

First reviewer's comments

This revision includes all the comments from Reviewer 1.

We have included several figures to provide a clearer explanation of the revisions and to better support the changes made in the manuscript. In addition, the four appendices, from Appendix A to Appendix D, have been included at the end of this revision document and the revised manuscript (after the Summary and Discussion section). Several additional analyses are reported in these appendices to support the statements made in the main text.

Please note that, following the changes made during the revision process, the line numbers in the manuscript have changed.

We would like to sincerely thank Reviewer 1 for the careful evaluation of our manuscript and for the detailed comments and suggestions. These comments helped us to improve the clarity, consistency, and scientific robustness of the manuscript.

Legend:

Black = reviewer's comments.

Blue = corrections (quoted text).

Green = response

In some cases, the text quoted in blue includes passages that are not limited strictly to the specific section being revised, but have been included to provide the necessary context and to better explain the concept requiring clarification or modification.

Some comments, especially for the minor issues, have been combined and addressed with a single response, because they refer to the same part of the text or to closely related aspects of the manuscript.

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.

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.

Response: we have revised the manuscript, now at lines: 45 – 59 to clarify the distinction between eddy detection and tracking, which are performed independently within each dataset, and the subsequent feature-matching and verification step used to determine whether an eddy detected in a reanalysis corresponds to an eddy observed in satellite altimetry.

“Several studies have investigated mesoscale eddy properties, such as population, duration, lifetime, and trajectories, in numerical models and reanalyses (de Vos et al., 2018; Mason et al., 2019; Xie et al., 2020; Bonaduce et al., 2021; Smith and Fortin, 2022). Eddy detection methods commonly used in these studies include approaches based on the Okubo-Weiss parameter, wavelet analysis of the relative vorticity field, and the geometry of streamlines (Souza et al., 2011). Another widely used approach, introduced by Chelton et al. (2011), relies on satellite-altimetry fields to identify coherent eddy structures. The py-eddy-tracker package developed by Mason et al. (2014) is also widely used to detect and track eddies at regional and global scales, as in de Vos et al. (2018) and López-Álzate et al. (2022). In the present study, we use py-eddy-tracker to perform the eddy detection and tracking process. This step is applied independently to each dataset and provides the eddy catalogues, including eddy position, amplitude, radius, polarity, and trajectories. However, evaluating the skill of an ocean reanalysis requires an additional step beyond detecting and tracking eddies independently in each dataset. In particular, it is necessary to determine whether an eddy detected in a reanalysis corresponds to the same eddy observed in satellite altimetry at a given time and location (a feature-matching and verification problem). To fill the gap, Smith and Fortin (2022) developed a verification method to evaluate the capability of ocean reanalyses in representing the eddy features at exact locations and time detected in satellite altimetry data. In this study we will adopt this method to evaluate the representation of eddies in two ocean reanalyses, thereby opening the door to a greater exploitation of ocean reanalyses.”

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.

Response: we have revised the manuscript, now at lines: 202 – 204.

“Here, our focus is to demonstrate the robustness of the verification approach and to assess the sensitivity of the reanalysis evaluation to the choice of satellite altimetry reference product by using two different satellite datasets.”

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).

Response: we have revised the manuscript, now at lines 224 – 228, to clarify the rationale for the selected amplitude threshold. The 0.4 cm value is not intended to represent a physically meaningful threshold for robust mesoscale eddies,

but rather the minimum algorithmic threshold resulting from the contour interval used in the detection procedure. We retain this permissive criterion to avoid imposing an a priori amplitude-based filtering of the detected population.

- *“The amplitude, computed as the difference between the contour extremum and the outermost valid contour, must be at least twice the contour interval. With the 0.2 cm contour interval used here, this gives a minimum algorithmic threshold of 0.4 cm. This low value is used only as a permissive detection criterion and not as a physical threshold for eddies; weak features are assessed later through lifetime filtering and amplitude-dependent verification statistics (see Sect. 3.2 and Appendix C for more details).”*

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

Response: we have revised the manuscript at lines 234 – 237, to clarify the normalization factors used in Equation 1.

“C =

$$\sqrt{\frac{(A_1 - A_2)^2}{A_1^2} + \frac{(R_1 - R_2)^2}{R_1^2} + \frac{(D)^2}{(125 \text{ km})^2}} \quad (1)$$

where the subscripts 1 and 2 denote, respectively, the satellite eddy and the candidate reanalysis eddy in the matching procedure. In the tracking procedure, they denote a detected eddy at time t and a candidate continuation at time t+1, respectively (see Appendix D for more details about the use of a different distance threshold of the matching cost function).”

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.

Response: we have revised the manuscript at lines 238 – 247 to clarify the rationale for adopting a 4-day minimum lifetime threshold and to distinguish this primary choice from the additional sensitivity analyses performed using 7-day and 28-day thresholds (Also following the suggestion of Reviewer 2, we have decided to include in Appendix A the two minimum lifetimes of the eddies, of 7 days and 28 days. Since the differences, in terms of average values of the verification metrics, are not large between 4-day and 7-day, we preferred also to include, following the Reviewer 2’s comment, the minimum 7-day lifetime in Appendix A).

“A minimum lifetime threshold is applied to reduce the identification of transient or spurious features, like in Mason et al. (2014). In this study, we adopt a 4-day minimum lifetime threshold, requiring each eddy to be detected on consecutive days without observational gaps. This relatively short threshold is used to verify the ability of the reanalyses to represent the full spectrum of mesoscale features detected in the satellite products, including short-lived eddies that may still affect the matching statistics. In addition, we have also introduced a calculation with a minimum lifetime threshold of 7 days and 28 days (see Appendix A, for more details). A lifetime threshold of 7 days is also used for certain evaluations to demonstrate the sensitivity to this choice (e.g 4 days). The 28-day threshold is useful for isolating persistent structures, but it is not our primary choice because the aim is also to evaluate the entire population, including short-lived eddies. Eddy properties such as effective radius, amplitude, trajectory, and lifetime are computed for each track. The effective radius is defined as the radius of a circle with the same area enclosed by the eddy contour.”

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.

Response: to address this point, we revised Figure 1 by overlaying the detected eddies as closed contours on the SSH/ADT background for all four datasets, so that their sizes and positions can be directly compared. Because Figure 1 now includes both the SSH fields and the detected eddies, Figure 3 became redundant and has therefore been removed. We also revised Figure 1 to use a single shared colorbar across all panels. We also replaced the label “ZOS” with “SSH (m)”. In the original figure, zos was the name of the NetCDF4 variable for the modeled sea surface height (it comes from the NEMO model). The revised caption and manuscript now distinguish between SSH in reanalysis products and Absolute Dynamic Topography (ADT) in satellite products; nevertheless, here they can be considered the same, because both of them describe the surface dynamic-topography field from which surface geostrophic currents are derived. Consequently, the same contour-based eddy-detection method can be applied consistently to both fields. We have revised the manuscript, now at lines: 248 – 275.

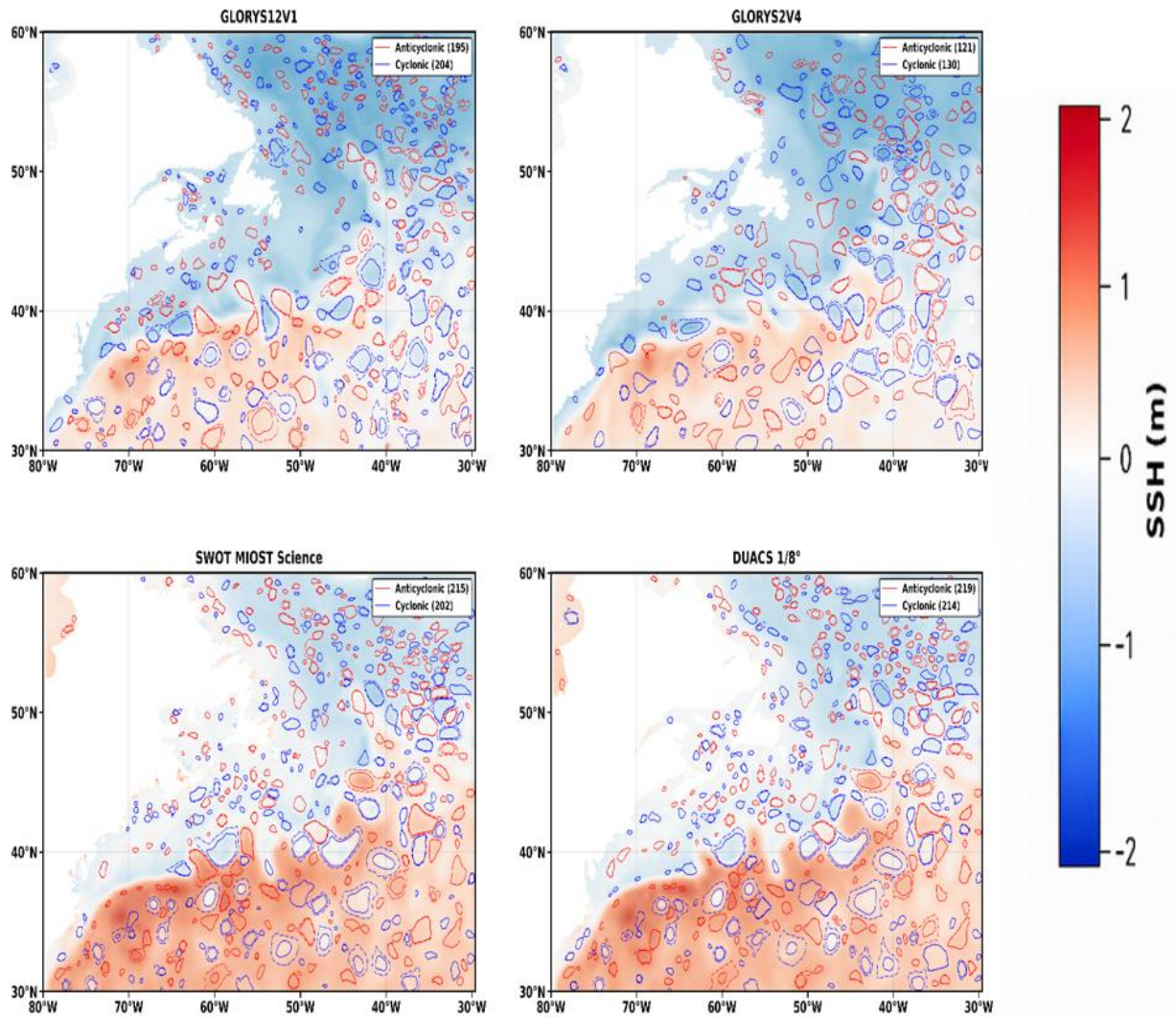


Fig.1. Daily sea-level fields on 1 January 2025 for GLORYS12V1, GLORYS2V4, SWOT MIOST Science, and AVISO DUACS 1/8°, after application of a high-pass Bessel filter with a cutoff wavelength of 500 km. The common colour bar is labelled SSH (m), with SSH used here as a general label for the sea-surface height field. Specifically, the colour shading represents modelled SSH for GLORYS12V1 and GLORYS2V4 and satellite-derived Absolute Dynamic Topography (ADT) for SWOT MIOST Science and AVISO DUACS 1/8°. Although SSH and ADT are not formally identical quantities, they are dynamically comparable and are treated consistently for contour-based eddy detection. Detected anticyclonic and cyclonic eddies are overlaid in red and blue, respectively. Solid contours indicate the contour of maximum circum-averaged geostrophic speed, whereas dashed contours indicate the effective contour used to calculate eddy area and effective radius. All fields were remapped onto the same regular 1/8° grid to allow direct comparison.

This consistent methodology for all datasets ensures that the detected eddies are comparable and that any differences in eddy representation between reanalyses or between them and observations are due to the intrinsic quality of the reanalysis systems and not to processing artifacts (Droghei et al., 2018; Ciani et al., 2019). First, an analysis of the power spectral analysis for the four products considered in this study, using 1 January 2025 as a representative test case. All datasets are remapped onto the same regular 1/8° grid, which is the native grid of both SWOT and DUACS. Figure 1 shows the SSH fields used in this study from GLORYS12V1, GLORYS2V4, SWOT MIOST Science and AVISO DUACS 1/8°. For the two reanalysis products, GLORYS12V1 and GLORYS2V4, SSH represents the modelled sea-surface height above the model geoid or reference equipotential surface, as described in Lellouche et al. (2021). For the satellite altimetry products used in this study, the corresponding SSH field is represented by Absolute Dynamic Topography (ADT), which is the sea-

surface height above the geoid and is obtained by adding the Mean Dynamic Topography (MDT) to the Sea Level Anomaly (SLA) (i.e. $ADT = SLA + MDT$, as explained in Mulet et al. (2021)). Although the reanalysis SSH and satellite ADT are not identical data products, since the former is produced by ocean models constrained by data assimilation and the latter is derived from satellite altimetry and an estimated MDT, they can be considered dynamically comparable, because they provide dynamically comparable representations of sea-surface dynamic topography, from which surface geostrophic currents are derived. As described in Mason et al. (2014), Mason et al. (2019) and Pegliasco et al. (2021), contour-based eddy detection and tracking methods, such as those implemented in *py-eddy-tracker*, can be consistently applied to both fields because they mainly rely on local extrema, closed contours, spatial gradients, and centre-to-edge height differences. The superimposition of the detected eddies on the SSH/ADT background also allows their sizes and positions to be evaluated in relation to the Gulf Stream pathway.”

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.

Response: following the reviewer’s suggestion (option 1), we have recomputed the spectra after projecting all datasets onto a common Lambert Azimuthal Equal-Area coordinate system and interpolating them onto a regular kilometre-scale grid. Figure 2 has been updated accordingly. We have revised the manuscript now at lines: 276 - 305.

“In analogy to the approach adopted in Droghei et al. (2018) and Ciani et al. (2019), we selected the North-West Atlantic region and computed one-dimensional wavenumber spectra using a fast Fourier transform (FFT) to quantify the spatial scales effectively represented by each dataset. To avoid the latitude-dependent distortion associated with zonal wavenumbers expressed in degrees, all SSH fields were first projected onto a Lambert Azimuthal Equal-Area coordinate system centred on the study domain. The fields were then interpolated onto a regular kilometre-scale grid with $dx = dy = 12.5$ km. The one-dimensional zonal and meridional spectra were computed on this projected grid using FFTs. The resulting wavenumbers are therefore expressed in cycles / km, with the corresponding wavelengths shown in kilometres.

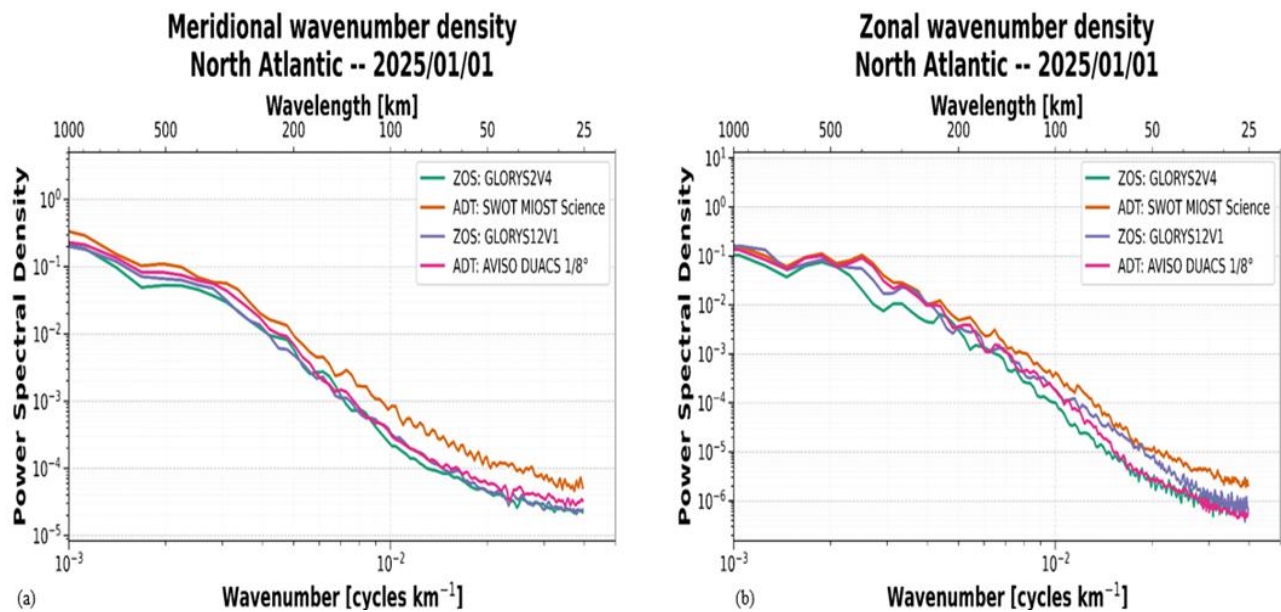


Fig.2. Power spectral density of SSH, corresponding to ADT for the satellite altimetry products, for the four datasets on 1 January 2025 over the selected North Atlantic region. All fields were projected onto a Lambert Azimuthal Equal-Area coordinate system and interpolated onto a regular kilometre-scale grid with $dx = dy = 12.5$ km, before computing the spectra. Panel (a) shows the meridional spectrum and panel (b) the zonal spectrum. Wavenumbers are expressed in cycles / km, with the corresponding wavelengths shown in kilometres on the upper x-axis.

Fig. 2 reports the comparative spectral analysis of the four SSH fields along the zonal and meridional directions on the projected grid. At large spatial scales, approximately larger than 200 km, all spectra exhibit comparable energy levels, indicating a consistent representation of the dominant mesoscale structures. At shorter wavelengths, differences become more evident. In the meridional direction, SWOT systematically exhibits the highest spectral energy over most of the resolved range, with a clearer separation from the other products below about 100 km. DUACS and GLORYS12V1 show intermediate and broadly similar energy levels, whereas GLORYS2V4 generally displays the weakest small-scale energy. In the zonal direction, the four spectra are also similar at large scales, but a progressive separation emerges toward smaller wavelengths, with SWOT again retaining the highest spectral energy. Over much of the mesoscale range, GLORYS12V1 and DUACS remain relatively close, while GLORYS2V4 decreases more rapidly. At wavelengths shorter than about 50 km, all spectra show a marked decay, but SWOT continues to preserve the largest amount of variance, and GLORYS12V1 generally retains more small-scale zonal energy than GLORYS2V4. Overall, these spectra indicate that the SWOT product better preserves fine-scale variability, while the higher-resolution GLORYS12V1 reanalysis captures small-scale mesoscale energy more effectively than GLORYS2V4.”

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.

Response: the different GLORYS12V1 and GLORYS2V4 eddy counts previously reported in Tables 2 and 3 resulted from the use of different evaluation periods. For the comparison with SWOT MIOST Science, the eddy counts had been calculated over a longer period, extending by an additional 59 days to 30 June 2025, whereas the DUACS comparison ended on 1 May 2025. We have now recalculated the values using the same common evaluation period, from 1 August 2023 to 1 May 2025, for all datasets. Consequently, the GLORYS2V4 and GLORYS12V1 eddy counts are now identical

in Tables 2 and 3. The time-series and histogram analyses are unchanged because they had already been calculated using the common evaluation period. Following the comments of both reviewers, the additional results obtained using minimum lifetime thresholds of 7 and 28 days have also been moved to Appendix A, while Tables 2 and 3 now report only the results for the primary 4-day minimum lifetime threshold. We have revised the manuscript, now at lines: 392 - 397.

<i>Dataset</i>	<i>Anticyclonic</i>	<i>% w.r.t. DUACS</i>	<i>Cyclonic</i>	<i>% w.r.t. DUACS</i>
<i>DUACS</i>	<i>125580 (4 days)</i>	<i>-</i>	<i>125229 (4 days)</i>	<i>-</i>
<i>GLORYS2V4</i>	<i>67961 (4 days)</i>	<i>54.12%</i>	<i>72953 (4 days)</i>	<i>58.26%</i>
<i>GLORYS12V1</i>	<i>97788 (4 days)</i>	<i>77.87%</i>	<i>106861 (4 days)</i>	<i>85.33%</i>

Table 2: total number of anticyclonic and cyclonic eddies identified in the region of study over the evaluation period for DUACS, GLORYS2V4, and GLORYS12V1, for a minimum lifetime threshold of 4 days. The percentage of eddies identified in GLORYS2V4 and GLORYS12V1 with respect to AVISO is also provided.

