-like and GRACE(-FO) products. Low resolution is 2° for GRACE-C-like and 3° for GRACE(-FO). Values are expressed in mm.

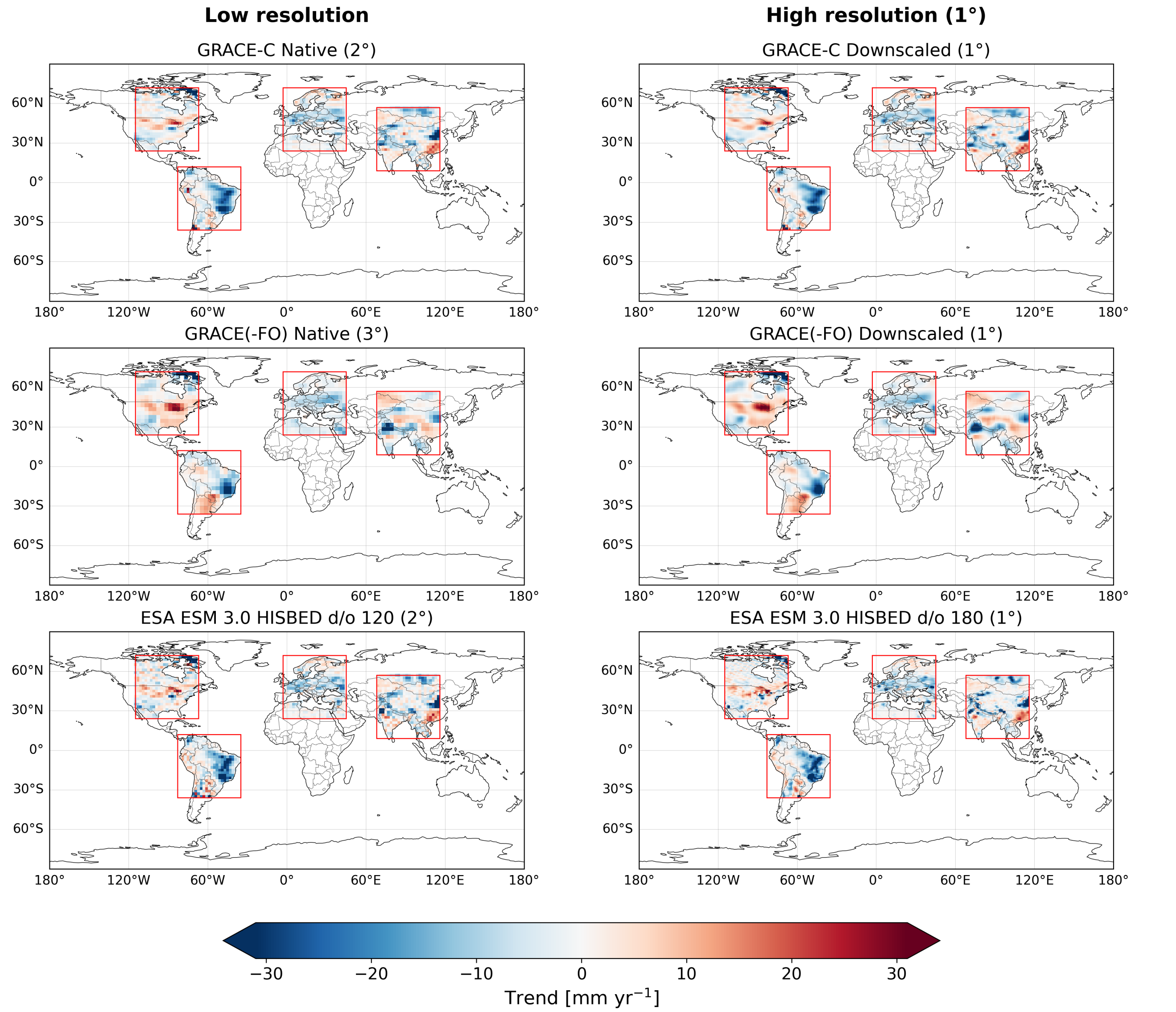
Product	GRACE-C-like	GRACE(-FO)
High resolution (1°)		
Downscaled	63.24 ± 487.13	80.29 ± 490.90
HISBED	63.24 ± 485.89	–
Low resolution		
Satellite	62.89 ± 479.91	76.96 ± 469.78
Downscaled	62.89 ± 479.64	76.94 ± 469.90
HISBED	62.93 ± 479.04	–

Figure 8. Calendar-month comparison of the WBE closure residuals over the Danube basin during the common 2007–2020 period. Points represent the mean residual for each calendar month, while horizontal error bars indicate the corresponding interannual standard deviation. The left panel compares GRACE-C-like NATIVE (2°), GRACE(-FO) NATIVE (3°), and ESA ESM 3.0 HISBED d/o 120 (2°) products. The right panel compares GRACE-C-like DS HR, GRACE(-FO) DS HR, and ESA ESM 3.0 HISBED d/o 180 at 1°. The overlap between the mission-specific intervals indicates consistency relative to the observed interannual variability.



https://doi.org/10.5194/egusphere-2026-4563
Preprint. Discussion started: 4 August 2026

Figure 9. Linear trends in TWSA over the common 2007–2020 period for GRACE-C-like, GRACE(-FO), and ESA ESM 3.0 HISBED low-resolution (left column) and high-resolution (right column) products. Trends are calculated independently for each grid cell using ordinary least squares (Hartmann et al., 2013) and are expressed in $\mathrm{mm\,yr^{-1}}$.



