Beyond GRACE: Evaluating the benefits of NGGM and MAGIC for precipitation estimation over Europe.

Author(s): Muhammad Usman Liaqat et al.

MS No.: egusphere-2025-3659

Reviewer 1

General Response

We would like to thank the reviewer for the careful evaluation of our manuscript and for the productive comments, which have helped us to improve the quality of the study. A detailed response to each comment is provided below. For clarity, the reviewers' comments (RC1) are reported in light blue, while the author's responses (**ACs)** are presented in black. The revised text in the track change version as well as below is highlighted in red to make it clear how the changes were made to address the reviewers' comments.

RC1: The use of ERA5-Land both to construct TWS (input) and as precipitation reference risks circularity. Validate against independent gauges/radar (ECA&D, national networks) or multi-product references (GPCC, MSWEP, IMERG) over representative regions/seasons.

ACs: We appreciate the reviewer's insightful comment about the possible circularity in using ERA5L both to (i) construct TWS used as input to SM2RAIN and (ii) serve as reference precipitation. We completely agree that, for an operational validation, independent precipitation datasets such as ECA&D, GPCC, MSWEP, or IMERG would be necessary to avoid relying on a single data source.

However, this study is not intended as an operational validation of a new precipitation product. Instead, it is designed as a controlled, synthetic experiment to evaluate how different future gravity mission configurations (NGGM and MAGIC) affect the sensitivity and performance of precipitation estimation. In this context, ERA5L serves as a self-consistent "proxy truth," providing physically coherent TWS and precipitation data. This setup allows us to:

- Focus on how temporal resolution and measurement errors influence results, without complications from inconsistencies between independent datasets.
- Ensure physical consistency between TWS variations and corresponding precipitation, which is crucial when assessing SM2RAIN's ability to infer rainfall from TWS for the first time.
- Create a synthetic environment in which differences between configurations (e.g., 5-day vs. 30-day temporal resolution, ± error levels) can be compared fairly.

Consequently, our results should be interpreted as relative assessments of sensitivity rather than absolute validation of metrics. The main objective is to understand the expected gains in precipitation estimation if future missions like NGGM or MAGIC meet their target accuracy and sampling goals, not validate an existing precipitation dataset. Moreover, we revised Section 2.4 and added a new subsection, "Limitations and Future Work," in the revised manuscript as well below, which explains the purpose of the synthetic experiments and outlines our plans for future validation using independent observational datasets (e.g., ECA&D, GPCC, MSWEP, IMERG).

"Limitations and Future Work"

This study conducts a controlled sensitivity experiment utilizing ERA5L as a self-consistent synthetic reference framework; hence, the results reflect relative performance assessment rather than independent validation against in situ observations. Importantly, these experiments provide a first, controlled synthetic assessment of the feasibility of retrieving precipitation from TWS. The added value of our study is not to validate the SM2RAIN algorithm (which has been done extensively using multiple observational precipitation datasets such as e.g., GPCC or E-OBS) but rather to assess its sensitivity to the sampling and accuracy expected from future gravity missions.

Measurement uncertainties are modeled as additive Gaussian noise with standard deviations derived from NGGM mission error specifications (Daras et al.,2023). This notice acknowledges that real satellite gravimetry errors may be temporally and spatially correlated and may be geographically dependent, with factors including latitude, topography and orbit geometry. While the Gaussian noise assumption provides a simplistic error model, it is sufficient in providing a first-order sensitivity of how measurement uncertainty may affect precipitation retrieval. In this work, our aim is not to replicate the full complexity of mission error characteristics but rather to study the relative effect of spatiotemporal sampling and measurement uncertainty on precipitation estimation from TWS. However, this comparison would demand mission specific error covariance models, the derivation of which is beyond the scope of this study.

Additionally, the SM2RAIN framework does not show evapotranspiration as independent fluxes, which could affect how storage dynamics change over weeks. Neglecting evaporation can lead to biased precipitation estimates, particularly in warm, vegetated, and semiarid regions where evapotranspiration is significant. Further, TWS changes generally include contributions from groundwater, but the current framework does not make a clear distinction between groundwater storage and other hydrological components, limiting process-level interpretation.

Future work will address these issues by integrating flux-consistent water balance constraints (incorporating evapotranspiration and runoff), parameter estimation using machine learning, and improving the attribution of groundwater dynamics with future gravity missions. This preserves the physical SM2RAIN structure while reducing calibration dependency, improving spatial coherence, and enhancing robustness in data scarce and nonstationary environments. The framework will also be extended to larger domains using more realistic, spatially, and temporally correlated error structures derived from advanced mission simulations.

RC1: The intro motivates higher cadence gravity but doesn't cleanly state why replacing SM with TWS is the key scientific gap. Explicitly articulate why it is expected to improve SM2RAIN (e.g., deeper storage sensitivity).

ACs: We thank the reviewer for the comment, and we revised the introduction to clarify why replacing surface soil moisture with terrestrial water storage represents the central scientific gap addressed in this study. The explanation is also provided below as well as in the main text from lines 69-79.

Gravimetry, on the other hand, observes TWS integrating water variations over the full soil column and shallow subsurface. This makes it less sensitive to surface saturation and vegetation impacts; however, the detection of precipitation signals in gravimetric observations is directly constrained by measurement uncertainty. As precipitation represents the primary source of TWS, its variations inherently reflect cumulative water inputs and losses at larger spatial scales (Zhong et al., 2025). These characteristics make TWS a suitable choice, potentially better than SM, for estimating accumulated precipitation at sub monthly scales. However, this potential has remained largely unexplored because TWS estimates derived from previous gravity missions, including the Gravity Recovery and Climate Experiment (GRACE, Tapley et al., 2004), Gravity Field and steady-state Ocean Circulation Explorer (GOCE, Drinkwater et al., 2003), and GRACE Follow-On (GRACE-FO, Landerer et al., 2020), are characterized by coarse spatial (~400 km) and, importantly, temporal (monthly) resolutions, which limit their ability to accurately capture rainfall variability, except in regions experiencing intense precipitation, such as the tropics. The next NGGM and MAGIC missions are projected to improve accuracy and temporal resolution, which makes this expansion of SM2RAIN possible for the first time. This clarification strengthens the scientific motivation of the study and better positions it within the context of future gravity mission applications.

RC1: Methods resample inputs to 100 km using bilinear interpolation while the resolution of GRACE km and NGGM/MAGIC are lower. Conduct a sensitivity analysis to motivate your selection.

ACs: We thank the reviewer for raising this important point regarding the choice of spatial resolution and the potential mismatch with the native resolution of GRACE and future NGGM/MAGIC gravity missions. Our objective is not to imply that gravity missions resolve hydrological processes at 100 km spatial scales. Rather, the 100 km grid is adopted as a common analysis grid that facilitates the application of the SM2RAIN framework and enables a consistent comparison across synthetic mission configurations while preserving the large-scale information content relevant for gravimetric observations. To support this choice, we conducted a set of scale-consistency and sensitivity analyses, which are now provided below in Fig. 1. The results in Fig. S1 show the spatial distribution of correlation coefficients (R) between ERA5-Land-derived TWS and GRACE TWS over Europe. The left panel shows correlations computed over the extended ERA5-Land period (2002–2022), while the right panel is restricted to the common GRACE observation period (2002–2015). Median correlation values increase from 0.66 to 0.79 when restricting the analysis to the overlapping period, indicating strong consistency between ERA5-Land and GRACE at large spatial scales. Higher correlations are observed across central and southern Europe, confirming that ERA5-Land captures gravity-relevant TWS variability when viewed at appropriate scales.