<i>Dataset</i>	<i>Anticyclonic</i>	<i>% w.r.t. SWOT</i>	<i>Cyclonic</i>	<i>% w.r.t. SWOT</i>
<i>SWOT</i>	<i>124154 (4 days)</i>	<i>-</i>	<i>125676 (4 days)</i>	<i>-</i>
<i>GLORYS2V4</i>	<i>67961 (4 days)</i>	<i>54.74%</i>	<i>72953 (4 days)</i>	<i>58.05%</i>
<i>GLORYS12V1</i>	<i>97788 (4 days)</i>	<i>78.76%</i>	<i>106861 (4 days)</i>	<i>85.03%</i>

Table 3: total number of anticyclonic and cyclonic eddies identified in the region of study over the evaluation period for SWOT MIOST, GLORYS2V4, GLORYS12V1, for a minimum lifetime threshold of 4 days. The percentage of eddies identified in GLORYS2V4 and GLORYS12V1 with respect to SWOT MIOST is also provided.

We have also added a discussion of the cyclonic/anticyclonic imbalance found in the reanalyses compared with the satellite products. The manuscript has been revised accordingly at lines 373 – 384.

“The number of identified eddies is sensitive to the choice of the minimum eddy lifetime threshold, as noted in Section 2.4. Tables 2 and 3 show the total number of anticyclonic and cyclonic eddies detected and tracked with a minimum lifetime of 4 days. The total number of identified and tracked eddies is higher in the satellite products than in GLORYS2V4 and GLORYS12V1, for both anticyclonic and cyclonic eddies. Consistent with its higher spatial resolution, GLORYS12V1 detects more eddies than GLORYS2V4. Tables 2 and 3 also show an almost balanced cyclonic/anticyclonic distribution in the satellite products, but a stronger cyclonic dominance in both reanalyses. This difference is likely related to the different nature of the products. Reanalyses combine assimilated observations with model dynamics and can therefore retain or generate small-scale structures, especially in the eddy-resolving GLORYS12V1 system. In contrast, DUACS and SWOT MIOST are Level-4 gridded satellite products derived from along-track altimetry, and their mapping and interpolation procedures may smooth or merge weak small-scale SSH features. Therefore, the cyclonic excess in the reanalyses should be interpreted with caution. It may reflect model dynamics, data-assimilation effects, and the sensitivity of the closed-contour detection method to weak cyclonic SSH minima, rather than a robust observed polarity asymmetry.”

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.

Response: We have removed the speculation about GLORYS2V4 anomalies, and we have revised the vague message of the following sentences. We have revised the manuscript, now at lines: 386 – 415.

“For each comparison, hits, misses, and false alarms are tabulated over the full evaluation period (1 year and 9 months for both comparisons). These counts tabulated, meaning that they are evaluated by adding up all the counts of hits, misses, and false alarms for the entire verification period, for both cyclonic and anticyclonic eddies are used to compute the Probability of Detection (POD) and the False Alarm Ratio (FAR) (following the Eqs. (2) and (3)). For the total counts of eddies, the results are reported separately for anticyclonic and cyclonic. The results for DUACS-based comparison are shown in Fig. 6(a), Table 2 and Table 4, while in Fig. 6(b), Table 3 and Table 5 for SWOT comparison results.

Dataset	Anticyclonic	% w.r.t. DUACS	Cyclonic	% w.r.t. DUACS
DUACS	125580 (4 days)	-	125229 (4 days)	-
GLORYS2V4	67961 (4 days)	54.12%	72953 (4 days)	58.26%
GLORYS12V1	97788 (4 days)	77.87%	106861 (4 days)	85.33%

Table 2: total number of anticyclonic and cyclonic eddies identified in the region of study over the evaluation period for DUACS, GLORYS2V4, and GLORYS12V1, for a minimum lifetime threshold of 4 days. The percentage of eddies identified in GLORYS2V4 and GLORYS12V1 with respect to AVISO is also provided.

Dataset	Anticyclonic	% w.r.t. SWOT	Cyclonic	% w.r.t. SWOT
SWOT	124154 (4 days)	-	125676 (4 days)	-
GLORYS2V4	67961 (4 days)	54.74%	72953 (4 days)	58.05%
GLORYS12V1	97788 (4 days)	78.76%	106861 (4 days)	85.03%

Table 3: total number of anticyclonic and cyclonic eddies identified in the region of study over the evaluation period for SWOT MIOST, GLORYS2V4, GLORYS12V1, for a minimum lifetime threshold of 4 days. The percentage of eddies identified in GLORYS2V4 and GLORYS12V1 with respect to SWOT MIOST is also provided.

Verification metric	GLORYS12V1	GLORYS2V4	Diff.	Diff. %
POD	0.628	0.460	0.168	36.52%
FAR	0.230	0.182	0.048	26.37%

Table 4: aggregated eddy verification statistics for the Probability of Detection (POD) and False Alarm Ratio (FAR) evaluated over the period, from 1st August 2023 to 1st May 2025, for the comparison with AVISO DUACS 1/8°, with a minimum lifetime threshold of 4 days.

Verification metric	GLORYS12V1	GLORYS2V4	Diff.	Diff. %
POD	0.624	0.455	0.169	37.14%
FAR	0.237	0.194	0.043	22.16%

Table 5: aggregated eddy verification statistics for the Probability of Detection (POD) and False Alarm Ratio (FAR) evaluated over the period 1st August 2023 to 1st May 2025, for the comparison with SWOT MIOST, with a 4-day minimum lifetime threshold.

There is certainly a slight decrease, when comparing the total number of anticyclonic eddies to that of cyclonic eddies, which is particularly evident for GLORYS12V1. Based on the observed data, the two satellite products show nearly balanced polarity counts, with differences not exceeding 1.3 %. In contrast, cyclonic daily detections exceed anticyclonic

detections by approximately 7.3 % in GLORYS2V4 and 9.3 % in GLORYS12V1. As shown in the eddy matching example in Fig. 5, there is a significant difference in terms of hits, misses, and false alarms between the two reanalyses. While the number of hits increases by about 37% when switching from GLORYS2V4 to GLORYS12V1, the number of misses decreases significantly, by about 31%, whereas the number of false alarms increases dramatically, around 77%. This demonstrates how GLORYS12V1 shows greater agreement with both observational datasets, but at the same time generates a significantly higher number of false alarms, which perfectly explains the increase in the FAR, as shown in Fig. 6. Furthermore, the daily POD and FAR fluctuate but remain stationary over time, with no apparent seasonal cycle as shown in Fig. 6(a) and Fig. 6(b). As noted above, the calculation is also applied to the minimum threshold of 7 and 28 days without any observational gaps for each eddy within this time interval, in Appendix A. This sensitivity test shows that short-lived eddies contribute to part of the variability in the verification metrics.”

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.

Response: We removed the previous conclusion that the additional eddies detected in SWOT are not detected by GLORYS2V4. The revised text now states that GLORYS2V4 matches a substantial proportion of the eddies identified in both satellite products, with only a small decrease in POD when SWOT MIOST Science is used as the reference. We have revised the manuscript, now at lines: 429 – 441.

“This indicates that, specifically regarding GLORYS2V4, the POD values are very similar in the DUACS and SWOT comparisons. There is a slight decrease when SWOT is used as reference. The decrease is about 0.005 for the 4-day threshold. GLORYS2V4 can match part of the eddies detected by SWOT, but with slightly lower efficiency than for DUACS. The FAR also slightly increases in the SWOT comparison, by about 0.012, showing a small increase in terms of false alarms. Overall, these differences are small, suggesting that the verification results for GLORYS2V4 are relatively insensitive to the choice between the two satellite reference products. However, the overall POD values, that are around 62%, for both comparisons with GLORYS12V1, show that a significant proportion of observed eddies (around 38%) remains unmatched. It highlights the presence of limitations in their high-resolution applications, such as difficulties in representing small or short-lived features and maintaining eddy-track continuity. As the minimum lifetime threshold increases, the POD improvement of GLORYS12V1 relative to GLORYS2V4 becomes smaller, whereas the reduction in FAR becomes more pronounced. GLORYS12V1 has a higher FAR than GLORYS2V4 for the 4-day and 7-day lifetime thresholds, likely because it detects a larger number of eddies, some of which do not find a corresponding match in the satellite datasets. At the 28-day threshold, this relationship reverses, with GLORYS12V1 showing a lower FAR than GLORYS2V4 (see Appendix A for further details).”

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.

Response: Following both of the reviewers’ suggestion, we have added Appendix B, which reports additional block-bootstrap sensitivity analyses using block lengths of 7, 20, and 30 days, as well as block lengths corresponding to the

mean eddy lifetimes estimated separately from the DUACS 1/8° and SWOT MIOST Science datasets. The mean satellite eddy lifetimes used to define these additional block lengths are also reported in Appendix B. We have revised the manuscript, now at lines: 442 – 460.

“To further assess the robustness of these findings, a block bootstrap resampling (95% confidence intervals) was performed using 4-day blocks, consistent with the imposed minimum lifetime threshold. Results for both SWOT and DUACS (Tables 6 and 7) confirm that GLORYS12V1 systematically outperforms GLORYS2V4 in terms of POD (0.618 - 0.620 vs 0.448 - 0.452), with a statistically significant difference (~ 0.17) that is insensitive to block length. This gain is consistently associated with a moderate increase in FAR (~ 0.045 - 0.049), confirming a robust trade-off between detection sensitivity and reliability. The agreement between the two observational references strengthens confidence in the statistical significance of the results, although it does not fully remove uncertainties related to the definition and detectability of small-scale or short-lived eddy structures.

Verification Metric	GLORYS12V1	GLORYS2V4	<i>Difference (12V1 – 2V4)</i>
POD (Mean)	0.618	0.448	0.170
POD (95% CI)	[0.609, 0.627]	[0.441, 0.454]	[0.160, 0.182]
FAR (Mean)	0.235	0.190	0.045
FAR (95% CI)	[0.229, 0.242]	[0.183, 0.198]	[0.038, 0.052]

Table 6. Bootstrap resampling for SWOT MIOST Science comparison, for a minimum lifetime threshold of 4 days, and a block length of 4 days.

Verification Metric	GLORYS12V1	GLORYS2V4	<i>Difference (12V1 – 2V4)</i>
POD (Mean)	0.620	0.452	0.168
POD (95% CI)	[0.610, 0.629]	[0.442, 0.463]	[0.154, 0.180]
FAR (Mean)	0.235	0.187	0.049
FAR (95% CI)	[0.230, 0.237]	[0.181, 0.192]	[0.040, 0.054]

Table 7. Bootstrap resampling for DUACS 1/8° comparison, for a minimum lifetime threshold of 4 days, and a block length of 4 days.

As additional sensitivity tests of the coupled block bootstrap, we also performed the bootstrap resample using 7-day, 20-day and 30-day blocks, and the mean lifetimes of both satellite eddies. The results are reported in Appendix B. The relative ranking of GLORYS12V1 and GLORYS2V4 and the main conclusions of the 4-day analysis remain consistent across the different block lengths, although the estimated confidence intervals vary with the choice of block length. Therefore, the 30-day test is included only as an additional robustness check. Very large block lengths may reduce the variability of the

resampling procedure and may partly weaken the main purpose of the block bootstrap, which is to account for temporal dependence in the data when the underlying statistical distribution is unknown.”

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.

Response: We have removed the previous statement that long-lived eddies are more robust and easier to match, as Figures 8 and 10 do not analyse the dependence of the verification metrics on eddy lifetime. We have added more detailed calculations in Appendix C. We have revised the manuscript, now at lines: 461 – 471.

“ In general, as the minimum eddy lifetime threshold increases, the FAR and POD values of each reanalysis products, for both satellite comparisons, tend to converge toward similar levels as described in Appendix A for longer-lived eddies (especially for 28-day minimum eddy lifetime threshold). Figures 7 and 9 show the dependence of the eddy matching statistics (hits, misses and false alarms) on the physical properties of the eddies, namely the effective radius and amplitude, respectively. These statistics are calculated to show how eddies are distributed across different ranges of radius and amplitude, and thus to identify the ranges with the highest clustering. For hits and misses, the radius and amplitude were taken from the satellite (the reference) product, because these eddies refer to observed features that are either detected or not detected by the reanalysis. For false alarms, the radius and amplitude are taken from the corresponding reanalysis product. Each bin of the histogram has a width of 10 km and 2 cm, respectively for eddy radius and eddy amplitude. The highest counts of misses and false alarms are associated with small eddies, specifically for those with radii below 50 km and amplitudes below 5 cm.

Figures 8 and 10 show the dependence of the verification metrics in terms of the same physical eddy features.”

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.

Response: we have revised the text preceding Figures from 7-12 (previously they are Figure 8-12) (paragraph above the following comments of the reviewer). We have also modified the caption of each Figure (from 7 to 12), in order to clarify this concept. Here, we show only an example for Fig.7. We have revised the manuscript, as noted in the previous comment, at lines: 461 - 477.

“In general, as the minimum eddy lifetime threshold increases, the FAR and POD values of each reanalysis products, for both satellite comparisons, tend to converge toward similar levels as described in Appendix A for longer-lived eddies (especially for 28-day minimum eddy lifetime threshold). Figures 7 and 9 show the dependence of the eddy matching statistics (hits, misses and false alarms) on the physical properties of the eddies, namely the effective radius and amplitude, respectively. These statistics are calculated to show how eddies are distributed across different ranges of radius and amplitude, and thus to identify the ranges with the highest clustering. For hits and misses, the radius and amplitude were taken from the satellite (the reference) product, because these eddies refer to observed features that are

either detected or not detected by the reanalysis. For false alarms, the radius and amplitude are taken from the corresponding reanalysis product. Each bin of the histogram has a width of 10 km and 2 cm, respectively for eddy radius and eddy amplitude. The highest counts of misses and false alarms are associated with small eddies, specifically for those with radii below 50 km and amplitudes below 5 cm.

Figures 8 and 10 show the dependence of the verification metrics in terms of the same physical eddy features.

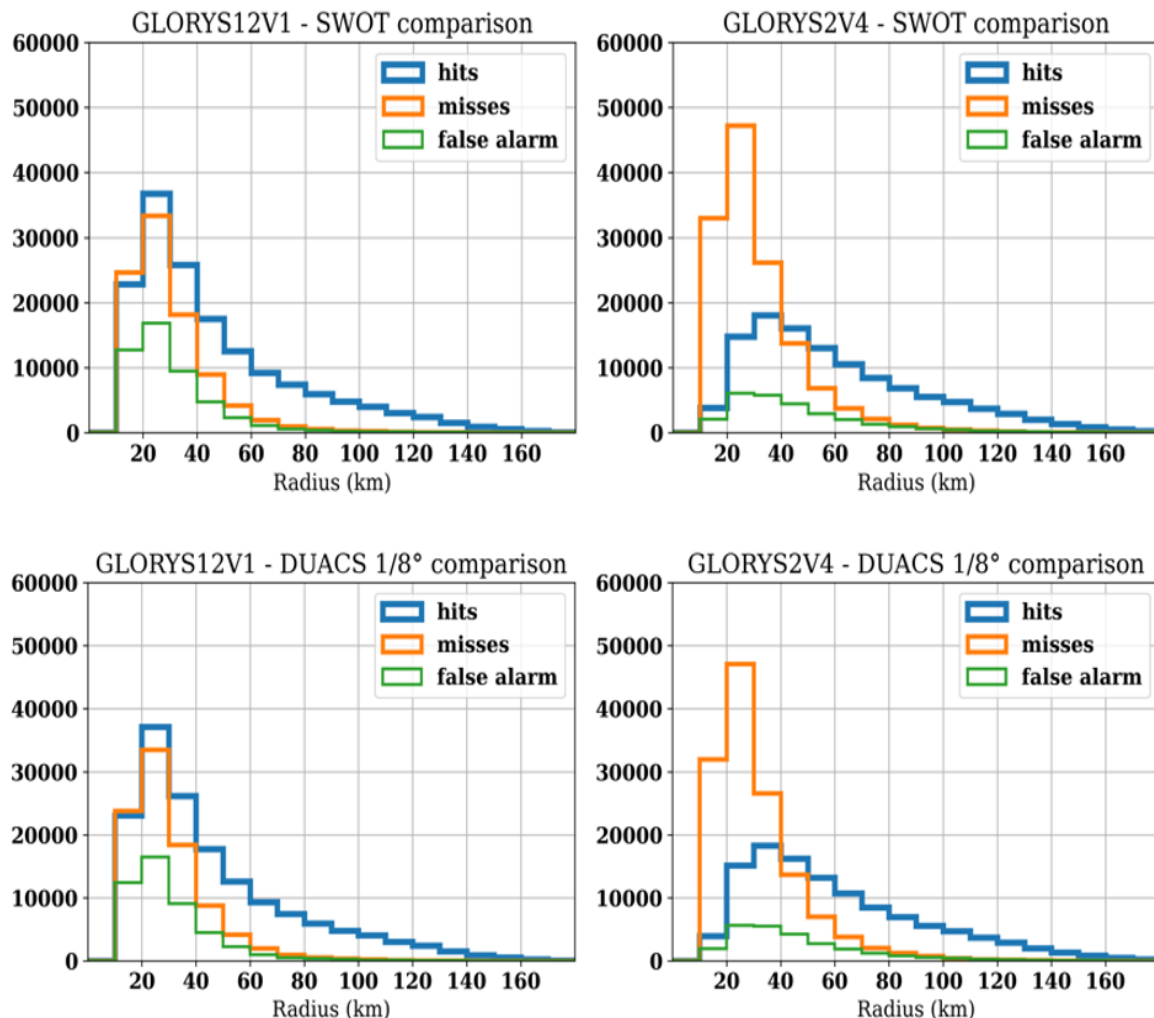


Fig. 7. On the top: eddy matching histograms for the total number of hits (blue), misses (orange), false alarms (green) for GLORYS12V1 (left) and GLORYS2V4 (right), compiled as a function of effective radius (km) over the evaluation period, for SWOT comparison, with a lifetime threshold of 4 days. In the bottom: eddy matching histograms compiled as a function of effective radius (km) over the same evaluation period, for comparison with DUACS, with a lifetime threshold of 4 days. The radius of hits and misses refers to the satellite (reference) eddies, while the radius of false alarms refers to those of unmatched eddies detected in the reanalysis.

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.

Response: We have corrected the figure references following the renumbering of the figures and removed the statement that false alarms dominate the smallest eddy classes. The revised text now states more precisely that the FAR is higher in

the smallest radius and amplitude bins and, where applicable, exceeds 30%. We have also revised the corresponding figure captions for clarity. We have revised at lines: 478 - 483 (for Fig. 8) and 509 - 513 (for Fig. 10).

“

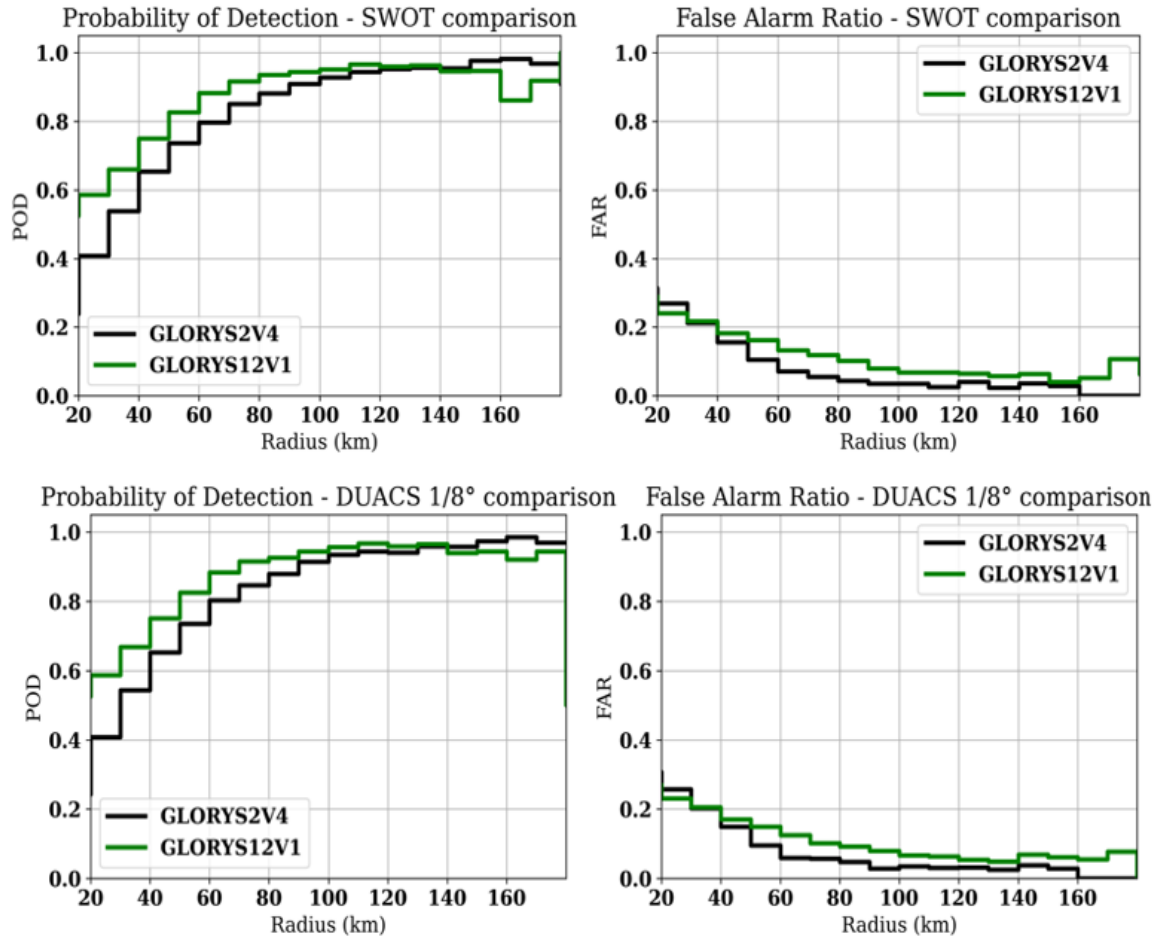


Fig.8. On the top: histograms of the probability of detection (POD; left) and the false alarm ratio (FAR; right), over the evaluation period, evaluated as a function of radius (km) for comparison with SWOT, with a lifetime threshold of 4 days. In the bottom: histograms of the POD and FAR as a function of radius (km), for the comparison with DUACS, with a lifetime threshold of 4 days. The radius of hits and misses refers to the satellite (reference) eddies, while the radius of false alarms refers to those of unmatched eddies detected in the reanalysis.