6 Conclusions

This study presented an unsupervised deep-learning framework for the spatial downscaling of terrestrial water storage anomalies (TWSA) from coarse satellite gravimetry products to a 1° target grid. The method relies on a U-Net convolutional neural network driven by hydro-climatic and topographic predictors, and is trained through a high-to-high strategy in which the

https://doi.org/10.5194/egusphere-2026-4563
Preprint. Discussion started: 4 August 2026

downscaled fields are aggregated back to the native satellite resolution and constrained to reproduce the original coarse-scale TWSA signal. Since no direct high-resolution TWSA reference is required for training, the framework can be applied both to simulated next-generation gravity mission products and to real GRACE(-FO) observations. The method has been tested in two settings: realistic end-to-end closed-loop simulations of future gravimetric observing configurations, including 5-daily products at 3° and monthly products at 2° resolution, and real monthly GRACE(-FO) TWSA products at 3° resolution. The simulated experiments allowed a direct evaluation against the ESA ESM 3.0 HISBED reference at 1°, while the real-data experiment assessed the applicability of the framework in the absence of a high-resolution ground truth. The composite score analysis shows that the quality of the input gravimetric products strongly affects the downscaled outputs. In the simulated scenarios, NGGM and MAGIC generally outperform the GRACE-C-like configuration at high resolution, while monthly products tend to achieve higher scores than the 5-daily products, consistently with their finer native resolution and lower effective noise. At the native satellite resolution, the downscaled products remain highly consistent with the original satellite fields, indicating that the downscaling process does not introduce relevant degradation of the large-scale gravimetric signal. The water balance equation analysis confirms that the downscaling generally preserves the basin-scale consistency of the input products, although the results depend on the temporal resolution and basin considered. In the simulated 5-daily case, the DS LR and DS HR products maintain WBE errors close to the corresponding NATIVE baselines in both basins. In addition, the improved observing scenarios produce lower errors that are retained after downscaling: over Congo, the DS HR RMSE decreases from 4.82 mm for GRACE-C-like to 3.32 mm for MAGIC, while over Danube it decreases from 14.47 to 7.10 mm. In the monthly case, the closure is well preserved over Danube and at the reaggregated satellite scale over Congo. However, the 1° Congo products show larger discrepancies with the high-resolution WBE reference, with DS HR RMSE values between 7.02 and 7.53 mm, compared with approximately 2–3 mm at the coarse resolution. For real GRACE(-FO), the aggregated downscaled products remain very close to the original satellite residuals, with WBE RMSE values of 4.35 mm over Congo and 3.03 mm over Danube, corresponding to NRMSE values of 0.014 and 0.007, respectively. Furthermore, the comparison between the downscaled GRACE-C-like simulations and real GRACE(-FO) observations over their common 2007–2020 period indicates broad consistency between simulations and real data in terms of both WBE closure and the spatial distribution of TWSA trends at 1°. Overall, the results indicate that unsupervised CNN-based downscaling can provide spatially enhanced TWSA fields that remain coherent with the native satellite observations. The approach is particularly promising for future gravity missions, since improved NGGM and MAGIC observations translate into more reliable downscaled products, while the same framework can also be transferred to existing GRACE(-FO) data. However, the analysis also highlights the intrinsic limitation of relying only on a coarse-scale consistency constraint: without independent high-resolution TWSA observations or additional physical constraints, the fine-scale redistribution within each coarse satellite cell is not uniquely determined. Therefore, the downscaled fields should be interpreted as data-driven reconstructions constrained by satellite gravimetry and auxiliary hydro-climatic predictors, rather than as direct observations of 1° TWSA. Future work will focus on the design of additional high-resolution terms in the training objective. In the current formulation, the loss function constrains the prediction only after aggregation to the native 2° or 3° satellite grid, while no explicit control is imposed on the solution at the 1° scale. Future developments will therefore investigate loss terms acting directly at the target resolution, for example through storage-evolution relationships in-

https://doi.org/10.5194/egusphere-2026-4563
Preprint. Discussion started: 4 August 2026

spired by land-surface and hydrological models. Such physics-informed constraints could help guide the network toward more realistic fine-scale TWSA distributions while preserving consistency with the original gravimetric observations. Finally, the downscaled products will be assessed in hydrologically relevant applications, such as the monitoring, detection and prediction of hydrological extremes. This will be an important step toward evaluating the practical added value of downscaled NGGM, MAGIC, and GRACE(-FO) TWSA products for satellite-gravimetry-based hydrological monitoring.

Code and data availability. The original simulated TWSA L3B products used in this study were generated within the ESA-funded SING project and are not yet publicly available, as they are currently being prepared for a dedicated data publication. The corresponding repository and DOI will be added to the manuscript as soon as they become available. For the simulated products, the released dataset includes only the outputs of the downscaling workflow, namely the 1° downscaled products and their 3° or 2° aggregated representations, while for GRACE(-FO) also the original L3 TWSA is included. It does not include the original SING simulated products used as input to the experiments. The downscaled TWSA dataset is available at [10.5281/zenodo.21669306]. The original public auxiliary datasets, including ERA5-Land, ETOPO2022, and ESA WorldCover, are available from the repositories cited in Sect. 3.3. The source code used for preprocessing, model training, and evaluation is not publicly archived but may be made available by the corresponding author upon reasonable request.

Appendix A: Structure-wise metrics formulation

This appendix reports the mathematical formulation of the structure-wise metrics used in the analysis, namely SSIM, GMSD, and FSS, together with their low-pass variants. For the low-pass structural diagnostics, the predicted and reference fields are first smoothed using a Gaussian low-pass filter with standard deviation $\sigma_{LP} = 3$ pixels. The filtering is implemented with the default SciPy (Virtanen et al., 2020) Gaussian kernel and nearest-neighbor boundary handling, so as to isolate the large-scale spatial component of the fields before computing LP-SSIM, LP-GMSD, and LP-FSS. For SSIM and FSS, multiple spatial scales are considered, including a global formulation for the SSIM, with the adopted physical and pixel-equivalent window sizes reported in Table A1. For SSIM and FSS, larger values indicate better structural agreement, whereas for GMSD lower values indicate better structural fidelity.

