

Toward Rigorous Benchmarking of InSAR-Derived Vertical Land Motion: Resolution, Validation, and Uncertainty Assessment Along the U.S. Gulf Coast

Manoochehr Shirzaei^{1,2,3*}, Leonard Ohenhen⁴, Carmen Atkins¹, Oluwaseyi Dasho¹, Nitheshnirmal Sadhasivam¹, Nivedita P. Kamaraj¹, Florence Onyike¹, Olasunkanmi Olorunsaye¹, Esther O. Oyedele¹ and Susanna Werth^{1,2}

¹Department of Geosciences, Virginia Tech, Blacksburg, VA, USA

²United Nations University Institute for Water, Environment and Health, Richmond Hill, ON, Canada

³UNESCO Land Subsidence International Initiative (laSII), Place de Fontenoy 7, Paris, France

⁴Department of Earth System Science, University of California, Irvine, Irvine, CA, USA

*Corresponding author (email: shirzaei@vt.edu)

Abstract

Li et al. (2026) compare two InSAR-derived surface-elevation change datasets, reported by Ohenhen et al. (2024; hereafter O24) and Wang et al. (2024; hereafter W24) for the central U.S. Gulf Coast and report negligible pixel-by-pixel agreement outside urban areas, concluding that InSAR is unreliable in vegetated coastal settings for rates below 5 mm/yr. InSAR reproducibility is a timely and consequential question, and we share the paper's interest in resolving it, but much of the reported disagreement can be traced to how the comparison was constructed. The two datasets span non-overlapping observational windows (2007–2020 versus 2017–2020) in a system with documented non-stationary subsidence rates; validating each product against GNSS over its own window versus the other product's window shifts residual standard deviations by 19–22%, indicating that epoch mismatch alone accounts for a meaningful share of the apparent disagreement. Three further methodological choices compound this effect: (1) spatial aggregation of the O24 dataset from its native 50 m to 1 km before comparison, a ~400x reduction in pixel density that discards the sub-kilometer spatial structure the dataset was designed to resolve; (2) a progressively filtered GNSS validation network of only ~20 stations concentrated in atypical, stable Pleistocene upland settings, versus O24's original validation across 157 stations spanning the full coastal domain; and (3) a 5 mm/yr caution threshold derived from inter-product disagreement rather than from principled, instrument-specific uncertainty quantification. We validate O24 at its native 50 m resolution against 88 GNSS stations from the Nevada Geodetic Laboratory within the Li et al. study domain, obtaining a residual standard deviation of 1.6 mm/yr, consistent with Ohenhen et al. (2024). We use this exchange as a starting point for a broader set of best practices for validating and benchmarking InSAR-derived deformation products more generally, covering native-resolution evaluation, validation-network representativeness, accuracy-precision separation, scale-dependent uncertainty, and principled threshold construction, so that inter-product disagreement is neither conflated with measurement failure nor permitted to drive policy-relevant conclusions without rigorous independent validation.

1. Introduction

Li et al. (2026) address a genuine and consequential challenge: two recent InSAR-derived surface elevation change (SEC) datasets for the central U.S. Gulf Coast, Ohenhen et al. (2024) and Wang et al. (2024) yield similar regional mean rates yet show negligible pixel-by-pixel spatial correlation ($R^2 = 0.05$) outside urbanized areas. Taken at face value, this finding would have serious implications for the use of InSAR in coastal subsidence monitoring and sea-level-rise impact assessments.

We agree that reproducibility is a critical issue in this field, and we support the paper's call for harmonized processing frameworks and systematic cross-validation. However, the analytical choices underlying the

comparison introduce confounds of sufficient magnitude to cast doubt on the paper’s primary conclusion, namely, that InSAR is presently unreliable for capturing rates below 5 mm yr⁻¹ in vegetated coastal settings. In what follows, we document an unacknowledged epoch mismatch between the two products together with three further methodological concerns, and identify statements in the paper that read more negatively than the evidence the authors themselves present would support.

One point of fact is worth stating up front, since it bears on several of the arguments below. O24 is not a single-sensor, C-band-only product: it combines C-band (Sentinel-1, 2015–2020) and L-band (ALOS, 2007–2011) observations with GNSS in a joint stochastic inversion, specifically to reduce the vegetation-driven decorrelation that limits C-band-only InSAR in densely vegetated coastal settings. Arguments that treat O24 as inheriting Sentinel-1’s canopy-penetration limits alone will overstate its vulnerability to land-cover-driven signal loss.

2. Spatial Resampling of O24: A Non-Neutral Transformation

The comparison between O24 and W24 is not conducted at the native resolution of the O24 dataset. Rather, Li et al. (2026) aggregate O24 from its native 50 m pixel spacing to the 1 km pixel spacing of W24 by computing median SEC values within a 500 m search radius around each W24 centroid. This yields a ~400-fold reduction in spatial information density: each 1 km² pixel in the resampled product represents the central tendency of about 400 original observations, each of which may reflect distinct land-cover types, subsidence drivers, or coherence conditions.

O24 was designed specifically to resolve fine-scale spatial heterogeneity in coastal subsidence, including localized signals from differential sediment compaction, fault-zone deformation, and fluid extraction. Its GNSS-constrained stochastic inversion framework exploits the spatial density of InSAR measurements precisely to differentiate these signals. Collapsing this information to a 1 km grid does not simply reduce resolution; it destroys the sub-kilometer spatial structure that distinguishes O24’s design from W24’s. The aggregation also mixes land-cover classes that O24 resolves separately, including urban-wetland boundaries that are among the most dynamically and geodetically interesting features in the coastal zone.