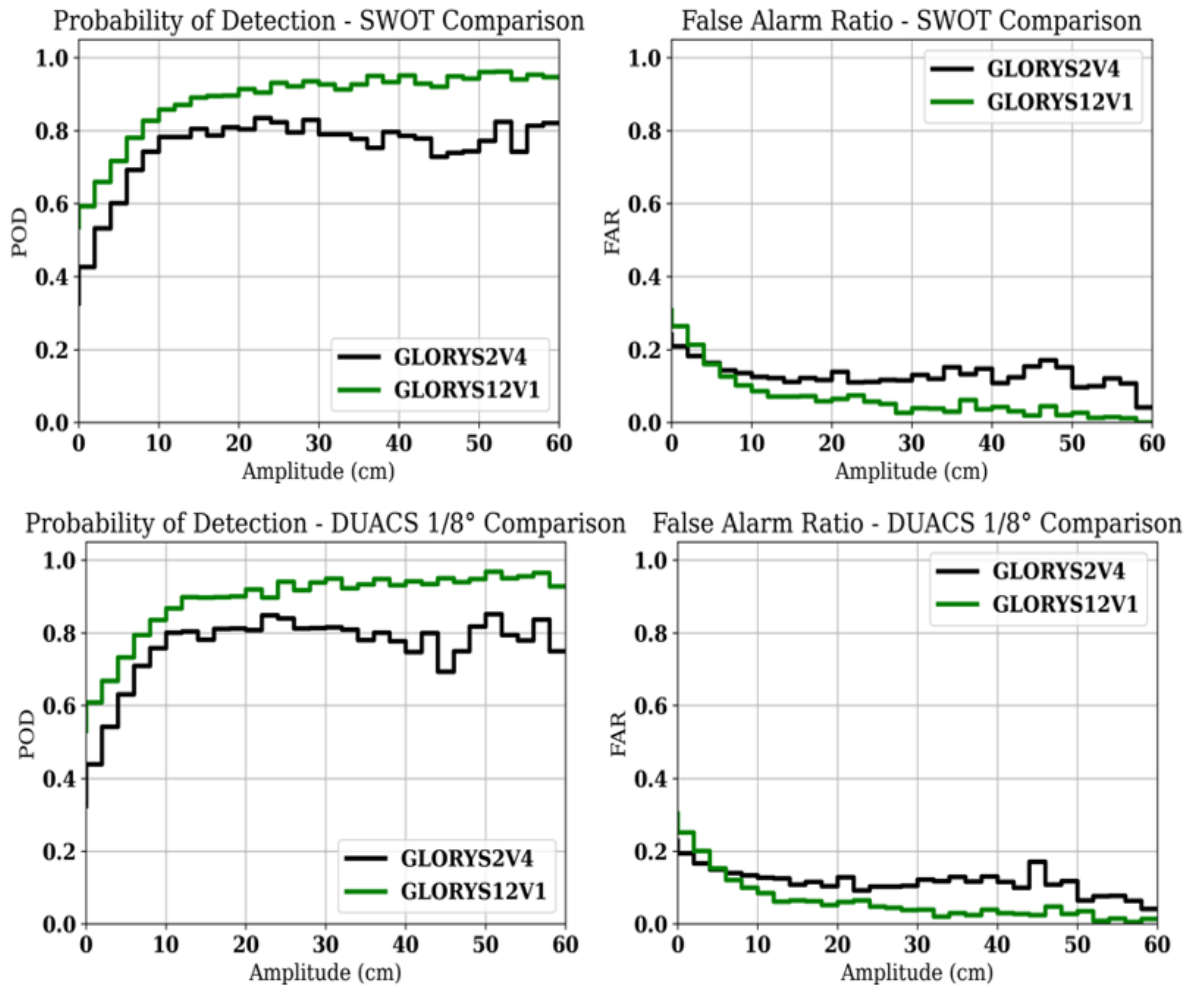


Fig. 10. On the top: histograms of the probability of detection (POD; left) and the false alarm ratio (FAR; right), over the evaluation period, evaluated as a function of amplitude (cm) for comparison with SWOT. In the bottom: histograms of the POD and FAR as a function of amplitude (cm), over the evaluation period, for the comparison with DUACS. The amplitude of hits and misses refers to the satellite (reference) eddies, while the amplitude of false alarms refers to those of unmatched eddies detected in the reanalysis..”

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.

Response: We have replaced the incorrect duplicated panels with the correct results for the DUACS and SWOT comparisons and corrected the corresponding panel titles. We have also revised the text to remove the incorrect statement that POD remains below 50% and FAR exceeds 30% for the smallest eddies, now at lines: 484 – 508.

“For the DUACS comparison, Figures 8 and 10 show that FAR is generally higher in the smallest effective-radius and amplitude bins, exceeding 30% in some of these classes and highlighting the greater difficulty in verifying the smallest features in this gridded product. The FAR remains close or higher than 20% until 50 km in terms of effective radius. At the same time, the POD of GLORYS12V1 is close to 60% even for small radius and amplitude values, showing a relevant fraction of small observed eddies is still detected by the reanalyses. The largest proportion of misses is consistently associated with small eddies, specifically those with radii below 50 km and amplitudes under 4 cm. The comparison with

SWOT shows a similar dependence on eddy size. The largest differences in the verification metrics are again found for small-radius and small-amplitude eddies. The POD of GLORYS12V1 remains close to 60%, while the GLORYS2V4 FAR is generally higher in the smallest bins. This indicates that the main difficulty is not only the detection of small observed eddies, but also the presence of unmatched (especially for the small eddies) produced by the reanalyses. However, for both satellite comparisons, the verification skills generally improve as eddy radius and amplitude increase. This suggests a better matching capability for larger and stronger mesoscale eddies, which are less affected by small-scale noise or detection uncertainty. The differences between the DUACS and SWOT comparison are most evident for small eddies, while they become smaller for larger eddies. For this reason, Appendix C includes additional analyses of the verification metrics using different amplitude-based subsets, including eddies with amplitudes greater than 5, 10, 15, and 20 cm, as well as eddies with amplitudes below 10 cm. Appendix C also presents the eddy matching histograms using a semi-logarithmic scale for the y-axis, which represents the number of eddies. These additional analyses allow us to assess more clearly how the verification results change when low-amplitude eddies are progressively excluded, when the analysis focuses on stronger mesoscale structures, and when the contribution of the most numerous small-amplitude eddies is examined in greater detail. ”

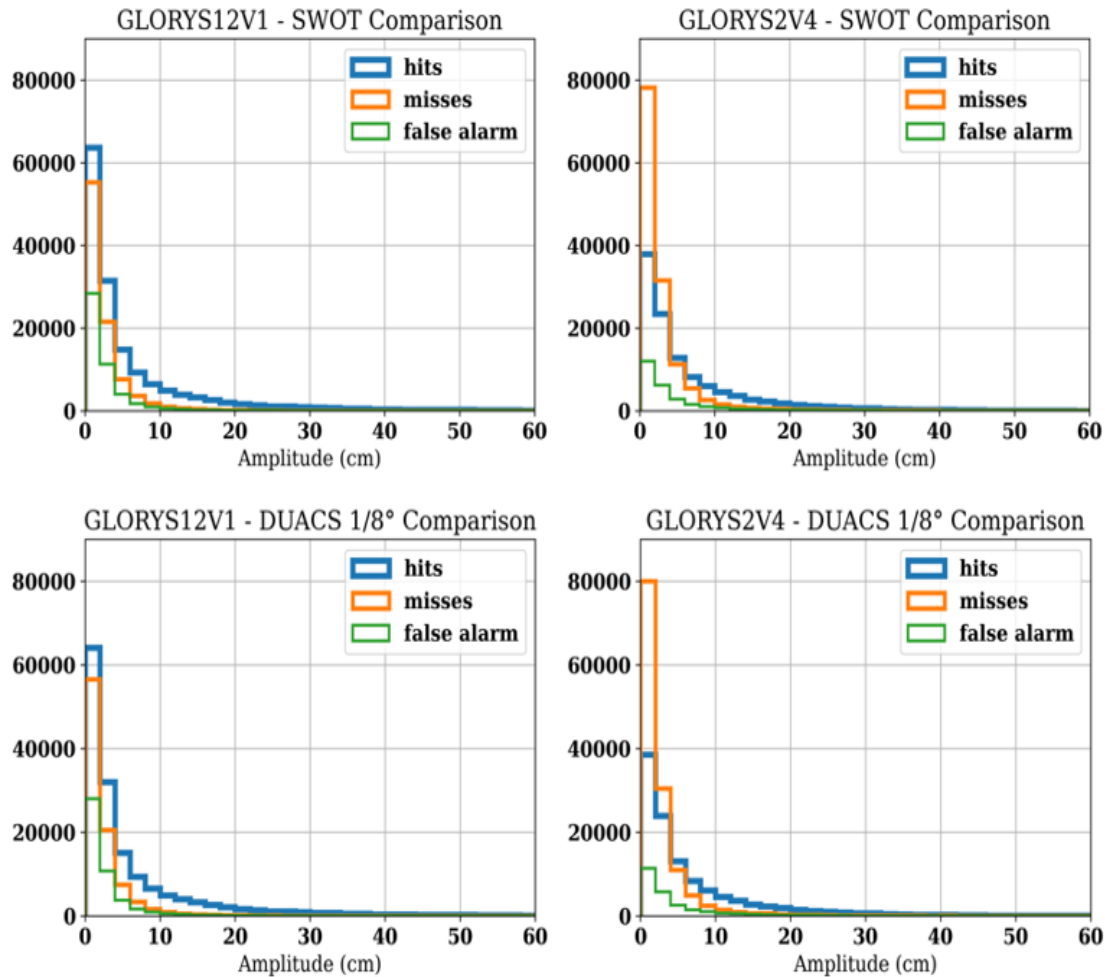


Fig.9. On the top: eddy matching histograms for the total number of hits (blue), misses (orange), false alarms (green) for GLORYS12V1 (left) and GLORYS2V4 (right), compiled as a function of amplitude (cm) over the evaluation period, for SWOT comparison. In the bottom: eddy matching histograms compiled as a function of amplitude (cm) over the evaluation period, for comparison with DUACS 1/8°. The amplitude of hits and misses refers to the satellite (reference) eddies, while the amplitude of false alarms refers to those of unmatched eddies, detected in the reanalysis.

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.

Response: we have revised the Figures in the manuscript: now they are Fig.9 and Fig. 10, at lines: 503 - 513.

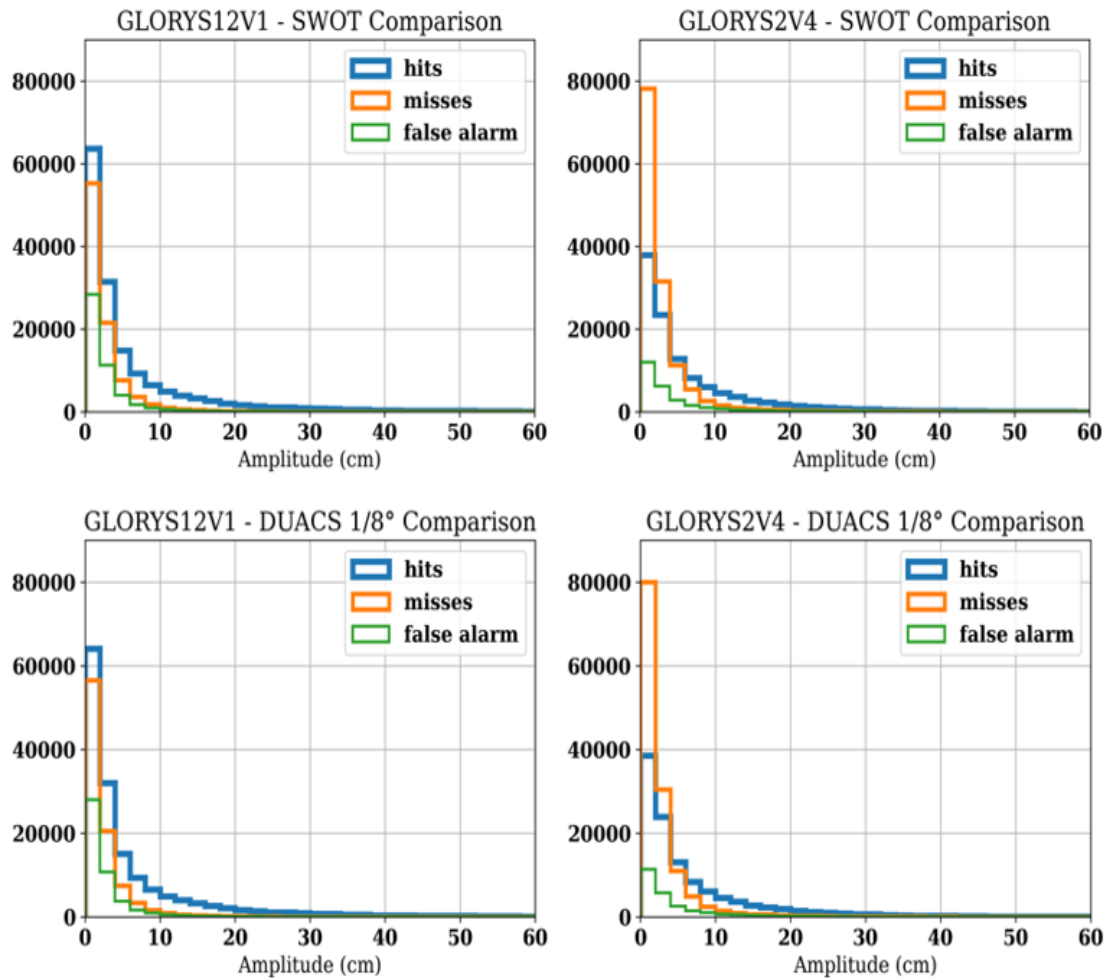


Fig.9. On the top: eddy matching histograms for the total number of hits (blue), misses (orange), false alarms (green) for GLORYS12V1 (left) and GLORYS2V4 (right), compiled as a function of amplitude (cm) over the evaluation period, for SWOT comparison. In the bottom: eddy matching histograms compiled as a function of amplitude (cm) over the evaluation period, for comparison with DUACS 1/8°. The amplitude of hits and misses refers to the satellite (reference) eddies, while the amplitude of false alarms refers to those of unmatched eddies, detected in the reanalysis.

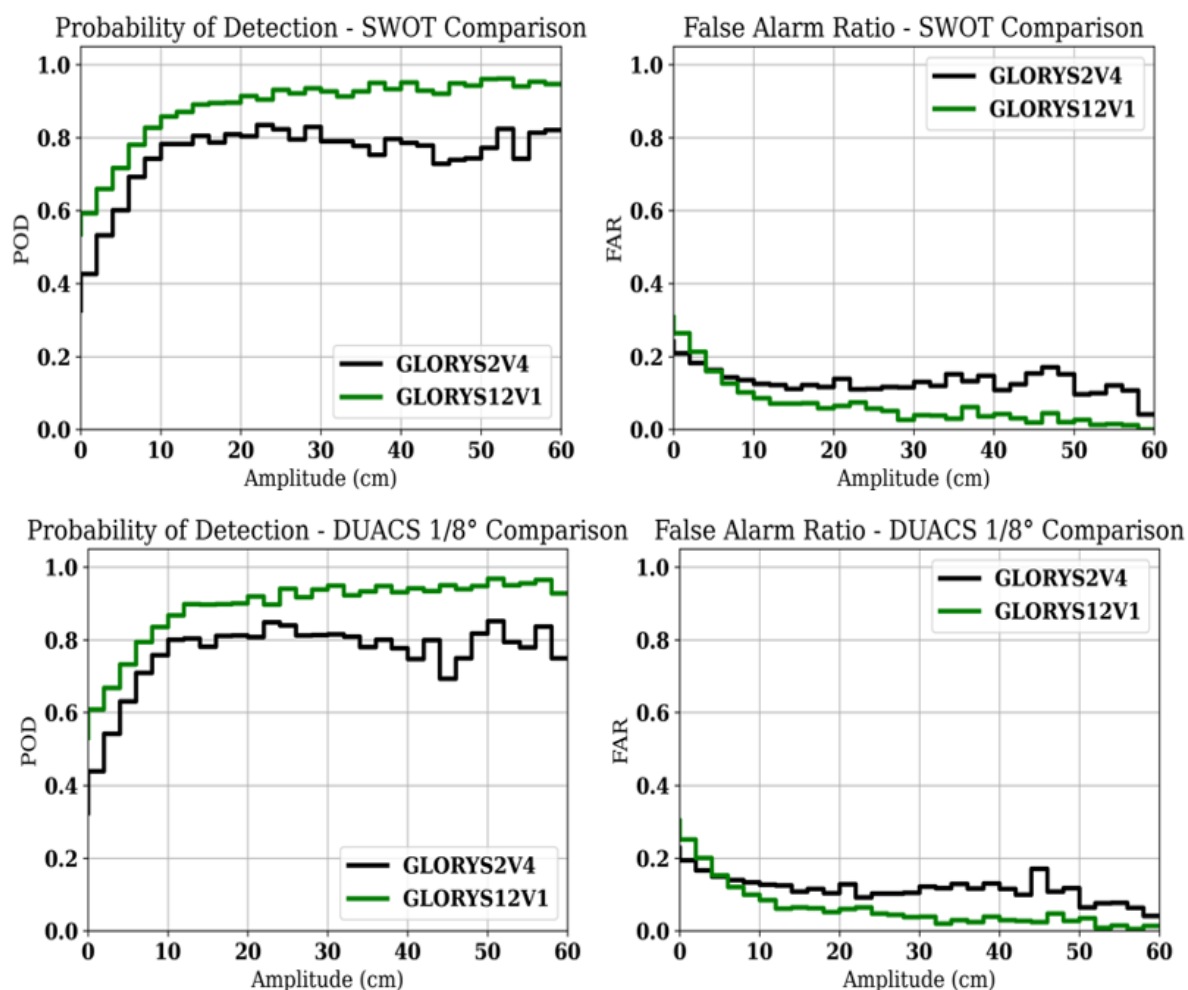


Fig. 10. On the top: histograms of the probability of detection (POD; left) and the false alarm ratio (FAR; right), over the evaluation period, evaluated as a function of amplitude (cm) for comparison with SWOT. In the bottom: histograms of the POD and FAR as a function of amplitude (cm), over the evaluation period, for the comparison with DUACS. The amplitude of hits and misses refers to the satellite (reference) eddies, while the amplitude of false alarms refers to those of unmatched eddies detected in the reanalysis.“

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.)

Response: The figures have now been corrected so that GLORYS12V1 is represented by the solid line and GLORYS2V4 by the dotted line, consistently with the corresponding cumulative distributions and the description in the manuscript. Following the renumbering of the figures, these are now Figures 11 and 12, at lines: 546 - 555.