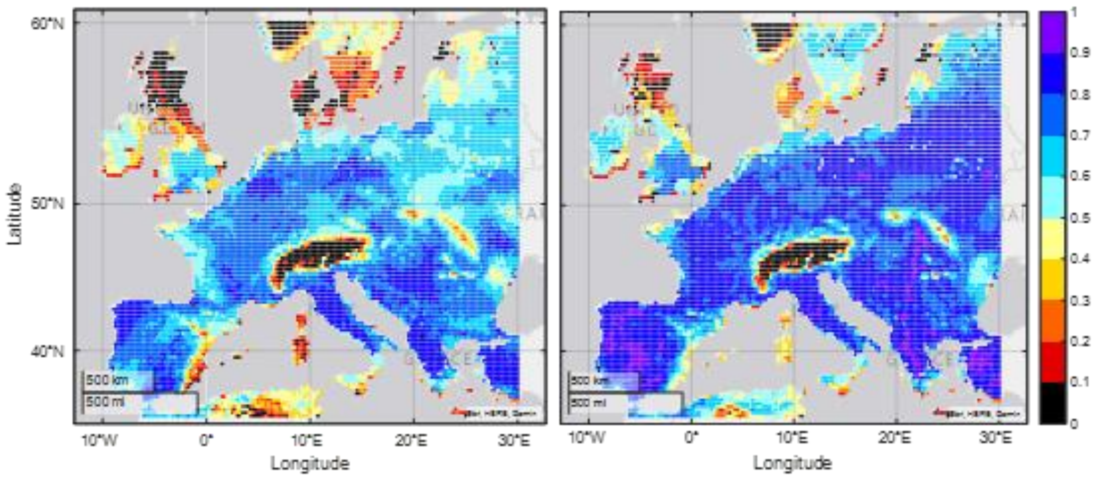

Figure 1. Spatial correlation between ERA5-Land TWS and GRACE TWS. Left panel shows correlations computed over the extended ERA5-Land period (2002–2022), while the right panel is restricted to the common GRACE observation period (2002–2015).

Additionally, Fig. 2 shows the comparison of TWS variability derived from ERA5-Land, ESM, and GRACE over Europe, illustrating the consistency of large scale signals across datasets with different native spatial resolutions. Panels include (i) spatial correlations between ERA5-Land TWS and GRACE TWS over different analysis periods and (ii) correlation maps between model-derived TWS (ESM and ERA5-Land) and GRACE solutions from two independent processing centers (CSR and JPL). Despite differences in native resolution and processing approaches, ERA5-Land and ESM exhibit coherent large-scale TWS patterns that closely match GRACE observations, with spatially consistent correlations approximately (R = 0.8) across large parts of Europe and across both GRACE solutions. Fine-scale variability present in model-based products is largely smoothed in GRACE, confirming that gravity-observable TWS signals are dominated by large-scale spatial variability.

These results demonstrate that resampling a common analysis grid preserves gravity relevant information and does not introduce artificial small-scale signals, supporting the methodological choice adopted in the synthetic experiments. Hence, these analyses demonstrate that the dominant terrestrial water storage (TWS) signals relevant for gravity missions are governed by large scale spatial variability, rather than by small scale features introduced through interpolation. We will clarify this rationale in the methods section of the revised manuscript and will add the corresponding analyses to the supplementary material. The main conclusions of the study are robust with respect to the chosen spatial resolution and should be interpreted as relative sensitivity assessments of future gravity mission configurations, rather than as claims about the native resolving power of NGGM or MAGIC. The above text and corresponding results are added to the supplementary material.

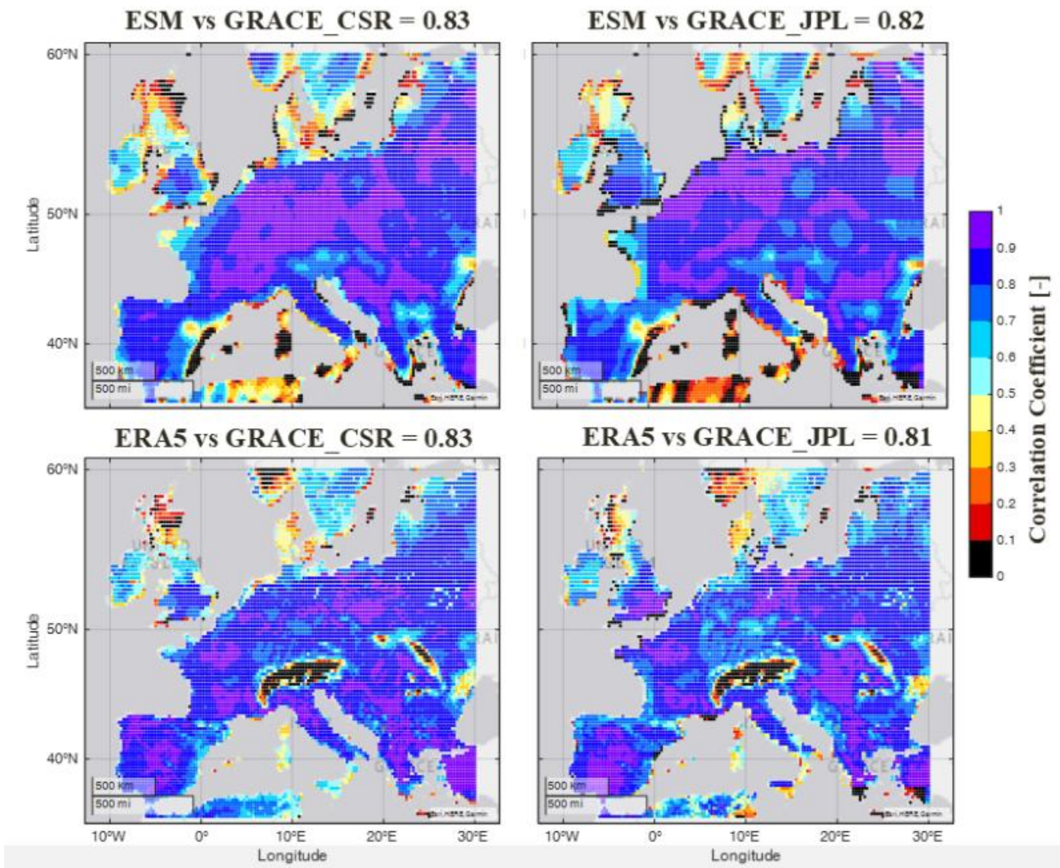

Figure 2. Scale consistency between model-derived and gravimetric terrestrial water storage over Europe.

RC1: Nash-Sutcliffe is mentioned but never reported. It is suggested to include it.

Authors: In this study, the Nash–Sutcliffe efficiency (NS) index is used as the objective function during the calibration of the SM2RAIN model parameters, while bias, R, and RMSE are used for performance evaluation. This is in line with, e.g., Knoben et al., 2019; Williams, 2025), who suggest using skill metrics only for calibration. The correction has been incorporated by ensuring consistency in the description of performance metrics throughout the revised manuscript text.

RC1: Specify calibration design (parameter bounds, train/validation split, spatial cross-validation).

ACs: Calibration is performed independently at each grid cell (point-by-point) using a gradient-based optimization function. The procedure estimates five parameters: Z*, a, b, T_{pot}, and T_{base}. The estimation of parameters a, b, and Z* is needed to compute the precipitation between two consecutive TWS* measurements. The filter parameters (T_pot and T_base) have narrow interquartile ranges, indicating consistent noise reduction across grid cells. Before running the algorithm, the noise in the TWS signal is reduced by applying the exponential filter as described by Brocca et al. (2019), where the characteristic time length parameter (T) is modeled as a function of soil moisture using a two-parameter power law relationship. The values of the two parameters (T_pot and T_base) of the modified exponential filter equation, and the a, b, and Z* parameter values are estimated through calibration against the ERA5L precipitation dataset. The parameter bounds are explicitly defined in the code as follows: Z∗ ranges between 10 and 1200, a between 0 and 160, b between 1 and 50, T_{pot} between 0.05 and 0.75, and T_{base} between 0.05 and 3.0. Optimization is carried out using MATLAB's fmincon function with the active set algorithm, allowing a maximum of 300 iterations and 500 function evaluations. The calibration is carried out by minimizing an objective function based on the Nash–Sutcliffe efficiency (NS, Nash and Sutcliffe, 1970), computed between simulated and reference precipitation (e.g., ERA5L). Regarding the validation strategy, the current implementation uses the full available time series at each grid cell for both calibration and performance assessment. An explicit temporal train/validation split or spatial cross validation was not applied. This design choice is consistent with the primary objective of the study, which is to conduct a controlled synthetic sensitivity assessment of precipitation retrieval performance under different NGGM/MAGIC sampling and error scenarios, rather than to provide operational out of sample validation of a precipitation product. These implementation details were added by the end of method section 2.3. Future work will explore independent validation strategies, including temporal holdout and spatial cross-validation.

RC1: Add some statistics for the calibrated parameters a, b and Z and discuss how they differentiate from the SM2RAIN other applications.

ACs: The spatial distributions of the calibrated SM2RAIN parameters are summarized in Table 1 as well as in the supplementary material. The storage capacity parameter "Z*" ranges from 38.85 to 1008.70, with a median value of 506.80 and an interquartile range (IQR) of 432.67–584.10. These values are substantially larger than those typically reported for surface soil moisture based SM2RAIN applications, reflecting the deeper and more integrated nature of terrestrial water storage signals derived from gravimetry.

