Manuscript: egushere-2026-2257
Title: Quantitative Evaluation of Mesoscale Eddies in the North Atlantic Using Satellite Altimetry and Ocean Reanalyses

This study evaluates two global ocean reanalysis products (GLORYS12V1 and GLORYS2V4) against satellite altimetry (AVISO DUACS 1/8° and SWOT MIOST) for their ability to represent mesoscale eddies in the North Atlantic. The topic is relevant, and the use of SWOT data as a reference is timely. However, the manuscript suffers from a large number of serious problems that undermine its scientific credibility. These include logical inconsistencies, factual errors contradicted by the figures, methodological flaws, unsupported claims, rebundant writing, and multiple figure errors. Also, the novelty of this study is not clear enough. Given the number and severity of these issues, I recommend major revision or, if the editor sees fit, rejection with encouragement to resubmit after thorough revision.

Major Issues:

Lines 42-54: I'm afraid the logical flow here doesn't hold up. You first mention Okubo-Weiss, wavelet, and py-eddy-tracker，all of which can be applied to satellite data to detect individual eddies at specific locations. Then you say "however, a feature detection method is needed to evaluate the accuracy of the eddy feature detected at specific locations." That doesn't make sense, because the methods you just listed already do exactly that.

What I think you mean is that those methods don't tell you whether an eddy in a reanalysis corresponds to the same eddy in satellite observations. That's a matching problem, not a detection problem. Smith and Fortin (2022) addressed matching, not detection. Please revise the text to make this distinction clear. As written, it reads like you're dismissing existing detection methods as inadequate for something they are actually capable of doing.

Line 172-173: You say your focus is to "assess the uncertainty associated with the DUACS 1/8° product". That contradicts the paper's stated goal of evaluating reanalysis products against satellite data. DUACS is a reference here, not the target of evaluation. Please clarify or remove.

Lines 192-193: Your minimum amplitude threshold of 0.4 cm (derived from a 0.2 cm contour interval) is unrealistic. This is far below the 2-4 cm noise floor of satellite altimetry. Please replace this with a physically justified amplitude threshold of at least 3-5 cm, as is standard in the literature, or explain the reason explicitly (I know you referenced Smith and Fortin, 2022, but you should have your own rationale by choosing this threshold).

Line 199: In equation 1, what are the normalization factors in the denominator? Are they based on satellite or reanalysis values? Please specify.

Lines 200-205: The logic is backwards. You describe the benefits of a 28-day threshold (reducing spurious eddies) but then adopt 4 days without justification. The following sentence about “the main impact” of a large threshold reads as if reducing small eddies is a problem, but for many studies it’s exactly the point. Please clarify why 4 days is appropriate for your objectives, and restructure the paragraph to flow from choice to rationale, not the other way around.

Figure 1 & 3: I wonder why detected eddies are not circled in Figure 1 (which has been provided in Figure 3)? Circling them would help readers compare their sizes, positions with the SSH background across these four datasets. Also, since the colorbar scale is the same for all panels, a single shared colorbar would be sufficient. And what is “ZOS (m)”? Please define it. In Figure 3, on the other hand, it’s difficult for readers to figure out where is the path of the Gulf Stream without SSH information.

Figure 2: The x-axis is wavenumber in units of cycles per degree. While this may be acceptable for the meridional direction (where 1° latitude ≈ 111 km), it is problematic for the zonal direction. Your study domain spans from 30°N to 60°N. At 30°N, 1° longitude ≈ 96 km; at 60°N, it is only about 55 km. Therefore, a given wavenumber in cycles/degree corresponds to very different physical scales at different latitudes. Averaging the zonal spectrum across this wide latitudinal band will inevitably distort the spectral energy distribution, particularly at higher wavenumbers (smaller scales).