“

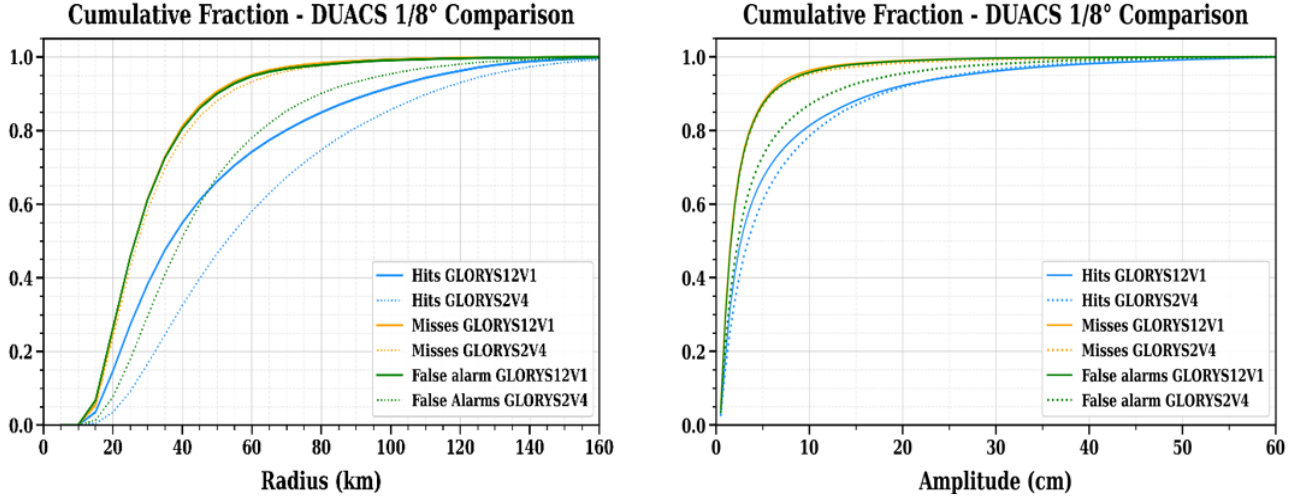


Fig. 11. Cumulative distribution of hits (blue), misses (orange) and false alarms (green) for GLORYS12V1 (solid) and for GLORYS2V4 (dotted) as a function of radius (left) and amplitude (right), for the comparison with DUACS, with a lifetime threshold of 4 days. The radius and the amplitude of hits and misses refers to the satellite (reference) eddies, while the radius and the amplitude of false alarms refers to the unmatched eddies detected in the reanalysis.

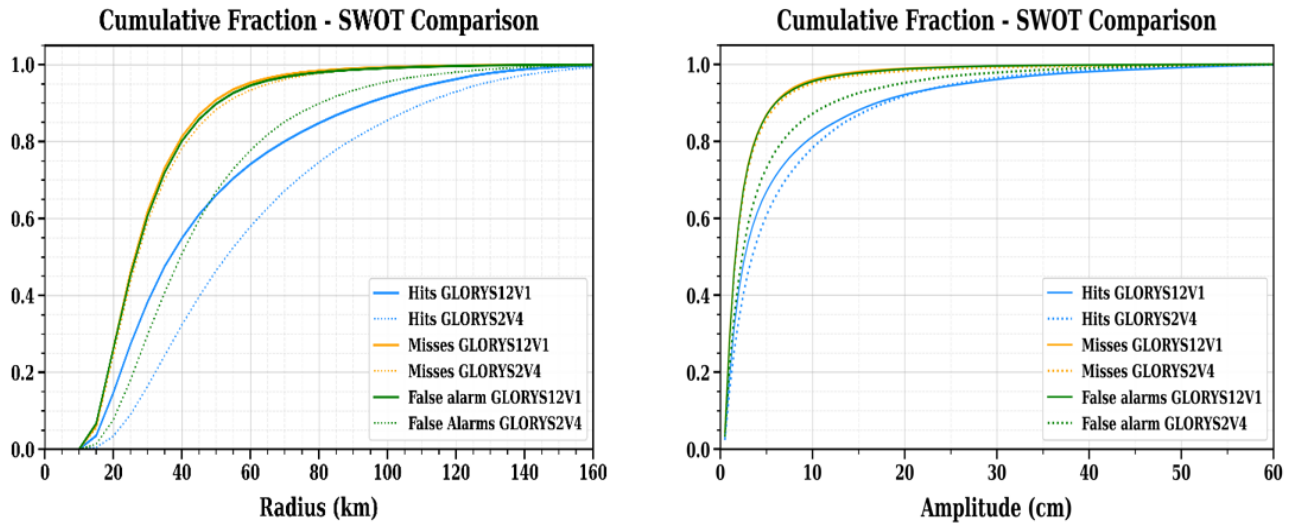


Fig.12. Cumulative distribution of hits (blue), misses (orange) and false alarms (green) for GLORYS12V1 (solid) and for GLORYS2V4 (dotted) as a function of radius (left) and amplitude (right), for the comparison with SWOT, with a lifetime threshold of 4 days. The radius and the amplitude of hits and misses refers to the satellite (reference) eddies, while the radius and the amplitude of false alarms refers to the unmatched eddies detected in the reanalysis.”

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.

Response: We thank the reviewer for pointing out this apparent inconsistency. After checking the calculations and the values reported in Tables 8 and 9, we found that the values previously reported for GLORYS12V1 were incorrect. We have corrected the tables and the corresponding discussion. The corrected values are consistent with the error distributions shown in the figure and reveal clear differences between the two reanalyses. We have revised the manuscript, now at lines: 566 – 618.

“Determining whether an observed eddy has a corresponding match in a reanalysis dataset is important, but it does not fully capture how accurately that eddy is represented. As illustrated in Fig. 5, and explained in Section 2.4, some matched eddies exhibit significant differences in terms of radius, amplitude, shape, position, compared to the satellite data, despite correctly identifying the same features (see Eq. 1). These discrepancies are expected to vary between the two reanalysis systems given their differing spatial resolutions, and thus their capacity to resolve mesoscale variability. So, we try to answer the question about the possibility to detect differences in the properties of matched eddies between the two reanalysis systems used in this study. To quantify these differences, we analyze the errors in radius, amplitude, and centre displacements for each “matched” eddy (hit). For each of these properties, we compute binned, normalized histograms that illustrate the distribution of radius error (difference between satellite and reanalysis effective radius), amplitude error (difference in SSH extremum relative to the outer contour between satellite and reanalysis matched eddy), separation distance between satellite and reanalysis matched eddy centroid.

Unlike the tracking cost function (Eq. 1), these quantities are presented without the normalization factors (i.e., not scaled by A_1 , R_1 , or the 125 km distance), to retain their physical units (km or cm). While this approach no longer allows a direct assessment of their relative contribution to the overall cost function, it provides a clearer, physically interpretable picture of where and how the reanalyses diverge from satellite data (facilitates a physical interpretation of the histograms). The resulting normalized histograms (normalized by the total number of hits) are shown in Fig. 13(a) and Fig. 13(b). The distributions of the cost function and associated errors for matched eddies reveal clear differences between GLORYS2V4 and GLORYS12V1, for all terms of the cost function (Eq. 1). Indeed, from Table 8, the mean cost function values, for DUACS comparison, are 1.76 for GLORYS2V4 and 1.31 for GLORYS12V1, indicating a clear reduction in the cost function for each matched eddy for the eddy-resolving (GLORYS12V1) system. Likewise, the mean errors in effective radius (24.62 km vs. 16.81 km), the mean amplitude errors (5.43 cm vs 3.55 cm) and centroid distance (57.49 km vs. 49.85 km) are quite different with general differences between the two reanalysis systems.

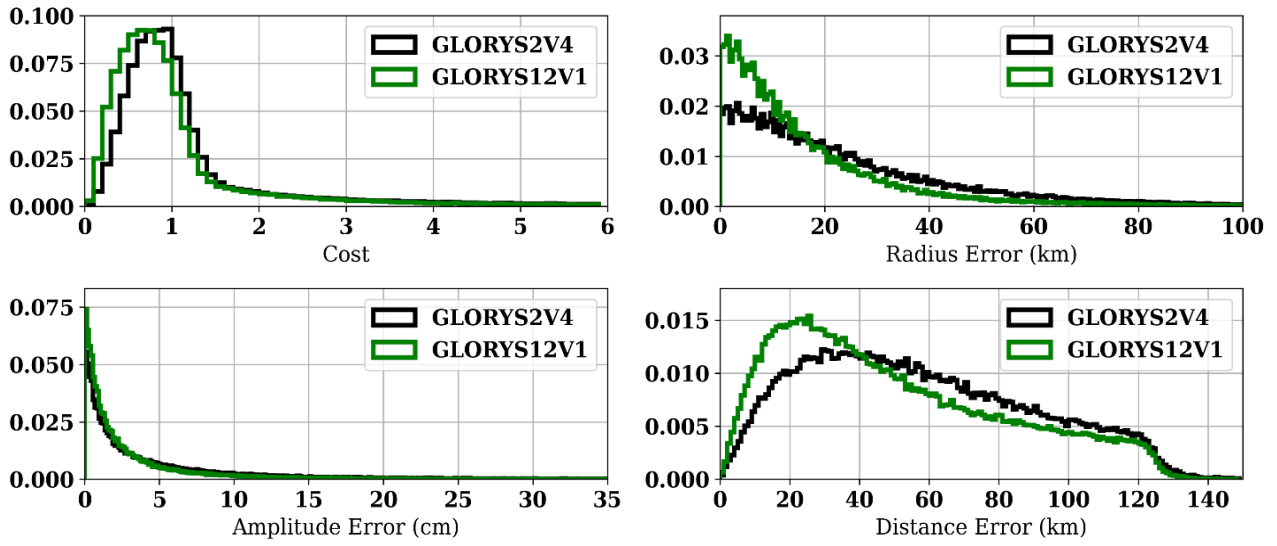
Mean error	Cost	Radius (Km)	Amplitude (cm)	Distance (Km)
GLORYS2V4	1.76	24.62	5.43	57.49
GLORYS12V1	1.31	16.81	3.55	49.85
% diff.	-25.57%	-31.72%	-34.62%	-13.29%

Table 8. Error contributions to the cost function for all hits evaluated over the period 1st August 2023 to 1st May 2025, for DUACS comparison. Mean values for the cost function as well as contributions from radius (km), amplitude (cm), and distance errors (km). The percentage difference crossing from GLORYS2V4 to GLORYS12V1 is provided in the bottom row.

Mean error	Cost	Radius (Km)	Amplitude (cm)	Distance (Km)
GLORYS2V4	1.79	24.92	5.53	57.86
GLORYS12V1	1.33	16.79	3.59	50.29
% diff.	-25.70%	-32.62%	-35.08%	-13.06%

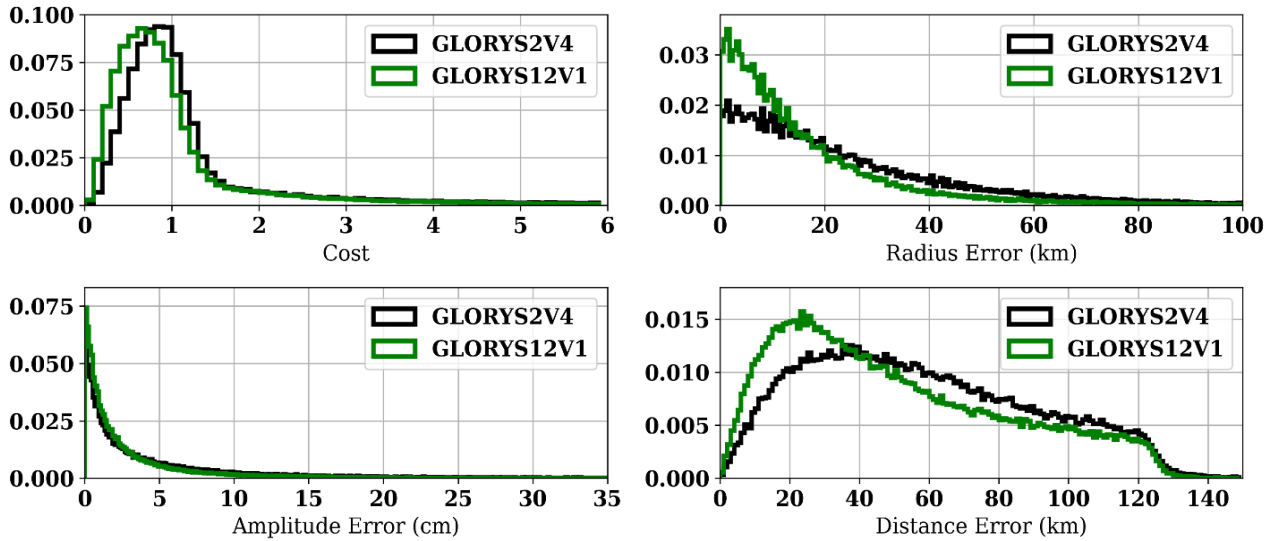
Table 9. Error contributions to the cost function for all hits evaluated over the period 1st August 2023 to 1st May 2025, for SWOT comparison. Mean values for the cost function as well as contributions from radius (km), amplitude (cm), and distance errors (km). The percentage difference crossing from GLORYS2V4 to GLORYS12V1 is provided in the bottom row.

Hits Cost - DUACS 1/8° comparison



(a)

Hits Cost - SWOT comparison



(b)

Fig. 13. In (a), normalized histograms of values for matched eddies between DUACS and GLORYS2V4 (black line) or GLORYS12V1 (green line). While in (b), the comparison between SWOT and GLORYS2V4 (black line) or GLORYS12V1 (green line). The distribution of values for the total cost is shown in the top panel on the left. Also shown are the normalized distributions of differences in terms of radius in the top on the right, amplitude in the bottom on the left, and distance between eddy positions, in the bottom on the right.

For SWOT comparison (see Table 9), the differences between the two reanalyses become somewhat more pronounced than in the DUACS comparison, for all terms of cost function (most notably in the error regarding the effective radius, but except for the distance between the centroid of eddies). The mean cost function is 1.79 for GLORYS2V4 and 1.33 for GLORYS12V1, corresponding to a -25.70% difference, crossing from GLORYS2V4 to GLORYS12V1. The amplitude error is 5.53 cm versus 3.59 cm (-35.08% in difference, from GLORYS2V4 to GLORYS12V1). The effective radius error is 24.92 km versus 16.79 km (-32.62% in difference) and the centroid distance error is 57.86 km vs 50.29 km (-13.06% in

difference). These results indicate that, for all terms of the cost function, GLORYS12V1 achieves nearly identical results in both comparisons (Table 8 and Table 9). Meanwhile, GLORYS2V4 shows slightly higher average values for all terms of the cost function when compared to SWOT. These results show that GLORYS12V1 has lower mean errors than GLORYS2V4 for all terms of the cost function in both satellite comparisons. The improvement is particularly evident for the effective radius and amplitude errors, while the reduction in centroid-distance error is smaller but still present. This confirms that the eddy-resolving reanalysis provides a better representation of the physical properties of matched eddies. The advantage of GLORYS12V1 becomes slightly more apparent when the comparison is performed against SWOT. However, the results for the SWOT and DUACS comparisons are very similar. This shows that the relative improvement of GLORYS12V1 over GLORYS2V4 is robust to the choice of satellite reference dataset.”

Minor Issues

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

Response: we have revised the abstract, at lines:12.

AVISO DUACS 1/8°

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

Response: we have revised the manuscript, now at lines: 60 – 76.

“A reference dataset is needed to validate and evaluate the performance of ocean numerical models in representing mesoscale eddies, also as done in previous feature-based eddy verification studies (Smith and Fortin, 2022). The finer spatial resolution of ocean reanalyses creates a problem when using the AVISO DUACS ¼° dataset as a reference, because this dataset cannot capture many small-scale features. Amores et al. (2018) show that signals smaller than about 50 km can be aliased into longer wavelengths in satellite altimetry. Ballarotta et al. (2019) also showed that the effective resolution of AVISO DUACS gridded altimetry maps is lower than their nominal grid spacing. This means that some of the small-scale variability found in high-resolution reanalyses may not be shown correctly in the AVISO DUACS product. For this reason, we did not use the ¼° AVISO DUACS product in this study. Instead, we used the AVISO DUACS dataset at 1/8°, which provides a finer nominal grid spacing and is consistent with the 1/8° grid of the SWOT MIOST product. Its use reduces the nominal grid-spacing mismatch with GLORYS12V1, but it does not overcome the effective-resolution limits of mapped altimetry, which depends on satellite sampling and the mapping procedure rather than on grid spacing alone. Small eddies close to the resolving capability of AVISO DUACS 1/8° are therefore interpreted with caution (Amores et al., 2018; Ballarotta et al., 2019).

The AVISO DUACS 1/8° satellite altimetry dataset, available from 1993 to the present at the global scale, has been widely used as a benchmark to validate and assess numerical models, ocean reanalyses, mesoscale eddy properties in many studies (Taburet et al., 2019; Mason et al., 2019; Xie et al., 2020). In this study, however, we use AVISO DUACS 1/8° only over the period overlapping with SWOT MIOST Science, from 1 August 2023 to 1 May 2025, in order to apply the same verification framework to both satellite reference products.”

Line 68 “the time coverage of SWOT data is limited”, so you mean this data can not be used for studies on decadal 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.

Response: We have revised the manuscript to address this comment together with the previous related comment concerning the limitations of SWOT and the independence of the satellite reference products. The discussion has been reorganized to remove repetition and to distinguish the two main limitations more clearly. The two satellite products are therefore treated as complementary evaluation references rather than as absolute observational truth. We have revised it, now at lines: 77 – 118.

“The availability of observations from the SWOT (Surface Water and Ocean Topography) wide-swath satellite altimetry mission provides another possible reference dataset for evaluating the capability of higher-resolution ocean reanalyses to identify and to track mesoscale eddies over the time in a regional area. SWOT was launched in December 2022 and offers wide-swath measurements of sea surface height, which allows observations of ocean features at much finer spatial scales than traditional altimetry. One of the objectives of SWOT is the observation of two-dimensional ocean surface topography and the characterization of mesoscale and submesoscale ocean circulation. Nevertheless, one important application of SWOT observations is their use in ocean analysis and forecasting systems through data assimilation, since wide-swath SSH observations can help better constrain mesoscale features and smaller-scale variability (Morrow et al., 2019; Fouchet et al., 2025). In the present study, however, we do not use the native SWOT swath observations directly. We use the experimental SWOT MIOST Science product, a global daily Level-4 gridded reconstruction of sea-surface height and derived geostrophic velocities produced by SSALTO/DUACS. MIOST stands for Multiscale Interpolation Ocean Science Topography. It merges irregularly sampled altimeter observations into fields that are continuous in space and time using a linear optimal-interpolation framework in which different covariance functions can be used to represent variability at different spatial and temporal scales. In the configuration considered here, the covariance functions are designed primarily to represent mesoscale geostrophic variability (Ubelmann et al., 2021; CLS and CNES, 2025). SWOT Level-4 sea-level products are not limited to the MIOST reconstruction used in this study. AVISO/DUACS also distributes experimental regional products generated using 4DVarNET and 4DVarQG. The 4DVarNET approach combines a variational data-assimilation formulation with a deep-learning framework to learn both the representation of ocean dynamics and the associated solver (Fablet et al., 2021; Beauchamp et al., 2023). In contrast, 4DVarQG uses a four-dimensional variational scheme constrained by a reduced-order quasi-geostrophic model, introducing an explicit dynamic prior into the reconstruction (Le Guillou et al., 2021). Comparisons over the North Atlantic indicate that 4DVarNET and 4DVarQG are capable of providing finer effective resolution and improving the reconstruction of the energetic and nonlinear structures of the Gulf Stream compared to MIOST, although their performance may vary spatially and neither method currently provides a global mapping solution (Ballarotta et al., 2025). MIOST is selected because it provides a spatially consistent global reconstruction (such as AVISO DUACS 1/8°) covering the intended validation domain, the North Atlantic, up to the upper mid-latitudes (60°N). The 4DVarNET and 4DVarQG experimental maps, as of the study's final implementation (October and November 2025), were currently regional products covering approximately 25–50° N and 80° W–10° W and therefore do not include the northern part of our study region. However, there are some important limitations that must also be considered. First, because the mission is very recent, the time coverage of SWOT data is limited. Therefore, SWOT data cannot yet be used to study decadal or long-term variability of

mesoscale activity. In this study, SWOT is used only for a short-period evaluation of the representation of mesoscale and smaller-scale eddy features in high-resolution reanalyses. Second, since similar nadir altimetry observations are assimilated in some ocean reanalysis products considered in this study, neither AVISO DUACS 1/8° nor SWOT MIOST Science can be regarded as a fully independent verification dataset. DUACS 1/8° is based on multi-mission nadir altimetry, while SWOT MIOST Science combines SWOT wide-swath observations with a multi-mission nadir-altimetry background used to constrain the larger-scale signal. Therefore, both satellite products share, at least partly, observational information with the reanalyses. However, SWOT MIOST Science also introduces additional fine-scale information from SWOT wide-swath measurements, which is not directly assimilated in the reanalysis products considered here. For these reasons, AVISO DUACS 1/8° and SWOT MIOST Science are used as complementary evaluation references rather than as absolute observational truth. The comparison with SWOT provides insight into how well the reanalyses represent fine-scale variability revealed by wide-swath altimetry, whereas the DUACS comparison provides continuity with the longer pre-SWOT altimetry record. This complementary approach allows us to assess the representation of mesoscale eddies in the reanalyses while explicitly acknowledging the limitations associated with the partial dependence of the reference datasets.”

