Referee #1

The paper describes a statistically based method to identify long and short term variability (trends and interannual variability) in TWS and GW. It also include analysis on the TWS components such as SWE, Soil moisture, surface water, and glacier. Th authors run their analysis on the G3P product. An already established framework that determines GWS from removing snow, soil moisture, etc. from GRACE/FO derived TWS estimates.

In general, I think the paper is a valuable contribution to the scientific community as it identifies trends and variabilities in GSW and TWS using the 21+ years of GRACE/FO record. The paper uses statistically robust methods to identify significance in the trends and compares their findings with previously defined trends.

We thank you for the constructive assessment of our manuscript. We have revised the manuscript and Supplement in response to the comments below.

More broadly, we have substantially revised the manuscript in response to all reviewers' comments. We clarified the construction and interpretation of the G3P groundwater product and its component-related limitations, we have strengthened the description of the statistical trend and uncertainty framework, refined the role and interpretation of the clustering analysis, and made the spatial aggregation, masking, and area-weighting procedures more explicit. We also expanded the sensitivity assessment, including the 95% significance analysis and the new all-cell aggregates, revised the interpretation of continental and global storage changes, and we further strengthened the comparison with previous global and regional studies. Corresponding changes were implemented consistently throughout the Methods, Results, Discussion, Supplement, tables, and conclusions. We have slightly adapted the paper title by adding "Continental" and "budgets", to better represent these prominent aspects of the paper.

I have a few recommendations, as noted next.

1. As a reader, I am lost to what the main novelty of the paper is as many previous references exist that determine long and short term variability using GRACE/FO data. I think it resides in determining trends using G3P, globally and for an extended time-period (21+years). I recommend the authors to simplify and reinforce what are the key novelties.

   We agree and have sharpened the novelty in the Abstract, Sect. 1.5, the opening of Sect. 3.2, Table 3, and the Conclusions. The contribution is now stated more directly as a first globally consistent assessment of statistically significant GWS trends from G3P over the extended 21.5-year GRACE/GRACE-FO record. The added value is the utilization of the G3P product in a consistent stochastic-memory-aware trend framework and the aggregation of significant grid-cell trends into continental and global groundwater gain, loss, and net storage-change budgets. We also distinguish now more clearly between our

original G3P-based results and the literature synthesis used for regional interpretation. This revision is related to Referee #2's request to clarify the meaning of an "uncertainty-aware" assessment.

2. The reviewer (as well as other readers) might not as familiar with many of the statistical terms/methods that are described in 2.2. I recommend streamlining or explain in a more simple words these as much as possible and try to the best of their possibilities to make these statistics methods more easily digestible.

   We agree and have revised Sect. 2.2 to make the workflow more accessible before introducing the formal equations. The section now explains, step by step, that for each grid-cell time series we estimate a trend and periodic terms, model the temporal correlation of the residuals, derive the trend uncertainty from the calibrated stochastic model, test whether the trend differs significantly from zero, and then aggregate significant trends volumetrically. Figure 1 provides an intuitive comparison of the considered stochastic processes. We also reduced statistical abbreviations that were not used repeatedly. OLS and GLS were retained because they are used throughout the spectral analysis and the explicit comparison in Sect. 3.4. These changes also address the related abbreviation and readability comments by Referee #2.

3. While I think the cluster analysis (fig 2) makes sence from a statistical prespective, I am finding hard to connect or link it to a physically meaningful hydrological meaning. Can the authors be more explicit about why a specifc trend in one ares can be represented the same way as the trend in an other area? If we then show higher resolution trends (Fig 4), whats the needs to report the cluster analysis? Again, it is not clear what is the added value of the cluster analysis, please clarify or perhaps move to a supplementary text?

   We agree that the clustering should not be interpreted as hydrological attribution. The revised Sect. 3.1.1 now defines it explicitly as an exploratory grouping of time-series structure. Regions with different physical drivers can therefore belong to the same cluster when their seasonality, interannual variability, or trend-like behaviour is statistically similar. The clustering is not used for significance testing, volumetric aggregation, or causal interpretation; these are based on the grid-cell trend analysis in Sect. 3.2. We shortened the main-text discussion, retained the GWS clustering as a compact overview of the dominant temporal regimes, and moved the detailed TWS clustering to Sect. S2.1 and Fig. S1. Descriptive cluster names were added, as also requested by Referee #2, and the link to the subsequent significance analysis is now stated explicitly.

4. In order to determine GWS, G3P removes single components of SWS, SWE, etc. While I think most of these are satellite derived, SWS is model based (according to 191-194) if so, how are human driven (diversions, dams, etc) fluxes/storages represented?

We have clarified this point in Sect. 2.1. The SWS component is derived from the OS LISFLOOD/GloFAS modelling framework and represents storage in lakes, rivers, wetlands, and reservoirs. The revised text does not imply that all managed processes are represented completely. Instead, it states that while reservoir operation and regulated river or lake storage are represented in the underlying models, they are certainly not perfectly represented but rely on the quality of the model input data and process representation. Because GWS is obtained residually, unrepresented or imperfectly represented managed surface-water changes can alias into the GWS estimate. We added a sentence in the Data section clarifying that. This limitation is also considered where relevant in the regional interpretation.