Table 1. Summary statistics of calibrated SM2RAIN parameters across all grid cells. Q1 and Q3 denote the first and third quartiles of the dataset, respectively. IQR represents the interquartile range (Q3 − Q1) while Min and Max indicate the minimum and maximum values

Parameter	Min	Q1	Median	Q3	IQR	Max
Z*	38.85	432.67	506.8	584.1	151.43	1008.7
a	0	0.73	1.95	3.29	2.56	42.29
b	1	1	1	6.84	5.84	50
Tpot	0.12	0.14	0.15	0.17	0.03	0.75
Tbase	0.05	0.09	0.12	0.18	0.09	3

The drainage coefficient "a" exhibits a median of 1.95 (IQR: 0.73–3.29), indicating relatively smooth storage discharge dynamics at a basin scale. The exponent parameter "b" shows a median of 1.00 (IQR: 1.00–6.84), with some grid cells reaching the imposed parameter bounds. The prevalence of lower "b" values suggests near linear drainage behavior for many regions, consistent with large-scale aggregated storage processes rather than highly nonlinear near-surface responses. The exponential filter parameters (Tpot and Tbase) display narrow interquartile ranges, indicating stable noise reduction behavior across grid cells. Overall, the calibrated parameter distributions demonstrate physically consistent behavior and support the applicability of the SM2RAIN framework to high-frequency TWS observations. The obtained parameter ranges are physically plausible and exhibit clear spatial variability across Europe, reflecting differences in climate regime and storage characteristics. Compared to typical SM2RAIN applications based on surface soil moisture, the calibrated Z^* values are generally larger, which is expected since terrestrial water storage integrates water variations over the full soil column and shallow subsurface rather than being limited to the near-surface layer. Similarly, the drainage-related parameters "b" tends to reflect smoother and slower storage discharge dynamics, consistent with the deeper and more integrated nature of TWS compared to surface soil moisture. These differences do not indicate a change in the conceptual structure of SM2RAIN but rather reflect the different hydrological control exerted by TWS relative to surface soil moisture. We added a short discussion of these parameter statistics and their interpretation to the revised manuscript as well as in detail to supplementary material.

RC1: While ERA5 provides groundwater and snow components, the authors calculate "TWS" only from soil moisture. This assumption may disqualify specific areas.

ACs: In this study, terrestrial water storage (TWS) was approximated using the four soil moisture layers available in ERA5L. We acknowledge that this representation does not explicitly include other storage components such as snow water equivalent, groundwater, surface water, or canopy interception, which are part of the total storage measured by GRACE and GRACE-FO. However, for the selected European domain in this study, the contribution of some of these components is relatively limited for the purposes of the present sensitivity experiments. . Although this ERA5L based TWS representation does not explicitly include additional storage components such as snow water equivalent, groundwater, surface water, or canopy interception - components included in the total signal observed by GRACE and GRACE-FO- it shows a strong correlation with GRACE based TWS (Fig S1). This agreement can be explained by the fact that, over the selected European domain, the omission of these components is expected to have a limited impact on the present sensitivity experiments. In particular, canopy interception and surface water storage are generally negligible at the ~100 km scale, while snow water equivalent may locally influence storage variability in northern and mountainous regions but has a limited overall contribution during the study period. Furthermore, groundwater storage is not explicitly represented in ERA5L and therefore cannot be included in the proxy dataset used for synthetic experiments. Since the goal of this study is to find out how possible and sensitive it is to estimate the precipitation under future gravity mission configurations. By isolating the soil moisture driven component of TWS, it allows for a clearer attribution between precipitation forcing and storage response. Nevertheless, we acknowledge that this simplification does not represent the full GRACE observed storage signal; future work will incorporate additional storage components when suitable datasets become available to assess their influence on precipitation retrieval and flux partitioning under realistic hydrological conditions. The following text is added to the manuscript from 145 to 152.

RC1: SM2RAIN neglects runoff and evaporation, which can be reasonable for short events but for 15-30 days, this can be considerable. These parameters should also be considered.

ACs: The SM2RAIN framework estimates precipitation based on changes in storage, without directly modeling runoff and evapotranspiration (ET) as distinct fluxes. The aim of this study is to perform a controlled sensitivity analysis of precipitation retrieval across various NGGM/MAGIC sampling scenarios, utilizing a physically coherent ERA5L framework. This enables us to isolate the effects of temporal resolution and measurement uncertainty on precipitation estimation. We acknowledge, however, that neglecting explicit runoff and ET terms may introduce limitations at longer aggregation scales. Future work will extend the framework to include a flux consistent water balance approach that includes constraints on evapotranspiration and runoff. We will incorporate these suggestions into the "Limitations and Future Work" section of the revised manuscript.

RC1: Clarify why a Gaussian noise model was chosen, how noise is injected (independent per sample? temporally spatially correlated?) and precisely how temporal resampling is implemented (averaging/endpoints). Also, how this resample is coupled with the exponential filter step.

ACs: The additive Gaussian noise introduced is a simplified error model. In reality, gravimetric mission errors may have spatial and temporal correlations, depend on orbital geometry of the mission, and be regionally dependent (e.g., topography), but the purpose of this noise model is to assess the sensitivity of the precipitation retrieval to measurement uncertainty. Thus, the Gaussian assumption must be understood as a linear approximation of the measurement uncertainty rather than the complete expected error structure of future gravity missions. Noise is generated using a normal distribution and is assumed to be temporally and spatially uncorrelated, providing a controlled baseline for sensitivity analysis (Landerer et al., 2020; Daras et al., 2023). The following text is added to the manuscript from 236 to 242. We acknowledge that real gravity mission errors may exhibit spatial or temporal correlation; however, the present study focuses on first-order sensitivity to error amplitude, and correlated error structures will be explored in future work. Temporal resampling is implemented using a moving average aggregation scheme. A fixed window (e.g., five samples for the 5-day scenario) is applied to both the TWS signal and its time axis. The smoothed values are then embedded between the first and last original observations to preserve endpoints, and the resulting series is linearly interpolated back to the original daily time grid. This ensures temporal alignment while emulating reduced mission sampling frequency. The exponential filter described by Brocca et al. (2019) is subsequently applied within the SM2RAIN framework to the resampled (and noise-perturbed) TWS signal. Therefore, the overall smoothing of the storage signal reflects both the moving average temporal aggregation and the adaptive exponential filtering step prior to precipitation inversion. This processing sequence has now been clarified and will be incorporated in the revised manuscript. The following text is added to the manuscript from 225 to 231.

RC1: Quantify performance stratified by Köppen–Geiger class and precipitation intensity, not only spatial maps. Relate weak regions (Alps, Baltics, coastal Norway) to process drivers (snowpack, orography, deep storage).

ACs: As per reviewer suggestion, we extended the analysis by stratifying model performance according to Köppen–Geiger climate classes and precipitation intensity by utilizing ERA5L reference precipitation as demonstrated in Table S2 and Table S3 (summarized below and in the revised manuscript and explained in the supplementary material).

Table S2. Performance of SM2RAIN derived precipitation estimates stratified by Köppen Geiger climate classes across the study domain. The table reports the number of grid cells (N), the mean correlation coefficient (R) between SM2RAIN-derived and ERA5L reference precipitation, and the corresponding root mean square error (RMSE) for 15D and 30D precipitation. Grid cells were assigned to climate classes using the Köppen–Geiger classification map. This stratification highlights the influence of regional hydroclimatic conditions on the relationship between terrestrial water storage variations and precipitation.

Class	N	R_mean_15D	R_mean_30D	RMSE_mean_15D	RMSE_mean_30D
B	137	0.786	0.792	18.505	29.227
Csa	241	0.972	0.976	6.544	10.162
Csb	172	0.958	0.959	9.039	14.459
Cfb	50	0.950	0.940	9.757	16.246
Dfb	200	0.785	0.773	15.693	24.854
Dfc	144	0.847	0.846	14.856	22.756
ET	142	0.868	0.857	13.626	21.835

For each climate class, we computed the mean correlation coefficient (R) and root mean square error (RMSE) across all grid cells belonging to that class. The results indicate that the highest performance is achieved in Mediterranean climates (Csa and Csb), where mean correlations exceed 0.95. These regions are dominated by rainfall-driven hydrological processes, allowing terrestrial water storage variations to respond more directly to precipitation inputs. Similarly, high performance is observed in oceanic climates (Cfb). In contrast, lower performance is observed in colder continental climates (Dfb) and arid regions (B), where mean correlations decrease to approximately 0.78. In cold regions, seasonal snow accumulation and freeze-thaw processes can delay the storage response to precipitation, weakening the relationship assumed by the SM2RAIN inversion. In mountainous regions such as the Alps and northern areas including the Baltics and coastal Norway, complex topography and snowpack dynamics further weaken the relationship between precipitation and TWS variations.

