Peer Review – The Cryosphere submission

Manuscript ID	egusphere-2026-1160 (https://doi.org/10.5194/egusphere-2026-1160)
Title	Two decades of GNSS-IR observations reveal an asymmetric decline in coastal sea ice phenology of the Beaufort Sea
Authors	Minfeng Song, Xiaolei Wang, Zhenyu He, Dongzhen Jia, Ruya Xiao, Xiufeng He (Hohai University)
Journal	The Cryosphere (TC)
Submission Version	Pre-print, discussion open 20th May 2026
Review Date	2026-08-02
Recommendation	Major revisions

Summary

The paper builds a continuous 20-year (2003-2023) record of coastal sea ice phenology at the TUKT ground station on the Beaufort Sea coast using GNSS Interferometric Reflectometry (GNSS-IR). The central methodological device is the Amplitude Integration Factor (AIF), which integrates reflected signal power across the 5–15° elevation range as a proxy for ice presence, together with a normalization scheme intended to bridge a 2015 receiver upgrade. Phenological transition dates (ice-out, freeze-up) are extracted from the calibrated AIF and used to derive interannual trends. The primary results are a shortening of the ice season of ~4.6 days per decade, an asymmetric decline dominated by delayed autumn freeze-up rather than advanced spring breakup, strong coupling of freeze-up to accumulated Freezing Degree-Days (R^2 = 0.74), and a systematic ~5.5-day lead of GNSS-IR over the 4-km IMS product at freeze-up, interpreted as detection of nearshore frazil ice. This review targets the use of the data, the calibration and detection methodology, the statistical inference, the cross-scale comparisons, and the conclusions drawn. I do not focus on the internal details of the GNSS-IR signal-processing chain (wavelet power extraction, Lomb-Scargle screening), which lie outside my primary expertise. The topic is a good fit for The Cryosphere and the dataset a continuous, high-completeness, sub-kilometre ground record over two decades. The physical reasoning is generally coherent and the multi-source 2019 case study is a nice demonstration.

Reservations stem from the verbosity of the written claims, and how strongly they are stated relative to the evidence. Several of the paper's central claims rest on a single significant p-value obtained after excluding one year, correlation windows selected to maximise fit, and detection rules that are asymmetric. Additionally, much more care should be taken to check for grammatical errors and consistency of reasoning, formatting and style, as it would benefit from a thorough editing pass for grammar, consistency and repetition before resubmission. The key questions to confront:

1. **Statistical framing of the trends** The abstract and conclusions lead with a "statistically significant shortening ... 4.63 days per decade ($p = 0.04$)", a value that applies only to one metric (corrected ice duration) and only with 2018 excluded; with 2018 retained the same metric gives $p = 0.06$, and every other trend in Table 4 is non-significant at 95% (ice-out $p = 0.35$, freeze-up $p = 0.18$, ice duration $p = 0.21$). Leading with the one number that crosses the threshold, without foregrounding that it required outlier removal, overstates the robustness. Please report all trends with confidence intervals up front, and consider a non-parametric test (Mann–Kendall / Theil–Sen) that is better suited to $n \approx 19$ with an extreme year and less sensitive to 2018.
2. **Asymmetric detection criteria** It should be investigated as to whether the asymmetry observed in the freeze up / ice-out is a product of the detection criteria. The freeze-up rule requires the AIF to stay above 3σ for at least 5 days, whereas ice-out is the latest date the AIF drops below 3σ. Asymmetric persistence requirements can themselves bias freeze-up and ice-out in opposite directions and could induce part of the very asymmetry that is the paper's central finding. A sensitivity analysis on the detection/persistence rules additionally to the ±10% threshold test already reported, plus a formal test that the freeze-up and ice-out trend magnitudes are different, would help establish that the asymmetry is physical rather than methodological.
3. **Validation of the decadal inter-calibration** Bridging the 2015-12-16 receiver upgrade is the core enabling claim, yet its validation is thin. How is the ~27% magnitude reduction derived, and is the assumption that the hardware change scales the ice signal and the summer baseline proportionally (so the ratio normalization cancels it) tested rather than asserted? An explicit homogeneity test showing no step change in the phenology residuals across the upgrade boundary would help resolve this question. For example, that pre-2016 and post-2015 transition dates are not systematically offset once calibrated.
4. **Selection of the degree-day windows** The FDD window for freeze-up (R^2 = 0.74) and the TDD window for ice-out ($r = -0.81$) are chosen by a 2-D search over accumulation-window centre and duration to maximise correlation (Fig. 10, "white stars denote maximum negative correlations"). The reported $p < 0.001$ is not corrected for that search, so the couplings look more robust than they are. Cross-validation, out-of-sample test, or a multiple-comparison correction may help back up these claims and enable presenting these as established thermodynamic couplings.
5. **Frazil-detection interpretation and the ice-out comparison** The claim that GNSS-IR resolves "initial nearshore frazil ice formation" rests on a single freeze-up event (2019) with Sentinel-2 imagery, then is generalised across two decades. Meanwhile the ice-out comparison against IMS shows little interannual correlation ($r = 0.05$, $p =$

0.85), which raises the question of how the ice-out retrieval is validated across the record. The paper also offers two distinct explanations for the 5.5-day freeze-up lead — spatial dilution of the coarse IMS pixel, and earlier freezing of shallow nearshore water — without disentangling them. Additional cloud-free transition cases and a clearer separation of these mechanisms would substantially strengthen the central claim.