5. Along the same line, if only root zone soil moisture is used (182-185) how are irrigation driven changes of the SURFACE soil moisture being accounted for? Please add caveats about this.

   The RZSM component used in G3P is derived from satellite surface-soil-moisture observations, followed by gap filling and propagation to the root zone. Irrigation-related surface wetting therefore enters the product where it is detected by the satellite observations. However, it may be under-represented (e.g., when there is a larger time gap between the irrigation event and the satellite overpass) and can be attenuated during root-zone propagation. We added this limitation to Sect. 2.1 and mention that missing or imperfectly represented irrigation signals can propagate into residual GWS trends, particularly in strongly irrigated regions. This complements the SWS clarification in our response to Comment 4.

6. Section 3.2 is the core of the paper in my opinion. I would emphasize it by adding a table to summarize also how the authors results agree or disagree with previously reported analysis. This will also help with framing the novelty point (above).

   We agree that Sect. 3.2 is central to the manuscript. To address the reviewer's comment we added a new Table 3 to the manuscript. It synthesizes the principal global and regional G3P-based findings, including the net and gross balances, the dominant Asian losses and compensating African gains, pronounced regional heterogeneity, and the main interpretations supported by the regional literature. It also summarizes the findings on trend detectability, component separation, spectral behaviour, and the OLS–GLS comparison. Because published estimates differ in period, spatial mask, GRACE processing, component separation, leakage and GIA treatment, and significance criteria, we do not present a direct numerical ranking. Instead, the discussion and table distinguish findings that are consistent with independent evidence from those whose origin remains unresolved. We additionally repeated the aggregation at the 95% confidence level as a sensitivity assessment.

7. Finally, I found some typos here and there, please re-read accordingly to identify any grammar issues. Example: 111: "This [IS] a strong.." 715-716: some words are missing perhaps? 725: fix parathesis in the

We have revised the manuscript and Supplement for grammar, terminology, unit consistency, and formatting. The specific examples identified by the referee were corrected: the missing verb was restored; the awkward expression “long-period-transient-type apparent-secular trends” was replaced by “apparent secular trends caused by long-period transient variability”; and the incomplete wording and parenthesis in the Africa subsection were corrected. We additionally revised awkward formulations in the Methods and clustering sections, harmonized storage-component names and units, and checked the manuscript–Supplement cross-references. This broader language revision also addresses the general language and nomenclature comment by Referee #2.

Referee #2

The article entitled "Water storage trends derived from the GRACE/-FO global gravity-based groundwater product (G3P)" discusses prospects and limitations of the newly available satellite-based global groundwater change assessment. I find the paper interesting and well within the scope of the journal. However, I also noted a few weaknesses that should be thoroughly addressed through a revision.

We thank you for the constructive assessment of our manuscript. We have revised the manuscript and Supplement in response to the comments below.

More broadly, we have substantially revised the manuscript in response to all reviewers’ comments. We clarified the construction and interpretation of the G3P groundwater product and its component-related limitations, we have strengthened the description of the statistical trend and uncertainty framework, refined the role and interpretation of the clustering analysis, and made the spatial aggregation, masking, and area-weighting procedures more explicit. We also expanded the sensitivity assessment, including the 95% significance analysis and the new all-cell aggregates, revised the interpretation of continental and global storage changes, and we further strengthened the comparison with previous global and regional studies. Corresponding changes were implemented consistently throughout the Methods, Results, Discussion, Supplement, tables, and conclusions. We have slightly adapted the paper title by adding “Continental” and “budgets”, to better represent these prominent aspects of the paper.

I think that it would be good to find indicative names for the clusters depicted in Fig. 2. Using the colors is not really helpful, since those can only be used with the Figure. There are common features in most of the clusters, and I strongly urge the authors to come up with a narrative to distinguish those groups.

We agree that references to clusters by colour or number alone were insufficient. We revised Fig. 2 byadding descriptive names based on the dominant time-series characteristics. The labels now distinguish, for example, background variability, high-latitude seasonal/interannual variability, moderate GWS drying, positive GWS

tendencies, strong GWS decline/cryosphere-adjacent signals, and Amazon high-amplitude seasonality. We also clarify that these are temporal-regime labels rather than hydrological attribution classes: regions with different physical drivers may belong to the same cluster if their GRACE/GRACE-FO time series are statistically similar. The clustering discussion was shortened, the detailed TWS clustering was moved to Sect. S2.1 and Fig. S1, and the connection to the subsequent trend-significance analysis was strengthened. This revision also addresses Referee #1, Comment 3.

There are quite a few inconsistencies in the nomenclature (e.g., identical units are abbreviated in different ways) and various rather sloppy formulations. Some examples are given below. A careful revision of the full manuscript in terms of language and style seems urgently necessary.

We performed a manuscript-wide consistency revision of terminology, storage-component names, unit notation, figure captions, table headings, and cross-references. Volume-change rates and equivalent-water-height trends are now reported consistently, and awkward or incomplete formulations were revised. The specific examples listed below were addressed individually. This revision also addresses Referee #1, Comment 7.