To further investigate the sensitivity of the method to rainfall magnitude, model performance was stratified by precipitation intensity using ERA5L reference precipitation (Table S3). The results indicate a strong dependence of performance on rainfall intensity. For very light precipitation events (0–2 mm day^{-1}), the correlation between SM2RAIN-derived precipitation and the reference dataset is relatively low (R = 0.13), reflecting the limited detectability of small storage variations in terrestrial water storage signals. Performance improves progressively with increasing precipitation intensity, reaching R = 0.46 for events between 5 and 10 mm day^{-1} and R = 0.77 for heavy precipitation events exceeding 10 mm day^{-1}. This behavior is consistent with the physical nature of gravimetric observations, which capture integrated water storage changes and therefore respond more clearly to large rainfall inputs than to weak precipitation events. The increase in RMSE with precipitation intensity reflects the larger absolute magnitude of precipitation amounts rather than a deterioration of relative performance. This reflects the physical nature of gravimetric observations, which are more sensitive to large storage changes associated with strong rainfall events, while weak precipitation signals can be masked by evaporation and measurement noise.

Table S3. Performance of SM2RAIN-derived precipitation estimates stratified by precipitation intensity using ERA5L reference precipitation. The table reports the number of samples (N), correlation coefficient (R), root mean square error (RMSE), and the mean reference precipitation within each intensity bin. Precipitation events were grouped into four intensity categories (0–2, 2–5, 5–10, and >10 mm day^{-1}). This analysis evaluates the sensitivity of precipitation estimation from terrestrial water storage signals to rainfall magnitude.

Intensity Bin	**R**	**RMSE**	**Mean Pobs**	**Interpretation**
0-2 mm/day	0.133	1.281	0.336	Very weak signal
2-5 mm/day	0.311	2.746	3.283	Slight improvement
5-10 mm/day	0.457	3.233	7.069	Moderate skill
>10 mm/day	0.774	6.099	16.871	Strong skill

RC11: Compare your model's performance with prior SM2RAIN studies (SM-based) and related inversions. Also, justify the choice of R≥0.7 as "satisfactory" (literature or application-driven).

Authors: We thank the reviewer for this important suggestion. A direct comparison between precipitation estimates derived from SM and TWS within the SM2RAIN framework is indeed very relevant. However, such a comparison requires a dedicated experimental design including consistent datasets, calibration strategies, and validation using real world observations; this is beyond the scope of this work, which aims to evaluate the feasibility of precipitation estimation from TWS and the sensitivity of the approach to the anticipated sampling and the accuracy characteristics of future gravity missions. A systematic comparison between SM based and TWS-based SM2RAIN approaches is therefore planned as part of future work using observational datasets. Regarding the second point, the threshold R ≥ 0.7 was used as an indicative benchmark of satisfactory performance based on previous SM2RAIN applications and satellite precipitation validation studies. As an example, global precipitation products derived using the SM2RAIN approach typically report correlations in the range 0.6–0.8 (e.g. with respect to reference precipitation datasets), which is deemed acceptable given the indirect nature of the inversion and uncertainties in the satellite observations. Brocca et al. (2019) reported similar performances in the global SM2RAIN precipitation dataset, where correlations above 0.7 were found to be representative of good precipitation retrievals. The following text is added to the manuscript from 286 to 293.

RC1: Study limitations are missing.

ACs: Limitations and future work for this work are explained already above and will be incorporated into the revised manuscript as a separate section 4.

RC1: It is suggested to incorporate discussion with results rather than conclusions to compare your results with prior studies and justify the results.

ACs: We thank the reviewer for this suggestion. We revised the manuscript by integrating comparison with prior SM2RAIN and gravity-based studies directly within the results and discussion sections, rather than limiting such comparisons to the conclusions. Further, the present work is intended as a preliminary proof of concept, i.e., a synthetic study designed to assess the potential of high resolution TWS observations from future gravity missions for precipitation retrieval. As such, the scope of available comparisons is limited to methodological and conceptual consistency with previous SM2RAIN and large scale hydrological applications, rather than full operational validation.

Editorial comments

RC1: Tighten the abstract starting with motivation and the specific temporal scales actually evaluated (15/30-day), state the calibration/reference clearly and report core numbers with uncertainty. (Minor but in abstract R=0.86 is mentioned as daily, which I think is typo).

ACs: The abstract is revised considering reviewer suggestions and incorporated in the revised version. Regarding the reported value of R = 0.86, this is not a typo error. The correlation coefficient of 0.86 corresponds to the daily resolution computation within the synthetic framework. Precipitation was first estimated at daily resolution using SM2RAIN to check model performance, followed by aggregation to 15 and 30 day scales for comparative analysis and clearer visualization of performance (as presented in Figure 2 in the main manuscript).

RC1: Expand the mission context in the introduction with a concise paragraph of NGGM/MAGIC specs (sampling cadence, effective resolution, target/threshold errors, current program status) and cite appropriately.

ACs: The introduction has been revised by summarizing the key characteristics of the NGGM/MAGIC mission framework, including the intended temporal sampling (approximately sub-weekly), effective spatial resolution (approximately 100 km), and target equivalent water height accuracy (~5–10 mm at regional scales), as well as threshold performance considerations. We also clarified the current program status, noting that NGGM is progressing through ESA's mission definition and Phase A extension activities under the Future Earth Observation program. Appropriate references (Daras et al., 2023, 2024; Haagmans and Tsaoussi, 2020) have been added to support the mission specifications. Within this framework, the European Next-Generation Gravity Mission (NGGM), coordinated with the NASA led GRACE-C component, is designed to drastically improve both the temporal sampling and effective spatial resolution of observations of terrestrial water storage relative to the GRACE and GRACE-FO missions. Current mission studies indicate a target temporal sampling on the order of sub-weekly days and an effective spatial resolution approaching approximately 100 km at regional scales, for equivalent water height accuracies of approximately 5–10 mm under target performance conditions. Together with further improvements in orbit and attitude modelling, these will enable the generation of fast-track gravity products on sub-weekly time-scales, with reduced temporal aliasing and enhanced signal to-noise characteristics. NGGM is currently progressing through mission definition activities (Phase A extension) funded under ESA's Future Earth Observation. The following information is added to the manuscript from 89 to 97. This addition strengthens the contextualization of the study and more clearly links the synthetic experiment design to the expected capabilities of next-generation gravity missions.

RC1: Add a short subsection “Evaluation framework” that upfront defines metrics (R, RMSE, bias, NS), aggregation windows, references used, comparisons performed (feasibility vs synthetic NGGM/MAGIC vs GRACE-FO-like), and the criteria for “satisfactory”.

ACs: We thank the reviewer for this suggestion. The information describing the evaluation framework was already distributed across Sections 2.2–2.4 of the manuscript. However, to improve clarity and readability, the revised manuscript now presents this framework more explicitly and consistently. In particular, we clarified the performance metrics used (R, RMSE, bias, and Nash–Sutcliffe efficiency), the aggregation windows (15-day and 30-day), the reference dataset (ERA5L), and the different experiment configurations considered in the study, including (i) feasibility experiments using ERA5L-derived TWS, (ii) synthetic NGGM/MAGIC sampling and error scenarios, and (iii) GRACE/GRACE-FO-like synthetic benchmark configurations. We also clarified that NS was used as the calibration objective function, while R, RMSE, and bias were used for performance evaluation.

RC1: Split Figure 1 into 1a (domain) and 1b (Köppen–Geiger); improve the caption accordingly and ensure all figures state grid characteristics (masking, number of time steps per pixel).

ACs: We thank the reviewer for the suggestion. Figure 1 has been revised in the main manuscript to clearly distinguish two panels: Figure 3a shows the study domain, and Figure 3b presents the Köppen–Geiger climate classification used in the analysis. The caption has been improved and extended to describe the grid characteristics of the analysis, including the spatial resolution (100 km) and the temporal coverage (2003–2012).

