

Review of manuscript: Homogenization of integrated water vapour time series using twin GNSS stations and vertical step offsets, by Tong Ning and Gunnar Elgered.

General comments

The manuscript concerns GNSS atmospheric measurement, processing-induced discontinuities, method intercomparison, validation against ERA5, and homogenization of long-term I WV records.

The study has a useful dataset and a potentially valuable result: co-located GNSS stations provide a much cleaner reference for estimating instrumental I WV offsets than ERA5. The observed relationship between estimated height and I WV discontinuities is also interesting.

The manuscript meets a basic scientific-quality threshold in that the problem is relevant, the observations are appropriate, and the principal empirical result is plausible. Nevertheless, important methodological information is missing, several conclusions are stronger than the evidence supports, and there are significant numerical and internal inconsistencies.

I recommend major revision rather than rejection on scope or fundamental-quality grounds. The central results are plausible and relevant, but the manuscript needs a reproducible statistical methodology, uncertainty analysis, sensitivity testing, correction of internal numerical inconsistencies, and a substantially more cautious interpretation of the height–I WV regression before its principal conclusions can be considered scientifically supported.

Major scientific issues

1. The homogenization method is not described reproducibly

Section 3.1 states that offsets are calculated from mean values before and after an intervention, but the procedure remains underspecified. The authors should provide equations defining:

- the GPS–GPS and GPS–ERA5 difference series;
- the subtraction order and therefore the sign convention;
- whether hourly, five-minute, or daily values are used;
- temporal matching and averaging;
- requirements concerning missing data and minimum data completeness;
- the exact definition of the pre- and post-intervention windows;
- treatment of seasonal cycles, trends and serial correlation;
- propagation of the offset correction through multiple interventions;
- uncertainty estimates for every offset.

Simply comparing two sample means may be biased if the windows contain trends, unequal seasonal sampling, or autocorrelated residuals. The uncertainty of each value in Table 2 is essential.

2. The “twin-station” assumption and equipment-change catalogue need verification

The GPS–GPS approach assumes that the companion station is stable at the intervention epoch. The manuscript should explain how the affected station is identified and how stability of the reference station is verified.

This is especially important because:

- both stations use the same processing strategy and external products;
- common processing changes could affect both stations;
- local multipath or environmental changes need not be common;
- the stations are separated by as much as 10.9 m vertically/horizontally as described, so small real atmospheric differences may remain.

More generally, the catalogue of station changes does not appear to have been constructed systematically. Table 1 is restricted to antenna and radome changes, whereas Table 2 and Figure 3 include a receiver change at BOR0. Receiver replacements, firmware upgrades, cable changes, and changes in tracking capability may also affect coordinate and tropospheric-delay estimates. The authors should compile all documented equipment and processing changes for both members of every twin-station pair and explain which events were tested, included, or excluded. Receiver changes should be evaluated using the same procedure as antenna and radome changes rather than considered only when an obvious offset is already visible.

Table 1 should therefore be expanded, or complemented by another table, to report station separation and height difference, all relevant equipment changes at both stations, the station to which each detected discontinuity is attributed, the available pre- and post-change observations, and the analysis windows. This information is necessary both to assess the completeness of the event catalogue and to verify that the companion station was stable at each analysed epoch.

3. The height–IWV relationship is partly a processing covariance

Height and ZTD are estimated simultaneously from the same GNSS observations. Their negative correlation is therefore expected from the observation geometry and parameter-estimation model. The relationship is not necessarily an independent physical confirmation of an IWV discontinuity.

The manuscript should distinguish between:

1. a real change in the antenna reference point or monument height;
2. an apparent coordinate displacement caused by antenna phase-centre offsets;
3. a ZTD–height covariance induced by the GNSS estimation process;
4. changes in multipath or other unmodelled errors affecting both parameters.

The statement that height offsets can directly correct IWV at arbitrary single stations is consequently too strong. The coefficient may depend on elevation cutoff, observation weighting, mapping function, satellite geometry, parameter constraints, antenna calibration, and processing software.

4. Regression analysis is incomplete

For Figures 5, 7 and 8, the authors should report:

- number of observations;
- fitted intercept and slope;
- standard errors or confidence intervals;

- p-values or bootstrap intervals for the correlations.

The scatterplots suggest that a few large coordinate jumps may strongly control the slope and correlation. A leave-one-out or robust-regression sensitivity analysis is needed.

The receiver intervention at BOR0 should also be treated separately from antenna/radome replacements, because it may represent a different error mechanism.

5. The choice of a 300-day or one-year window is not demonstrated adequately

The claim that approximately one year is required appears to be inferred visually from Figure 6. A systematic sensitivity analysis should compare offset bias or dispersion as a function of window length (for example: 30, 90, 180, 300, 365 and 730 days), using interventions for which long windows are available. This would provide direct evidence for the proposed minimum rather than confounding window length with intervention date, site and season.

Also clarify the inconsistency between “one year” and “greater than 300 days.” These are not identical criteria.

