

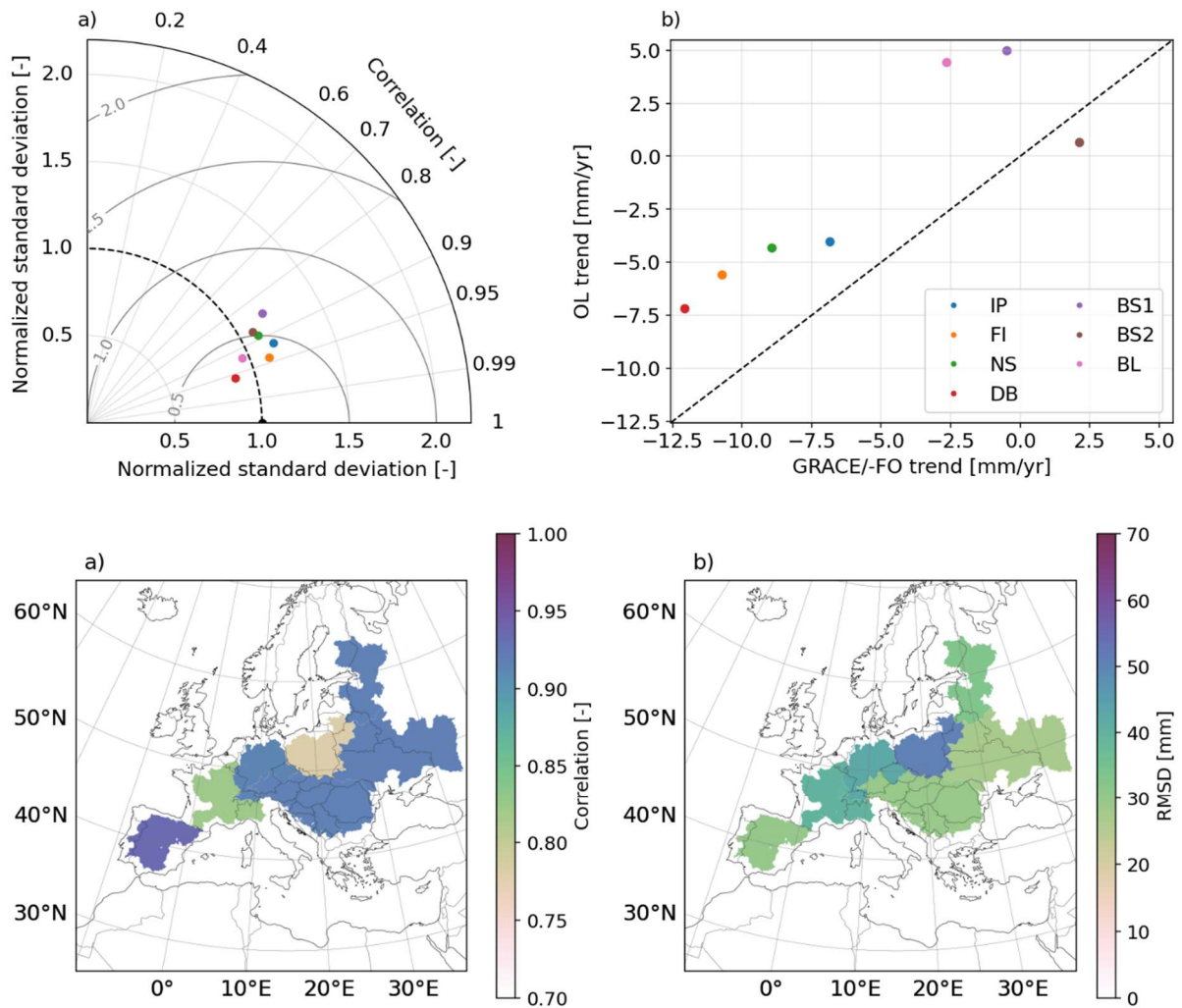
Reviewer 2 (Laura Jensen)

We would like to thank the reviewer for the positive evaluation of our manuscript and the very favorable reception of our dataset. Below, we have addressed the specific remarks and comments.