Specific Comments

- **Abstract** – "statistically significant shortening ... 4.63 days per decade ($p = 0.04$)". The abstract does not state that this is the corrected-duration metric with 2018 excluded. State the metric explicitly and give the with-2018 value ($p = 0.06$) alongside it, so the headline is not read as an unconditional result.
- **Introduction** – the context is thorough, but the framing is often heavily promotional. For TC I would tone the language to match the strength of the evidence given $n \approx 19$ and the wide confidence intervals. It leans heavily on intensifiers and evaluative adjectives (examples throughout the text - "profound," "extreme," "rigorously," "true," "genuine", "anomalously", "exceptionally", "exact", "clearly", "remarkable", "crucially", "disruptive", "resounding", "ultimate", "conclusively", "robust", "highly reliable") that overstate. Given the modest sample size and wide confidence intervals, this undercuts credibility. The results would read as more authoritative in plain concise language using figures and statistics in place of adjectives.
- **Introduction** – the review of GNSS-R sea-ice work is almost entirely ground-based / near-nadir. A short acknowledgement of the broader (including spaceborne) GNSS-R sea-ice literature would better position the specific contribution of a fixed ground station, and clarify what ground-based GNSS-IR adds that the spaceborne approaches cannot (fixed footprint, high temporal continuity at one nearshore point).
- **Section 2.2 / 2.3** – data completeness is quoted as 97.5% in the text but 97% in Table 1; reconcile.
- **Section 3.1 (physical basis; Fig. 2, Fig. 3)** — The technique operates at 5–15° elevation, i.e. 75–85° incidence, which is a grazing-angle geometry, but this is never stated and the term "grazing" does not appear. Figure 2 compounds this by depicting satellites high overhead with steep reflections, which misrepresents the near-horizon geometry used; a schematic drawn at representative low elevation would serve the reader better. More substantively, the physical mechanism in Section 3.1.2 is mis-attributed. Seawater ($\varepsilon \approx 70$) is a far stronger reflector than sea ice ($\varepsilon \approx 3$) on dielectric contrast alone, so "stark dielectric contrast" as written predicts the opposite of the ordering shown in Fig. 3 (green above red). The ordering in Fig. 3 is in fact defensible, but for a reason the manuscript does not give: at grazing incidence the relevant quantity is the co-polar RHCP reflection coefficient, whose angular behaviour is set by the pseudo-Brewster angle. Seawater's pseudo-Brewster (~83° incidence, ~7° elevation) falls inside the 5–15° operating window, suppressing the RHCP-reflected component from open water there, whereas sea ice's (~60° incidence, ~30° elevation) does not — so ice retains the higher RHCP coefficient across the operational range. This grazing-angle polarization effect, not dielectric contrast, is what underpins the AIF's ice/water discrimination, and it should be explained as such.
- **Section 3.1.2** – the reflectivity numbers are internally inconsistent. "At a 5° elevation angle, the reflectivity for ice is ~23 versus ~0.22 for seawater, representing a ~45% increase" — 23 vs 0.22 is a factor of ~100, not a 45% increase, and it does not match the "factor of 2–3" amplitude ratio stated in the next sentence. Please correct these figures and make the quantities (permittivity vs reflection coefficient vs amplitude) explicit.
- **Section 3.2** – please clarify a few AIF definitions for reproducibility: what $\sigma\_Rh = 0.2$ m represents the choice of $n = 800$ interpolated elevation samples, and the physical units of Λ (a dimensionless "calibrated AIF"?). These would help future users reuse the metric.
- **Section 3.3.1** – the QC retains ~90% of observations. How do the discarded 10% differ from the retained set — are they biased by season (e.g. are noisier summer tracks preferentially removed, which would affect the ice-free baseline), by satellite, or across the pre/post-2015 boundary? A short characterisation would reassure that screening does not itself shape the baseline or the trend.
- **Section 3.3.1** – the reference-period normalization sets DOY 200–250 to zero-mean as the ice-free baseline. Please confirm this window is genuinely ice-free in every year of the record (it appears safe given a mean ice-out of DOY 173 and freeze-up of DOY 284, but this should be stated and evidenced) and note how many tracks/satellites typically contribute per day to that baseline.
- **Section 3.3.2** – see key question 2 on the asymmetric detection rules. Extending the sensitivity analysis beyond ±10% on the threshold to include the persistence criterion would directly test whether the reported asymmetry survives.
- **Section 4.1** (2019 case) – the five-phase interpretation is a nice narrative, but the physical attributions (e.g. the Phase II "rebound" as bare ice / melt ponds, Phase I decline as snow densification) are asserted with strong language for a single year. Please temper "physically represents" / "perfectly coincides" to reflect that these are plausible interpretations of one season's co-timing.
- **Section 4.2.2** (Table 4, Fig. 8) – the exclusion of 2018 is justified by the authors because it "dampens" the trend, which reads as motivated. The authors should cite precedent for excluding samples this way. It is stated to be "extreme noise" with no evidence. Define an objective, a-priori outlier criterion and present the with/without-2018 results even-handedly rather than foregrounding the version that reaches significance. Figure 8's caption says "2004 through 2022" while Table 4 and the text say 2004–2023 — reconcile the end year.