A1 Structural similarity index measure (SSIM)

The structural similarity index measure (SSIM) between a predicted field x and a reference field y is defined as

$$\mathrm{SSIM}(x,y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}, \tag{A1}$$

where μ_x and μ_y are the local means of the two fields over the selected window, σ_x^2 and σ_y^2 are the corresponding local variances, and σ_{xy} is the local covariance. For the global SSIM, these statistics are computed over all jointly valid grid cells of the predicted and reference maps.

https://doi.org/10.5194/egusphere-2026-4563
Preprint. Discussion started: 4 August 2026

The constants C_1 and C_2 are introduced to avoid numerical instability when the denominator approaches zero, and are defined as

$$C_1 = (K_1 L)^2, \qquad C_2 = (K_2 L)^2, \tag{A2}$$

where $K_1 = 0.01$ and $K_2 = 0.03$ are dimensionless constants, and L is the dynamic range of the compared fields. In this implementation, L is estimated robustly as

$$L = P_{99}(x \cup y) - P_1(x \cup y), \tag{A3}$$

where P_1 and P_{99} denote the 1st and 99th percentiles of the valid values from the two fields.

A2 Gradient magnitude similarity deviation (GMSD)

For the gradient magnitude similarity deviation (GMSD), let X denote the predicted field and Y the reference field. The horizontal and vertical gradients are first computed, and the gradient magnitude at each spatial location (i,j) is defined as

$$m_X(i,j) = \sqrt{g_{x,X}(i,j)^2 + g_{y,X}(i,j)^2}, \qquad m_Y(i,j) = \sqrt{g_{x,Y}(i,j)^2 + g_{y,Y}(i,j)^2}, \tag{A4}$$

where $g_{x,X}$ and $g_{y,X}$ are the horizontal and vertical gradients of X, and $g_{x,Y}$ and $g_{y,Y}$ are the corresponding gradients of Y. In this implementation, the gradients are estimated using Sobel operators (Sobel and Feldman, 1968).

A local gradient magnitude similarity map is then computed as

$$\mathrm{GMS}(i,j) = \frac{2 m_X(i,j)\, m_Y(i,j) + c}{m_X(i,j)^2 + m_Y(i,j)^2 + c}, \tag{A5}$$

where $c = 0.0026$ is a positive constant introduced for numerical stability.

The final GMSD score is defined as the standard deviation of the local GMS map over the valid spatial domain:

$$\mathrm{GMSD} = \sqrt{\frac{1}{N} \sum_{(i,j)\in\Omega} \left[\mathrm{GMS}(i,j) - \overline{\mathrm{GMS}}\right]^2}, \tag{A6}$$

with

$$\overline{\mathrm{GMS}} = \frac{1}{N} \sum_{(i,j)\in\Omega} \mathrm{GMS}(i,j), \tag{A7}$$

where Ω is the set of valid spatial locations and N is the number of valid pixels.

A3 Fractions skill score (FSS)

The Fractions Skill Score (FSS) is a neighborhood-based verification metric that measures the spatial agreement between the local fractions of an event in predicted and reference fields. In this study, it is applied to evaluate the spatial agreement of dry and wet TWSA events, defined through quantile-based thresholds of the reference distribution. The FSS can be interpreted as

https://doi.org/10.5194/egusphere-2026-4563
Preprint. Discussion started: 4 August 2026

a measure of the ability of the downscaled product to reproduce the approximate location and spatial extent of extreme TWSA conditions, while being less sensitive than pixel-wise metrics to small displacement errors.

For each region-frequency combination, two event thresholds are derived exclusively from the complete reference field y, considering all valid grid cells and timesteps:

$$\tau_{\text{dry}} = Q_{0.10}(y), \qquad \tau_{\text{wet}} = Q_{0.90}(y), \tag{A8}$$

where $Q_{0.10}$ and $Q_{0.90}$ denote the 10th and 90th percentiles, respectively. The thresholds are fixed for the corresponding regional experiment and are applied identically to the predicted and reference fields. They are not estimated independently from the prediction or recomputed at each timestep.

For an event type $e \in \{\text{dry}, \text{wet}\}$, the binary event indicators are defined as

$$I_x^{\text{dry}}(j,t) = \mathbf{1}[x(j,t) \le \tau_{\text{dry}}], \qquad I_y^{\text{dry}}(j,t) = \mathbf{1}[y(j,t) \le \tau_{\text{dry}}], \tag{A9}$$

$$I_x^{\text{wet}}(j,t) = \mathbf{1}[x(j,t) \ge \tau_{\text{wet}}], \qquad I_y^{\text{wet}}(j,t) = \mathbf{1}[y(j,t) \ge \tau_{\text{wet}}]. \tag{A10}$$

For a neighborhood $\mathcal{N}_i$ centered at grid cell i, the predicted and reference event fractions are

$$f_x^e(i,t) = \frac{\sum_{j\in\mathcal{N}_i} m_x(j,t)\, I_x^e(j,t)}{\sum_{j\in\mathcal{N}_i} m_x(j,t)}, \qquad f_y^e(i,t) = \frac{\sum_{j\in\mathcal{N}_i} m_y(j,t)\, I_y^e(j,t)}{\sum_{j\in\mathcal{N}_i} m_y(j,t)}, \tag{A11}$$

where m_x and m_y are validity masks equal to one for finite values and zero otherwise. Neighborhoods containing no valid values are excluded from the calculation.

