

Responses to reviewer's comments

We appreciate the constructive comments from Reviewer #2. In the revised manuscript, we have incorporated changes (highlighted in blue) to address all comments presented below in *italics*. Our responses to each comment are provided in **bold**.

1. Issues with scale: Have the authors thought about GRACEs coarse footprint? The basin, measuring about 23,000 km², is significantly less than GRACEs native resolution of around 90,000 to 300,000 km². Why is there the implication that interpolating GRACE to 0.25°–0.5° would add new detail? The ms explains that the "downscaled" Δ TWS on 0.25°–0.5° does not significantly change the results, since it correlates strongly with $r \approx 0.85$ –1 with the initial 3° mascon and whole-Borneo signal. That is, the authors are basically looking at the same broad signal. Can we discuss whether this adds to our knowledge, or could this section be reworded or revised for simplicity? Could explain why these finer grids matter, as they are not independent signals.

R1: We thank the reviewer for these critical comments. We fully acknowledge that the study area is substantially smaller than the effective resolution of GRACE. We do not claim that the downscaled (0.25°–0.5°) GRACE data introduce new independent spatial information. The finer-grid analyses were used solely to: (1) examine internal consistency across commonly used Level-3 GRACE products at their distributed resolutions; (2) evaluate the degree of correlation between sub-mascon grids and larger-scale signals to test the limits of GRACE applicability in small coastal basins; and (3) quantify potential leakage effects from adjacent oceanic masses at different spatial scales. We have made revisions (lines 88-95, 268-272, 284-290 – see below) to clarify that the higher-resolution GRACE products do not provide independent Δ TWS signals but serve instead to assess spatial coherence and leakage behaviour across scales.

Lines 88-95: "This study examines whether GRACE can provide physically meaningful signals of groundwater storage variability in the small, coastal, and data-scarce Lower Kutai Basin (LKB), recognizing that its spatial scale lies below GRACE's effective resolution and is highly susceptible to ocean leakage. Specifically, we (1) compare Δ TWS across multiple GRACE products, (2) evaluate the plausibility of GRACE-derived Δ GWS against limited piezometric data, and (3) assess whether large-scale climate drivers, particularly ENSO, are detectable. The aim is to evaluate the limits of GRACE

in this challenging setting and to demonstrate that without a substantially expanded piezometric network, neither GRACE nor in situ observations alone can provide robust groundwater storage assessments to support the development of climate-resilient groundwater management strategies, particularly in rapidly urbanizing regions such as Nusantara where water security is a growing concern.”

Lines 268-272: “Terrestrial water storage anomaly (Δ TWS) values were classified into three spatial scales (Fig. 1b): (1) the study area grids, (2) a single 3° mascon grid, and (3) the entire Borneo Island. These classifications enable a diagnostic assessment of GRACE performance across spatial scales, recognizing that the finer (0.25°-0.5°) products are not independent of the native data. The analyses do not add spatial detail but rather test internal consistency among datasets and evaluate how leakage from adjacent ocean grids may influence Δ TWS estimates in a small coastal basin that is below GRACE’s effective footprint.”

Lines 284-290: “Despite the differences, Δ TWS in the study area remains highly correlated with both single-mascon and values for Borneo Island ($r = 0.78-1$; RMSE = 0.8-4.4 cm, Table S3). The study area tracks the single mascon closely ($r = 0.93-1$; RMSE = 0.8-2.5 cm) but divergence is more conspicuous relative to Borneo Island, particularly for COST-G ($r = 0.78$; RMSE = 1.3 cm). Applying leakage corrections increases coherence with the single mascon ($r = 0.92$ for GFZ, 0.94 for COST-G) yet reduces consistency with Borneo Island, especially for GFZ ($r = 0.64$; RMSE = 3.7 cm). The strong correlations among sub-mascon grids, the single mascon, and Δ TWS values for Borneo Island suggest that the apparent fine-scale structure may not represent true spatial heterogeneity, underscoring the scale limitation of GRACE for basins smaller than its native resolution.”