- **Section 4.4** (Table 5) – the ice-out row has essentially zero correlation with IMS ($r = 0.05$), yet ice-out trends are still reported and interpreted. Please discuss what, if anything, independently validates the ice-out dates across the 20-year record beyond the single 2019 Sentinel-2 case. Also the sign convention is confusing: the caption states negative = earlier detection by GNSS-IR, but the ice-out offset of −1.4 is labelled "(later)". Please make the convention consistent.
- **Section 4.5** (Table 6) – the benchmarking against pan-Arctic passive-microwave trends is a useful addition, and the honest discussion of epoch and sensing-regime differences is appreciated. The agreement of the GNSS-IR trend (−4.19 ± 5.43) with the concurrent IMS trend (−4.23 ± 1.8) is described as "resounding cross-validation", but the GNSS-IR interval is ~3× wider and IMS is not an independent ground truth at these scales. I would present the agreement as reassuring rather than conclusive.
- **Section 5** (Conclusions) – the statement that the method "requires no localized calibration" sits awkwardly next to Section 3.3, which is entirely an inter-calibration procedure. Please clarify what is meant (presumably: no external/auxiliary calibration data are needed because the summer baseline is self-referencing) so the two statements do not appear to contradict.

Smaller Issues

- The dataset is cited with two different Zenodo DOIs: 10.5281/zenodo.18452514 in Section 6 (and on the preprint landing page) versus 10.5281/ZENODO.18212535 in the reference list (Song et al., 2026). Please reconcile to a single archived version.
- Typos: double full stop in "polar night (Li et al., 2023).."; stray space in "records remain virtually absent ,"; "1.85°C total temperature increases" should read "increase". Missing space in line 174 – please read through for grammar issues.
- The IMS window is described as an 11×11 pixel (~44×44 km) region but the comparison is framed against a "4-km product" / "16 km^2 pixel" elsewhere; a sentence clarifying that the offset statistics use the single central pixel vs the window would avoid confusion.
- Table 1 lists a period for IMS/ERA5/NAVCAN but leaves Sentinel-2's period blank; add the years/scenes used.
- Figure 1 – Remove the legend with unused symbols, I cannot see the pink square utilised in the figure.
- Analysis in caption of Figure 3. Analysis should be kept for the text, and captions remain as information as to what is contained in the figure. This is also a problem in figure 6
- Repeated paragraphs – eg. Line 212 – 215 repeated from line 114 onward; line 412 – 415 results repeated from above
- Figure 6 is very hard to follow – put images in 3x2 with DOY marked in the top corner in white additionally a coastline would help.
- Figure 7 - grey shading hard to pick out on top of the purple. This is the key figure, the time series should extend the width of the page, the purple points maybe need to be smaller in order that they are more easily visible, and the whole thing a lot bigger.
- Figure 7 caption "a filtered AIF" what filter? A rolling mean or something else?
- Figure 8 – orange border looks a lot like the red negative anomaly bars – find a more unique colour to avoid confusion.
- Line 387 "Crucially…cycle" is repetition from 354
- Line 390 "decade) is a genuine, persistent climatological shift rather than a statistical artifact driven by outlier years." – Large claim please evidence, see Point 1 above.
- Several conclusions are stated causally without support — please evidence or soften:
  - Line 399 "It is increasingly delayed by a potent "thermal memory" effect…" remove unless you can include proof of the driver for the delay
  - Line 464 "validates" much too strong, suggest "supports" or "indicates" – additionally what does "validates the temporal accuracy of the GNSS-IR record" mean? Implies we have doubts on the time axis uncertainty, which was thus far not mentioned.
  - Line 494 "indicate that this mean offset is not a systematic lead, but rather a reflection of extreme spatial heterogeneity" seems like a big claim – evidence or remove. Not being a function of time is not sufficient to back up this claim.
  - Line 500 "Ultimately, these timing differences represent physical detection thresholds rather than measurement errors."
- LaTeX has leaked into the text: "strict statistical significance ($p < 0.05$)" in Section 4.2.2
- ERA5 vs in-situ agreement is reported both as "R = 99.1%" and "r = 0.99". Report a correlation coefficient (r = 0.99), not a percentage, and use it consistently.
- Figure 9 caption should be made more concise to aid legibility of figures. Long sentence descriptions delay comprehension of figures.
- Figure 10 caption unclear. Subplots have headings below and repeated above but only for some of them. Stated as Heatmaps when they are scatter plots (or lines charts). White stars referenced in caption but only coloured stars present in figures.

- Table 6 – seems unnecessary to list all 4 results of this study in the review summary table. Authors should choose one to stand behind or at least guide the reader as to their favourite of the four (maybe in bold?)