The authors justify this procedure by noting that search radii from 100–400 m and 1000 m produced only minor differences. This confirms the aggregation is internally stable, but it does not validate the approach: consistency across aggregation radii simply means the spatial averaging has already erased the sub-kilometer variability at the smallest radius tested. The appropriate comparison would preserve O24 at 50 m resolution while spatially interpolating W24 to the finer grid, or, equivalently, report the correlation as a function of spatial scale to determine at which resolution, if any, the two products agree.

The consequence for the headline result is significant. The $R^2 = 0.05$ figure, the evidentiary cornerstone of the paper’s most far-reaching claims, may be partly or substantially an artifact of the asymmetric resampling, not a property of InSAR coherence in vegetated landscapes. Until the comparison is conducted at native resolution, or at a series of spatial scales, this cannot be determined.

3. GNSS Validation Network: Filtering, Scope, and Representativeness

Li et al. (2026) establish a ‘background VLM’ benchmark by progressively filtering 110 candidate GNSS sites to a final set of 36, of which only approximately 20 fall within the region covered by both InSAR datasets at 1 km resolution. The filtering logic, excluding Holocene-deposit sites, all Texas sites, and stations with elevated fluid extraction activity, is defensible in principle: the goal is to isolate GIA from anthropogenic and compaction signals. In practice, however, the resulting network raises representativeness concerns that the paper does not adequately address.

The exclusion of all 30 Texas GNSS sites is notably strict. The stated rationale is that 80% of Texas sites have nearby active wells. The remaining 20%, six sites without proximate extraction activity, located on the Pleistocene surface, are discarded without explanation. A more rigorous approach would apply extraction-volume-weighted or distance-weighted criteria to distinguish sites by degree of anthropogenic influence, rather than imposing a state-level binary exclusion. Retaining even a subset of unaffected Texas stations would extend the network westward and improve spatial coverage across the study domain.

More fundamentally, the 20 sites used to evaluate InSAR performance are concentrated in Pleistocene upland settings, by design, the most stable, least dynamic parts of the landscape. InSAR-derived SEC is most consequential in the adjacent Holocene coastal lowlands, where subsidence is fastest, vegetation is densest, and decorrelation noise is most severe. Its limitations are most relevant to policy. Demonstrating that InSAR does not closely match GNSS at stable upland sites does not constitute a validation failure in the coastal lowland context, where the datasets are primarily intended to be used. The physical processes, signal magnitudes, and noise environments differ substantially between these two settings.

The contrast with O24's own validation framework is instructive. Ohenhen et al. (2024) validated their dataset using 157 GNSS sites distributed across the full U.S. Gulf Coast, including settings spanning the full range of land-cover and subsidence conditions, yielding a standard deviation of 1.5 mm/yr for the difference between O24 and GNSS datasets. The narrower, filtered subset in Li et al. (2026) and the broader domain-wide network used here answer different, complementary questions: the former offers a conservative bound on long-wavelength systematic bias in a geologically simple setting, while the latter assesses performance across the range of environments the datasets are actually applied to. Generalizing conclusions from the narrower scope to the full coastal domain without noting that difference is a restriction of scope that the paper does not adequately acknowledge.

We recommend that the authors supplement their Pleistocene upland comparison with available Holocene-setting benchmarks, including rod surface-elevation table-marker horizon (RSET-MH) records, continuously operating tide gauge records, and additional GNSS sites, to assess whether the conclusions hold across the full range of coastal environments relevant to the paper's claims.

4. The 5 mm yr⁻¹ Caution Threshold: Derivation and Internal Consistency

Li et al. (2026) recommend that InSAR-derived vertical velocities below 5 mm yr⁻¹ be "interpreted with utmost caution" in regions lacking dense GNSS ground-truth. This threshold is derived from the 95th percentile of absolute SEC differences between O24 and W24 in medium-to-high-developed urban areas, the land-cover class where InSAR coherence is most reliable and inter-product agreement is strongest. The paper then extrapolates this urban-derived bound to non-urban vegetated environments where the noise environment is known to be considerably worse. This extrapolation direction is counterintuitive: if anything, the urban 95th-percentile figure represents a lower bound on uncertainty in vegetated settings, not an appropriate central estimate.

The threshold derivation has a second, more fundamental problem: it conflates disagreement between two methodologically dissimilar products with measurement uncertainty. O24 and W24 differ across four independent dimensions: sensor configurations (ALOS-1 + Sentinel-1 + GNSS vs. Sentinel-1 only), 3D displacement inversion vs. 1D line-of-sight-to-vertical conversion, temporal coverage (2007–2020 vs. 2017–2020), reference frame strategies, and native spatial resolution. Their disagreement at any given pixel, therefore, reflects a mixture of true observational differences (including legitimate temporal and spatial variability in SEC), methodological choices, and measurement noise. Treating the full inter-product spread as an uncertainty envelope assigns the consequences of methodological divergence equally to both products, regardless of which is closer to the ground truth.