2. Data and uncertainty: The delGWS estimates are based on the subtraction of soil and surface water models from GRACE TWS and thus inherit all related uncertainties. You are using several GLDAS models and WGHM and reporting their differences. But it would be helpful to discuss the unreliability of those in this situation. For example, you noticed that one of the GLDAS runoff datasets was “implausible,” and you discarded it – this says that model results can not be completely trusted. Did you compare GLDAS soil moisture or WGHM surface water with any local observations? If that is the scenario, it would be helpful if this could be mentioned. Your technique for producing 36 GWS “realizations” through combinations of GRACE and model ensembles is resourceful. Nevertheless, the selection to eliminate some 30% of them as “implausible”. Have you thought about reducing model biases

instead? You are correct to have noted that these arithmetic anomalies mainly appear when ΔTWS is smaller than ΔSMS plus ΔSWS in the wet months, or the reverse during the dry months. Overlooking those months might affect your trends. It would be useful to have a numerical estimate of the effect of the filtering on the results. You show that removing the outliers improves the TWS–GWS correlation. It might be helpful to provide error estimates for ΔGWS or to note that the "mean plausible" series is just one model of noisy data? Assumptions of the model: Along these lines, it should be pointed out that GLDAS and WGHM don't account for groundwater pumping.

There appear to be signs that some wells are experiencing decline due to abstraction, but the GWS calculations don't account for this. If water pumping is stopped from the basin (and not simply recharging local surface water), GRACE would find a deficit. But the average GRACE ΔGWS trend indicates a very gentle positive change, even when one well went significantly deeper and had a local decrease. How do you resolve that? The conclusion blames the discrepancy on "withdrawals," but it would be nice to supply some quantification or at least investigate whether the pumping is too localized to be seen in GRACE. Are there other factors to take into account, such as groundwater draining from the basin by rivers or recharge assumptions?

R2: We appreciate the comments. We agree that uncertainties in both GRACE ΔTWS and the simulated water storage components (ΔSMS , ΔSWS) may propagate to groundwater storage change (ΔGWS) estimates. We have noted in the revised text that GLDAS and WGHM are not independently calibrated for our study area, as ground-based observations are largely unavailable (lines 601-605).

In the revised manuscript, we generated 54 ΔGWS realizations because we included two GRACE spherical harmonic products (GFZ and COST-G). We agree that discarding months with implausible ΔGWS estimates can influence long-term trends. Therefore, we computed and found that the unfiltered ensemble mean ΔGWS values produce a linear trend of 0.02 cm/year, whereas the filtered plausible values produce a larger trend of 0.2 cm/year (lines 425-427).

We agree that the ensemble mean plausible ΔGWS estimates should not be interpreted as true values as validation using spatially distributed piezometric data and storage coefficients is necessary (lines 500-510). We also acknowledge that GLDAS and WGHM do not explicitly represent groundwater abstraction in the study area (lines 610-612).

Because spatially distributed storage coefficients that are required to convert groundwater level changes to ΔGWS are unavailable, we can only estimate ΔGWS using a range of possible storage coefficient values (lines 500-510). The effect of groundwater pumping on ΔGWS , however, would only be apparent if it is spatially extensive and volumetrically large enough to influence basin-scale mass detected by GRACE. In our study area, abstraction is spatially concentrated south of Balikpapan City and at depths within specific screened intervals (lines 529-531). We consider that the observed declines in some piezometers are likely localized.

Lines 425-427: “The ensemble mean of plausible ΔGWS ranges from -4.5 to 7.2 cm with an annual trend of 0.2 cm/year ($p < 0.05$), whereas incorporating implausible values results in ΔGWS ranges between -5.6 and 8.7 cm with an annual trend of 0.02 cm/year ($p > 0.05$).”

Lines 500-510: “In the study area, groundwater abstraction is concentrated within ~20 km of coast, particularly south of Balikpapan City (Fig. S24). We compared ensemble mean plausible GRACE-derived ΔGWS with groundwater level changes (ΔGWL) as the storage coefficient (S or S_y) is not well constrained. Lithological logs reveal heterogeneous interbedded sand and clay units (Fig. S25; Arifin et al. (2024)). Piezometer depths range from 21 to 135 m; shallow screens up to ~40 m may represent unconfined aquifers, whereas deeper screens may tap confined aquifers. In similar deltaic settings such as the Mekong and Indo-Gangetic deltas, storage coefficients mostly vary from ~0.08 to 0.25 (mean ~0.15) for unconfined aquifers and from $\sim 10^{-5}$ to 8×10^{-4} (mean $\sim 5 \times 10^{-4}$) for confined aquifers (BGS and DPHE, 2001; Bonsor et al., 2017; Pechstein et al., 2018; Van et al., 2023). This uncertainty translates into a range of possible ΔGWS values (Fig. 8). On average, confined aquifers yield ~0.5 mm of storage loss per metre of decline ($S = 5 \times 10^{-4}$), whereas unconfined aquifers yield ~15 cm ($S_y = 0.15$). Although correlations are unaffected by this uncertainty, the amplitude of GRACE-derived ΔGWS remains highly uncertain without reliable storage coefficients.”