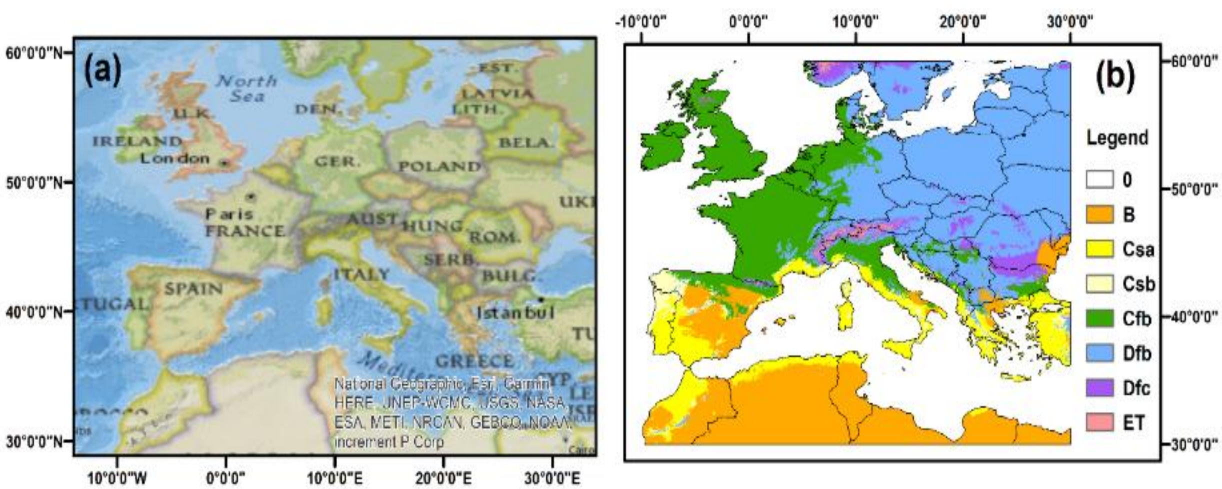

Figure 3. (a) Study domain covering the European region used in this analysis. The spatial grid corresponds to the ERA5L dataset resampled to approximately 100 km spatial resolution. The analysis period spans 2003–2012, corresponding to 3653 daily time steps used in the SM2RAIN experiments. The base map is derived from OpenStreetMap data © OpenStreetMap contributors (available under the Open Database License). **Figure 1. (b)**

Köppen–Geiger climate classification across the study domain used to assess the influence of climatic conditions on the performance of precipitation estimation from terrestrial water storage signals. The major climate classes include arid (B), Mediterranean climates (Csa, Csb), temperate oceanic climates (Cfb), cold continental climates (Dfb, Dfc), and tundra climates (ET).

RC1: Describe how the 100 random points and the eight countries were selected (random seed, spatial/climatic stratification) to ensure it cover everything.

ACs: The set of 100 grid cells was selected by randomly sampling from all valid land pixels within the study domain that contained complete TWS and precipitation records (i.e., no missing values). To ensure reproducibility of the sampling procedure, the random selection was performed using a fixed random seed as shown in Fig. 4. The following text is added to the manuscript from 396 to 400.

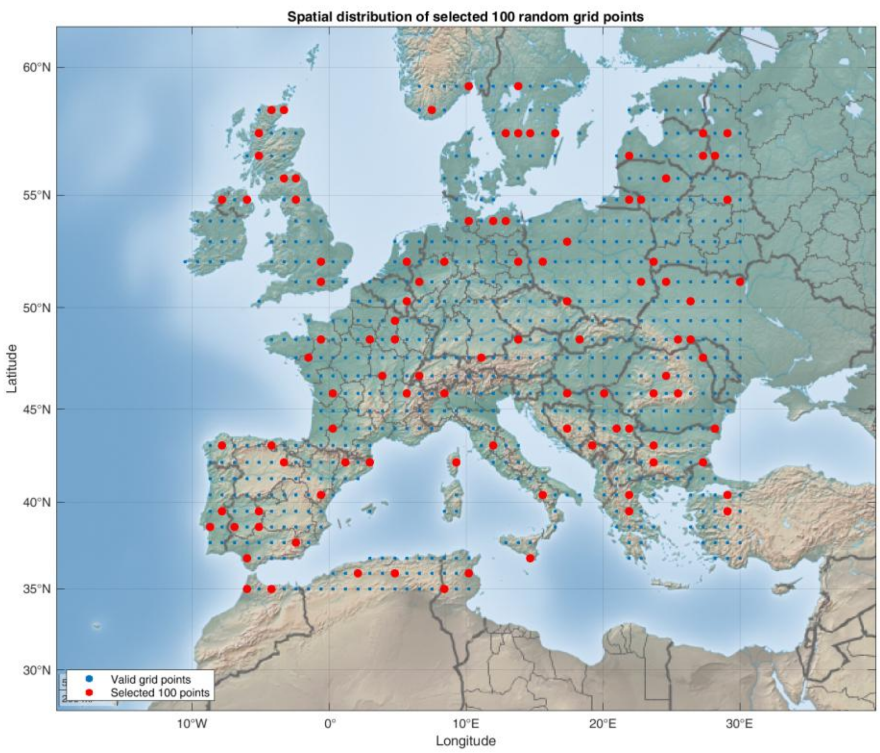

Figure 4 Spatial distribution of the 100 randomly selected grid points (red) used in the sensitivity analysis, overlaid on all valid grid cells (blue) across Europe. Sampling was performed using a fixed random seed to ensure reproducibility.

The purpose of this experiment was to provide a representative sensitivity analysis to evaluate the impact of temporal aggregation and measurement error on precipitation estimation using SM2RAIN. Therefore, no explicit climatic or spatial stratification was imposed in this step. The eight countries shown in the regional analysis were selected intentionally to provide broad geographic and hydro climatic coverage across Europe, including southern, central, northern, and eastern regions, and to represent contrasting environments such as Mediterranean, temperate, continental, and colder climates. They were therefore used as illustrative regional examples rather than as a statistically stratified sample of all European conditions. To improve transparency and demonstrate the spatial representativeness of the random sampling, we will also include Fig. 4 added as a supplementary figure in the revised manuscript showing the spatial distribution of the selected 100 grid points across the study domain.

RC1: Clarify the GRACE-FO-like benchmark (≈30-day cadence, ≈25 mm error) as a synthetic reference rather than a run driven by actual GRACE fields and keep these numbers together wherever it appears.

ACs: In this study, we used terrestrial water storage (TWS) from ERA5-Land as a proxy dataset. We did not use real GRACE or GRACE-FO observations in our analysis. Rather, we used a proxy data set and degraded its temporal sampling and measurement errors to mimic different mission configurations, in which the GRACE-FO-like configuration serves as a synthetic benchmark that approximates the nominal sampling of current satellite gravimetry missions. The measurement error levels used in the synthetic experiments are derived from the ESA Next-Generation Gravity Mission (NGGM) Mission Requirements Document (MRD), which specifies expected uncertainty levels for different spatial and temporal scales of future gravity observations (ESA, 2023). Based on these specifications, we defined a reference configuration representing present-day gravimetry capability with approximately 30-day temporal sampling and ~25 mm equivalent water height uncertainty, which we refer to as a GRACE-FO-like synthetic benchmark. This configuration serves only as a reference scenario to evaluate the potential improvements achievable with future missions such as NGGM and MAGIC. To avoid ambiguity, the manuscript will be revised to explicitly state that this benchmark is synthetic and not derived from actual GRACE observations, and the values describing the configuration (30-day cadence and ~25 mm error) are now consistently reported together wherever the GRACE-FO-like configuration is mentioned. The following text is added to the manuscript from 247 to 254.

RC1: Prefer quantitative phrasing in conclusions (e.g., deltas in R/RMSE relative to the GRACE-FO-like case under ≤10 mm error and 5-day sampling) and answer the stated research questions directly.

ACs: We thank the reviewer for this suggestion. In the revised manuscript, the conclusions section is revised by incorporating the verification of SM2RAIN leads to designing NGGM/MAGIC configurations that clearly show how temporal resolution and error levels have an impact on the performance for precipitation estimation. The results indicate high correlation values (between 0.61 and 0.88) and low RMSE values (between 28 and 50 mm) for short temporal aggregation (5 days) and for retrieval errors lower than 20 mm (see Figure 6). For longer temporal aggregation (e.g. 30 days, which is representative of GRACE-FO-like conditions), correlation values drop significantly ranging from 0.40 and 0.60 for retrieval errors lower than 20 mm. For error levels lower than 10 mm, and temporal aggregation lower than 15 days, the performance deterioration is limited; whereas for error levels larger than 20 mm and 30-day aggregation the performance drop is highly significant. These results highlight the importance of minimizing the retrieval errors in future gravity missions.