Please either (1) interpolating the data onto an equal-area or equal-distance grid (e.g., km-scale) before computing the spectra, or (2) using spherical harmonic analysis to properly account for the curvature of the Earth. At the very least, the authors should quantify the potential bias introduced by using degrees for the zonal direction. As it stands, the comparison of zonal spectra across products (especially the claimed superior performance of SWOT at small scales) may be partly an artifact of the chosen units.

Table 2 and 3: Why does the same GLORYS12V1 product show different eddy counts in these tables? A fixed reanalysis has a fixed number of detected eddies. If the two tables cover different time periods (due to SWOT's shorter record), this must be stated explicitly. The current presentation is misleading. Also, please discuss why the cyclonic/anticyclonic imbalance is much larger in the reanalyses than in the satellite data.

Lines 335-345: This paragraph has several problems. First, you attribute the FAR peak in GLORYS2V4 (mid-2024) to “changes in its data assimilation configuration or observation sampling strategy” but neither Figure 7 nor any earlier section provides evidence for such changes. Please either cite documentation confirming a configuration change in that period, or remove the speculation. Second, the last sentence is unclear. What does “tend to plateau and reduce the differences” mean? Looking at Tables 4 and 5, the POD and FAR values do change between the 4-day and 7-day thresholds, but the implications are not stated. Are you suggesting that short-lived eddies are the main source of disagreement between the two reanalyses? Please clarify the take-home message. The current phrasing is too vague to be useful.

Lines 365-368:The claim that “the additional eddies detected in SWOT do not include any eddies that are also detected in GLORYS2V4” does not follow from the observation that GLORYS2V4’s POD is similar across the two comparisons. If GLORYS2V4 missed all SWOT-detected additional eddies, its POD would decrease (hits unchanged, denominator larger). A constant POD implies hits increased proportionally, meaning GLORYS2V4 does detect at least some of the new eddies. This conclusion is logically backwards.

Lines 400-405: If you ran the 20-day block bootstrap, show the results, even in supplementary material. A hand-waving claim of “quite similar” with no data is not acceptable. The 20-day analysis could also support your convergence argument in the next paragraph. I strongly suggest you to add it to the supplement.

Lines 407-408: The text concludes that “long-lived eddies are more robust and easier to match” and states “This is shown in Fig. 8 and Fig. 10.” But these figures display POD/FAR as functions of radius and amplitude, not of eddy lifetime. Please revise the text to correctly reference the appropriate analysis or figures.

Fig. 8 and related Figures 9-13: Please specify what x-axis “radius” and “amplitude” refer to in these figures. This distinction matters. If they are taken from the satellite product, then the figure shows: for observed eddies of a given radius or amplitude, how many are matched (hit), missed, or falsely reported by the model. That is interpretable. But if the radius is taken from the reanalysis for false alarms and from the satellite for misses, the physical meaning becomes inconsistent. Please clarify in the caption and/or methods section.

Lines 430-432: “FAR exceeding 30%” is shown in Figs. 9 and 11, not in Figs. 8 and 10. Also, Figs. 8 and 10 show hits (blue) dominate all bins, so “false alarms dominate” is incorrect. Please correct both the figure references and the claim.

Line 433 & Figure 9: First, check Figure 9. For GLORYS12V1 at radii <30 km, POD >0.5 and FAR <0.3, contradicting your statement “POD remains below 50% and FAR exceeds 30%.” Please correct. Second, the titles for the bottom panels in Figure 9 are wrong, it should be “Probability of Detection - SWOT comparison”. Third, the top and bottom rows are identical in this figure (this may be why their titles are the same). This is likely an error, as they are meant to show results for AVISO and SWOT separately. Please replace with the correct panels.

Figures 10 &11: The caption and the figure content appear inconsistent regarding which row corresponds to AVISO vs SWOT. Please verify and correct. As it stands, the reader cannot trust which dataset is being shown.