Lines 529-531: “Groundwater abstraction is not regional but largely concentrated near the southern coast of Balikpapan City (Fig. S24), likely contributing to localized declines in groundwater levels.”

Lines 601-605: “Since ΔGWS is not directly measured by satellites but inferred through hydrological balance calculations, its accuracy depends on the reliability of ΔTWS estimates and the simulated water components in the study area, primarily ΔSMS and

ΔSWS, which can propagate errors into ΔGWS estimates. However, these simulated components are not locally calibrated as soil moisture, river stage, and lake volume observations are largely unavailable.”

Lines 610-612: “Another major source of uncertainty in GRACE-derived ΔGWS calculations is the representation of surface water storage and anthropogenic influences in the study area, including groundwater abstraction within the study area in the GLDAS and WGHM datasets.”

3. ENSO correlations: The moderate values of $\Delta TWS/\Delta SMS$ versus ENSO indices seem to make sense and are in line with results from other studies, but care should be taken with the short time series. Did you test for statistical significance or account for autocorrelation? Has it been possible for you to investigate correlating detrended or deseasoned data? It might improve your case if you think about stripping the seasonal cycle before associating it with ENSO. In some way or other, the physical explanation works out: large El Niño events dry the basin, which in turn amplifies the GRACE ΔGWS drops. But it is really the soil moisture changes that are having an effect on TWS or it seems that ΔGWS is much less sensitive? It may be beneficial to emphasize how the soils and floodplains contribute significantly to the GRACE signal here.

R3: We thank the reviewer for these suggestions. In the revised manuscript, we have added *p-values* to indicate statistical significance (lines 453-467), all of which are statistically significant ($p < 0.05$). We have also computed correlations between climate indices and precipitation with deseasonalized and detrended ΔTWS , ΔSMS , and ΔGWS . The resulting correlations are weaker than those obtained from the raw time series (lines 467-470), suggesting that ENSO-driven rainfall variability primarily influences seasonal changes in near-surface water storage components.

Lines 453-467: “Across GRACE products, ΔTWS shows weak-to-moderate correlations (*p-values* <0.05) with ENSO indices, with *r* values ranging from -0.46 to -0.56 for MEI and -0.41 to -0.48 for ONI. The ensemble mean ΔTWS yields correlations of -0.52 (MEI) and -0.46 (ONI), whereas weaker correlations are observed with the IOD (DMI, *r* = -0.26) and PDO (*r* = -0.25). These results indicate that ENSO is the dominant driver of ΔTWS variability in the study area compared to IOD and PDO.

ΔSMS exhibits slightly stronger ENSO sensitivity than ΔTWS. Catchment and Noah LSMs yield *r* values with MEI and ONI ranging from -0.56 to -0.62, whereas the ensemble mean shows correlations of -0.60 (MEI) and -0.61 (ONI) with *p-values* <0.05. As with

ΔTWS, correlations with DMI (-0.31 to -0.35) and PDO (-0.2 to -0.22) are weaker. These suggest that soil moisture anomalies are relatively sensitive to ENSO events. Monthly precipitation also correlates moderately with ΔTWS ($r = 0.47-0.52$) and ΔSMS ($r = 0.46-0.58$).

In contrast, correlations between ensemble mean plausible ΔGWS in this study with climate indices or precipitation remain consistently weaker than those for ΔTWS or ΔSMS, with values of -0.41 (MEI), -0.33 (ONI), -0.18 (DMI), -0.23 (PDO), and 0.38 (precipitation) with p -values < 0.05 . The ensemble mean plausible ΔGWS estimates from GLWS datasets are similar, with values of -0.44 (MEI), -0.48 (ONI), -0.17 (DMI), -0.26 (PDO), and 0.29 (precipitation). In contrast, the GLDAS dataset exhibits substantially stronger correlations with MEI (-0.6) and ONI (-0.63), whereas correlations with other indices are relatively comparable: -0.32 (DMI), -0.21 (PDO), and 0.43 (precipitation)."