The most acute problem, however, is internal inconsistency. The GNSS-derived background VLM rate that anchors the paper's GIA analysis, approximately -1.2 mm yr^{-1} , lies well within the range the paper simultaneously declares unresolvable. If InSAR-derived rates below 5 mm yr⁻¹ are to be treated with "utmost caution," the same epistemological standard should apply consistently. The paper does not explain why GNSS is exempt from this caution at the -1.2 mm yr^{-1} level while InSAR is not. The answer that GNSS achieves sub-millimeter-per-year precision under favorable conditions, whereas InSAR does not, is correct, but it is precisely the kind of instrument-specific uncertainty characterization that the paper fails to provide. A rigorous uncertainty analysis would separately quantify GNSS and InSAR precision at these signal amplitudes, rather than applying a single inter-product-disagreement-derived threshold across both measurement systems.

5. Statements in Li et al. (2026) and Their Quantified Basis in the O24 Evidence

Beyond the methodological concerns above, several statements in Li et al. (2026) describe O24's performance in stronger terms than the paper's own results support. Table 1 lists these statements alongside the relevant O24 evidence, the specific methodological choice or framing decision responsible for the gap, and our assessment of the likely impact on the paper's conclusions.

Table 1. Summary of statements in Li et al. (2026) that are inconsistent with O24 evidence or with the paper's own analysis, with classification and likely impact on their conclusions.

Li et al. (2026) Statement	O24 Evidence	Classification	Technical Concern	Impact on Conclusions
Negligible spatial correlation between datasets ($R^2 = 0.05$).	O24 was developed and validated at 50 m resolution using GNSS-constrained inversion.	Methodological choice, quantified effect	O24 was aggregated 50 m $\rightarrow$ 1 km prior to comparison. This $\sim 400\times$ reduction in pixel density is not a neutral transformation.	May substantially exaggerate inter-product disagreement and understate O24 performance.
InSAR is 'only reasonably robust in densely urbanized settings.'	O24 was validated broadly along the U.S. coast and applied beyond dense urban areas with low reported errors.	Framing stronger than evidence supports	The analysis does not isolate InSAR limitations from confounds: differing sensors, temporal windows, referencing strategies, and spatial support.	May lead readers to incorrectly conclude that coastal InSAR is broadly unreliable outside cities.
"Neither InSAR dataset fully captures the background VLM rate."	O24 demonstrates site-level GNSS correlation with $RMSE < 1 \text{ mm yr}^{-1}$ — acknowledged in the paper itself.	Framing stronger than evidence supports	The authors' own results show measurable agreement; 'does not fully capture' mischaracterizes imperfect agreement as failure.	Overstates the negative interpretation of the comparison.
"InSAR data are presently unable to capture this rate."	O24 validation demonstrates sensitivity to mm/yr-scale deformation signals.	Framing stronger than evidence supports	The evidence shows imperfect agreement, not inability. The claim is stronger than the data support.	May incorrectly imply a fundamental failure of InSAR methodology.
GNSS subset (110 $\rightarrow$ 36 $\rightarrow$ ~ 20 sites) used to establish background VLM benchmark.	O24 validation used a substantially broader GNSS framework across the full study domain.	Methodological choice, quantified effect	Progressive filtering excludes all Texas sites and concentrates the network on atypical stable upland locations.	Background VLM benchmark may not be representative of the broader coastal region.
5 mm yr⁻¹ uncertainty threshold recommended for non-urban InSAR.	O24 uncertainty estimates were derived from independent GNSS validation, not inter-product disagreement.	Methodological choice, quantified effect	Threshold is derived from disagreement between two methodologically dissimilar products rather than from a rigorous uncertainty model.	May substantially overestimate uncertainty, leading to unwarranted discounting of valid observations.
Inter-product comparison interpreted as a reliability test.	O24 and W24 differ in sensors (ALOS-1 + Sentinel-1 vs. Sentinel-1 only), methodology,	Methodological choice, quantified effect	Comparison conflates methodological and framework differences with accuracy. Disagreement may	May attribute substantive measurement differences to error rather than to

temporal coverage (13 vs. 4 years), referencing strategy, and resolution.	reflect legitimate observational differences, not error.	differing measurement frameworks.
--	--	---

Taken together, these statements illustrate how the methodological choices discussed in Sections 2–4 propagate into the paper’s language. Where O24’s own results show measurable GNSS agreement and low site-level RMSE, the paper describes this as a failure to capture background VLM rather than as imperfect agreement within a quantifiable margin; and where inter-product differences plausibly reflect methodological divergence rather than error, the paper attributes the full spread to uncertainty. Tightening this language to match the quantified evidence in Table 1 would not change the paper’s core empirical contribution, but it would bring its conclusions more closely into line with what the data support.

6. Re-evaluating O24 and W24 Performance within Li et al. (2026)’s Study Area

To provide an independent assessment of O24 performance, we obtained daily vertical displacement time series in the IGS14 reference frame at 88 GNSS stations from the Nevada Geodetic Laboratory (NGL; <https://geodesy.unr.edu/>) within the Li et al. (2026) study domain. Following standard processing procedures, including removal of offsets, outliers, and common-mode errors, we estimated vertical velocities and associated uncertainties for two periods: 2007–2020 (corresponding to the O24 observational window) and 2017–2020 (corresponding to W24).