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

Response: we have revised the manuscript, now at lines: 40 - 44.

“The first aim of this study is to evaluate the ability of ocean reanalyses with different spatial resolutions (1/4° and 1/12°) to represent mesoscale eddies in the North Atlantic. Second, we assess how the verification results change when using two different satellite reference products, AVISO DUACS 1/8° and SWOT MIOST Science, in order to understand whether the additional fine-scale information from SWOT modifies the evaluation of reanalysis performance.”

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.

Response: We have revised the manuscript to provide the DOI and access/download date for all satellite and reanalysis products used in this study. This information has been added at lines 150–156 for the satellite products (Section 2.1) and at lines 177–180 for the reanalysis products (Section 2.2).

149 - 155: “The AVISO DUACS 1/8° product used in this study corresponds to the Copernicus Marine Service product SEALEVEL_GLO_PHY_L4_MY_008_047, dataset cmems_obs-sl_glo_phy-ssh_my_allsat-l4-duacs-0.125deg_P1D, DOI: 10.48670/moi-00148. This product was accessed and downloaded on 31 October 2025.

The SWOT MIOST Science product used in this study corresponds to the AVISO/DUACS Experimental Multimission Gridded Level-4 Sea Surface Heights and Velocities product including SWOT observations, generated using the MIOST mapping method, version 2.0.1, DOI: 10.24400/527896/A01-2025.001. This product was accessed and downloaded from the v_2_0_1/miost/science directory on 1 November 2025. “

176-179: “GLORYS12V1 was obtained from the Copernicus Marine Service product GLOBAL_MULTIYEAR_PHY_001_030, DOI: 10.48670/moi-00021. The product was accessed and downloaded on 1 November 2025.

GLORYS2V4 was obtained from the Copernicus Marine Service product GLOBAL_MULTIYEAR_PHY_ENS_001_031, DOI: 10.48670/moi-00024. The product was accessed and downloaded on 1 November 2025. ”

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

Response: we have tried to check similar mistakes for the whole 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.

Response: We have revised and shortened Sections 2.1 as much as possible, while also taking into account the comments of both reviewers. In particular, Reviewer 2 requested additional information on the different SWOT Level-4 products and on the fact that SWOT-based gridded products are not limited to MIOST Science, but also include other experimental reconstructions, such as 4DVarNET and 4DVarQG. For this reason, we considered it necessary to retain a concise description of these alternative SWOT products and to explain why MIOST Science was selected for the present study. We nevertheless removed or reduced information that was not directly relevant to the analysis and tried to avoid repetition, especially concerning the limitations and characteristics of the satellite products. We have revised the manuscript, now at lines: 128 - 155.

“2.1 Satellite Altimetry Products

The AVISO DUACS global 1/8° daily gridded product provides sea level anomaly (SLA), absolute dynamic topography (ADT), and geostrophic velocities derived from multi-mission satellite altimetry. The product is generated by the CNES/CLS DUACS system and distributed by the Copernicus Marine Service (Taburet et al., 2019; Ballarotta et al., 2019). Although its nominal grid spacing is 1/8° (approximately 12.5 km), its effective resolution is lower because the optimal-interpolation procedure smooths part of the small-scale variability. Its effective resolving capability is typically between 50 and 100 km, depending on latitude and the observing constellation, which limits the representation of the smallest mesoscale and submesoscale features (Chelton et al., 2011b; Ballarotta et al., 2019; CMEMS Sea Level Product User Manual, 2025).

To complement the standard DUACS product, we also use the experimental global 1/8° SWOT MIOST Science gridded sea-level dataset distributed by AVISO/DUACS (CLS and CNES, 2025). This product combines multi-mission nadir altimetry with SWOT observations and uses the MIOST framework to reconstruct fields that are continuous in space and time through a multiscale optimal-interpolation approach designed to represent mesoscale geostrophic variability (Ubelmann et al., 2015; Rogé et al., 2017; CLS and CNES, 2025). However, SWOT MIOST remains a daily Level-4 mapped product and should not be interpreted as a direct representation of the native KaRIn observations. Its smallest-scale variability is still affected by observational noise, filtering, interpolation, and the spatial and temporal sampling of the observing system. In addition, its relatively short temporal coverage limits its use for investigating decadal or longer-term variability.

The SWOT Level-4 products over the North Atlantic are not limited to MIOST. There are, as noted in Introduction, other experimental regional products, such as 4DVarNET and 4DVarQG, are also available and use different data-driven or

dynamically constrained mapping approaches (Le Guillou et al., 2021; Beauchamp et al., 2023; Ballarotta et al., 2025). As explained in Introduction, we select MIOST because it provides a spatially consistent global reconstruction covering the complete verification domain, differently from these other products that don't include the whole study area.

The AVISO DUACS 1/8° product used in this study corresponds to the Copernicus Marine Service product SEALEVEL_GLO_PHY_L4_MY_008_047, dataset cmems_obs-sl_glo_phy-ssh_my_allsat-l4-duacs-0.125deg_P1D, DOI: 10.48670/moi-00148. This product was accessed and downloaded on 31 October 2025.

The SWOT MIOST Science product used in this study corresponds to the AVISO/DUACS Experimental Multimission Gridded Level-4 Sea Surface Heights and Velocities product including SWOT observations, generated using the MIOST mapping method, version 2.0.1, DOI: 10.24400/527896/A01-2025.001. This product was accessed and downloaded from the v_2_0_1/miost/science directory on 1 November 2025.

.”

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.

Response: for Section 2.2, we shortened the description of the reanalysis datasets and removed information that was not directly relevant to the present analysis, including details on variables and model components not used in this study. For completeness, even though it wasn't required by the reviewer's comment here, we decided to include Table 1 as well. We have revised the manuscript, now at lines: 156 – 197.

“2.2 Global Ocean Reanalysis products

To evaluate the impact of the spatial resolution on reanalysis skills, we used two Global Ocean Reanalysis products, distributed by CMEMS: GLORYS2V4 and GLORYS12V1. These products are part of the CMEMS multi-year reanalysis portfolio and are designed to deliver consistent, observation-constrained reconstructions of the ocean state over the satellite altimetry era (1993–present).

GLORYS12V1 is an eddy-resolving reanalysis, developed by Mercator Ocean International. It provides a global ocean reanalysis at 1/12° horizontal resolution (~8 km), with 50 vertical levels. The system is based on the NEMO ocean general circulation model (v3.1) coupled with the LIM2 sea-ice model and uses Arakawa C-grid discretization, as explained in Madec, G., & the NEMO System Team (2024). The reanalysis assimilates along-track delayed-time sea-level anomaly (SLA) from all altimetric satellites, AVHRR-only Reynolds SST (1/4°), in-situ temperature and salinity profiles from the CORA v4.1 database (Szekely et al., 2019), and sea-ice concentration from CERSAT (Ezraty et al., 2007).

For both comparisons, the study requires the use of the interim extension for GLORYS12V1 (cmems_mod_glo_phy_myint_0.083deg_P1D-m) : daily mean sea surface height fields from the interim archive, matching the same period used for satellite datasets.

GLORYS2V4 is a lower-resolution global ocean reanalysis included in the Copernicus Marine Global Ocean Ensemble Physics Reanalysis product. It is based on the NEMO v3.1 model framework coupled with the LIM2 thermodynamic–dynamic sea ice model, configured on the ORCA025 grid with a nominal horizontal resolution of 1/4° and 75 vertical levels. For the present study, the temporal coverage selected for GLORYS2V4 also matches the period used for the satellite datasets, in order to allow a direct comparison with GLORYS12V1. All details on the main parameters of the two reanalysis products are included in Table 1.

GLORYS12V1 was obtained from the Copernicus Marine Service product GLOBAL_MULTIYEAR_PHY_001_030, DOI: 10.48670/moi-00021. The product was accessed and downloaded on 1 November 2025.

GLORYS2V4 was obtained from the Copernicus Marine Service product GLOBAL_MULTIYEAR_PHY_ENS_001_031, DOI: 10.48670/moi-00024. The product was accessed and downloaded on 1 November 2025.

Parameters	GLORYS12V1	GLORYS2V4
Model version + grid version	LIM2 EVP NEMO 3.1, driven at the surface by ECMWF ERA Interim reanalysis until end of 2018 and ERA5 reanalysis thereafter + ORCA Grid (1/12°)	LIM2 EVP NEMO 3.1 + ORCA1/4°
Domain	Global	Global
Horizontal resolution	1/12°: 4320 X 2041 grid points	1/4°: 1440 X 1021 grid points
Vertical sampling	50 z-levels	75 z-levels
Numerical technique	Primitive equations with finite differences: on Arakawa C native grid	Same as GLORYS12V1
Time integration	Leapfrog explicit, barotropic/baroclinic splitting; baroclinic step: 300 s	Leapfrog explicit, barotropic/baroclinic splitting; baroclinic step: 450 s
Prognostic variables	3D currents, potential temperature, salinity, turbulent kinetic energy (TKE), sea surface height (SSH)	Same variables as GLORYS12V1
Geophysical variables	Bathymetry from ETOPO2 plus some local smoothing to accommodate large tidal currents	Same as GLORYS12V1
Horizontal diffusion	horizontal Laplacian on tracer variables, bi-Laplacian on momentum variables	Same as GLORYS12V1
Surface scheme	ERA-Interim until 2018, ERA5 from 2019, bulk CORE formula with corrections for precipitation and radiative fluxes	No surface nudging precipitation, flux correction Climatological runoff + ice shelf and iceberg melting
Turbulent mixing	TKE scheme with mixing length set to 10 m	Same as GLORYS12V1
Assimilated Satellite observations	Reynolds 0.25° AVHRR-only SST, Delayed Time SLA from all altimetric satellites, in situ T/S profiles from Copernicus Marine CORAv4.1 database, CERSAT Sea Ice Concentration	SST, SLA, T/S profiles, SIC

Table 1. Parameters used in GLORYS12V1 and GLORYS2V4. Both the Reanalysis dataset used the same assimilated scheme: SAM2, a large-scale bias correction 7-day assimilation window Merge MDT (observations + model) Reynolds SST, CORA.

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

Response: yes, we have removed the first one (“The number of enclosed grid cells must be between 5 and 2000, corresponding to an effective radius of at least 25 km”).

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?

Response: We have cutted the repetition. We clarify what is “amplitude error”. We have revised the manuscript, now at lines: 573 – 576.

“For each of these properties, we compute binned, normalized histograms that illustrate the distribution of radius error (difference between satellite and reanalysis effective radius), amplitude error (difference in SSH extremum relative to the outer contour between satellite and reanalysis matched eddy), separation distance between satellite and reanalysis matched eddy centroid.”

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

Response: we have corrected the Figure 1 caption, at lines: 248 – 257.

“Fig.1.Daily sea-level fields on 1 January 2025 for GLORYS12V1, GLORYS2V4, SWOT MIOST Science, and AVISO DUACS 1/8°, after application of a high-pass Bessel filter with a cutoff wavelength of 500 km. The common colour bar is labelled SSH (m), with SSH used here as a general label for the sea-surface height field. Specifically, the colour shading represents modelled SSH for GLORYS12V1 and GLORYS2V4 and satellite-derived Absolute Dynamic Topography (ADT) for SWOT MIOST Science and AVISO DUACS 1/8°. Although SSH and ADT are not formally identical quantities, they are dynamically comparable and are treated consistently for contour-based eddy detection. Detected anticyclonic and cyclonic eddies are overlaid in red and blue, respectively. Solid contours indicate the contour of maximum circum-averaged geostrophic speed, whereas dashed contours indicate the effective contour used to calculate eddy area and effective radius. All fields were remapped onto the same regular 1/8° grid to allow direct comparison.”

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.

Response: we have revised the manuscript, now the Fig. 3 and the text are modified, at lines: 306 – 319. We have removed “60 km”.

“In Fig.3, we see the dependence between radius and amplitude, in which we note that eddies display similar properties in terms of effective radius and amplitude. Many small eddies are detected with a radius below 50 km and amplitude less than 5 cm. In Fig. 4, most eddies have amplitudes lower than 10 cm. However, their mean lifetime is lower than 50 days (see Appendix C for more details about mean lifetime and eddy amplitude thresholds). However, features with amplitudes of only a few millimetres, should not be interpreted as physically robust eddies. The 0.4 cm value corresponds only to the permissive algorithmic threshold used during the py-eddy-tracker detection step, and is not used here as a physical threshold to define the robust eddy population. It can be noted, generally, that larger amplitude eddies are also found to have larger radius (Fig. 3). The eddy identification and tracking procedure are independently applied for each product and produce plots about physical features and time series of anticyclonic and cyclonic eddies.

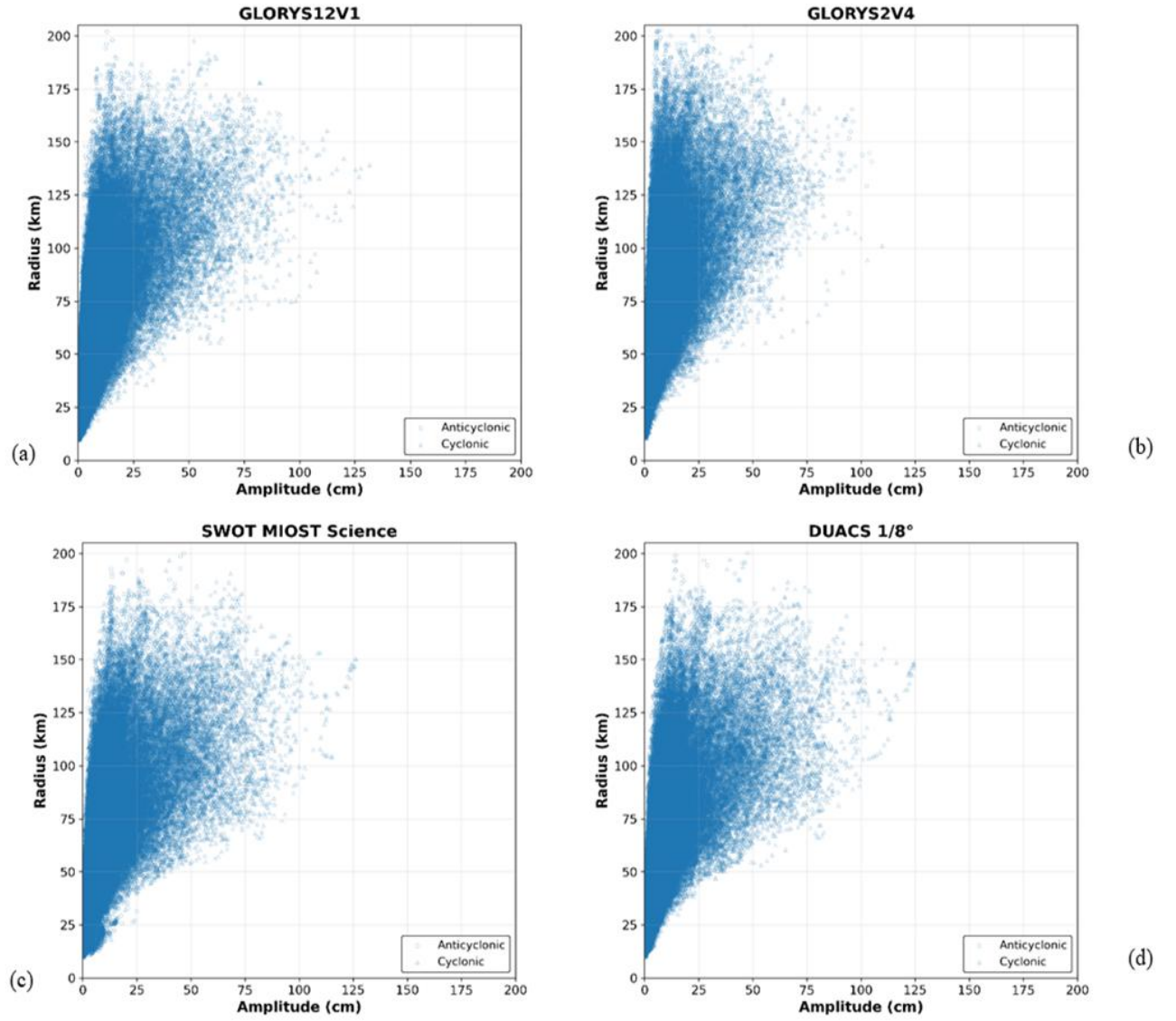


Fig. 3. Scatter plots of detected eddies, with a minimum lifetime threshold of 4 days, between Radius and Amplitude, from 2023-08-01 to 2025-05-01, for (a) GLORYS12V1 (bilinear remapped onto DUACS and SWOT grid), (b) GLORYS2V4 (bilinear remapped onto DUACS and SWOT grid), (c) SWOT MIOST Science, and (d) AVISO DUACS 1/8°. The eddy amplitude and the eddy radius are calculated at the track level, by averaging the amplitude and radius over the entire lifetime of the eddy.

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

We have revised the manuscript, now at lines: 336 – 345.

“It is interesting to consider what fraction of the observed eddies are captured (not excluded) by reanalysis products, using the Probability of Detection (POD), as introduced and explained in Smith and Fortin (2022), defined as:

$$POD = \frac{hits}{(hits + misses)} \quad (2)$$

The POD is the number of matched eddies over the total number of eddies present in the satellite product.

Similarly, one can consider what fraction of the predicted eddies in the model analyses did not occur using the False Alarm Ratio (FAR), as described in Smith and Fortin (2022), defined as:

$$FAR = \frac{\text{false alarms}}{(\text{false alarms} + \text{hits})} \quad (3)$$

Both POD and FAR have values between $[0,1]$ with perfect values of 1 and 0, respectively. “

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

Response: we have corrected the manuscript, now at lines: 361 – 363.

“The properties of identified eddies for the four datasets are quite similar (see Fig. 3) over the entire period. In Fig. 3, we see similar trends between radius and amplitude. Many small eddies, with a radius lower than 60 km (see Fig.3 and Fig.4) not far from the threshold of detection, and an amplitude lower than 5 cm, have been identified.”

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

Response: we have revised the manuscript, now at lines: 368 – 373.

“In Fig. 4 we see that most have amplitudes below 10 cm, but at the same time there is a moderate positive correlation between amplitude and lifetime, whereby as amplitude increases, so does the average survival time of the eddies. Considering anticyclonic and cyclonic eddies together, the Spearman correlation coefficients are $\rho = 0.422$ for GLORYS12V1, $\rho = 0.518$ for GLORYS2V4, $\rho = 0.464$ for SWOT MIOST Science, and $\rho = 0.455$ for DUACS 1/8°, with $p < 0.001$ in all cases. These results indicate that larger-amplitude eddies tend to persist longer, although a moderate correlation and the amplitude alone does not fully explain the variability in eddy lifetime.”

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

Response: we have revised, now at lines: 374 – 375:

“Tables 2 and 3 show the total number of anticyclonic and cyclonic eddies detected and tracked with a minimum lifetime of 4 days.”

Here, we have added the four Appendices (in the revised manuscript, after the Summary and Discussion section: A, B, C, D).

As already mentioned at the beginning of this response document, we have added some calculations that had already been carried out and were previously included in the main text. In addition, we have included new calculations to further support and complete the statements made in the main text.

“Appendix A

In this appendix we aim to assess how the verification metrics, used to evaluate the ability of global ocean reanalyses to detect eddies observed in satellite datasets over a highly eddy-populated region, change under these different methodological choices.

Eddy detection-tracking and matching results for 7 and 28 minimum lifetime thresholds.

To assess the sensitivity of the verification results to the lifetime criterion, the detection, tracking, and matching analyses were repeated using minimum-lifetime thresholds of 7 and 28 days. Tables A1, A2, A5, and A6 report the numbers of anticyclonic and cyclonic eddies retained in each catalogue, whereas Tables A3, A4, A7, and A8 report POD, FAR, and the differences between GLORYS12V1 and GLORYS2V4. Fig. A1 shows the corresponding time series and power spectra. Increasing the minimum lifetime substantially reduces the number of retained eddies, especially at 28 days. This reduction does not, however, lead to lower FAR values: relative to the 4-day analysis, POD decreases and FAR increases for both reanalyses and both observational references. In Fig. A1, panel (a) shows the time series of the verification metrics over the full study period for both comparisons, while panel (b) shows the power spectra associated with each time series. The aim for the calculation of the power spectra is assessing how the variability of the verification metrics changes as the minimum eddy lifetime threshold is progressively increased.