Major comments:

1. The authors apply the same benchmarking procedure as proposed in Klos et al. (2023), where GRACE and GLWS data were used as the benchmarking data sets. Here, they rely on another hydrological model, eCLM. However, I am wondering if the authors have evaluated the suitability of this model for representing TWS variability, or if there are any references that assess the quality of eCLM w.r.t. TWS. I assume that the benchmarking results heavily depend on the quality of the hydrological model data. Thus, I would also recommend to investigate to which extent the results are robust against using another hydrological model in the benchmarking procedure. This part of the manuscript should be extended.

Response: Many thanks for this important question. Indeed, we evaluated our model extensively for a range of different variables, including a comparison with GRACE/-FO-derived terrestrial water storage anomalies (TWSA). At the catchment scale, the modeled TWSA shows very good skill in terms of correlation and RMSD, with correlation coefficients above 0.9 for most catchments, and around 0.8 for the Eastern European catchments. RMSD ranges between approximately 30 and 50 mm (see figures below). We have also performed extensive experiments with GRACE/-FO data assimilation. Comparisons of the open-loop and assimilated simulations with different observation-based datasets already showed a good representation of hydrological extremes in the standard model run. In addition, we evaluated the model against in-situ soil moisture observations and superconducting gravimeter measurements and found that the model dynamics were remarkably well represented. Based on these results, we are confident that the model is a reasonable choice for our application. In addition, we decided to use this model because it provides higher spatial resolution than most global models and includes a sophisticated soil hydrology scheme. Nevertheless, we agree that it is worthwhile to consider alternative model products when evaluating modeled TWSA. We would expect the models to show broadly similar benchmarking results on seasonal temporal scales, while more pronounced differences may emerge on shorter and longer temporal scales. We therefore plan to perform a more detailed analysis based on an ensemble of different hydrological and land-surface models. Such a comprehensive multi-model benchmark, however, is beyond the scope of the present paper. We have added a sentence to the paper to clarify this aspect.



2. Section 3.1 discusses the results for 11 regions separately. Line 210 – 212 mentions that the regions are characterized by a similar amplitude of power-law noise and a similar spectral index, and refers to the Supplementary material for further information. However, I did not understand from that material either, how the region outlines are defined? There is only one sentence about it: “By combining both these estimates, we group the GPS stations regionally into 11 regions, ...” How are the estimates combined, how is the grouping conducted?

Response: We agree that this description is currently too brief. We combined the power-law noise amplitude and spectral indices estimated for a set of GPS displacements using “k-means clustering” approach. We added this information to the Supplementary material.

3. In Section 3.1 the reflection of droughts and floods in the three data sets are objectively described for each region. Periods of droughts and floods are marked in blue and orange in the figure, but it is not clear to me from which data source these blue/orange periods come? Is it from the literature or from an index data set? Please clarify this. Furthermore, I miss a critical discussion on the reasons why for some time periods the different data sets agree well and for some not. Specifically, it is often mentioned that some variations are not captured by SLR and DORIS, but no further conclusions are derived from that. In my opinion, the description of the

individual regions could be shortened, in favor for a more extensive discussion of the overall conclusions to be drawn from this analysis.

Response: The blue and orange shading periods in Figure 2 do not result from any of the three plotted datasets (GPS, SLR+DORIS, soil moisture), nor from a single hydrological index. These are taken from independent, region-specific literature. For each of the 11 regions we described, we collected papers mentioning these regions and reporting the timing of major dry and wet events. These papers are cited in the text describing temporal changes in a given region, e.g. Arheimer and Lindström (2015) for the 1995 floods in Sweden; Spinoni et al. (2015) for pan-European drought indices; Ingver et al. (2010), Walz et al. (2018), Ducros et al. (2025), Berbel and Esteban (2019) and the other region-specific references already cited throughout Section 3.1.