Lines 467-470: "In addition, we repeated the correlation analysis using deseasonalized and detrended components of ΔTWS, ΔSMS, and ΔGWS which resulted in generally weaker correlations than those of the raw data (Fig. S23). This reduction indicates that much of the ENSO-related signal is expressed through modulation of the seasonal cycle of water storage changes rather than non-seasonal anomalies."

4. Piezometer comparisons: I do like the attempt to incorporate groundwater wells, but please be careful. In the first place, matching GRACE ΔGWS (in cm over the whole basin) with a small set of point measurements is of course an approximation. Your conclusion shows a "moderate" correlation, which is reasonable. Claiming, however, that the time series "align with groundwater-level" (Abstract) goes too far. For example, GRACE indicates a small upward trend in general, while the deep well has a downward trend. How are the authors confirming that these are actually deep observation wells? Furthermore, head to storage conversion demands a certain yield values, which has not been discussed. Thus, the comparison is qualitatively of the rise/fall type instead of being concerned with quantitative volume. Please the range of specific yield you are assuming, or flagging this as a possible source of uncertainty? without knowing aquifer properties, its not easy to relate a GRACE-derived cm change to an observed m depth change.

R4: We thank the reviewer for these comments. We agree that matching basin-scale GRACE-derived ΔGWS with a limited number of point-scale groundwater level observations is inherently approximate. We have revised the abstract accordingly. As in R2 (lines 500-510), we have reported the depths of screened intervals in each piezometer to indicate that observations are generally from deep. We agree that

converting head changes to storage change requires spatially distributed storage coefficients, which are not available in this basin. We have estimated ΔGWS from piezometric data using plausible storage coefficients derived from comparable deltaic aquifers (lines 500-510).

Abstract: "Groundwater is considered a climate-resilient source of freshwater yet its long-term response to climate variability remains poorly understood in environments with limited ground-based monitoring networks. In the Lower Kutai Basin where Indonesia's new capital (Nusantara) is under development, we examine evidence from Gravity Recovery and Climate Experiment (GRACE) satellite data, global-scale models, precipitation records, and in situ piezometric observations to investigate groundwater storage changes (ΔGWS) over the last two decades. GRACE-derived terrestrial water storage anomalies (ΔTWS) exhibit strong seasonal and interannual variability that are consistent across different spatial scales ($r = 0.78-1$) and are dominated by changes in soil moisture storage (ΔSMS). Paired land-ocean grid analyses show product-dependent residual correlations after detrending and deseasonalizing with values of up to 0.68, suggesting potential ocean leakage. Across 54 realizations, 21-60% of ΔGWS estimates per realization are plausible with ensemble mean values ranging from -4.5 to 7.2 cm. Agreement between GRACE-derived ΔGWS and groundwater-level anomalies (ΔGWL) varies by site and depth, with correlations that are generally weak and reflect discrepancies in scale between GRACE's basin-scale signals and localized aquifer dynamics influenced by heterogeneity and groundwater abstraction. Statistical analyses show weak-to-moderate coupling of ΔTWS and ΔSMS with ENSO indices ($r = -0.4$ to -0.6), whereas ΔGWS is less responsive. The strongest 2015-2016 El Niño is a notable example, associated with ΔTWS deficits (-2.4 to -4.6 cm/month) and ΔGWS declines (-1.1 cm/month). High-frequency (hourly) groundwater-level observations indicate that episodic, high-intensity rainfall events (>90th percentile) disproportionately contribute to groundwater recharge. These findings underscore the need for expanded in situ monitoring and accurate storage coefficients to validate GRACE-derived ΔGWS , particularly in regions such as Nusantara where water security is a growing concern."

5. Monitoring groundwater: Filtering for the minimum daily level is an able strategy and seems to track recharge. As a reader, I am interested in the strength of that: can pump-off intervals be incorrectly interpreted, or can the lowest daily level nevertheless be influenced by slower pumping rates? Can you indicate how well these loggers are calibrated? A mention of data quality control, would be helpful.