Dataset	Anticyclonic	% w.r.t. DUACS	Cyclonic	% w.r.t. DUACS
DUACS	116644	-	116295	-
GLORYS2V4	62479	53.56%	67486	58.03%
GLORYS12V1	86976	74.57%	95836	82.40%

Table A1. Total number of anticyclonic and cyclonic eddies identified in the region of study over the evaluation period for DUACS, GLORYS2V4, and GLORYS12V1, for eddies with a minimum lifetime threshold of 7 days. The percentage of eddies identified in GLORYS2V4 and GLORYS12V1 with respect to DUACS is also provided.

Dataset	Anticyclonic	% w.r.t. SWOT	Cyclonic	% w.r.t. SWOT
SWOT	114015	-	116036	-
GLORYS2V4	62479	54.79%	67486	58.15%
GLORYS12V1	86976	76.28%	95836	82.59%

Table A2. Total number of anticyclonic and cyclonic eddies identified in the region of study over the evaluation period for SWOT, GLORYS2V4, and GLORYS12V1, for eddies with a minimum lifetime threshold of 7 days. The percentage of eddies identified in GLORYS2V4 and GLORYS12V1 with respect to SWOT is also provided.

Verification metric	GLORYS12V1	GLORYS2V4	Diff.	Diff.%
POD	0.608	0.449	0.159	35.41%
FAR	0.225	0.196	0.029	14.80%

Table A3. aggregated eddy verification statistics for the Probability of Detection (POD) and False Alarm Ratio (FAR) evaluated over the period, from 1st August 2023 to 1st May 2025, for the comparison with DUACS, with a minimum lifetime threshold of 7 days.

Verification metric	GLORYS12V1	GLORYS2V4	Diff.	Diff.%
POD	0.606	0.447	0.159	35.57%
FAR	0.237	0.209	0.028	13.4%

Table A4. Aggregated eddy verification statistics for the Probability of Detection (POD) and False Alarm Ratio (FAR) evaluated over the period, from 1st August 2023 to 1st May 2025, for the comparison with SWOT, with a minimum lifetime threshold of 7 days.

Dataset	Anticyclonic	% w.r.t. DUACS	Cyclonic	% w.r.t. DUACS
DUACS	67686	-	69657	-
GLORYS2V4	39218	57.94%	43705	62.74%
GLORYS12V1	41768	61.70%	49935	71.69%

Table A5. Total number of anticyclonic and cyclonic eddies identified in the region of study over the evaluation period for DUACS, GLORYS2V4, and GLORYS12V1. The percentage of eddies identified in GLORYS2V4 and GLORYS12V1 with respect to DUACS is also provided, with a lifetime threshold of 28 days.

<i>Dataset</i>	<i>Anticyclonic</i>	<i>% w.r.t. SWOT</i>	<i>Cyclonic</i>	<i>% w.r.t. SWOT</i>
<i>SWOT</i>	67089	-	68842	-
<i>GLORYS2V4</i>	39218	58.45%	43705	63.49%
<i>GLORYS12V1</i>	41768	62.26%	49935	72.54%

Table A6. Total number of anticyclonic and cyclonic eddies identified in the region of study over the evaluation period for SWOT, GLORYS2V4, and GLORYS12V1. The percentage of eddies identified in GLORYS2V4 and GLORYS12V1 with respect to SWOT is also provided, with a lifetime threshold of 28 days.

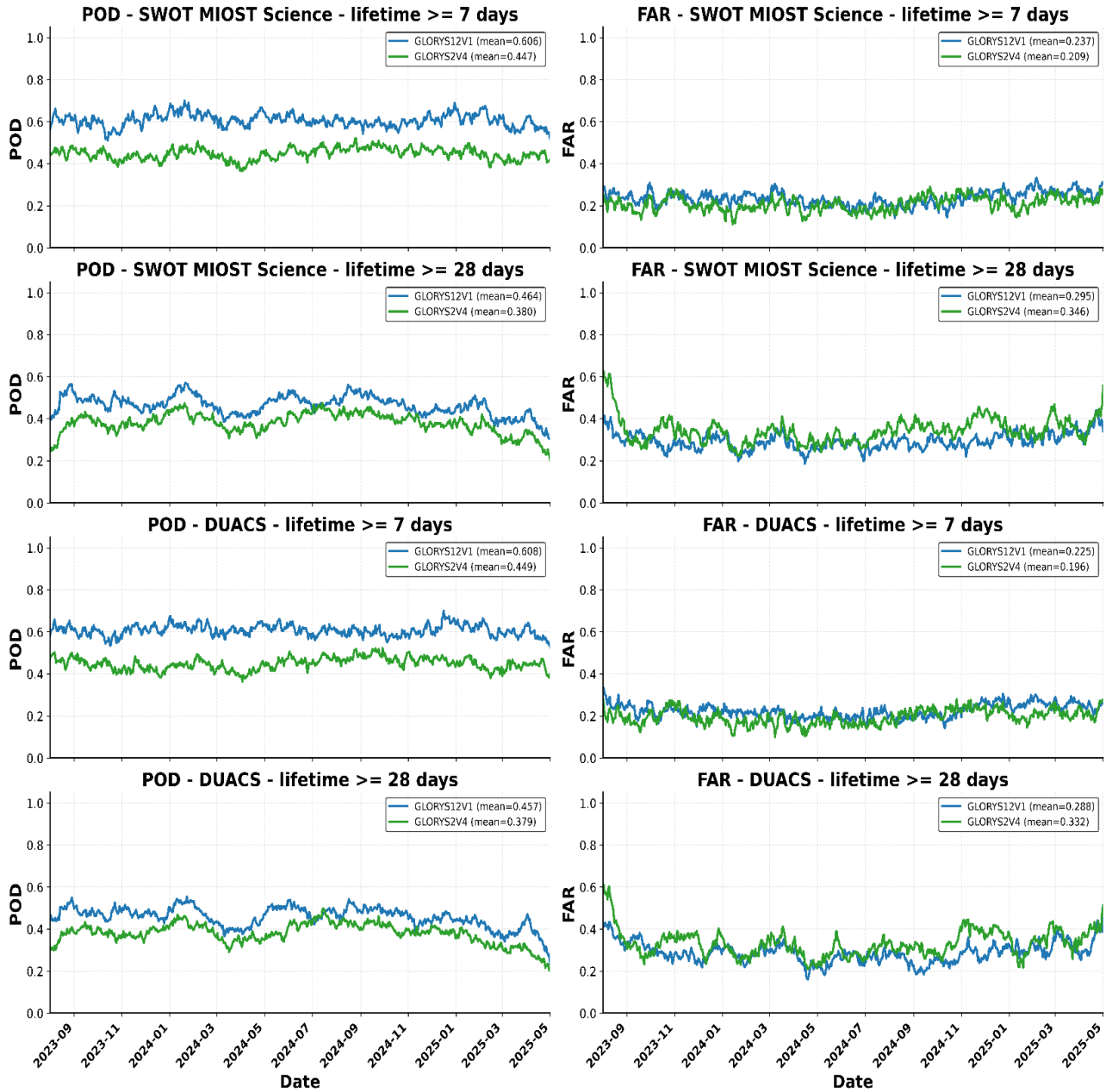
<i>Statistics</i>	<i>GLORYS12V1</i>	<i>GLORYS2V4</i>	<i>Diff.</i>	<i>Diff.%</i>
<i>POD</i>	0.464	0.380	0.084	22.1%
<i>FAR</i>	0.295	0.346	-0.051	-14.7%

Table A7. aggregated eddy verification statistics for the Probability of Detection (POD) and False Alarm Ratio (FAR) evaluated over the period, from 1st August 2023 to 1st May 2025, for the comparison with DUACS, with a minimum lifetime threshold of 28 days.

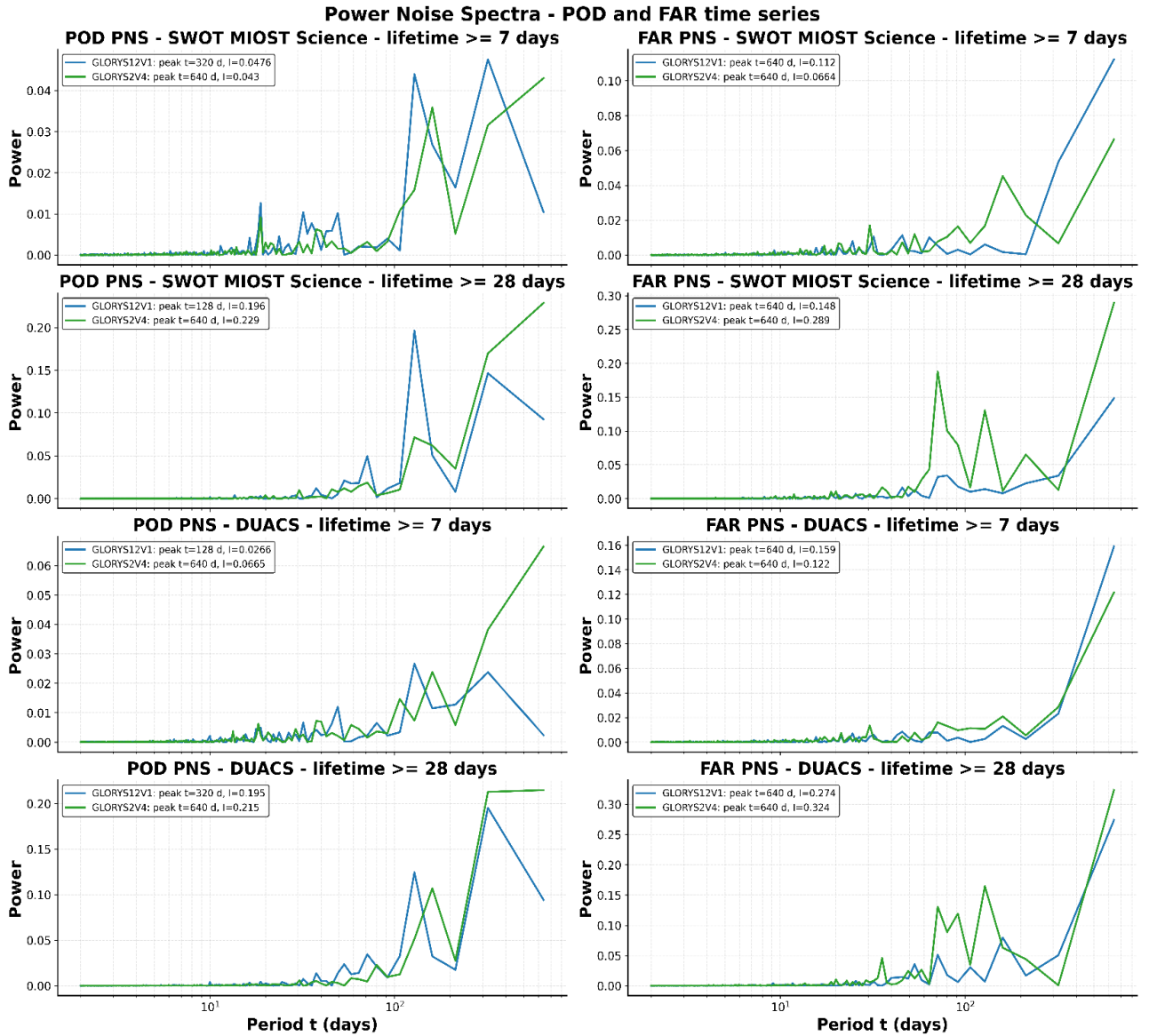
<i>Statistics</i>	<i>GLORYS12V1</i>	<i>GLORYS2V4</i>	<i>Diff.</i>	<i>Diff.%</i>
<i>POD</i>	0.457	0.379	0.078	20.5%
<i>FAR</i>	0.288	0.322	-0.034	-10.5%

Table A8. aggregated eddy verification statistics for the Probability of Detection (POD) and False Alarm Ratio (FAR) evaluated over the period, from 1st August 2023 to 1st May 2025, for the comparison with SWOT, with a minimum lifetime threshold of 28 days.

POD and FAR time series - SWOT MIOST Science and DUACS



(a)



(b)

Fig. A1. Time series of matching statistics for GLORYS2V4 and GLORYS12V1, compared with SWOT and DUACS (a), over the period from 1st August 2023 to 1st May 2025, with a minimum lifetime of 7 and 28 days. In (b), the power spectra of each time series, for both products and comparisons.

At the 7-day threshold, POD decreases for both reanalyses. FAR increases for GLORYS2V4, whereas for GLORYS12V1 it decreases slightly against DUACS and remains essentially unchanged against SWOT. At the 28-day threshold, POD decreases and FAR increases relative to the 4-day configuration, while GLORYS12V1 retains higher POD and attains lower FAR than GLORYS2V4 for both references. This condition highlights not only the greater ability of GLORYS12V1 to capture observational eddies compared to GLORYS2V4, for both satellite datasets, but also its lower error rate in detecting and tracking false alarms. The GLORYS12V1 FAR becomes lower than that of GLORYS2V4 when only long-lived eddies are considered. However, the application of a longer lifetime threshold also leads to a substantial reduction in the total number of detected and tracked eddies compared to the lower lifetime cases, as shown in Tables A1, A2, A5, and A6. This reduction is stronger for the reanalyses than for the satellite datasets, producing an overall decrease in POD and an overall increase in FAR for both comparisons and for both reanalysis products with respect to the results discussed in Sect. 3.2 and reported in Tables 4 and 5. This behaviour indicates that, although the 28-day threshold isolates more stable eddy structures, it also exposes a significant discrepancy between the reanalyses and the observational datasets. In particular, some of the eddies matched in the reanalyses at lower lifetime thresholds appear to be less

temporally stable than those observed in the satellite products. As a result, the number of hits clearly tends to decrease as the minimum lifetime threshold increases.

This sensitivity test therefore suggests that temporally more stable eddies tend to reduce the resolution-related gap between the two reanalysis products. Nevertheless, this does not lead to an increase in POD or to a systematic decrease in FAR. This confirms that both ocean reanalyses still show limitations in capturing observed eddies and in avoiding the generation of unmatched eddy-like structures. More generally, these results show that one of the main challenges is to identify a methodology able to remove spurious and unstable structures, for which it is not always clear whether they can be considered robust mesoscale eddies.

In panel (b), the power spectra show that, at short temporal scales, the variability of the verification-metric time series is almost similar across the different configurations. The spectral differences are mainly concentrated at the longest resolved periods. For FAR, GLORYS12V1 has the larger peak with the 7-day threshold, whereas GLORYS2V4 has the larger peak with the 28-day threshold. Because the analysed record is only 640 days long, peaks at periods of 320–640 days should be interpreted as low-frequency variability rather than as evidence of robust periodic or seasonal signals. It should also be noted that the axes do not have the same numerical range across all panels. In general, the numerical range of the FAR power spectra is larger than that of POD, indicating that the strongest spectral peaks are generally associated with FAR.

Appendix B

Sensitivity of bootstrap resampling results for verification metrics to the block length

To assess the sensitivity of the confidence intervals to temporal dependence, the coupled block-bootstrap analysis was repeated using block lengths of 7, 20, and 30 days, as well as block lengths derived from the mean lifetimes of the satellite-observed eddies. Results are reported for the 4-day and 7-day minimum-lifetime thresholds. Across the tested block lengths, GLORYS12V1 maintains higher POD than GLORYS2V4 for both observational references. The FAR difference also remains positive, although it becomes smaller for the 7-day minimum-lifetime threshold.

Verification Metric	GLORYS12V1	GLORYS2V4	<i>Difference (12V1 – 2V4)</i>
POD (Mean)	0.603	0.442	0.161
POD (95% CI)	[0.594, 0.611]	[0.431, 0.453]	[0.148, 0.173]
FAR (Mean)	0.234	0.200	0.034
FAR (95% CI)	[0.224, 0.245]	[0.189, 0.213]	[0.020, 0.048]

Table B1. Bootstrap resampling for SWOT, for a minimum lifetime threshold of 7 days and a block length of 7 days.

Verification Metric	GLORYS12V1	GLORYS2V4	<i>Difference (12V1 – 2V4)</i>
---------------------	------------	-----------	--------------------------------

POD (Mean)	0.599	0.438	0.161
POD (95% CI)	[0.586, 0.610]	[0.430, 0.446]	[0.147, 0.174]
FAR (Mean)	0.241	0.208	0.034
FAR (95% CI)	[0.233, 0.250]	[0.197, 0.219]	[0.024, 0.043]

Table B2. Bootstrap resampling for DUACS comparison, for a minimum lifetime threshold of 7 days and a block length of 7 days.

Verification Metric	GLORYS12V1	GLORYS2V4	Difference (12V1 – 2V4)
POD (Mean)	0.620 0.603	0.455 0.444	0.165 0.159
POD (95% CI)	[0.608 , 0.632] [0.591 , 0.614]	[0.442 , 0.468] [0.429 , 0.460]	[0.149 , 0.183] [0.142 , 0.177]
FAR (Mean)	0.236 0.235	0.187 0.200	0.049 0.034
FAR (95% CI)	[0.228 , 0.244] [0.222 , 0.250]	[0.175 , 0.202] [0.182, 0.221]	[0.031 , 0.065] [0.011 , 0.054]

Table B3. Block-bootstrap estimates of POD and FAR using 20-day blocks for the comparison with DUACS. Values are reported for GLORYS12V1, GLORYS2V4, and their difference, together with 95% confidence intervals. In the first row, there is the calculation for the 4 days minimum lifetime threshold. In the second row, there is the calculation for 7 days minimum lifetime threshold.

Verification Metric	GLORYS12V1	GLORYS2V4	Difference (12V1 – 2V4)
POD (Mean)	0.617 0.598	0.450 0.440	0.167 0.158
POD (95% CI)	[0.603 , 0.633] [0.582, 0.614]	[0.440 , 0.459] [0.429, 0.451]	[0.151 , 0.187] [0.140, 0.178]
FAR (Mean)	0.236 0.242	0.192 0.207	0.044 0.035
FAR (95% CI)	[0.230, 0.241] [0.233, 0.251]	[0.180 , 0.207] [0.193, 0.225]	[0.031 , 0.056] [0.021, 0.046]

Table B4. Block-bootstrap estimates of POD and FAR using 20-day blocks for the comparison with SWOT MIOST. Values are reported for GLORYS12V1, GLORYS2V4, and their difference, together with 95% confidence intervals. In the first row, there is the calculation for the 4 days minimum lifetime threshold. In the second row, there is the calculation for 7 days minimum lifetime threshold.

Verification Metric	GLORYS12V1	GLORYS2V4	Difference (12V1 – 2V4)
POD (Mean)	0.620	0.452	0.168

	0.603	0.441	0.162
POD (95% CI)	[0.604, 0.634] [0.589, 0.615]	[0.437, 0.466] [0.425, 0.459]	[0.148, 0.190] [0.142, 0.185]
FAR (Mean)	0.235 0.234	0.186 0.201	0.049 0.033
FAR (95% CI)	[0.227, 0.245] [0.223, 0.250]	[0.173, 0.200] [0.184, 0.217]	[0.031, 0.066] [0.011, 0.055]

Table B5. Block-bootstrap estimates of POD and FAR using 30-day blocks for the comparison with DUACS. Values are reported for GLORYS12V1, GLORYS2V4, and their difference, together with 95% confidence intervals. In the first row, there is the calculation for the 4 days minimum lifetime threshold. In the second row, there is the calculation for 7 days minimum lifetime threshold.

Verification Metric	GLORYS12V1	GLORYS2V4	<i>Difference (12V1 – 2V4)</i>
POD (Mean)	0.618 0.616	0.448 0.460	0.170 0.166
POD (95% CI)	[0.600, 0.636] [0.602, 0.626]	[0.438, 0.456] [0.448, 0.484]	[0.149, 0.194] [0.131, 0.173]
FAR (Mean)	0.236 0.240	0.191 0.206	0.045 0.035
FAR (95% CI)	[0.231, 0.240] [0.236, 0.245]	[0.180, 0.202] [0.194, 0.217]	[0.032, 0.056] [0.023, 0.046]

Table B6. Block-bootstrap estimates of POD and FAR using 30-day blocks for the comparison with SWOT. Values are reported for GLORYS12V1, GLORYS2V4, and their difference, together with 95% confidence intervals. In the first row, there is the calculation for the 4 days minimum lifetime threshold. In the second row, there is the calculation for 7 days minimum lifetime threshold.

For the sake of completeness, we also calculated the average lifetimes of the observed eddies for each satellite dataset using two different minimum lifetimes (4 and 7 days). The idea is to use these average lifetimes as the block lengths for the coupled bootstrap resampling. The mean eddy lifetime is calculated from the number of daily observations associated with each individual trajectory, using the “nb_obs_by_track” property provided by py-eddy-tracker. Tracks with zero observations are excluded.