Reference

Brocca, L., Filippucci, P., Hahn, S., Ciabatta, L., Massari, C., Camici, S., Schüller, L., Bojkov, B. and Wagner, W., 2019. SM2RAIN–ASCAT (2007–2018): Global daily satellite rainfall data from ASCAT soil moisture observations. Earth System Science Data, 11(4), 1583-1601.

Daras, I., March, G., Pail, R., Hughes, C., Braitenberg, C., Güntner, A., Eicker, A., Wouters, B., Heller-Kaikov, B., and Pivetta, T.: Level-2a simulated gravity field solutions of ESA's science support study to Mass change And Geosciences International Constellation (MAGIC) Phase AV 1.0., https://doi.org/10.5880/icgem.2023.005, 2023.

Daras, I., March, G., Pail, R., Hughes, C. W., Braitenberg, C., Güntner, A., Eicker, A., Wouters, B., Heller-Kaikov, B., and Pivetta, T.: Mass-change And Geosciences International Constellation (MAGIC) expected impact on science and applications, Geophys J Int, 236, 1288–1308, https://doi.org/10.1093/gji/ggad472, 2024.

ESA (2023). Next-Generation Gravity Mission (NGGM) Mission Requirements Document (MRD), Version 1.0. European Space Agency, Future Earth Observation Programme.

Haagmans, R. and Tsaoussi, L.: Next Generation Gravity Mission as a Mass-change And Geosciences International Constellation (MAGIC) Mission Requirements Document, Earth and Mission Science Division, European Space Agency, https://doi.org/10.5270/esa.nasa.magic-mrd.2020, 2020.

Knoben, W.J., Freer, J.E. and Woods, R.A., 2019. Inherent benchmark or not? Comparing Nash–Sutcliffe and Kling–Gupta efficiency scores. Hydrology and Earth System Sciences, 23(10), pp. 4323-4331.

Nash, J.E. and Sutcliffe, J.V., 1970. River flow forecasting through conceptual models part I—A discussion of principles. Journal of hydrology, 10(3), pp.282-290.

Williams, G.P., 2025. Friends Don't Let Friends use Nash-Sutcliffe Efficiency (NSE) or KGE for Hydrologic Model Accuracy Evaluation: A rant with data and suggestions for better practice. Environmental Modelling & Software, p. 106665.

Zhong, Y., Tian, B., Kim, H., Yuan, X., Liu, X., Zhu, E., Wu, Y., Wang, L. and Wang, L., 2025. Over 60% precipitation transformed into terrestrial water storage in global river basins from 2002 to 2021. Communications Earth & Environment, 6(1), p.53.

Beyond GRACE: Evaluating the benefits of NGGM and MAGIC for precipitation estimation over Europe.

Author(s): Muhammad Usman Liaqat et al.

MS No.: egusphere-2025-3659

Reviewer 2

General Comments:

This ms addresses an important question, which is whether the next generation of gravity missions (NGGM/MAGIC) will deliver sufficient accuracy and temporal resolution to enable precipitation estimation via the SM2RAIN algorithm. The ms contributes to applying SM2RAIN for the first time, using TWS as input rather than soil moisture. However, several methodological choices require better justification before potential publication in HESS. The use of ERA5L as both the input proxy and the reference benchmark creates insufficient independence between the forcing and evaluation, which is not adequately acknowledged. The synthetic error framework needs stronger physical grounding, and the discussion of regional performance disparities (Alps, Baltic coast, Scandinavia) remains shallow. Furthermore, the ms could be improved by considering a deeper mechanistic analysis of failure modes, clearer separation of the validation and synthetic experiment frameworks, and more rigorous attention to equifinality in parameter calibration.

General Response:

We appreciate the reviewer's helpful comments, which improved overall quality of manuscript. In response, the evaluation framework has been made clearer, and the limitations of using ERA5-Land as both proxy input and reference dataset are now explicitly acknowledged. The synthetic experiment design has been strengthened by explicitly connecting the imposed measurement errors to the ESA NGGM Mission Requirements Document and discussing the assumptions underlying the Gaussian noise model. In addition, the analysis has been extended by stratifying results by Köppen–Geiger climate classes and precipitation intensity to better explain regional performance differences. The structure of the manuscript has also been changed to make it clear how the baseline feasibility experiment is different from the synthetic mission simulations and to make some parts of parameter calibration clearer. These changes make the methods clearer and the results easier to understand thoroughly, and we thank the reviewer again for their constructive feedback. A detailed response to each comment is provided below. For clarity, the reviewers' comments (RC1) are reported in light blue, while the author's responses (**ACs)** are presented in black. The revised text in the manuscript as well as below is highlighted in red to make it clear how the changes were made to address the reviewers' comments.

Specific Comments:

RC2 Around line 20: The study uses ERA5L soil moisture layers as a TWS proxy and ERA5L precipitation as the benchmark for validation. This impacts the independence of the evaluation framework, as mentioned above, a limitation that is not adequately discussed in the framework, where the algorithm is essentially optimized and tested against the same underlying data model. The authors must explicitly acknowledge this limitation and discuss how it may inflate performance metrics. An independent validation using, e.g., GPCC or E-OBS precipitation data, could be considered, even if only for a subset of stations or a sub-period.

ACs: The objective of this study is not to perform an operational validation of precipitation estimates but rather to conduct synthetic experiments to assess the feasibility of retrieving precipitation from TWS and to evaluate the sensitivity of the SM2RAIN algorithm to the temporal sampling and accuracy expected from future gravity missions such as NGGM and MAGIC. For this purpose, the use of a single physically consistent dataset is common practice in mission simulation studies, as it ensures closure of the water balance and allows the impact of temporal resolution and measurement uncertainty to be isolated. We agree that independent validation using observational precipitation datasets (e.g., GPCC or E-OBS) would be valuable for assessing the performance of TWS based precipitation retrievals under real world conditions. Such an analysis would require additional calibration strategies and consistent treatment of observational uncertainties and is therefore planned as part of future work using independent observations. The manuscript is revised to explicitly acknowledge this limitation as well as explained below:

Limitations and Future Work

This study conducts a controlled sensitivity experiment utilizing ERA5L as a self-consistent synthetic reference framework; hance, the results reflect relative performance assessment rather than independent validation against in situ observations. Importantly, these experiments therefore provide a first, controlled synthetic assessment of the feasibility of retrieving precipitation from TWS. The added value of our study is not to validate the SM2RAIN algorithm (which has been done extensively using multiple observational precipitation datasets such as (e.g., GPCC or E-OBS) but rather to assess its sensitivity to the sampling and accuracy expected from future gravity mission.

Measurement uncertainties were modeled as additive Gaussian noise with standard deviations given by the NGGM mission error specifications from (Daras et al.,2023). This notice acknowledges that real satellite gravimetry errors may be temporally and spatially correlated and may be geographically dependent, with factors including latitude, topography and orbit geometry. While the Gaussian noise assumption provides a simplistic error model, it is sufficient in providing a first-order sensitivity of how measurement uncertainty may affect precipitation retrieval. In this work, our aim is not to replicate the full complexity of mission error characteristics but rather to study the relative effect of spatiotemporal sampling and measurement uncertainty on precipitation estimation from TWS. However, this comparison would demand mission specific error covariance models, the derivation of which is beyond the scope of this study.

Further, the SM2RAIN framework does not show evapotranspiration as independent fluxes, which could affect how storage dynamics changes over weeks. Neglecting evaporation can lead to biased precipitation estimates, particularly in warm, vegetated, and semiarid regions where evapotranspiration is significant. The SM2RAIN parameterization is also expected to be improved by using machine learning approaches. Machine learning can be used to learn how the SM2RAIN parameters depend on TWS dynamics, climate indicators, and land surface characteristics. This preserves the physical SM2RAIN structure while reducing calibration dependency, improving spatial coherence, and enhancing robustness in data scarce and nonstationary environments. Finally, TWS changes generally include contributions from groundwater, but the current framework does not make a clear distinction between groundwater storage and other hydrological components. Future work will address these issues by integrating flux-consistent water balance constraints (incorporating evapotranspiration and runoff), parameter estimation using machine learning an process based ground water depletion attribution with future gravity missions. The future work will also extend to larger domains using more realistic, spatially, and temporally correlated error structures derived from advanced mission simulations.

RC2 Around line 85: The third stated objective refers to "adding Gaussian error on simulated precipitation estimates," but the Gaussian errors are actually added to the TWS input, not to the precipitation estimates. Please, revise this to avoid confusion about the experimental design.