At each timestep, the event-specific FSS is computed as

$$\text{FSS}^e(t) = 1 - \frac{\sum\limits_{i\in\Omega_t} \left[f_x^e(i,t) - f_y^e(i,t)\right]^2}{\sum\limits_{i\in\Omega_t} \left[f_x^e(i,t)^2 + f_y^e(i,t)^2\right]}, \tag{A12}$$

where Ω_t denotes the set of neighborhood centers for which both predicted and reference fractions are valid. If no event occurs in either field, such that the denominator is zero, the corresponding timestep-level FSS is treated as undefined and excluded from the temporal aggregation. FSS values corresponding to wet and dry events are evaluated separately at neighborhood scales of 3° and 5°, yielding 2 event types $\times$ 2 neighborhood scales $= 4$ standard FSS realizations and, analogously, 4 LP-FSS realizations. For the LP-FSS, the thresholds are derived from the complete low-pass reference field.

https://doi.org/10.5194/egusphere-2026-4563
Preprint. Discussion started: 4 August 2026

Table A1. Physical window sizes used for SSIM and FSS. Pixel kernels are obtained by converting each physical scale into the nearest valid odd kernel according to the grid spacing. Metrics evaluated over the entire grid do not require a local kernel.

Metric	Grid resolution	Physical scale	Effective kernel [px]
SSIM	$1°$	$3°$	3×3
SSIM	$1°$	$5°$	5×5
SSIM	$1°$	$7°$	7×7
SSIM	$1°$	$11°$	11×11
SSIM	$1°$	$15°$	15×15
SSIM	$1°$	Entire grid	
SSIM	$2°$	$7°$	3×3
SSIM	$2°$	$11°$	5×5
SSIM	$2°$	$15°$	7×7
SSIM	$2°$	Entire grid	
SSIM	$3°$	$7°$	3×3
SSIM	$3°$	$11°$	3×3
SSIM	$3°$	$15°$	5×5
SSIM	$3°$	Entire grid	
LP-SSIM	$1°$	Entire grid	
FSS	$1°$	$3°$	3×3
FSS	$1°$	$5°$	5×5
LP-FSS	$1°$	$3°$	3×3
LP-FSS	$1°$	$5°$	5×5

Appendix B: Composite score implementation

This appendix provides the detailed implementation of the composite score introduced in Sec. 4.1.

The purpose of the composite score is to combine multiple evaluation metrics into a single scalar indicator while preserving comparability across metric families, aggregation views, and experiments.

B1 Metric instances and families

Let $k = 1, \ldots, M$ denote the metric families included in the composite score. In the present implementation, these families are Pearson correlation, NSE, variance ratio, bias, nRMSE, nMdAE, SSIM, GMSD, and FSS. Each family groups one or more metric realizations sharing the same diagnostic meaning, even when they differ in spatial scale, low-pass formulation, or aggregation view. The assignment of the individual metrics to each family is reported in Table B1.

https://doi.org/10.5194/egusphere-2026-4563
Preprint. Discussion started: 4 August 2026

Table B1. Assignment of the individual metrics to the composite score families. Metrics belonging to the same family are first normalized separately and then averaged within family.

Metric	Family
r	Pearson correlation
NSE	NSE
β variance ratio	Variance ratio
nRMSE	nRMSE
nMdAE	nMdAE
Bias	Bias
SSIM (global)	SSIM
SSIM (multiscale, all window sizes)	SSIM
LP-SSIM	SSIM
GMSD	GMSD
LP-GMSD	GMSD
FSS (multiscale, all window sizes)	FSS
LP-FSS (multiscale, all window sizes)	FSS

To keep the notation general, let i denote an individual metric instance. Each instance i is associated with:

- a metric family $k(i)$,
- an aggregation view $v(i) \in \{\mathrm{TS}, \mathrm{ST}\}$,
- and, when applicable, a specific spatial scale or metric variant (e.g. window size or low-pass version).

The original value of metric instance i is denoted by S_i.

B2 Orientation of the metrics

Since the metrics entering the composite score do not share the same optimization direction, each metric is first converted into an oriented value S_i^{ori} such that higher values always correspond to better performance.

For metrics that are naturally maximized, the oriented value coincides with the original one:

$$S_i^{\mathrm{ori}} = S_i. \tag{B1}$$

This applies to positively oriented metrics such as correlation r, NSE, SSIM, and FSS.

For metrics that are naturally minimized, the orientation is reversed:

$$S_i^{\mathrm{ori}} = -S_i. \tag{B2}$$

https://doi.org/10.5194/egusphere-2026-4563
Preprint. Discussion started: 4 August 2026

This applies to metrics such as nRMSE, nMdAE, and GMSD.

For metrics whose optimum is attained at a target value rather than at an extremum, a target-centered transformation is used. In particular, for the variance ratio β, whose optimal value is $\beta = 1$,

$$S_{\beta}^{\mathrm{ori}} = -|\beta - 1|. \tag{B3}$$

and for the bias, whose optimum is $\mathrm{bias} = 0$,

$$S_{\mathrm{bias}}^{\mathrm{ori}} = -|\mathrm{bias}|. \tag{B4}$$

With these transformations, the best attainable value within each metric type corresponds to the largest oriented score. For example, $\beta = 1$ maps to an oriented value of 0, while departures from the optimum produce increasingly negative values.