In the revised version of the manuscript, we decided to keep the detailed regions description but we added a paragraph discussing the possible sources of the discrepancies we found between the displacement time series derived in different measurement methods as follows:

“Discrepancies between the time series shown in Figure 2 may result, among other things, from: (i) different spatial resolutions represented by the data, (ii) different temporal resolutions of the data, (iii) the detection threshold for a given event by a specific system, or (iv) the presence of non-hydrological signals. The SLR+DORIS gravimetric solution is limited to spherical harmonic degrees and orders up to 60, which gives it an effective resolution of several hundred kilometers, similar to that of the GRACE satellites. Mass changes confined to a smaller area, such as local flash floods, agricultural droughts within sub-basins, or anomalies in individual watersheds, are smoothed out; however, regionally averaged GPS displacements are sensitive to such signals with limited spatial extent. Short-duration events lasting from a few days to a few weeks, e.g., flash floods, short-term droughts caused by heat waves, are strongly suppressed or undergo a phase shift when averaged into the monthly product. Consequently, SLR and DORIS techniques, as well as GRACE, may fail to detect them, whereas daily GPS measurements do. Furthermore, the SLR and DORIS techniques are most sensitive to mass anomalies with a wide spatial extent and large amplitudes, whereas regional signals with smaller amplitudes may remain undetectable but are measurable using GPS. On the other hand, some anomalies detected by GPS that have no corresponding SLR/DORIS or SM data series may contain a residual non-hydrological component, e.g., local station-specific effects that were not fully removed during preliminary processing. These are consistent with specific discrepancies we have already reported, such as the drought of 2006-2007 in Sweden (Region 1), the floods of 1997 and 2010 in Poland (Region 7), and the droughts of 1995-1996 and 2004-2005 in the United Kingdom and Spain (Regions 9 and 11).”

References:

Arheimer, B., and Lindström, G.: Climate impact on floods: changes in high flows in Sweden in the past and the future (1911–6352100). Hydrol. Earth Syst. Sci., 19, 2, 771–784, <https://doi.org/10.5194/hess-19-771-2015>, 2015.
Berbel, J., and Esteban, E.: Droughts as a catalyst for water policy change. Analysis of Spain, Australia (MDB), and California. Glob. Environ. Change, 58, 101969, <https://doi.org/10.1016/j.gloenvcha.2019.101969>, 2019.

Ducros, G., Tiggeloven, T., Ma, L., Daloz, A.S., Schuhen, N., Claassen, J., and de Ruiter, M.C.: Multi-hazards in Scandinavia impacts and risks from compound heatwaves, droughts and wildfires. *Nat. Haz. Earth Syst. Sci.*, 25, 11, 4693-4712, <https://doi.org/10.5194/nhess-25-4693-2025>, 2025.

Ingver, A., Tamm, I., Tamm, U., Kangor, T., and Koppel, R.: The characteristics of spring cereals in changing weather in Estonia. *Agronomy Research*, 8(III), 533-562, 2010.

4. Line 478 – 481: “we clearly see” To be honest, I don’t clearly see the noise reduction from Fig 4. I suggest to add a direct comparison between the benchmarked and the original (full) data set for each of the components (short-term, seasonal, long-term), maybe in an additional figure. This could go along with my comment above, the usage of a different hydrological benchmarking data set. This would also further highlight the advantage of applying the benchmarking procedure and emphasize the quality of the new data set promoted in this study in comparison to using the full GPS station data.

Response: We agree with the Reviewer that Figure 4 alone may not clearly indicate the noise-reduction argument sufficiently visible, since panel (a) uses the full/no-benchmarking dataset while panels (b)-(d) each show only the benchmarked result at one temporal scale. We have softened the wording at lines 478-481 (removing “we clearly see”).

However, we decided not to add any new figure, because plotting the original dataset with its individual components would hide the changes we want to highlight. In addition, it’s worth noting the y-axis scale: for the short-term temporal scale, the y-axis scale is significantly smaller, so in a figure with the original time series, the short-term changes wouldn’t be visible at all.

Regarding the term “noise reduction,” we would like to clarify that by “noise” we mean all changes that do not belong to a given temporal scale; thus, when discussing, for example, the seasonal temporal scale, “noise” refers to both the long-term and short-term temporal scales.

5. In the summary, I miss a few sentences about the limitations (or accuracy) of the data set.

Response: We agree and added a paragraph to Summary explicitly discussing the main limitations of the dataset, which includes: (i) the reliance on a single hydrological model for the benchmark classification, (ii) the reduced reliability of the short-term temporal scale, (iii) the heterogeneous and generally low station density before 2005, and (iv) the sensitivity of the long-term trend estimates to the length of time series.