6. Trend validation and the interpretation of Figure 9 are insufficient

Correlation between the ERA5 and GPS trends is used as the principal measure of performance, but correlation alone does not establish agreement. Figure 9 appears to show two additional features that require quantitative analysis. First, by visual inspection, the mean GPS trend is approximately 0.25 kg m^{-2} per decade, whereas the mean ERA5 trend is closer to 0.5 kg m^{-2} per decade. Second, the station-to-station spread of the GPS trends appears substantially larger than that of ERA5. A correlation can improve while a systematic mean difference or an incorrect regression slope remains.

For every homogenization case, the authors should report the individual GPS and ERA5 trends and their uncertainties, the paired GPS-minus-ERA5 trend differences, mean and median difference, standard deviation or MAD, RMSE, regression slope and intercept, and confidence intervals. They should test whether the mean trend difference and the difference in spread are statistically significant. Trend uncertainties must account for temporal autocorrelation, and the trends should be calculated from identical temporal samples with the treatment of missing data clearly specified.

The larger GPS spread could reflect genuine local or regional variability that is smoothed in ERA5, but it could also indicate residual station-specific inhomogeneities, uncertain offset corrections, or other GNSS-processing effects. The manuscript currently does not distinguish between these explanations. The twin-station design offers a direct diagnostic: the trend difference between the two co-located GPS stations should be close to zero if both records are homogeneous, while their mean trend provides a less station-dependent value for comparison with ERA5.

The manuscript notes that ERA5 trends are generally larger than the GPS trends, but the present analysis cannot determine which dataset is responsible. Treating ERA5 as the validation reference implicitly assumes that it is closer to the true climatic trend. This assumption is not justified, particularly because ERA5 pressure and weighted mean temperature are used to retrieve GPS IWV, making the comparison partly non-independent. The evaluation of corrections

derived from GPS–ERA5 differences against ERA5 trends is even less independent and should be interpreted cautiously.

Attribution would require an independent third dataset or additional sensitivity tests. Where possible, the authors should compare with homogenized radiosonde, microwave-radiometer, or VLBI observations; repeat the analysis with other reanalyses; compare GNSS and modelled ZTD trends before conversion to IWV; or recompute GPS IWV using independent pressure and weighted mean temperature. Sensitivity to the beginning and end of the relatively short analysis period should also be evaluated. Spatial smoothing could be investigated by comparing the spatial variance or pairwise differences of GPS pair-mean and ERA5 trends sampled at the same sites.

Until these analyses are provided, the results should be described as showing improved statistical consistency with ERA5, rather than demonstrating that the corrected GPS trends are more accurate or that the remaining mean difference originates in ERA5.

7. Outlier removal is potentially too aggressive

A 1σ winter filter will reject a substantial fraction of normally distributed observations and may alter seasonal signals or estimated discontinuities. The authors should:

- define precisely whether the threshold is applied iteratively and to which residual model;
- state how many observations are removed at each station;
- demonstrate that genuine geophysical variations are retained;
- provide sensitivity tests using, for example, 2σ , 3σ , and a robust MAD-based criterion.

The phrase “seasonal sensitivity” is not a sufficient methodological description.

8. ERA5 processing needs more detail

The vertical adjustment of ERA5 IWV to antenna height must be described and referenced. The manuscript should specify:

- whether interpolation uses the nearest grid point or spatial interpolation—the manuscript currently states both;
- which ERA5 variables and levels are used;
- how surface pressure is adjusted to the antenna;
- how weighted mean atmospheric temperature is calculated;
- how IWV above/below the model surface is handled;
- the temporal interpolation method;
- whether ERA5 pressure and IWV are extracted consistently.

The pressure comparison at ONSA/ONS1 also requires a more quantitative assessment. Figure 2 appears to contain low-frequency variations or possible level changes of approximately 0.2–0.3 hPa, but the density of the hourly points makes these features difficult to evaluate. The claim that there are no concerning systematic effects or jumps is therefore not sufficiently demonstrated. The authors should examine the ERA5-minus-observed pressure differences for a trend, seasonal dependence, and possible changepoints, and discuss how any time-dependent pressure bias would propagate into GPS-derived IWV. In addition, a comparison at one site cannot establish the temporal stability of ERA5 pressure throughout the network.

9. Data availability is inadequate

“Will be registered and archived upon acceptance” limits reproducibility during review. The data underlying Tables 2 and Figures 2–9, together with the intervention catalogue and processing scripts or equations, should be made available through a persistent repository, subject to journal policy.