For DUACS, we calculate the average lifetime of eddies:

- Anticyclonic eddies with a minimum lifetime threshold of 4 days: 18.03 days.
- Cyclonic eddies with a minimum lifetime threshold of 4 days: 18.23 days.
- Anticyclonic eddies with a minimum lifetime threshold of 7 days: 22.69 days.
- Cyclonic eddies with a minimum lifetime threshold of 7 days: 23.05 days.

Therefore, considering both anticyclonic and cyclonic eddies, we calculate an average eddy lifetime of: 18.13 days (minimum lifetime of 4 days) and 22.87 days (minimum lifetime of 7 days). Therefore, for the block bootstrap resample, we rounded up to the next whole day, setting a value of 19 days for a minimum lifetime of 4 days, and a value of 23 days for a minimum lifetime of 7 days.

For SWOT, we calculate the average lifetime of eddies:

- Anticyclonic eddies with a minimum lifetime threshold of 4 days: 17.51 days.
- Cyclonic eddies with a minimum lifetime threshold of 4 days: 17.58 days.
- Anticyclonic eddies with a minimum lifetime threshold of 7 days: 22.69 days.
- Cyclonic eddies with a minimum lifetime threshold of 7 days: 22.38 days.

Therefore, considering both anticyclonic and cyclonic eddies, we calculate an average eddy lifetime of: 17.54 days (minimum lifetime of 4 days) and 22.54 days (minimum lifetime of 7 days). Therefore, for the block bootstrap resample, we rounded up to the next whole day, setting a value of 18 days for a minimum lifetime of 4 days, and a value of 23 days for a minimum lifetime of 7 days.

Verification Metric	GLORYS12V1	GLORYS2V4	Difference (12V1 – 2V4)
POD (Mean)	0.620	0.452	0.168
	0.602	0.443	0.159
POD (95% CI)	[0.608, 0.632]	[0.438, 0.466]	[0.153, 0.186]
	[0.589, 0.614]	[0.426, 0.460]	[0.143, 0.180]
FAR (Mean)	0.235	0.186	0.049
	0.235	0.200	0.034
FAR (95% CI)	[0.228, 0.243]	[0.175, 0.199]	[0.035, 0.062]
	[0.223, 0.250]	[0.182, 0.221]	[0.011, 0.055]

Table B7. Block-bootstrap estimates of POD and FAR using 18 days and 23 days blocks for the comparison with DUACS. Values are reported for GLORYS12V1, GLORYS2V4, and their difference, together with 95% confidence intervals. In the first row, there is the calculation for the 4 days minimum lifetime threshold. In the second row, there is the calculation for 7 days minimum lifetime threshold.

Verification Metric	GLORYS12V1	GLORYS2V4	Difference (12V1 – 2V4)
POD (Mean)	0.618	0.448	0.170
	0.598	0.441	0.157
POD (95% CI)	[0.601, 0.632]	[0.440, 0.457]	[0.153, 0.188]
	[0.576, 0.613]	[0.430, 0.452]	[0.139, 0.176]
FAR (Mean)	0.236	0.192	0.044
	0.241	0.207	0.034
FAR (95% CI)	[0.229, 0.244]	[0.180, 0.204]	[0.032, 0.056]
	[0.234, 0.250]	[0.195, 0.220]	[0.019, 0.047]

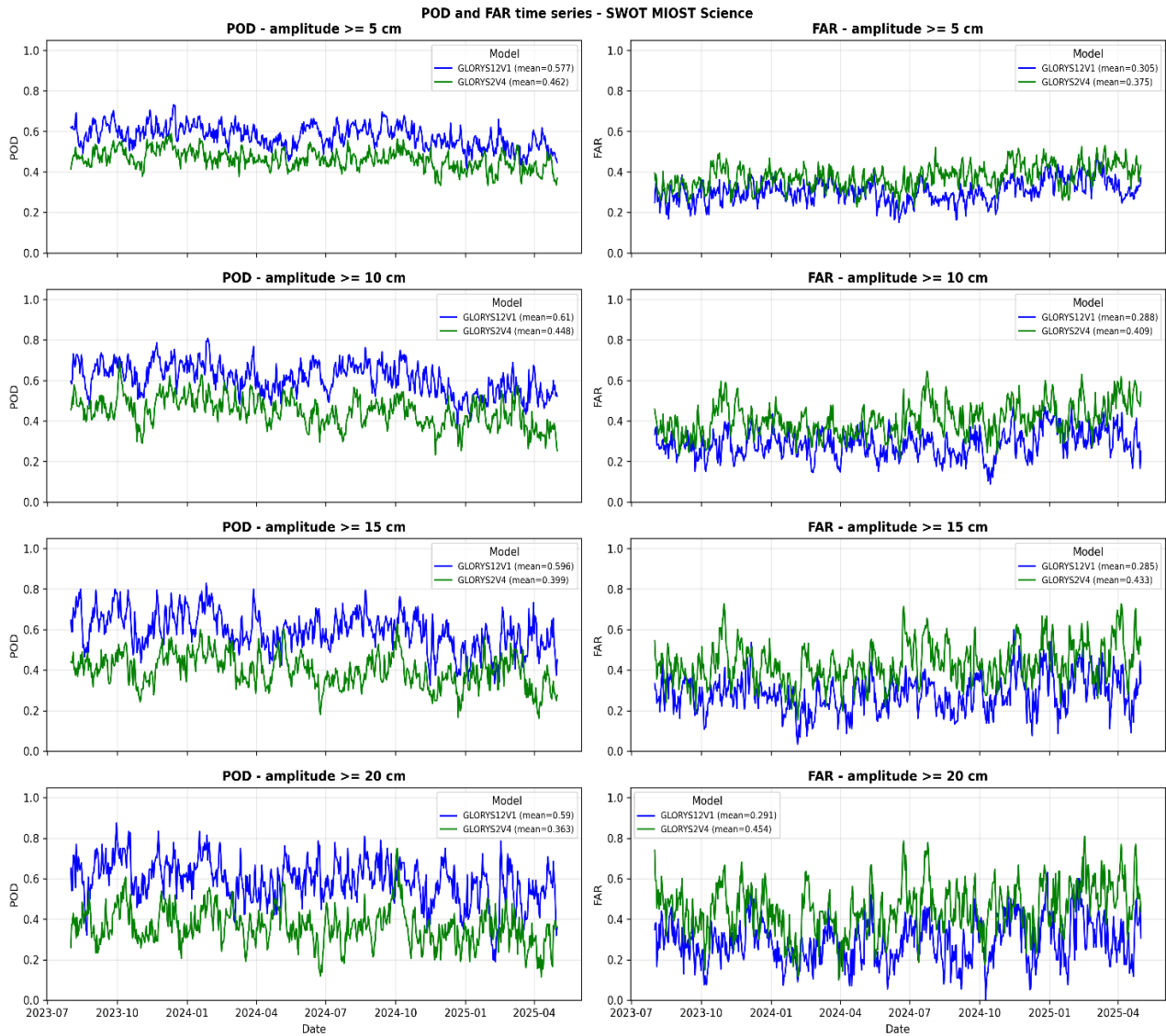
Table B8. Block-bootstrap estimates of POD and FAR using 19 days and 23 days blocks for the comparison with SWOT. Values are reported for GLORYS12V1, GLORYS2V4, and their difference, together with 95% confidence intervals. In the first row, there is the calculation for the 4 days minimum lifetime threshold. In the second row, there is the calculation for 7 days minimum lifetime threshold.

Appendix C

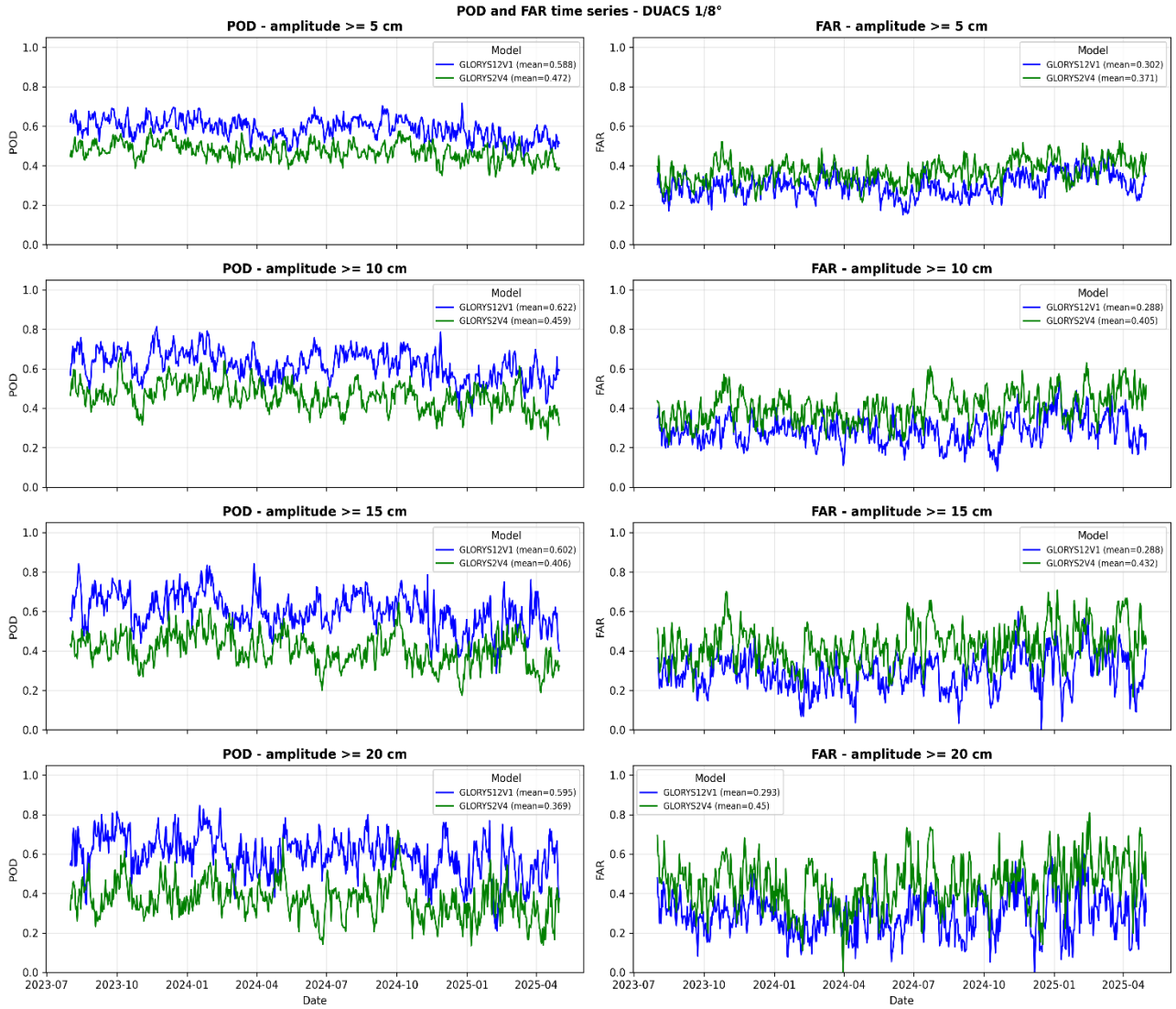
Sensitivity of verification metrics to eddy amplitude thresholds.

To assess the sensitivity of the verification results to eddy amplitude, POD and FAR were recomputed using minimum amplitude thresholds of 5, 10, 15, and 20 cm. We also considered the complementary subset of eddies with amplitudes

lower than or equal to 10 cm to isolate the contribution of low-amplitude features. The 10 cm value is used only as a diagnostic criterion and not as a physical boundary between different eddy regimes.



(a)

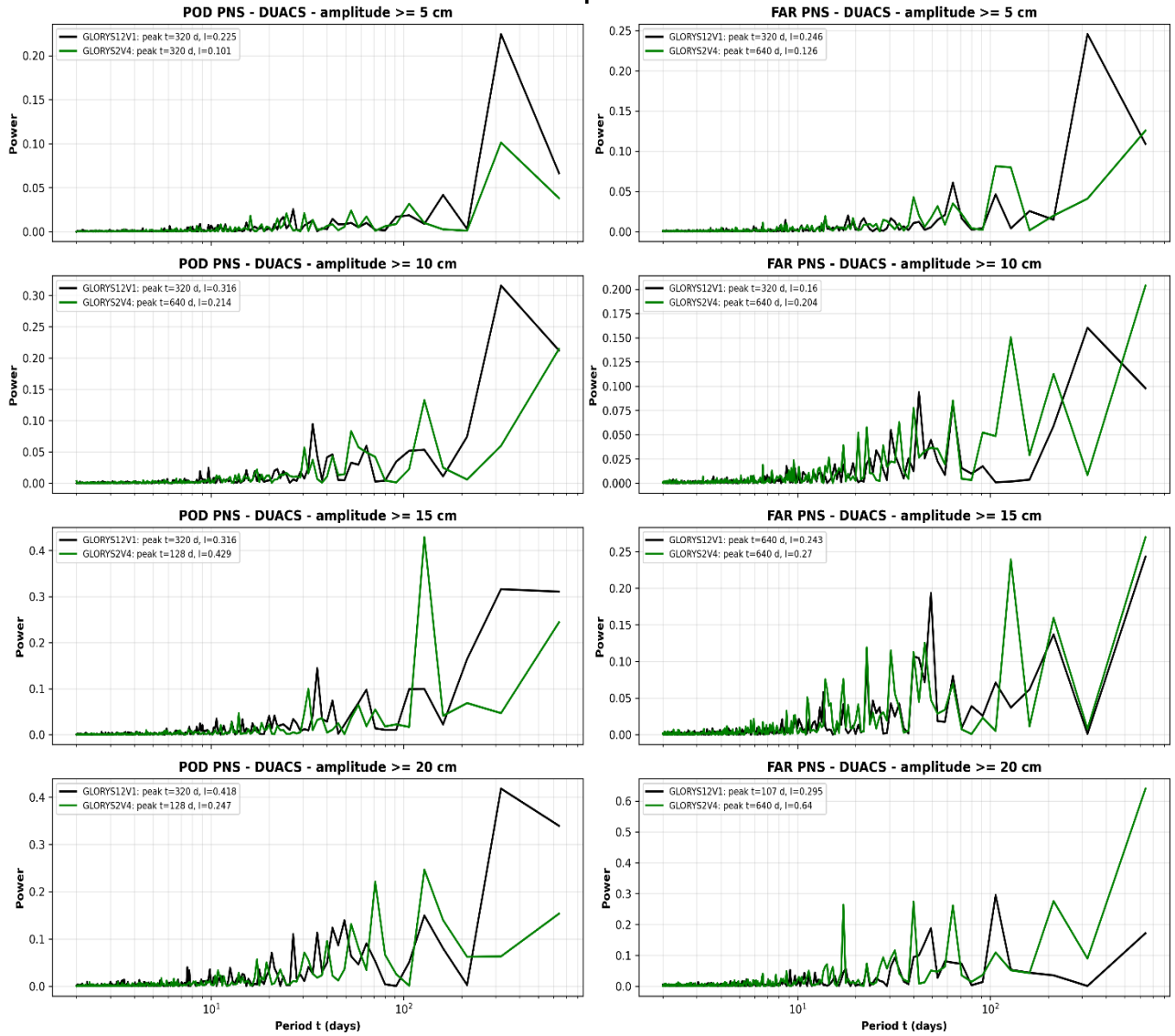


(b)

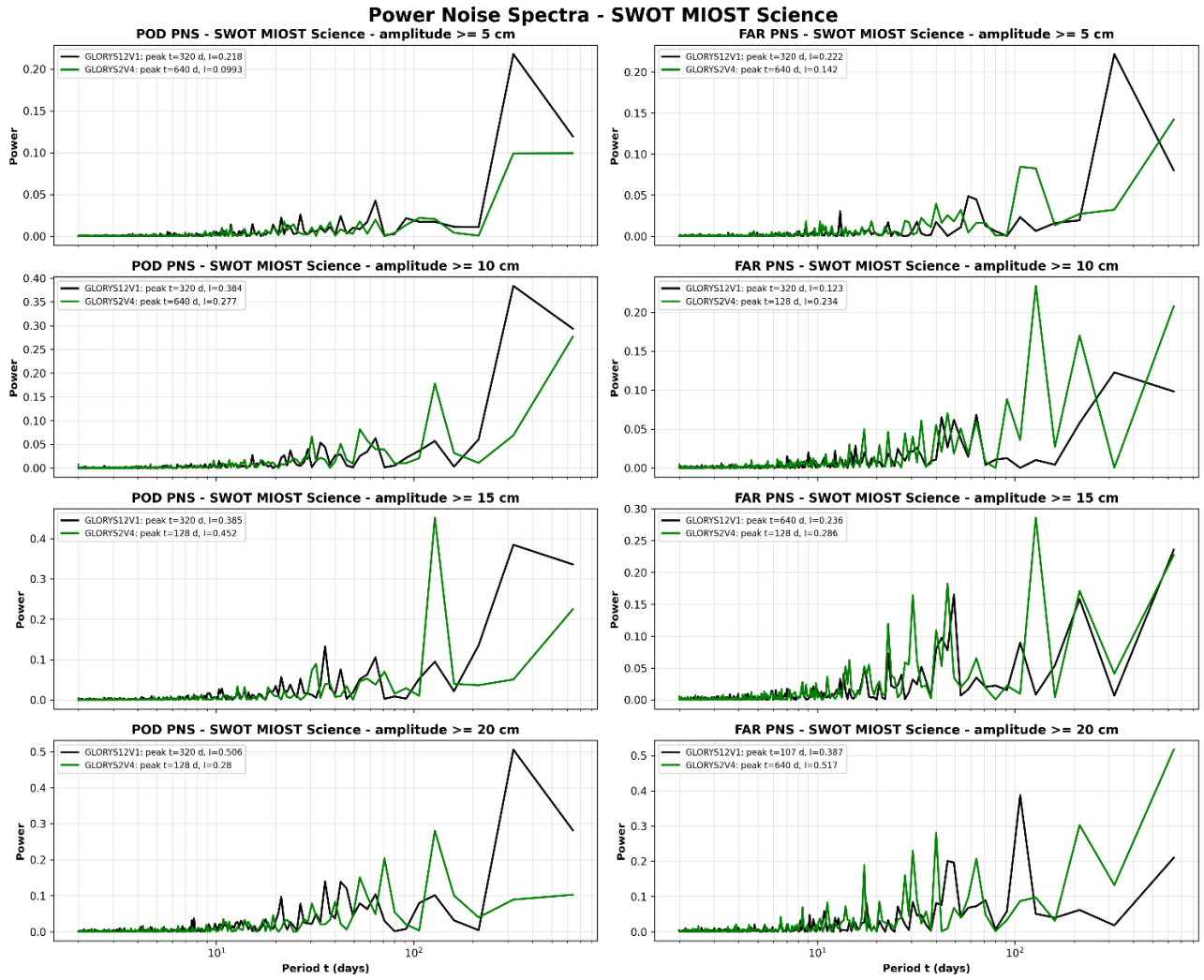
Fig. C1. Time series for POD & FAR by amplitude threshold: 5 cm, 10 cm, 15 cm, 20 cm, for SWOT comparison (a) and for DUACS 1/8° comparison (b), with a minimum lifetime threshold of 4 days.

The verification metrics do not improve monotonically as low-amplitude eddies are progressively excluded. For GLORYS12V1, POD remains approximately between 0.58 and 0.62, while FAR remains between about 0.27 and 0.31. For GLORYS2V4, POD decreases from approximately 0.46 to 0.47 at the 5 cm threshold to about 0.36 – 0.37 at 20 cm, while FAR increases from about 0.37–0.38 to approximately 0.45. The higher thresholds also retain fewer eddies, producing larger day-to-day variability. Therefore, these results do not indicate a uniform improvement in the representation of stronger eddies.