ACs: We thank the reviewer for noting this inconsistency. The Gaussian errors were added to the TWS input signal, not directly to the precipitation estimates. We have revised the objective statement in the manuscript "to analyse the impact of different synthetic mission configurations by adding Gaussian error into simulated TWS observations used as input to the SM2RAIN algorithm" to avoid this confusion and to clarify that the third objective is to assess the impact of adding Gaussian error to the simulated TWS observations used as input to the SM2RAIN algorithm.

RC2 Around line 115: The TWS estimate (Eq. 1) is computed exclusively from the four soil moisture layers of ERA5L and does not account for snow water equivalent, groundwater, surface water, or canopy interception, all of which are components of the full GRACE-measured TWS signal. The authors should either redefine their variable explicitly as "soil water storage" (SWS) or justify why this simplification is appropriate for the European domain and the 2003–2012 period. This distinction has direct implications for interpreting how faithfully ERA5L-TWS proxies GRACE-based TWS.

ACs: In this study, terrestrial water storage (TWS) was approximated using the four soil moisture layers available in ERA5L. We acknowledge that this representation does not explicitly include other storage components such as snow water equivalent, groundwater, surface water, or canopy interception, which are part of the total storage measured by GRACE and GRACE-FO. However, for the selected European domain in this study, the contribution of some of these components is relatively limited for the purposes of the present sensitivity experiments. In particular, the contributions of canopy interception and surface water storage are generally small at the spatial scale considered (100 km). Snow water equivalent may locally influence storage variability in northern and mountainous regions, but its contribution is comparatively limited across most of the study's domain during the analyzed period. Furthermore, groundwater storage is not explicitly represented in ERA5L and therefore cannot be included in the proxy dataset used for the synthetic experiments. The following text is added to the manuscript from 145 to 152.

Given the objective of this study to assess the sensitivity of precipitation retrieval to temporal sampling and measurement errors expected from future gravity missions, we considered the integrated soil water storage from ERA5L as a suitable proxy for TWS variability. Nevertheless, we acknowledge that this simplification does not represent the full GRACE observed storage signal. The manuscript is revised to clarify this limitation, and future work will extend the analysis to larger domains and incorporate additional storage components when suitable datasets become available.

RC2 Around lines 135: The SM2RAIN parameters (a, b, Z*) are calibrated point-by-point against ERA5L precipitation. No discussion is provided about parameter sensitivity, equifinality, or the spatial coherence of calibrated parameters across Europe. A supplementary figure showing the spatial distribution of calibrated parameters would greatly strengthen the methodology and allow readers to assess whether unrealistic parameter values emerge in certain regions (e.g., the Alps, Nordic regions).

ACs: We thank the reviewer for this valuable suggestion. In the revised manuscript, we expanded the discussion on the calibration of SM2RAIN parameters. The spatial distributions of the calibrated SM2RAIN parameters are summarized in Table 1 as well as in the supplementary material.

The parameters (Z*, a, b) are estimated independently for each grid cell through point-by-point calibration against ERA5-Land precipitation. To assess the robustness and spatial consistency of the calibrated parameters, we computed summary statistics of their distributions across the European domain. The results show physically reasonable parameter ranges without unrealistic values in regions with complex hydrology such as the Alps or northern Europe. The spatial distributions of the calibrated SM2RAIN parameters are summarized in Table 1. The storage capacity parameter "Z*" ranges from 38.85 to 1008.70, with a median value of 506.80 and an interquartile range (IQR) of 432.67–584.10. These values are substantially larger than those typically reported for surface soil moisture based SM2RAIN applications, reflecting the deeper and more integrated nature of terrestrial water storage signals derived from gravimetry.

Table 1. Summary statistics of calibrated SM2RAIN parameters across all grid cells. Q1 and Q3 denote the first and third quartiles of the dataset, respectively. IQR represents the interquartile range (Q3 − Q1) while Min and Max indicate the minimum and maximum values

Parameter	Min	Q1	Median	Q3	IQR	Max
Z*	38.85	432.67	506.8	584.1	151.43	1008.7
a	0	0.73	1.95	3.29	2.56	42.29
b	1	1	1	6.84	5.84	50
Tpot	0.12	0.14	0.15	0.17	0.03	0.75
Tbase	0.05	0.09	0.12	0.18	0.09	3

The drainage coefficient "a" exhibits a median of 1.95 (IQR: 0.73–3.29), indicating relatively smooth storage discharge dynamics at a basin scale. The exponent parameter "b" shows a median of 1.00 (IQR: 1.00–6.84), with some grid cells reaching the imposed parameter bounds. The prevalence of lower "b" values suggests near linear drainage behavior for many regions, consistent with large-scale aggregated storage processes rather than highly nonlinear near-surface responses. The exponential filter parameters (Tpot and Tbase) display narrow interquartile ranges, indicating stable noise reduction behavior across grid cells. Overall, the calibrated parameter distributions demonstrate physically consistent behavior and support the applicability of the SM2RAIN framework to high-frequency TWS observations. The obtained parameter ranges are physically plausible and exhibit clear spatial variability across Europe, reflecting differences in climate regime and storage characteristics. Compared to typical SM2RAIN applications based on surface soil moisture, the calibrated values are generally larger, which is expected since terrestrial water storage integrates water variations over the full soil column and shallow subsurface rather than being limited to the near-surface layer. Similarly, the drainage-related parameters "b" tends to reflect smoother and slower storage discharge dynamics, consistent with the deeper and more integrated nature of TWS compared to surface soil moisture. These differences do not indicate a change in the conceptual structure of SM2RAIN but rather reflect the different hydrological control exerted by TWS relative to surface soil moisture. We added a short discussion of these parameter statistics and their interpretation to the revised manuscript as well as in the supplementary material. This analysis helps to assess potential equifinality and provides an overview of parameter variability across the study domain. We agree that a more detailed sensitivity analysis would be valuable, but this falls outside the scope of the present study and will be considered in future work.

RC2. Around line 155: The synthetic errors (1.9, 4.2, 19, 42 mm) are introduced as additive Gaussian noise drawn from the NGGM mission error table (Daras et al., 2023). However, real mission errors are not necessarily Gaussian in space and time since they may be correlated, anisotropic, and orographically structured. The authors should discuss the limitations of the Gaussian error assumption and whether correlated or spatially structured noise could materially change the conclusions.

ACs: We thank the reviewer for raising this point. In this study, measurement uncertainties were represented by additive Gaussian noise with standard deviations derived from the NGGM mission error specifications reported by Daras et al. (2023). We acknowledge that real satellite gravimetry errors may exhibit spatial and temporal correlations and may also depend on geographical factors such as latitude, orbit geometry, and topography. Therefore, the Gaussian noise assumption used here represents a simplified error model intended to provide a first-order sensitivity assessment of how measurement uncertainty may influence precipitation retrieval. The objective of this work is to investigate the relative impact of temporal sampling and measurement uncertainty on precipitation estimation from TWS rather than to reproduce the full complexity of mission error characteristics. A more realistic representation of correlated or spatially structured noise would require mission specific error covariance models and is therefore beyond the scope of the present study. We added this information in the manuscript from 237-244.

We also added a discussion of this limitation in the manuscript and clarify that the Gaussian noise assumption should be interpreted as a simplified representation of measurement uncertainty. We also highlighted that future work will extend this analysis to larger domains using more realistic, spatially and temporally correlated error structures derived from advanced mission simulations.

RC2. Around line 120: The bilinear interpolation from ERA5L's native 0.1° resolution to 100 km is described briefly but without justification. How does this resampling affect the TWS signal and the precipitation estimates, particularly in areas with complex topography such as the Alps or the Scandinavian mountains? The authors should verify that no significant smoothing bias is introduced by this step. Furthermore, a realistic GRACE-TWS signal would require a spherical harmonic representation up to a degree and order corresponding to NGGM.

ACs: We thank the reviewer for this comment. The 100 km grid was adopted as a common analysis scale to apply the SM2RAIN framework consistently across the European domain and to compare different synthetic mission configurations within a unified spatial framework. This resolution does not represent the native resolving capability of GRACE, GRACE-FO, or future NGGM/MAGIC observations. We acknowledge that bilinear interpolation smooths small scale spatial variability and may reduce local gradients in both precipitation and terrestrial water storage, particularly in regions with complex topography such as the Alps. This spatial smoothing may partially contribute to the reduced model performance observed in these regions. The objective of this study is to investigate the relative sensitivity of precipitation retrieval to temporal sampling and measurement error, rather than to reproduce the full spatial resolution of gravity observations. A fully realistic simulation of gravity mission observations would require spherical harmonic representation and mission-specific filtering consistent with NGGM/MAGIC retrieval chains. Such an end to end mission simulation is beyond the scope of this preliminary sensitivity study but represents an important direction for future research. The above text is added to the manuscript from 153 to 162.