R5: We appreciate this comment and agree that the filtering approach may involve some uncertainties. In the revised manuscript, we have added a brief description of the procedure used to ensure data quality (lines 707-716).

Lines 707-716: “The recorded data were visually inspected for noise, gaps, and abrupt shifts that may indicate sensor malfunction. The pressure transducers were factory-calibrated and manual depth-to-water checks were carried out at deployment and retrieval for calibration. Barometric pressure corrections were applied to all measurements. By filtering out data recorded during pumping periods and retaining only the minimum daily groundwater levels from recovery phases, an approximated near-static groundwater level time series was extracted from the pumping well (Fig. 9b). These suggest that with appropriate data filtering techniques, pumping wells can serve as a viable and cost-effective alternative to traditional monitoring wells, offering a scalable solution for groundwater monitoring in Indonesia. However, reliability requires periodic manual validation of piezometric sensors, such as monthly checks during the first few months of deployment and semi-annual checks thereafter, in order to mitigate instrumental issues such as sensor drift and calibration errors that can bias long-term records at seasonal or interannual timescales.”

6. Clarity and flow: The writing is generally easy to follow, although there are spots where its a bit dense. The methods section is very thorough, although it could do with more obvious labelling. It may be a good idea to divide the GRACE discussion from the GLDAS/WGHM. Some of the sentences are also very long – try breaking them up to make them easier to read, especially in the introduction and methods sections.

R6: We appreciate the positive feedback. We have reviewed and revised the prose of the manuscript to improve its clarity and flow.

7. This paper shows a worthy effort in tackling a difficult problem, however, I am concerned about the 23,000 km² study area. The authors have used a wide range of data and monitoring methods. I would recommend significant revision, moderating any very assertive conclusions, especially about delGWS match with wells etc, and being forthright about the large uncertainties involved. I look forward to reading a revised version.

R7: We thank the reviewer for the constructive comments. We acknowledge that the study area is small relative to GRACE’s native footprint. Accordingly, we have reviewed

all comments and made revisions to present a balanced interpretation consistent with GRACE's scale and uncertainty limitations.

References

- Arifin, Shamsudduha, M., Ramdhan, A. M., Reksalegora, S. W., & Taylor, R. G. (2024). Characterizing deep groundwater using evidence from oil and gas exploration wells in the Lower Kutai Basin of Indonesia. *Hydrogeology Journal*, 32, 1125-1144. <https://doi.org/10.1007/s10040-024-02776-0>
- BGS, & DPHE. (2001). Arsenic contamination of groundwater in Bangladesh. In D. G. Kinniburgh & P. L. Smedley (Eds.), *British Geological Survey Technical Report WC/00/19*. British Geological Survey.
- Bonsor, H. C., MacDonald, A. M., Ahmed, K. M., Burgess, W. G., Basharat, M., Calow, R. C., Dixit, A., Foster, S. S. D., Gopal, K., Lapworth, D. J., Moench, M., Mukherjee, A., Rao, M. S., Shamsudduha, M., Smith, L., Taylor, R. G., Tucker, J., van Steenbergen, F., Yadav, S. K., & Zahid, A. (2017). Hydrogeological typologies of the Indo-Gangetic basin alluvial aquifer, South Asia. *Hydrogeology Journal*, 25(5), 1377-1406. <https://doi.org/10.1007/s10040-017-1550-z>
- Pechstein, A., Hanh, H. T., Orilski, J., Nam, L. H., & Manh, L. V. (2018). *Detailed Investigations on the hydrogeological situation in Ca Mau Province, Mekong Delta, Vietnam*. https://www.recyclingrohstoffe-dialog.de/EN/Themen/Wasser/Projekte/abgeschlossen/TZ/Vietnam/techn_repIII-5_en.pdf?__blob=publicationFile&v=3
- Van, T. D., Zhou, Y., Stigter, T. Y., Van, T. P., Hong, H. D., Uyen, T. D., & Tran, V. B. (2023). Sustainable groundwater development in the coastal Tra Vinh province in Vietnam under saltwater intrusion and climate change. *Hydrogeology Journal*, 31(3), 731-749. <https://doi.org/10.1007/s10040-023-02607-8>