The paper is heavily relying on abbreviations that are introduced at some point but later hardly used again. This applies in particular to abbreviations of certain mathematical methods, which have hardly any value apart from expert readers who might recognize the abbreviations from earlier encounters outside this manuscript. I think that this should be changed to make the article better readable.

We reduced abbreviations that were not required for repeated use or reproducibility and wrote out restricted maximum likelihood and the stochastic-model descriptions where appropriate. Core hydrological abbreviations (TWS, GWS, RZSM, SWE, and SWS) were retained because they are used throughout the manuscript. OLS and GLS were also retained because they are introduced in the spectral analysis and compared explicitly in Sect. 3.4. Their meaning is now explained at first use. The stepwise overview added to Sect. 2.2 further reduces the dependence on specialist terminology, as described in our response to Referee #1, Comment 2.
Additionally, we addressed the following individual points:

line 29: "close to persistent" is not very precise wording

The expression was removed and replaced by direct wording that distinguishes statistically significant trends from negative or near-threshold tendencies.

line 58 and 69: different ways for volume change units are given without an obvious reason. Please harmonize throughout the article.

We harmonized the notation for volume-change rates and equivalent-water-height trends throughout the main text, tables, and figure captions, and revised the corresponding notation in the Supplement.

line 84: There is an extensive body of (recent) literature from Donald Argus and colleagues on the use of GRACE data in California which should be cited here.

We added relevant California literature in Sect. 1.5, including Argus et al. (2022), together with Thomas et al. (2017), when introducing established GRACE/GRACE-FO evidence for groundwater and water-storage loss in California.

line 109: "long-period-transient-type apparent-secular trend" is a very much German construct. Please revise.

The sentence was revised to refer to “apparent secular trends caused by long-period transient variability in records of limited duration”. The following sentence was also corrected to “This provides a strong motivation…”. This is the same passage noted by Referee #1, Comment 7.

line 148: I am still not sure what an "uncertainty aware" assessment should be. Please elaborate and explain better.

We clarified this term in the Abstract, Sect. 1.5, Sect. 2.2, and the Conclusions. In the revised manuscript, “uncertainty-aware” means that trends and periodic signals are estimated together with a calibrated stochastic model of the residual temporal covariance. The resulting trend uncertainties account for persistent seasonal and interannual variability and hydrological memory and are used in the significance test and the significance-filtered storage-change budgets.

line 152: Two times "limited" in a single line

The sentence was rewritten and the repetition removed.

line 155: demonstrated scalability has not been achieved, I believe. Please re-consider your statement.

The scalability claim was removed. The revised text describes the application to the global grid without presenting computational scalability as a demonstrated scientific result.

line 165: I think that this is rather a sequential than a hierarchical approach.

We agree. Sect. 2.1 now describes the GRACE/GRACE-FO and G3P processing as a sequential chain from level-2 gravity fields to level-3 TWS and subsequently to level-4 residual GWS.

line 199: It should rather read w.r.t. the temporal mean over the reference period. Same applies to line 204 (and elsewhere).

The revised Sect. 2.1 now states explicitly that all storage variables are treated as anomalies with respect to their temporal mean over the reference period.

line 236: pre-analysis sounds wrong here

The term “pre-analysis” was removed. The clustering and spectral work are now consistently described as exploratory analyses that characterize the data and motivate the stochastic models.

line 238: I think that overtones of the annual signal are meant here. Harmonics basically can have any frequency.

The Methods now state that the functional model contains the annual signal and its first five overtones, i.e. the second to sixth harmonics, with frequencies f_j = j/365.25 d^{-1} for j = 1, …, 6.

line 266: "gaps as-is" is a non-information. Please explain or remove.

The treatment of missing observations is now stated explicitly: missing monthly GRACE/GRACE-FO solutions, including the mission gap, are retained as missing epochs and are not interpolated. The analysed period and the number of available monthly observations are also specified.

line 277 and following: Various abbreviations listed here could be omitted since those are never used again: GLS, RMLE, MLE, OLS, PLWN.

Single-use abbreviations were removed or written out. OLS and GLS were retained because they are used repeatedly and form the basis of the sensitivity comparison in Sect. 3.4. The stochastic-model combinations are described in words at first use.

line 301: space are missing in the equation.

The mathematical notation and spacing in the significance-test equations were corrected.

line 311: "surface is calculated geodetically" sounds quite blatant for a trivial latitude

weighting.

The aggregation description now states that grid-cell areas are computed using the WGS84 ellipsoid, which provides the area weighting used in the continental and global aggregation.

line 313: what is a "direct physical integration"?

The wording was replaced by an explicit description of the spatial summation: significant grid-cell trends in equivalent water height are multiplied by their grid-cell

areas, converted to $km^3\ yr^{-1}$, and summed separately for gains and losses; their algebraic sum defines the net change.

line 625: why do you use bold font in the subsection on Asia only?

The inconsistent bold emphasis was removed and subsection formatting was harmonized across the regional discussion.