B3 Quantile-based normalization

After orientation, all metric instances are mapped to a common $[0,1]$ range through a quantile-based min–max normalization. Here, m identifies an individual metric realization, including its aggregation view, variant, and spatial scale. For each metric m, let $q_{m,\mathrm{low}}$ and $q_{m,\mathrm{high}}$ denote the lower and upper empirical quantiles of its oriented-value distribution. In the present implementation, these bounds correspond to the 5th and 95th percentiles:

$$q_{m,\mathrm{low}} = Q_{0.05}\left(S_m^{\mathrm{ori}}\right), \qquad q_{m,\mathrm{high}} = Q_{0.95}\left(S_m^{\mathrm{ori}}\right), \tag{B5}$$

where S_m^{ori} denotes the collection of all available oriented values for metric realization m across the configurations included in its normalization pool. The pool includes all available oriented values of the metric realization across regions, observing configurations, and temporal and spatial resolutions. Separate normalization bounds are computed only for distinct metric realizations, including their aggregation view, spatial scale (for SSIM and FSS), and standard or low-pass formulation. This choice preserves the comparability of composite scores across all evaluation settings, including comparisons among observing configurations, between 5-daily and monthly products, and between high-resolution evaluations against the HISBED reference and low-resolution evaluations against the native satellite products. If the normalization bounds were defined separately for each comparison setting—for example, using distinct pools for the 5-daily and monthly products—the resulting scores would express only relative performance within that setting and would therefore not be directly comparable. The same applies to comparisons among the different observing configurations, such as GRACE-C-like and MAGIC.

The normalized score of instance i is then computed as

$$\widetilde{S}_i = \mathrm{clip}\left(\frac{S_i^{\mathrm{ori}} - q_{m,\mathrm{low}}}{q_{m,\mathrm{high}} - q_{m,\mathrm{low}}}, 0, 1\right), \qquad m = m(i), \tag{B6}$$

where $\mathrm{clip}(x, 0, 1)$ truncates values smaller than 0 to 0 and values larger than 1 to 1.

This transformation ensures that extreme outliers do not dominate the scaling, while keeping all normalized metrics on a comparable interval. By construction, values close to 1 indicate relatively strong performance within the corresponding metric realization, whereas values close to 0 indicate comparatively weak performance.

https://doi.org/10.5194/egusphere-2026-4563
Preprint. Discussion started: 4 August 2026

If the normalization interval becomes undefined or collapses, i.e.

$$q_{m,\mathrm{high}} - q_{m,\mathrm{low}} = 0 \tag{B7}$$

or cannot be evaluated because of insufficient valid data, the implementation assigns a neutral fallback score:

$$\widetilde{S}_i = 0.5. \tag{B8}$$

This prevents degenerate metric realizations from artificially favoring or penalizing a given experiment.

B4 Aggregation within a metric family

Some metric families are represented by multiple realizations. This occurs, for instance, when SSIM or FSS are computed at multiple spatial scales, or when the same family is available under both TS and ST aggregation views. For instance, the SSIM family includes the global SSIM, the windowed SSIM realizations at the selected spatial scales, and the low-pass SSIM diagnostic; similarly, the FSS and GMSD families include both their standard and low-pass variants.

For a given family k and aggregation view $v \in \{\mathrm{TS}, \mathrm{ST}\}$, let $J_{k,v}$ be the number of available normalized realizations in that family-view combination. The family-view score is defined as

$$F_{k,v} = \frac{1}{J_{k,v}} \sum_{j=1}^{J_{k,v}} \widetilde{S}_{k,j,v}, \tag{B9}$$

where $\widetilde{S}_{k,j,v}$ denotes the normalized score of the j-th realization of family k under aggregation view v.

This means that all realizations belonging to the same metric family and aggregation view contribute equally. For example, if SSIM is evaluated at five window sizes, each window contributes one fifth of the corresponding SSIM family-view score.

If both aggregation views are available for family k, they are combined with equal importance:

$$F_k = \frac{F_{k,\mathrm{TS}} + F_{k,\mathrm{ST}}}{2}. \tag{B10}$$

If only one aggregation view is available, the corresponding value is used directly:

$$F_k = F_{k,\mathrm{TS}} \qquad \text{or} \qquad F_k = F_{k,\mathrm{ST}}. \tag{B11}$$

Thus, F_k denotes the final normalized score of family k, obtained after averaging across all its valid realizations and, when applicable, across both aggregation views.

B5 Renormalization of the family weights

Not all metric families are available for every experiment, comparison setup, or aggregation mode. To avoid penalizing experiments with partially missing families, the user-defined raw weights w_k are renormalized over the subset of available families.

Let $\mathcal{A}$ denote the set of metric families available in a given experiment. The effective weight assigned to family $k \in \mathcal{A}$ is then

$$\widehat{w}_k = \frac{w_k}{\sum_{\ell \in \mathcal{A}} w_\ell}. \tag{B12}$$

https://doi.org/10.5194/egusphere-2026-4563
Preprint. Discussion started: 4 August 2026

By construction,

$$\sum_{k\in\mathcal{A}} \widehat{w}_k = 1, \qquad \widehat{w}_k \geq 0. \tag{B13}$$

580 Therefore, the contribution of each family remains proportional to its intended importance, while the total weight is redistributed only across the families that are effectively defined in the considered case. For example, not all metric families are available in every comparison setting. FSS, GMSD, their low-pass variants, and the LP-SSIM realization are only defined for the high-resolution evaluation against the reference field. Consequently, in the low-resolution comparisons against the native satellite products, the corresponding metric realizations are absent. The raw weights are therefore redistributed through renor-
585 malization over the subset of metric families that are actually available, as reported in Table B2. Let $\mathcal{A}_{\mathrm{LR}}$ denote the subset of metric families available in the low-resolution evaluation. Their raw weights sum to

$$\sum_{k\in\mathcal{A}_{\mathrm{LR}}} w_k = 1 - w_{\mathrm{GMSD}} - w_{\mathrm{FSS}} = 1 - 0.11 - 0.11 = 0.78. \tag{B14}$$

Accordingly, for each $k \in \mathcal{A}_{\mathrm{LR}}$, the effective weight is

$$\widehat{w}_k = \frac{w_k}{0.78}. \tag{B15}$$

Table B2. Raw and effective metric-family weights used for the low-resolution DS_{LR} evaluations defined as in Sect. 4.1. The effective weights are obtained by renormalizing the raw weights over the available metric families. GMSD and FSS are not available in these comparison scenarios and therefore do not receive an effective weight.