The left panels of Figures 12 and 13: The legend appears to be reversed. Based on the physical expectation that a higher-resolution model (GLORYS12V1) should have its false alarms concentrated at smaller radii, the solid line (reaching 90% at ~50-55 km) should be GLORYS12V1, and the dashed line (reaching 90% at ~80-85 km) should be GLORYS2V4.

However, the current legend shows the opposite. Please verify and correct the legend. (The text description in lines 490-495 correctly follows the physical expectation, so only the figure needs correction.)

Tables 8-9 and Figure 14: GLORYS12V1 and GLORYS2V4 have nearly identical errors in radius, amplitude, and distance for matched eddies, despite large differences in POD. This is a striking result that needs explanation. Why does higher detection skill not translate into better accuracy for the eddies that are detected? Does averaging over all matched eddies mask scale-dependent differences? Please discuss. Currently the paper is silent on this apparent contradiction.

Minor issues:

Line 12：It should be AVISO DUACS 1/8° rather than 1/4°.

Line 63: Please provide references as evidence for the “benchmark” claim.

Line 68 “the time coverage of SWOT data is limited”, so you mean this data can not be used for studies on dacadal or long-term variability of mesoscale activity? Please clarify.

Lines 65-79:The discussion of SWOT’s limitations (lines 65-72) is repetitive. Points 2 and 3 both argue non-independence, and point 2 is logically weak anyway. Then you repeat the same non-independence issue for AVISO. Consider consolidating: state the two real limitations once, acknowledge that both products share this issue, then give your conclusion. The current structure is confusing and over-long.

Line 86: You stated “the first aim of this study” in Line 40, what about the second and third aims, if any?

Line 100: Please provide the DOI of the AVISO DUACS global 1/8° product and the date of access/download. The same information is needed for SWOT MIOST and the reanalysis data products.

Lines 107, 438 & 446: change “50 Km” to “50 km”. Please check similar mistakes throughout the manuscript.

Section 2.1: This section is overly long, especially the part on SWOT MIOST. The level of detail is disproportionate. Do we really need the launch date, collaborating agencies, and KaRIn wavelength requirements in a Methods section? Much of this belongs in the Introduction, if anywhere. More importantly, the same points are repeated multiple times: effective resolution, limitations, independence issues. The “limitations” list at the end (lines 130-135) mostly repeats what was already said earlier in the paragraph and in the previous section (lines 65-79). Suggest cutting this down significantly. Focus on what the product is, why it’s used, and its key limitations for this study. Everything else is extraneous.

Section 2.2: This section is also too wordy. It’s not necessary to provide all the data information including the parts not used in this study, such as atmospheric forcing data (precipitation and radiative components), detailed vertical grid information both in the text and Table 1. You also repeated three times of study period in this section.

Line191: This criterion is the same as the second one, please remove one of them.

Line 207 & 531-532: The definition of the amplitude has already been provided in Line 192. Please avoid repetition. Also, clarify what “amplitude error” means: is it the absolute difference between reanalysis and satellite amplitude?

Line 220: Should be “SWOT MIOST (c) and DUACS 1/8° (d)”?

Figure 4: The linear regression and correlation coefficients do not salvage this analysis. Linear model is not justified (heteroscedasticity is visible). Confidence intervals for correlations are missing. Most importantly, averaging over the full lifetime obscures life-stage dynamics. Please either drop the "dependence" framing and describe the figure descriptively, or provide a properly justified analysis with life-stage stratification and uncertainty quantification.

Lines 252 & 305: “a lifetime lower than 21 days” and “60 km”, these thresholds are mentioned but not visible in the figures or justified in the text. Please clarify or remove.

Lines 274-278 and equations 2 and 3: Please add the reference Smith and Fortin (2022).

Line 304: “(see Fig. 4)) ” should be “(see Fig. 4)”

Line 307: Please provide statistic evidence for “a positive correlation”, for example, correlation coefficient with significance level.

Line 310: It should be “Table 2 and Table 3” rather than “Table 4 and Table 5”.