Figure 1 compares these period-specific GNSS velocity estimates. The mean difference between the two periods is -0.3 mm yr^{-1} , with a 95th-percentile absolute difference of 3.7 mm yr^{-1} . Applying Li et al. (2026) own logic to this result is instructive: if inter-measurement spread at the 95th percentile defines a caution threshold, as the authors argue in deriving their 5 mm yr^{-1} recommendation, then the equivalent threshold derived from this GNSS-to-GNSS comparison is 3.7 mm yr^{-1} , below which the authors’ own GIA estimate of -1.2 mm yr^{-1} falls unambiguously. This *reductio ad absurdum* illustrates that the 5 mm yr^{-1} threshold, far from being a principled uncertainty bound, is an artifact of conflating inter-product methodological divergence with measurement error. No physically meaningful caution threshold can be derived by this approach without simultaneously invalidating the benchmark it is meant to protect.

To assess O24 accuracy directly, we identified all O24 pixels within a 50 m radius of each GNSS station and computed the median InSAR-derived VLM value at each site. Figure 2 shows the resulting comparison between O24 and GNSS rates across the study domain. The standard deviation of the residuals is 1.6 mm yr^{-1} (mean residuals of -0.1 mm yr^{-1}), closely consistent with the 1.5 mm yr^{-1} validation uncertainty reported in Ohenhen et al. (2024) and well within the uncertainty bounds acknowledged by Li et al. (2026) for the original 50 m product. These results affirm that O24, evaluated at its native resolution and against a spatially representative GNSS network, performs robustly across the study area, in direct contrast to the characterization offered by Li et al. (2026).

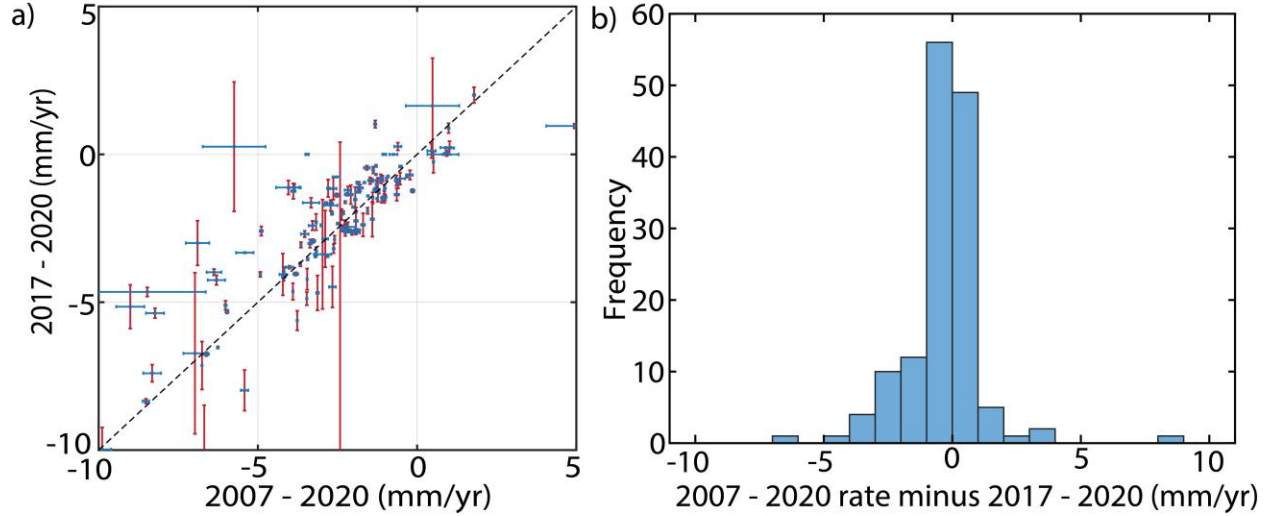


Figure 1. Consistency of GNSS-derived vertical velocities between the 2007–2020 and 2017–2020 observation windows. (a) Bivariate comparison of period-specific rates at each station; error bars denote 1σ velocity uncertainties, and the dashed line marks 1:1 agreement. (b) Histogram of per-station rate differences (2007–2020 minus 2017–2020), with a mean difference of -0.3 mm/yr and a 95th-percentile absolute difference of 3.7 mm/yr.