Metric family	Raw weight w_k	Effective weight $\widehat{w}_k$
Pearson correlation r	0.18	0.230769
β variance ratio	0.17	0.217949
NSE	0.15	0.192308
SSIM	0.11	0.141026
GMSD	0.11	–
FSS	0.11	–
nMdAE	0.07	0.089744
nRMSE	0.06	0.076923
Bias	0.04	0.051282
Total	1.00	1.000000

590 ## B6 Final composite score

The final composite score is computed as the weighted sum of the family scores:

$$\mathrm{CS} = \sum_{k\in\mathcal{A}} \widehat{w}_k F_k. \tag{B16}$$

https://doi.org/10.5194/egusphere-2026-4563
Preprint. Discussion started: 4 August 2026

By construction, this formulation ensures that:

1. all metrics are converted to a common $[0, 1]$ scale;
2. all metrics are positively oriented, so that larger values always indicate better performance;
3. multiple realizations within the same metric family contribute equally to the corresponding family score;
4. TS and ST aggregations contribute equally whenever both are available;
5. missing metric families do not induce artificial penalties, since the weights are renormalized only over the available set $\mathcal{A}$.

The resulting score provides a compact ranking index while preserving traceability to the individual metric families listed in Table B1.

Author contributions. GG conceptualized the study, developed the methodology and software, performed the analyses, produced the visualizations, and prepared the original manuscript. ID contributed to the scientific interpretation of the satellite gravimetry products, the downscaled outputs, and the study results, and contributed to the structure of the manuscript. NL supervised the study, contributed to the methodology, and reviewed the manuscript. All authors discussed the results and approved the final manuscript

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. OpenAI Codex was used to assist with the development and refinement of Python scripts for data visualization, table generation, and the graphical representation of the results. ChatGPT was used to assist with language editing and the improvement of clarity and readability in selected passages of the manuscript. The underlying data, numerical results, scientific interpretations, and conclusions were independently produced by the authors. All AI-assisted code and text were reviewed, tested, and revised by the authors, who take full responsibility for the final manuscript.

https://doi.org/10.5194/egusphere-2026-4563
Preprint. Discussion started: 4 August 2026

References

Ahmed, M., Sultan, M., Becker, R., Yan, J., and Milewski, S.: Improving the Resolution of GRACE Data for Spatio-Temporal Groundwater Storage Assessment, Remote Sensing, 11, 2103, https://doi.org/10.3390/rs11182103, 2019.

Ali, S., Khan, M. S., Usman, M., Tufail, M., and Abid, M. A.: Machine learning downscaling of GRACE-GRACE-FO data to capture spatial-temporal drought effects on groundwater storage at a local scale under data-scarcity, Journal of Hydrology, 617, 129 039, https://doi.org/10.1016/j.jhydrol.2023.129039, 2023.

Boergens, E., Dobslaw, H., and Dill, R.: COST-G GravIS RL02 Terrestrial Water Storage Anomalies, Version 0001, https://doi.org/10.5880/COST-G.GRAVIS_02_L3_TWS, 2025.

Chen, J., Wang, L., Chen, C., and Peng, Z.: Downscaling and Gap-Filling GRACE-Based Terrestrial Water Storage Anomalies in the Qinghai–Tibet Plateau Using Deep Learning and Multi-Source Data, Remote Sensing, 17, 1333, https://doi.org/10.3390/rs17081333, 2025.

Cohen, J.: Statistical Power Analysis for the Behavioral Sciences, Lawrence Erlbaum Associates, Hillsdale, New Jersey, 2 edn., https://doi.org/10.4324/9780203771587, 1988.

Copernicus Climate Change Service: ERA5-Land hourly data from 1950 to present, https://doi.org/10.24381/cds.e2161bac, 2019.

Dahle, C., Boergens, E., Sasgen, I., Döhne, T., Reißland, S., Dobslaw, H., Klemann, V., Murböck, M., König, R., Dill, R., Sips, M., Sylla, U., Groh, A., Horwath, M., and Flechtner, F.: GravIS: mass anomaly products from satellite gravimetry, Earth System Science Data, 17, 611–631, https://doi.org/10.5194/essd-17-611-2025, 2025.

Daras, I., March, G., Joint Mass Change Mission Expert Group, Wiese, D., Blackwood, C., Forman, F., Loomis, B., Luthcke, S., Peralta-Ferriz, C., Sauber-Rosenberg, J., Eicker, A., Panet, I., Visser, P., Meyssignac, B., Wouters, B., Bamber, J., Pail, R., and Flechtner, F.: Next Generation Gravity Mission (NGGM) Mission Requirements Document, Mission Requirements Document Issue 1.0, European Space Agency, https://doi.org/10.5270/ESA.NGGM-MRD.2023-09-v1.0, 2023.