RC2. Around line 175–185: The ms identifies systematic underperformance in the Alpine region of northern Italy, in Estonia and Latvia, and along the Norwegian coast, but does not provide a mechanistic explanation. Are these failures related to orographic precipitation, snow processes, coastal sea-level effects on TWS, or inherent limitations of SM2RAIN? Please consider exploring this issue, and a dedicated paragraph examining the physical drivers of regional failure would improve the scientific depth of the results section.

ACs: In the revised manuscript, we extended the discussion to investigate the weaker performance observed in the Alpine region of northern Italy, the Baltic countries, and coastal Norway. The reduced performance in some areas (for instance, coastal Norway, the Alpine region of northern Italy and the Baltic countries) can be attributed to a number of hydrological factors. For instance, the strong orographic precipitation gradients and pronounced sub-grid variability of mountainous regions such as the Alps are hard to represent at the 100 km spatial resolution used in our study. Furthermore, in northern latitudes (for example, Estonia, Latvia and coastal Norway), seasonal snow accumulation and melt introduce lags between precipitation input and changes in terrestrial water storage that also weaken the direct relationship assumed by the SM2RAIN inversion. In coastal areas, maritime precipitation variability and groundwater dynamics may complicate the precipitation storage relationship (Trautmann et al., 2018). These process based explanations are consistent with the climate stratified analysis shown in the Table S2 and Table S3, which shows that model skill is lower in the colder, more continental, snow dominated climate regimes. The above text is added to the manuscript from 276 to 284.

RC2. Around line 200: In Section 3.1, ERA5L precipitation serves as the reference; in Section 3.2, the SM2RAIN-derived output from Section 3.1 is used as the reference for the synthetic experiments. This might imply that any systematic errors in the Section 3.1 output are inherited and compounded in Section 3.2. The authors should discuss how errors in the first step propagate into the second and explicitly state this design choice.

ACs: The precipitation simulated in the first step serves as the reference for the synthetic experiments. This design choice makes sure that the experiments are consistent with each other and lets the analysis separate the effects of temporal sampling and measurement uncertainty in the TWS signal. Although systematic errors from the initial step may carry over to the subsequent step; nevertheless, the purpose of the synthetic experiments is to assess the comparative sensitivity of precipitation retrieval to NGGM/MAGIC sampling configurations, rather than to provide an independent validation. Using the SM2RAIN derived precipitation as a reference ensures internal consistency of the experimental framework and allows the analysis to focus specifically on the impact of TWS sampling and measurement uncertainty. It is explained in detail between lines 231-235.

RC2. Section 3.2 appears twice, first for "Performance of NGGM synthetic TWS for precipitation estimation" (p. 7) and again for "Performance of NGGM\MAGIC mission with respect to GRACE-FO" (p. 10). This should be corrected throughout.

ACs: We thank the reviewer for pointing out these numbering inconsistencies. Corrections have been carried out.

RC2. Around line 275: The authors define R > 0.7 as the threshold for "satisfactory" performance, but this threshold is neither justified in the text nor referenced to any established community standard. Given that the application context is precipitation estimation for hydrological use, the justification should be tied to downstream impacts (e.g., flood forecasting, drought detection), or a sensitivity analysis should show how conclusions change under alternative thresholds such as 0.6 or 0.8.

ACs: The threshold R > 0.7 was used as an indicative benchmark of satisfactory performance based on previous SM2RAIN applications and satellite precipitation validation studies (Brocca et al., 2015, 2019). As an example, global precipitation products derived using the SM2RAIN approach typically report correlations in the range 0.6–0.8 (e.g. with respect to reference precipitation datasets), which is deemed acceptable given the indirect nature of the inversion and uncertainties in the satellite observations. Brocca et al. (2019) reported similar performances in the global SM2RAIN precipitation dataset, where correlations above 0.7 were found to be representative of good precipitation retrievals. Considering the inherent uncertainties in precipitation estimation from terrestrial water storage and the large spatial scales analyzed in this study, correlations above 0.7 generally indicate that the main temporal variability of precipitation is well captured. The following discussion is incorporated into section 3.1 between 287-294.

RC2. Around line 280: The statement that the GRACE-FO error is 25 mm and its temporal resolution is 30 days is used as the baseline for comparison. Please add a citation or derivation that supports these specific values in the main text. The source and spatial scale at which this 25 mm figure applies must be explicitly cited and contextualized, as the GRACE-FO error itself might be resolution-dependent.

ACs: In this study, we used terrestrial water storage (TWS) from ERA5-Land as a proxy dataset. We did not use real GRACE or GRACE-FO observations in our analysis. Rather, we used a proxy data set and degraded its temporal sampling and measurement errors to mimic different mission configurations, in which the GRACE-FO-like configuration serves as a synthetic benchmark that approximates the nominal sampling of current satellite gravimetry missions. The measurement error levels used in the synthetic experiments are derived from the ESA Next-Generation

Gravity Mission (NGGM) Mission Requirements Document (MRD), which specifies expected uncertainty levels for different spatial and temporal scales of future gravity observations (ESA, 2023). Based on these specifications, we defined a reference configuration representing present-day gravimetry capability with approximately 30-day temporal sampling and ~25 mm equivalent water height uncertainty, which we refer to as a GRACE-FO-like synthetic benchmark. This configuration serves only as a reference scenario to evaluate the potential improvements achievable with future missions such as NGGM and MAGIC. The manuscript is revised to clarify that this configuration is synthetic, to explicitly cite the ESA MRD, and to report the temporal cadence and error magnitude together wherever the GRACE-FO-like configuration is discussed.

RC2. Section 4, Discussion and Conclusion: This section reads as a summary of the results without engaging with the literature on gravity-based hydrological applications or competing approaches. The authors should discuss how their findings offer any complementary advantage over SM2RAIN-SM, which is briefly mentioned but deserves more elaboration.

ACs: We thank the reviewer for this suggestion. We added this explanation in the conclusion section of the revised manuscript discussing the potential complementarity between precipitation retrieval from terrestrial water storage (TWS) and the more widely used SM2RAIN approach based on soil moisture (SM). In particular, the use of TWS in respect to standard SM2RAIN applications based on satellite soil moisture observations (e.g., Brocca et al., 2014, Brocca et al., 2019) provides a more integrated representation of the terrestrial water cycle by combining contributions from soil moisture, groundwater, and surface water storage. While soil moisture measurements typically represent only the near-surface layer, gravity-based TWS observations capture deeper storage components and large-scale water mass redistribution. Moreover, precipitation estimation from TWS may be complementary to soil moisture based approaches for large spatial scales and over regions where surface soil moisture observations are sparse and/or strongly impacted by local processes (e.g. densely vegetated areas and frozen/snow-covered soils).

References

Brocca, L., Massari, C., Ciabatta, L., Moramarco, T., Penna, D., Zuecco, G., Pianezzola, L., Borga, M., Matgen, P., and Martínez-Fernández, J.: Rainfall estimation from in situ soil moisture observations at several sites in Europe: an evaluation of the SM2RAIN algorithm, Journal of Hydrology and Hydromechanics, 201–209, https://doi.org/10.1515/johh-2015-0016, 2015.

Brocca, L., Filippucci, P., Hahn, S., Ciabatta, L., Massari, C., Camici, S., Schüller, L., Bojkov, B. and Wagner, W., 2019. SM2RAIN–ASCAT (2007–2018): Global daily satellite rainfall data from ASCAT soil moisture observations. Earth System Science Data, 11(4), pp.1583-1601.

Daras, I., March, G., Pail, R., Hughes, C., Braitenberg, C., Güntner, A., Eicker, A., Wouters, B., Heller-Kaikov, B., and Pivetta, T.: Level-2a simulated gravity field solutions of ESA's science support study to Mass change And Geosciences International Constellation (MAGIC) Phase AV 1.0., https://doi.org/10.5880/icgem.2023.005, 2023.

ESA (2023). Next-Generation Gravity Mission (NGGM) Mission Requirements Document (MRD), Version 1.0. European Space Agency, Future Earth Observation Programme.

Brocca, L., Ciabatta, L., Massari, C., Moramarco, T., Hahn, S.,Hasenauer, S., Kidd, R., Dorigo, W., Wagner, W., and Levizzani, V.: Soil as a natural rain gauge: Estimating global rainfall from satellite soil moisture data, Journal of Geophysical Research: Atmospheres, 119, 5128–5141, https://doi.org/10.1002/2014JD021489, 2014.