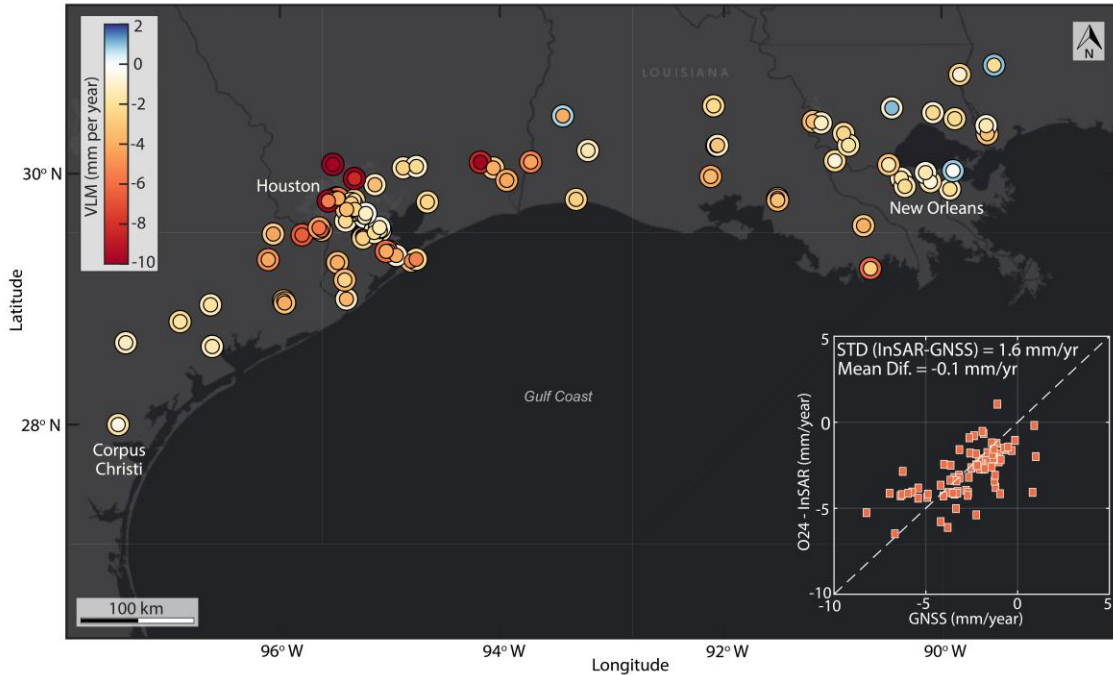


Figure 2. Comparison of Ohenhen et al. 2024 (O24) rate (filled circles) against GNSS vertical rate (open circles) for the 2007–2020 period. The inset shows a bivariate plot comparing GNSS vertical rates with O24 InSAR-derived rates. The standard deviation of the difference between the two datasets is 1.6 mm/year.

We extended this validation to W24 over their 1 km native resolution and 2017–2020 temporal window using the closest GNSS station to each pixel (Figure 3), obtaining a residual standard deviation of 2.3 mm yr^{-1} (mean residuals 1.1 mm yr^{-1}). When each InSAR product is instead validated against the GNSS temporal window it does not share, performance degrades. Validating W24 against GNSS rates over the

2007–2020 period raises the residual standard deviation from 2.3 to 2.8 mm yr⁻¹ (a ~22% increase) and shifts the mean difference to 1.6 mm yr⁻¹; conversely, validating O24 against the 2017–2020 period raises its residual standard deviation from 1.6 to 1.9 mm yr⁻¹ (a ~19% increase) and shifts the mean difference to -0.6 mm yr⁻¹. Each product, therefore, agrees most closely with GNSS over its own observational window and degrades against the mismatched temporal window.

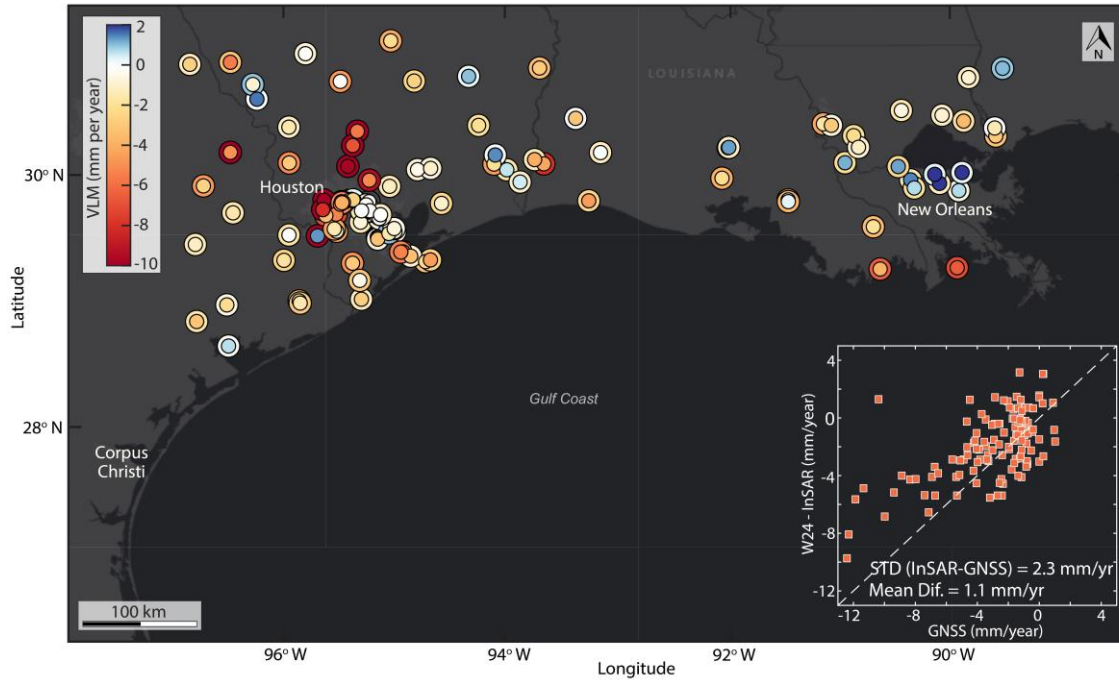


Figure 3. Comparison of Wang et al. 2024 (W24) rate (filled circles) against GNSS vertical rate (open circles) for the 2017-2020 period. The inset shows a bivariate plot comparing GNSS vertical rates with W24 InSAR-derived rates. The standard deviation of the difference between the two datasets is 2.3 mm/year.

7. Discussions

The exchange above concerns a single product and a single comparison, but the underlying methodological failures are not specific to Li et al. (2026); they recur throughout the InSAR literature whenever high-resolution products are evaluated against coarser or more limited benchmarks. We therefore use this discussion to set out general principles for validating and benchmarking geodetic deformation products, in the hope that they prove useful beyond the present exchange.