Daras, I., March, G., Pail, R., Hughes, C. W., Braitenberg, C., Güntner, A., Eicker, A., Wouters, B., Heller-Kaikov, B., Pivetta, T., and Pastorutti, A.: Mass-change and Geosciences International Constellation (MAGIC) Expected Impact on Science and Applications, Geophysical Journal International, 236, 1288–1308, https://doi.org/10.1093/gji/ggad472, 2024.

Dobslaw, H., Bergmann-Wolf, I., Dill, R., Forootan, E., Klemann, V., Kusche, J., and Sasgen, I.: The updated ESA Earth System Model for future gravity mission simulation studies, Journal of Geodesy, 89, 505–513, https://doi.org/10.1007/s00190-014-0787-8, 2015.

European Space Agency (ESA): ESA SING Project: Data Generation and End-to-End Simulations, https://www.esa-sing.org/data-generation, accessed: 2025-11-07, 2025a.

European Space Agency (ESA): ESA SING Project: Studying the Impact of NGGM and MAGIC Missions, https://www.esa-sing.org, accessed: 2025-11-07, 2025b.

Fatolazadeh, F., Eshagh, M., Goïta, K., and Wang, S.: A New Spatiotemporal Estimator to Downscale GRACE Gravity Models for Terrestrial and Groundwater Storage Variations Estimation, Remote Sensing, 14, 5991, https://doi.org/10.3390/rs14235991, 2022.

Flechtner, F., Wiese, D., Webb, F., Landerer, F., Gross, M., Snopek, K., Fischer, S., and Dahle, C.: Global Gravity and Mass Change Observations beyond GRACE-FO: Updates on the Upcoming GRACE-Continuity Mission, in: GRACE/GRACE-FO Science Team Meeting, Potsdam, Germany, https://doi.org/10.5194/gstm2024-21, 8–10 October 2024, abstract GSTM2024-21, 2024.

GFZ German Research Centre for Geosciences: ESA Earth System Model for Gravity Mission Simulation Studies, Information System and Data Center (ISDC), https://isdc-data.gfz.de/future_missions/esaesm/, data repository, accessed 24 July 2026, 2026.

Gou, J. and Soja, B.: Global high-resolution total water storage anomalies from self-supervised data assimilation using deep learning algorithms, Nature Water, 2, 139–150, https://doi.org/10.1038/s44221-024-00194-w, 2024.

Hartmann, D. L., Klein Tank, A. M. G., Rusticucci, M., Alexander, L. V., Brönnimann, S., Charabi, Y., Dentener, F. J., Dlugokencky, E. J., Easterling, D. R., Kaplan, A., Soden, B. J., Thorne, P. W., Wild, M., and Zhai, P. M.: Observations: Atmosphere and Surface, in: Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change, edited by Stocker, T. F., Qin, D., Plattner, G.-K., Tignor, M., Allen, S. K., Boschung, J., Nauels, A., Xia, Y., Bex, V., and Midgley, P. M., Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, supplementary Material, Section 2.SM.3.2, Equations 2.SM.1–2.SM.3, 2013.

International Union of Geodesy and Geophysics: Resolution 2: Sustained Terrestrial Water Storage (TWS) Monitoring by Dedicated Gravity Satellite Constellations, Resolution adopted by the International Union of Geodesy and Geophysics, https://iugg.org/wp-content/uploads/2023/09/2023_IUGG-GA-Resolutions.pdf, collated Resolutions, final version dated 18 July 2023, 2023.

Jensen, L., Dill, R., Balidakis, K., Grimaldi, S., Salamon, P., and Dobslaw, H.: Global 0.05° water storage simulations with the OS LISFLOOD hydrological model for geodetic applications, Geophysical Journal International, 241, 1840–1852, https://doi.org/10.1093/gji/ggaf129, 2025.

Knijff, J. M. V. D., Younis, J., and Roo, A. P. J. D.: LISFLOOD: a GIS-based distributed model for river basin scale water balance and flood simulation, International Journal of Geographical Information Science, 24, 189–212, https://doi.org/10.1080/13658810802549154, 2010.

Kusche, J., Strohmenger, C., Gerdener, H., Uebbing, B., Springer, A., Ewerdwalbesloh, Y., Eicker, A., Braitenberg, C., Pastorutti, A., Pail, R., Zingerle, P., Schlaak, M., Reguzzoni, M., Rossi, L., Migliaccio, F., and Daras, I.: Benefit of MAGIC and multipair quantum satellite gravity missions in Earth science applications, Geophysical Journal International, 242, ggaf195, https://doi.org/10.1093/gji/ggaf195, 2025.

Lehner, B. and Grill, G.: Global river hydrography and network routing: baseline data and new approaches to study the world's large river systems, Hydrological Processes, 27, 2171–2186, https://doi.org/10.1002/hyp.9740, 2013.

MacFerrin, M., Amante, C., Carignan, K., Love, M., and Lim, E.: The Earth Topography 2022 (ETOPO 2022) global DEM dataset, Earth System Science Data, 17, 1835–1849, https://doi.org/10.5194/essd-17-1835-2025, 2025.

Muñoz-Sabater, J., Dutra, E., Agustí-Panareda, A., Albergel, C., Arduini, G., Balsamo, G., Boussetta, S., Choulga, M., Harrigan, S., Hersbach, H., Martens, B., Miralles, D. G., Piles, M., Rodríguez-Fernández, N. J., Zsoter, E., Buontempo, C., and Thépaut, J.-N.: ERA5-Land: a state-of-the-art global reanalysis dataset for the land surface, Earth System Science Data, 13, 4349–4383, https://doi.org/10.5194/essd-13-4349-2021, 2021.

NOAA National Centers for Environmental Information: ETOPO 2022 15 Arc-Second Global Relief Model, https://doi.org/10.25921/fd45-gt74, 2022.