Important numerical and internal inconsistencies

These should be corrected before publication:

- Line 80: The statement that a pressure uncertainty of ± 0.2 hPa corresponds to only ± 0.014 kg m⁻² in IWV should be checked. Using a ZHD sensitivity of approximately 2.277 mm hPa⁻¹ and a typical ZWD-to-IWV conversion factor of 0.15 kg m⁻² per mm, the corresponding IWV uncertainty is approximately 0.07 kg m⁻²; a pressure variation of 0.3 hPa would correspond to approximately 0.10 kg m⁻². The value given in the manuscript appears too small by about a factor of five.
- Lines 171–172 give a regression slope of approximately -30.4 mm per (kg m⁻²). Its inverse is approximately -0.0329 kg m⁻² per mm.
- Line 200 incorrectly states “0.03 mm step offset in IWV”, but IWV is not expressed in mm in the rest of the paper; this should be approximately 0.03 kg m⁻² per mm of height, or even better 0.0329 kg m⁻² per mm of height, retaining more digits.
- Figure 9 caption repeats the incorrect unit and truncated value: “-0.03 mm IWV per 1 mm vertical displacement.” Replace it with -0.0329 kg m⁻² per mm of height.
- Lines 159–160 refer to a “relative ZTD error of 0.3 mm.” A quantity in millimetres is an absolute error, not a relative error.
- The Discussion reports a GPS–ERA5 correlation of +0.34 around line 225, whereas Section 3.3 and Figure 5 report +0.22. One value is incorrect.
- The manuscript alternates between a threshold of 300 days and one year. Use the exact applied criterion consistently.
- The stated analysis interval ends in December 2024, but Table 1 includes equipment changes in March 2025. Explain whether these entries are contextual only or whether observations extend into 2025.
- KIRO’s first change is dated 2016-10-31 in Table 1 but 2016-08-23 in Table 2.
- KIRO’s later change is dated 2023-10-17 in Table 1 but 2023-10-20 in Table 2.
- Table 1 uses “SEK8,” whereas Table 2 uses “SKE8.”
- Table 1 contains the impossible date “2022-03-031” for VIL0.
- “VAN0 1995-5-22” should use the same ISO date format as the other entries: 1995-05-22.
- Table 2 footnote d says “GPS-ERA”; it should be “GPS–ERA5.”

All intervention dates should be checked against station logs, not only typographically reconciled.

Figures and tables

- Figure 1: station labels are crowded and difficult to distinguish. Show twin stations as paired symbols or use one labelled site symbol with an inset/table giving station pairs.

- Table 1: it spans several pages without repeated column headings and is difficult to follow. Separate station metadata from intervention records, repeat headers, and clearly identify which interventions enter the analysis.
- Figure 2: superpose monthly means or medians, preferably with an interquartile or uncertainty band, so that the low-frequency behaviour and possible level changes can be assessed. If observations from both ONSA and ONS1 are included, distinguish the two series. The time period and number of matched observations should also be stated.
- Figure 3: define every plotted difference and its sign in the caption. Label intervention types and report the estimated offsets with uncertainties.
- Figure 4: the dense daily series makes individual removed observations difficult to assess. A residual panel or winter close-up would better demonstrate the effect of the filter.
- Table 2: include uncertainties, window lengths, number of observations, intervention type, and regression inclusion/exclusion flags.
- Figure 5: give regression equations, confidence intervals, sample size and intercept treatment. Identify the most influential points.
- Figure 6: a scatterplot of absolute GPS–GPS/GPS–ERA5 disagreement versus window length, with a 300-day reference line, would test the stated conclusion more directly.
- Figures 5 and 7: the axis orientation (height on the ordinate and IWV on the abscissa) makes the subsequently quoted conversion from height to IWV require inversion of the plotted regression. Either regress IWV on height or explain the inversion and its uncertainty.
- Figure 7 caption has a missing closing parenthesis and calls 300 days “approximately one year.”
- Figure 8: the two panels should use identical axis limits in both directions.
- Figure 9: define the orange and blue lines directly in the caption. Report the number of stations, trend period, regression equations, and mean and standard deviation of both sets of trends in each panel. Trend uncertainties should either be shown as error bars or reported in a supplementary table. A companion plot or table of paired GPS-minus-ERA5 trend differences would be more informative than correlation alone.

Language and typographical corrections

The terminology concerning station changes and discontinuities should be revised throughout. A documented “station change”, “equipment change”, or, if retained, “intervention” is an event at a known epoch; a “changepoint” or “discontinuity” is a feature detected in the time series; and a “step offset” is the estimated magnitude of that discontinuity. The manuscript currently conflates these concepts, for example when stating that interventions are referred to as step offsets or when referring to “correcting interventions”. “Intervention” may be retained for a documented human action, but it should be defined once and used consistently. Otherwise, the more direct term “equipment change” would generally be clearer.

Other examples include:

- Abstract: “step offsets ... significantly improve” → “correction of step offsets ... significantly improves.”
- Line 12: “global navigational satellite systems” → “global navigation satellite systems.”
- Line 22: “a shift ... that obscure” → “shifts ... that obscure” or “a shift ... that obscures.”

- Line 39: “such approach” → “such an approach.”
- Line 61: “based on that there is” → “based on the presence of.”
- Line 68: format the JPL URL correctly and cite the product/version.
- Lines 81–83: “no alarming systematics or jumps neither” → “no concerning systematic effects or jumps in either...”
- Line 85: “differences exhibits” → “differences exhibit.”
- Line 157: “antenna height of station” → “station antenna height.”
- Line 187: “This in done” → “This is done.”
- Lines 208: “is subjected to” → “is subject to.”