7.1. Best Practices for Validation and Benchmarking of InSAR-Derived Deformation Products

A first and largely unremarked distinction is between validation and benchmarking. Validation asks whether a product agrees with an independent source of ground truth, such as GNSS, precise leveling, or extensometer records, whereas benchmarking asks whether two or more products agree with one another. The literature, including Li et al. (2026), routinely substitutes the second for the first. This substitution is not innocuous: two products can disagree substantially while each remains correct within its own stated uncertainty, and two products can agree closely while both are biased relative to the true deformation field. A benchmarking exercise that never touches independent ground truth can establish that two products differ, and by how much, but it cannot on its own establish which product, if either, is right.

A second principle concerns spatial resolution. Products should be evaluated at their native resolution wherever possible, because resampling to a coarser grid changes the spatial variability, noise characteristics, signal amplitude, and correlation structure of the field being measured (see Hanssen (2001), on the scale-

dependence of InSAR error budgets). Aggregation therefore does not preserve the quantity under evaluation; it substitutes a different, smoother quantity for it, and any degradation attributed to the finer product following such aggregation may in fact be an artifact of the aggregation itself. Where aggregation cannot be avoided, its effect should be quantified explicitly, for instance with reference to the variogram range or the power spectral density of the field (Cressie, 2015), rather than justified by visual impressions of smoothness, which is not a resolution metric in any rigorous sense.

Related to resolution is the question of what a benchmark is for. There is no universal accuracy standard against which all products should be judged: a 1-km product may be entirely adequate for GIA assessment while being unsuitable for coastal subsidence mapping or infrastructure monitoring, where deformation gradients occur over tens of meters. The appropriate benchmark, given resolution, accuracy target, and validation network alike, should therefore be defined relative to the scientific application, not applied uniformly across applications for which it was never designed.

Validation data must also be independent of the product under test, in the sense that calibration stations should not be reused for validation and that shared processing assumptions or reference frames should not be allowed to artificially inflate agreement. Independence alone is not sufficient, however: the validation network must also be representative of the range of conditions the product is intended to characterize. A network restricted to a single, atypical setting, such as stable Pleistocene uplands, for instance, in place of the full validation network of Ohenhen et al. (2024), answers a narrower question than the one posed, and will systematically misstate performance in the coastal, wetland, and urban regimes that the product was built to resolve.

The literature also tends to conflate accuracy and precision, and more broadly repeatability, reproducibility, and uncertainty, treating them as interchangeable when they are not. A product can be precise, in close numerical agreement with another product, while systematically biased against ground truth, or comparatively imprecise while unbiased. These properties should be evaluated and reported separately. Uncertainty itself is rarely a single number, and it depends on coherence, land cover, temporal sampling, atmospheric delay, viewing geometry, topography, surface roughness, and the local deformation gradient, so that a single RMSE computed across millions of pixels can obscure more than it reveals. A coherence-weighted, spatially resolved uncertainty estimate, of the kind used in O24, is the more defensible standard, and should be treated as such rather than as an optional refinement. Scale dependence compounds this problem, as errors at 50 m, 500 m, and 1 km are not the same quantity measured with different amounts of noise, but different quantities, and any accuracy claim should state explicitly the scale at which it holds rather than being extrapolated from a coarse-grid comparison to characterize a fine-grid product.

A further requirement, and one frequently treated as optional rather than foundational, is that per-pixel uncertainty be propagated from the underlying phase observations through to the final deformation product, rather than estimated post hoc from residual scatter. Interferometric phase variance is a function of coherence, and this relationship is well characterized (Bamler and Hartl, 1998; Zebker and Villasenor, 1992). As coherence degrades, phase noise increases, and this noise propagates through unwrapping, atmospheric correction, time-series inversion, and, where applicable, LOS-to-vertical decomposition, compounding at each stage rather than canceling. A product's reported uncertainty is only meaningful if it reflects this full error chain, coherence-derived phase variance carried forward through every processing step to the final velocity estimate, rather than a single global RMSE assigned after the fact. O24 illustrates the alternative by its coherence-weighted stochastic inversion, which propagates phase-derived variance directly into the final vertical land motion estimate, yielding a spatially explicit uncertainty surface rather than a single summary statistic. Without a formally propagated, spatially varying uncertainty of this kind, a stated precision figure is not a measurement of confidence but an assumption, and any comparison against another product inherits that assumption uncritically. This is, in a strict sense, a precondition for benchmarking rather than a refinement of it. Two products cannot be meaningfully compared against a common uncertainty threshold if neither product's own uncertainty was propagated from first principles, since the comparison would then rest on two unknown and potentially incommensurate error budgets rather than on two independently quantified ones.

A further, closely related failure mode is the derivation of deformation thresholds from inter-product disagreement rather than from principled uncertainty analysis. A threshold constructed in this way, as in Li et al.'s proposed 5 mm/yr criterion, is circular by construction, since it measures how much two products disagree rather than how confidently either one can be trusted, and it becomes self-defeating when it is tighter than the paper's own independently derived quantities, such as a GIA correction of -1.2 mm/yr. Thresholds of this kind should instead be derived from confidence intervals and probability-of-detection analysis specific to the application at hand.