Pail, R., ed.: Observing Mass Transport to Understand Global Change and to Benefit Society: Science and User Needs—An International Multi-Disciplinary Initiative for IUGG, no. 320 in Deutsche Geodätische Kommission, Reihe B: Angewandte Geodäsie, Verlag der Bayerischen Akademie der Wissenschaften, Munich, Germany, ISBN 978-3-7696-8599-2, https://dgk.badw.de/fileadmin/user_upload/Files/DGK/docs/b-320.pdf, 2015.

Pail, R., Landerer, F. W., Eicker, A., Reager, J. T., Braitenberg, C., Sauber-Rosenberg, J., Brocca, L., Rodell, M., Panet, I., Han, S.-C., Bruinsma, S., and Peralta-Ferriz, C.: Report of MAGIC Science and Applications Workshop 2023, Tech. rep., European Space Agency and National Aeronautics and Space Administration, https://doi.org/10.5270/MAGIC-Workshop-2023, 2024.

Pfeffer, J. and the SING consortium: Studying the Impact of the NGGM and MAGIC Future Satellite Gravity Missions for Scientific Applications and Operational Services, in: EGU General Assembly 2026, Vienna, Austria, https://doi.org/10.5194/egusphere-egu26-18908, 2026.

Ram, A. P.: Unsupervised Representation Learning of GRACE Improves Groundwater Predictions, Water, 14, 2947, https://doi.org/10.3390/w14192947, 2022.

Sahour, H.: Statistical Downscaling Techniques to Enhance the Spatial Resolution of the GRACE Satellite Data and to Fill Temporal Gaps, Ph.D. thesis, Western Michigan University, https://scholarworks.wmich.edu/dissertations/3634, doctoral dissertation, Western Michigan University, 2020.

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 Sensing, 11, 824, https://doi.org/10.3390/rs11070824, 2019.

Shihora, L., Schlaak, M., Klemann, V., Jensen, L., Dill, R., Tanaka, Y., Sasgen, I., Wouters, B., Han, S.-C., Sauber-Rosenberg, J., Braitenberg, C., Javed, M. T., Maurizio, G., Lecomte, H., and Dobslaw, H.: The Updated ESA Earth System Model for Future Gravity Mission Simulation Studies: ESA ESM 3.0, in: EGU General Assembly 2026, Vienna, Austria, https://doi.org/10.5194/egusphere-egu26-9665, abstract EGU26-9665, presented 3–8 May 2026, 2026.

Sobel, I. and Feldman, G.: A 3x3 Isotropic Gradient Operator for Image Processing, in: Stanford Artificial Intelligence Project, pp. 271–272, 1968.

Tao, H., Al-Sulttani, A. H., Salih, S. Q., Mohammed, M. K., Khan, M. A., Beyaztas, B. H., Ali, M., Elsayed, S., Shahid, S., and Yaseen, Z. M.: Development of high-resolution gridded data for water availability identification through GRACE data downscaling: Development of machine learning models, Atmospheric Research, 291, 106 815, https://doi.org/10.1016/j.atmosres.2023.106815, 2023.

Tourian, M. J., Saemian, P., Ferreira, V. G., Sneeuw, N., Frappart, F., and Papa, F.: A copula-supported Bayesian framework for spatial downscaling of GRACE-derived terrestrial water storage flux, Remote Sensing of Environment, 295, 113 685, https://doi.org/10.1016/j.rse.2023.113685, 2023.

Uz, M., Akyilmaz, O., and Shum, C.: Deep learning-aided temporal downscaling of GRACE-derived terrestrial water storage anomalies across the Contiguous United States, Journal of Hydrology, 645, 132 194, https://doi.org/10.1016/j.jhydrol.2024.132194, 2024.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat, İ., Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P., and SciPy 1.0 Contributors: SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python, Nature Methods, 17, 261–272, https://doi.org/10.1038/s41592-019-0686-2, 2020.

Vishwakarma, B. D., Zhang, J., and Sneeuw, N.: Downscaling GRACE total water storage change using partial least squares regression, Scientific Data, 8, 95, https://doi.org/10.1038/s41597-021-00862-6, 2021.

Wang, J., Shen, Y., Awange, J., Tabatabaeiasl, M., Song, Y., and Liu, C.: A novel generative adversarial network and downscaling scheme for GRACE/GRACE-FO products: Exemplified by the Yangtze and Nile River Basins, Science of the Total Environment, 969, 178 874, https://doi.org/10.1016/j.scitotenv.2025.178874, 2025.

Welch, B. L.: The Generalization of Student's Problem when Several Different Population Variances Are Involved, Biometrika, 34, 28–35, https://doi.org/10.1093/biomet/34.1-2.28, 1947.

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, Journal of Geophysical Research: Atmospheres, 123, 5973–5987, https://doi.org/10.1029/2017JD027468, 2018.

https://doi.org/10.5194/egusphere-2026-4563
Preprint. Discussion started: 4 August 2026

Zanaga, D., Van De Kerchove, R., Daems, D., De Keersmaecker, W., Brockmann, C., Kirches, G., Wevers, J., Cartus, O., Santoro, M., Fritz, S., Lesiv, M., Herold, M., Tsendbazar, N. E., Xu, P., Ramoino, F., and Arino, O.: ESA WorldCover 10 m 2021 v200, https://doi.org/10.5281/zenodo.7254221, 2022.

Zhong, D., Wang, S., and Li, J.: Spatiotemporal Downscaling of GRACE Total Water Storage Using Land Surface Model Outputs, Remote Sensing, 13, 900, https://doi.org/10.3390/rs13050900, 2021.