Specific comments

1. Line 25: I would I would avoid using references in the abstract. Half-sentence “even for the period of 8 years prior to GRACE” is kind of redundant to the period 1994 to 2023 mentioned four lines above. I suggest to remove it.

Response: Removed.

2. Line 71: “see Copernicus Climate Change Service” is a rather unspecific reference, please make it more specific to substantiate the statement.

Response: Corrected.

3. Line 183/184: I don’t understand “summing the remaining components”. Isn’t it just Level 9 that is remaining for the long-term signal? Please clarify.

Response: Yes, indeed. Thank you for pointing this out. Corrected.

4. Line 189/190: I don’t understand this step. I thought that all stations with data coverage less than one year were already excluded in step 1? Why does the number of stations further reduce now?

Response: We mean here a coverage between the GPS displacements and the CLM5 model. The GPS displacements may be longer than 1 year, but if they cover less than 1 year of a common time frame with CLM5, this station is removed. This is why the number of stations is further reduced.

5. Line 199: I was confused by the reference until I realized that this is the data publication belonging to this study. Maybe mention it in a sentence here, e.g.: “These data sets are made publicly available in Klos and Bogusz (2026).”

Response: Added.

6. Figure 2: It would be nice if the map with the region outlines could be increased in size.

Response: A larger version is presented in Supplementary materials, Fig. S1. Since that may not have been clear in the previous version, we have also added a comment in the caption of Fig. 2.

7. Line 448/449: Do you have any idea why for the long-term variability the correlations are so inhomogeneous? There are also many strong (probably significant) negative correlations right next to strong positive correlations. This phenomenon should be further discussed.

Response: We thank the reviewer for this comment. We attribute the variability in the correlation coefficients between the long-term temporal scales primarily to differences in the length of the common period between GPS and InSAR at individual stations. The InSAR dataset spans about 7 years (2017-2024), but the length of GPS displacements changes from station to station, which means that for two neighbouring stations, the length may be completely different, and it will also influence the correlation coefficient.

8. Line 510: “a heat wave” I wonder if it was just a single heat wave or more than one? If only one, please specify the year and provide a more specific reference.

Response: There were several heat waves that affected the Alpine region. As shown by Jacquemart et al. (2024), the frequency and intensity of summer heatwaves and droughts have increased, as observed in 2003, 2015, 2018, and 2022, even at higher elevations, which enhances the climate changes signal. We added this comment to the manuscript.

Reference: Jacquemart M., Weber S., Chiarle M., Chmiel M., Cicoira A., Corona C., Eckert N., Gaume J., Giacona F., Hirschberg J., et al. (2024): Detecting the impact of climate change on alpine mass movements in observational records from the European Alps, Earth-Science Reviews, 258, 104886, <https://doi.org/10.1016/j.earscirev.2024.104886>.

9. Line 556 – 568: This paragraph is a bit unclear to me. “do not provide them separately” Do you mean in the publication of the data set? The following sentences about different regions: Iceland, etc, do they refer to the sub-periods or to the full period? The last sentence does not seem to be connected to the others. Furthermore, the paragraph ends quite abrupt. I suggest to integrate this paragraph into the previous paragraphs about the four periods and to end the section with a clear conclusion.

Response: Thank you for this comment. Indeed, the phrase “do not provide them separately” sounded a bit unclear. It is now reformulated into: “For this reason, we do not provide trends for the subperiods in a separate file within the benchmark dataset, but we do show their consistency with the GRACE and GLWS trends over the years.” We also integrated the last paragraph into the earlier paragraphs, and believe, that the text is more clear now.

10. Line 561: “mismodelling of GIA effect”. I suggest to apply the same GIA model to the GRACE, GLWS and GPS data sets. Then the trend might still be affected by a suboptimal GIA model, but at least it could be excluded that the differences in the three data sets come from the GIA model.

Response: The same GIA model was applied to all datasets we use. The comment about mismodelling concerns the mismodelling in general, and not the mismodelling due to applying different GIA models.

Technical corrections:

Response: All technical corrections were included.

1. Line 156: how are the additional offsets estimated?

Response: the additional offsets are estimated via manual checking of the displacement time series.

2. Line 203: why is the analysis preliminary?

Response: We always called it preliminary, because this is not the main analysis we do, and because it only results in removing unnecessary effects or correcting for unwanted phenomena.