Finally, reproducibility is not incidental to benchmarking but a precondition for it. Processing parameters, masks, validation datasets, comparison code, and uncertainty calculations should accompany any benchmarking claim so that the result can be independently reproduced rather than taken on faith.

We note, finally, that what should count as a genuinely like-with-like comparison, such as matched epochs, matched resolution, matched validation scope, is itself an evolving part of the discipline's methodology rather than a settled checklist. The principles above are offered as a starting point for that conversation, not its final word, and we expect them to be refined as the community's experience with cross-product benchmarking grows.

7.2. A statistical framework for rigorous comparison

Much of the InSAR benchmarking literature relies on scatter plots, correlation coefficients, and RMSE as if these were sufficient to establish that one product is more accurate than another, or that two products differ meaningfully at all. These are descriptive statistics, so they characterize the magnitude of disagreement but do not test whether an observed difference is statistically significant, attributable to an identifiable source of uncertainty, or large enough to matter for the application at hand. A more rigorous comparison combines descriptive, inferential, agreement-based, and spatial analyses, each of which addresses a different question.

Descriptive statistics, such as mean bias, mean and median absolute error, RMSE, the standard deviation of residuals after bias removal, Pearson and Spearman correlation, the coefficient of determination, and normalized RMSE, remain useful as a first characterization of disagreement, provided their limits are respected. The central limitation is that correlation measures association, not agreement. So two datasets can correlate almost perfectly while differing systematically by several millimeters per year, exactly the pattern at issue in the comparison with Li et al. (2026).

Whether an observed difference could plausibly have arisen by chance is a separate question, addressed by inferential statistics rather than descriptive ones. For paired observations, such as collocated InSAR pixels or GNSS comparisons, the paired t-test is appropriate when residuals are approximately normal, with the Wilcoxon signed-rank test or the sign test as non-parametric alternatives. Variances are more robustly compared using Levene's test than the classical F-test, since the former is less sensitive to departures from normality (Levene, 1960). Whole distributions can be compared with the Kolmogorov–Smirnov or Anderson–Darling tests, and comparisons across more than two products can be handled with repeated-measures Analysis of Variance (ANOVA) or, where normality is doubtful, the Friedman test. These tests establish whether a difference is statistically real; they say nothing about whether it is scientifically important.

That second question, whether two products agree within their expected uncertainty, rather than merely whether they are correlated, is the one most often left unasked in the InSAR literature, and it requires a different family of methods. Bland–Altman analysis (Bland and Altman, 1986), Lin's concordance correlation coefficient (Lawrence and Lin, 1989), the intraclass correlation coefficient, and Deming or orthogonal-distance regression all treat both datasets as containing uncertainty, in contrast to ordinary least-squares regression, which implicitly assumes that one dataset is error-free.

Spatial structure introduces a further complication that is easy to overlook given the pixel counts typical of InSAR products. Residuals between deformation fields are rarely spatially independent, and treating them as if they were, inflates the effective sample size and overstates the significance of any test performed on them. Moran's I (Moran, 1950) and Geary's C (Geary, 1954) provide global measures of spatial autocorrelation; semi-variograms and spatial correlograms characterize its range and structure; and block

or spatial bootstrap methods (Efron and Tibshirani, 1994) yield confidence intervals that account for it. With millions of pixels but far fewer statistically independent observations, this step cannot be treated as optional. Uncertainty should in general be reported as an interval, such as a 95% confidence interval, a bootstrap interval, or a propagated uncertainty on both bias and RMSE, rather than as a bare point estimate. A final distinction, and one with direct bearing on the present exchange, is between statistical and practical significance. With millions of observations, a mean difference of 0.1 mm/yr can register as highly significant in a hypothesis test while being entirely negligible for any practical application. Effect size measures, such as Cohen's d (Cohen, 2013), the standardized mean difference, or the ratio of the observed difference to the combined measurement uncertainty, should accompany any significance test, so that a difference can be judged not merely as nonzero but as large enough to matter.

Taken together, these principles suggest a straightforward analytical sequence as follows: establish spatial and temporal consistency between the datasets being compared; compute descriptive statistics; test the assumptions, such as normality, heteroscedasticity, spatial independence, that any subsequent inferential test requires; apply the appropriate hypothesis test; quantify agreement directly with Bland–Altman analysis, Lin's concordance correlation coefficient, or Deming regression; assess the spatial autocorrelation of the residuals; report confidence intervals and propagated uncertainty; and interpret the result in terms of its practical significance for the application at hand, rather than resting on a p -value alone. Much of what is wrong with the comparison in Li et al. (2026), namely the unacknowledged epoch mismatch, the atypical validation network, and the disagreement-derived threshold, can be traced to skipping one or more of these steps, and a benchmarking literature that adopted this sequence as standard practice would be considerably less prone to mistaking comparison artifacts for genuine product deficiencies.

8. Conclusions

Li et al. (2026) address a legitimate and important question about InSAR reproducibility in vegetated coastal settings. Their call for harmonized processing frameworks, transparent methodological documentation, and systematic cross-validation reflects sound scientific priorities, and we support these recommendations unreservedly.

However, the specific conclusions of the paper, that InSAR is presently unreliable for SEC rates below 5 mm yr⁻¹ and that coastal subsidence maps should not yet be used for policy guidance, rest on an analysis that compares two products observed over non-overlapping windows spanning nearly a decade, compares O24 at approximately 400 times coarser than its native resolution, validates against a small and geographically narrow network of stable-upland GNSS sites, applies an urban-derived uncertainty bound to non-urban environments, and describes O24 performance in stronger terms than the paper's own results justify. Each of these choices has a quantifiable effect that shifts the analysis in the same direction, toward a more pessimistic assessment of InSAR capability. Our independent validation of O24 within the Li et al. (2026) study area, conducted at native 50 m resolution against 88 GNSS stations drawn from the full NGL network, yields a residual standard deviation of 1.6 mm yr⁻¹, consistent with the uncertainty reported by Ohenhen et al. (2024) and with the performance the paper's own uncertainty bounds would lead one to expect for the 50 m product. When evaluated at native resolution, over its own observational window, and against a spatially representative network, O24 (and W24, over its own window) performs robustly; the apparent discrepancy documented by Li et al. (2026) is, in substantial part, a consequence of the comparison's construction rather than of either product's underlying accuracy.

We further demonstrate that the 5 mm yr⁻¹ caution threshold is not a principled uncertainty bound but an artifact of conflating inter-product methodological divergence with measurement error. Applying the same logic to a GNSS-versus-GNSS period comparison within the study area yields a threshold of 3.7 mm yr⁻¹, below which the paper's own GIA estimate of -1.2 mm yr⁻¹ falls, rendering it, by the authors' reasoning, equally unreliable. This internal contradiction exposes the threshold for what it is: a framework that cannot consistently adjudicate between signal and noise, and that should not be the basis for policy-facing conclusions about the state of coastal InSAR.

The appropriate response to genuine concerns about InSAR reproducibility is not a blanket caution threshold derived from a methodologically mismatched comparison, but rather a commitment to robust analyses that can distinguish between competing sources of inter-product disagreement. A comparison conducted at native resolution, with a geographically representative validation network, and with uncertainty estimates grounded in instrument-specific noise characterization rather than inter-product spread, would either substantiate Li et al.'s conclusions or reveal that InSAR performance in vegetated coastal settings is considerably more nuanced than the current analysis suggests. Either outcome would constitute a more durable and credible contribution to the field. The InSAR community would be better served by investing in that rigorous intercomparison infrastructure than by adopting caution thresholds that have not yet been derived from first principles, given the risk of discounting valid, policy-relevant datasets on the basis of a comparison that does not yet separate measurement error from methodological divergence.

9. Acknowledgments

This work was supported by the U.S. Department of Defense. We thank Roland Burgmann and Robert Nicholls for their valuable comments and insightful discussions.

10. Author contributions

M.S. conceived the study and wrote the manuscript. L.O. created the figures. All co-authors contributed to discussing the results and revising the manuscript.

11. Code availability

This study did not involve developing new code. We performed all analyses using standard, publicly available statistical methods in MATLAB and Python, as described in the manuscript.

12. Data availability

This study did not generate any new data. All datasets used are publicly available from their original publications: the Ohenhen et al. (2024) and Wang et al. (2024) coastal subsidence datasets, and GNSS station data from the Nevada Geodetic Laboratory (<https://geodesy.unr.edu/>).

12. Competing interests

The authors declare no competing interests.

References

- Bamler, R. and Hartl, P.: Synthetic aperture radar interferometry, *Inverse problems*, 14, R1–R54, 1998.
- Bland, J. M. and Altman, D.: Statistical methods for assessing agreement between two methods of clinical measurement, *The lancet*, 327, 307–310, 1986.
- Cohen, J.: *Statistical power analysis for the behavioral sciences*, routledge2013.
- Cressie, N.: *Statistics for spatial data*, John Wiley & Sons2015.
- Efron, B. and Tibshirani, R. J.: *An introduction to the bootstrap*, CRC press1994.
- Geary, R. C.: The contiguity ratio and statistical mapping, *The incorporated statistician*, 5, 115–146, 1954.
- Hanssen, R. F.: *Radar interferometry, data interpretation and error analysis*, Kluwer Academic Publishers, 328 pp.2001.
- Lawrence, I. and Lin, K.: A concordance correlation coefficient to evaluate reproducibility, *Biometrics*, 255–268, 1989.
- Levene, H.: Robust tests for equality of variances, *Contributions to probability and statistics*, 278–292, 1960.

439 Li, G., Törnqvist, T. E., and Chen, J.: Evaluating InSAR-derived rates of surface-elevation change along
 440 the central US Gulf Coast, *Earth Observation*, 1, 1–13, 2026.
 441 Moran, P. A.: Notes on continuous stochastic phenomena, *Biometrika*, 37, 17–23, 1950.
 442 Ohenhen, L. O., Shirzaei, M., Ojha, C., Sherpa, S. F., and Nicholls, R. J.: Disappearing cities on US coasts,
 443 *Nature*, 627, 108–115, <https://doi.org/10.1038/s41586-024-07038-3>, 2024.
 444 Wang, K., Chen, J., Valseth, E., Wells, G., Bettadpur, S., Jones, C. E., and Dawson, C.: Subtle land
 445 subsidence elevates future storm surge risks along the Gulf Coast of the United States, *Journal of*
 446 *Geophysical Research: Earth Surface*, 129, e2024JF007858, 2024.
 447 Zebker, H. and Villasenor, J.: Decorrelation in interferometric radar echoes, *IEEE Trans. Geosci. Rem.*
 448 *Sensing*, 30, 950–959, 1992.
 449