Power Noise Spectra - DUACS



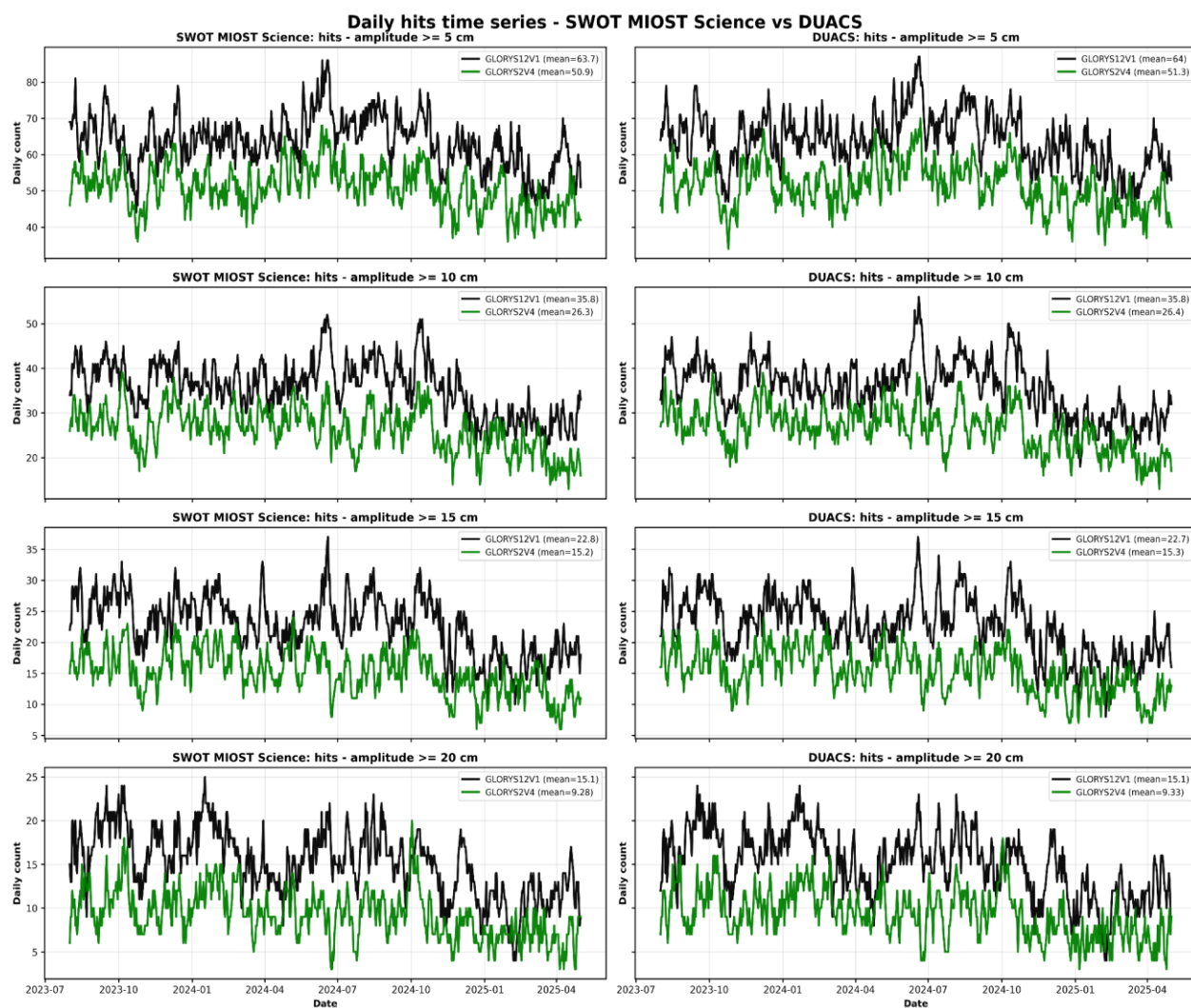
(a)



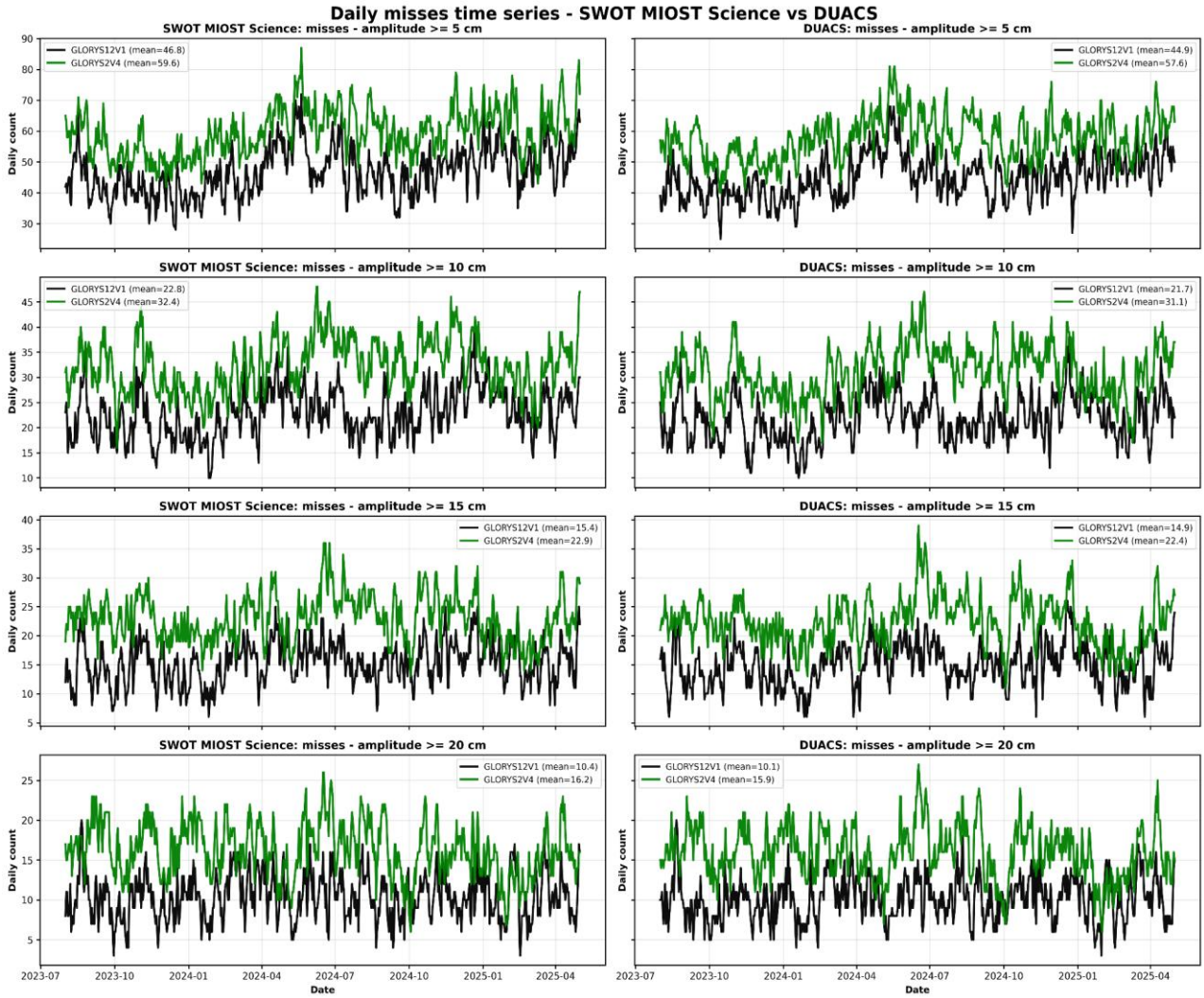
(b)

Fig. C2. Power Spectra of time series of both verification metrics for each reanalysis products, for DUACS (a) and SWOT MIOST Science comparison (b). It is shown that the agreement does not improve significantly compared with the more general case, namely that extended to all eddies. Generally, for both observational datasets, we observe that the agreement between the two reanalyses is never markedly better. Only in terms of FAR is it observed that GLORYS2V4 shows a net deterioration in agreement, as there is a clear increase in FAR, particularly at high amplitudes.

For higher amplitude thresholds, the metric values show much larger variations over time. On some days, the POD values are close to 1 and the FAR values are close to 0, while on other days they are much farther from these ideal values. This variability is mainly caused by the smaller number of eddies with amplitudes above 10 or 20 cm. For this reason, the POD and FAR time series can show large variations and can rapidly move towards values close to 0 or 1. These results suggest that the imperfect matching between the reanalyses and the observational datasets mainly concerns small ocean eddies. However, this improvement does not mean that the complete eddy field is better represented.



(a)

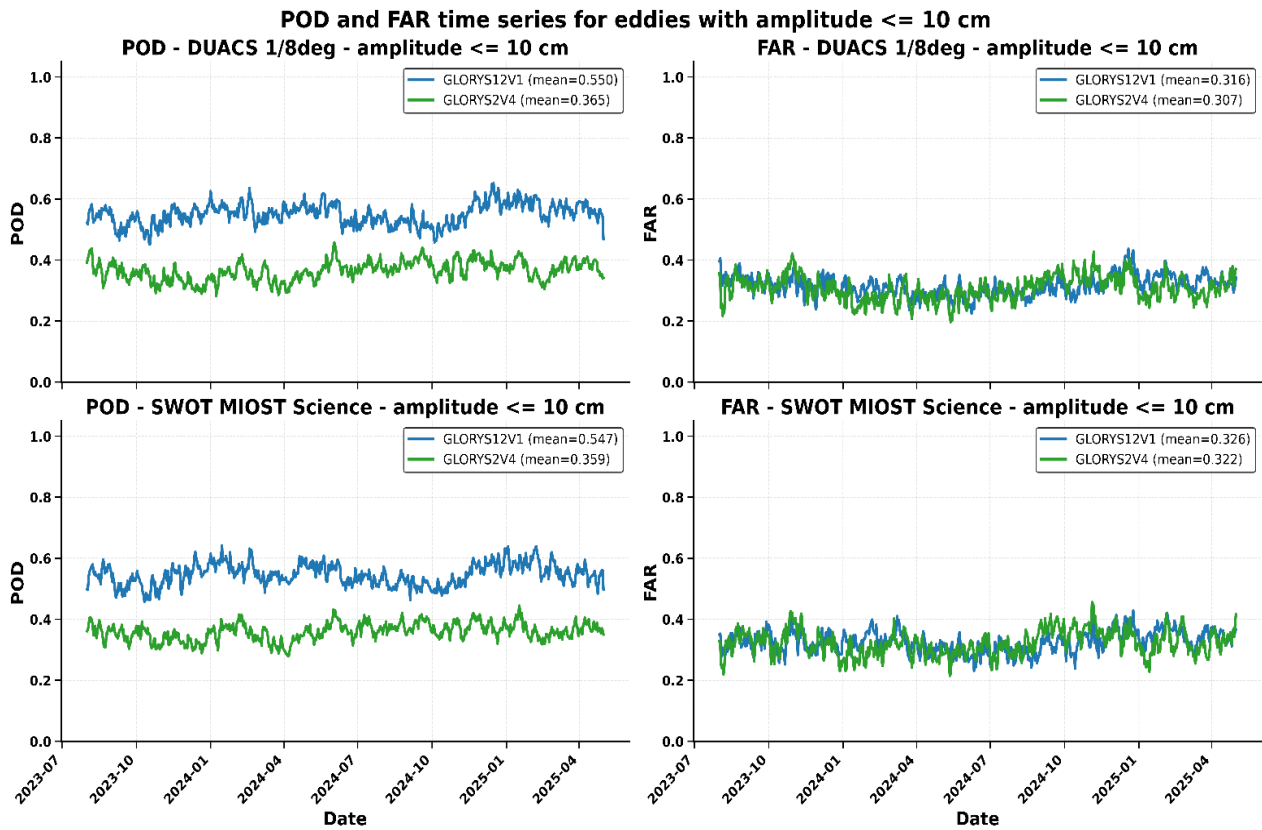


(b)

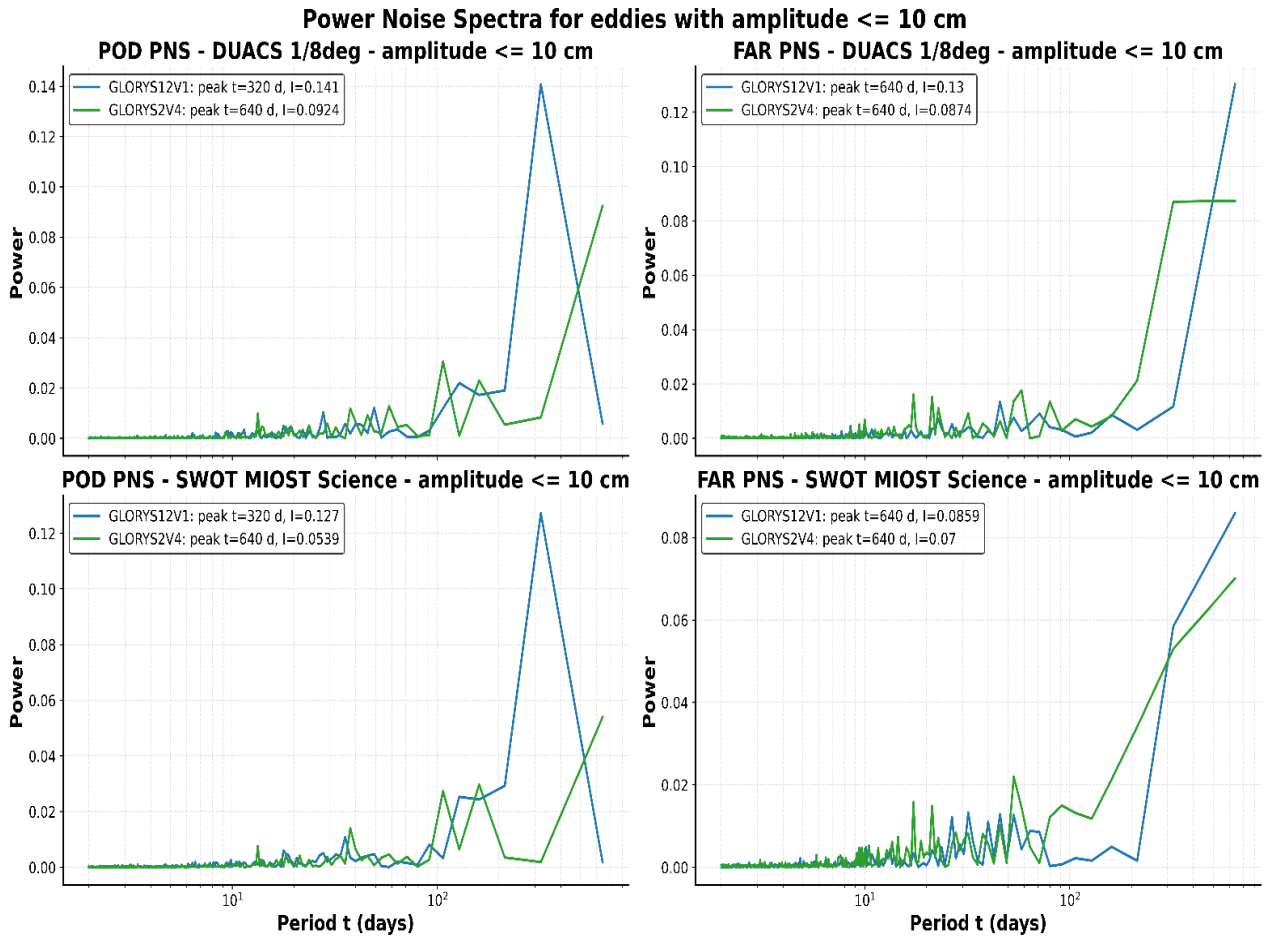
Fig. C3. Time series for daily number of hits (a), misses (b), for both satellite data compared to the reanalyses. On the left, the comparison with SWOT MIOST Science, while on the right, the comparison with DUACS.

As the minimum amplitude threshold increases, the absolute numbers of both hits and misses decrease because fewer eddies satisfy the criterion. For GLORYS12V1, the mean number of hits remains higher than the number of misses at all thresholds. For GLORYS2V4, misses remain more numerous than hits and their relative importance increases at higher thresholds. This behaviour is consistent with the reduction in POD observed for GLORYS2V4. The power spectra in Fig. C2 shows that temporal variability generally increases with the minimum amplitude threshold, mainly at longer periods, while differences remain limited at short periods. GLORYS12V1 generally shows larger POD spectral peaks, whereas GLORYS2V4 generally shows larger FAR peaks, with exceptions at 15 cm for POD and 5 cm for FAR. Fig. C3 shows that both hits and misses decrease in absolute number as the threshold increases because fewer eddies are retained. However, GLORYS12V1 maintains more hits than misses, whereas misses remain more numerous than hits for GLORYS2V4, indicating a greater reduction in matching performance for GLORYS2V4 at higher amplitude thresholds. We also evaluated POD and

FAR for low-amplitude eddies (≤ 10 cm), whose weak SSH signatures make their detection and matching more sensitive to mapping, filtering, and background variability. This subset is therefore expected to show lower POD and higher FAR because small differences between observational and reanalysis fields can alter eddy position and contours or generate unmatched features.



(a)



(b)

Fig. C4. (a) POD and FAR time series for eddies with amplitudes ≤ 10 cm for the comparisons with DUACS 1/8° and SWOT MIOST Science. (b) Corresponding power spectra. A 4-day minimum-lifetime threshold is used.

Fig. C5 shows the distributions of hits, misses, and false alarms as functions of eddy radius and amplitude.

We further introduce the histograms of the eddy matching statistics, namely hits, misses, and false alarms, as a function of both eddy radius and eddy amplitude. The aim is to provide a more detailed view, similar to a zoomed-in analysis, of how the matching statistics vary with the two main physical eddy features. This additional representation is useful because, on a linear scale, the histograms are not always able to clearly show the distinct behaviour of each matching statistic as eddy amplitude and radius increase. A logarithmic y-axis is used to distinguish the matching categories in bins containing fewer eddies. This is shown in Fig. C5.

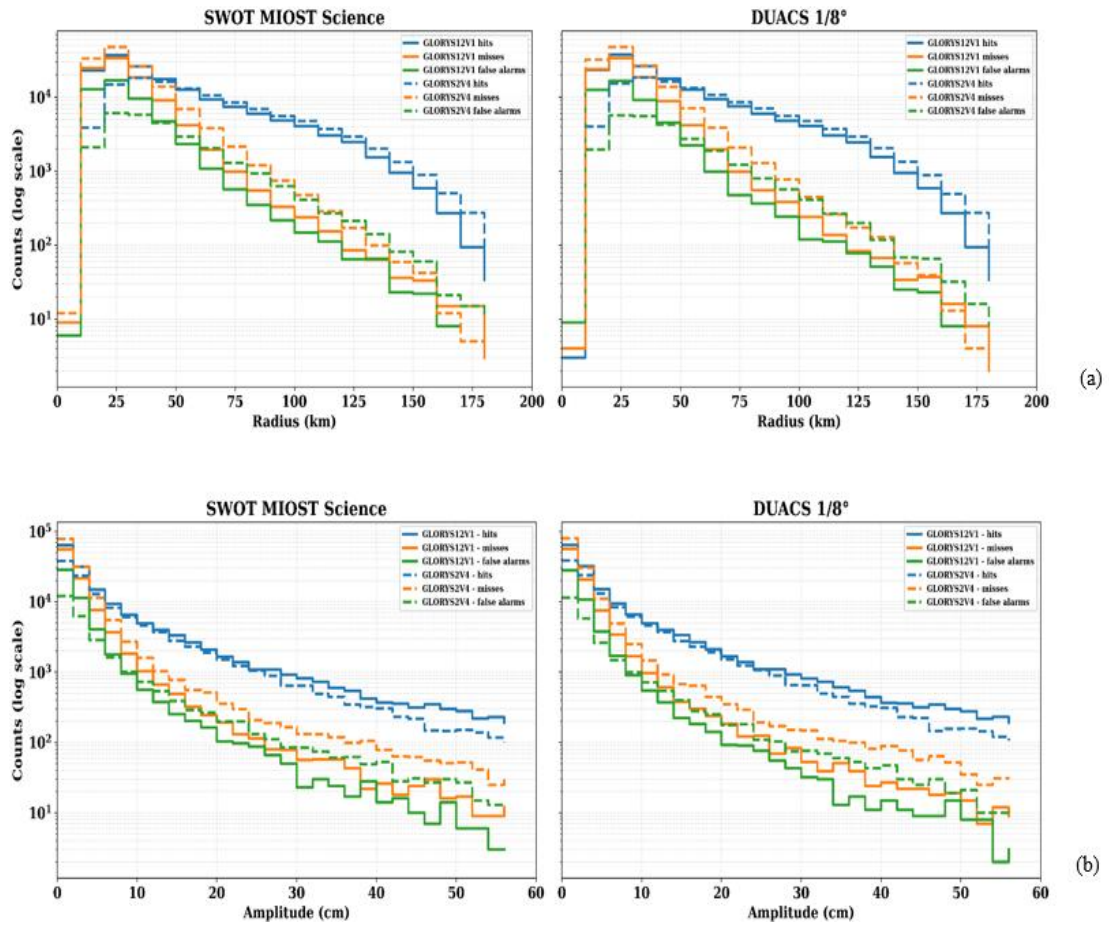


Fig. C5. Histograms for eddy matching statistics on a logarithmic scale, for both comparisons, for eddy radius (a) and eddy amplitude (b).

At the smallest radii and amplitudes, misses are particularly prominent, especially for GLORYS2V4, confirming its greater difficulty in representing small eddies. GLORYS12V1 produces more hits in these classes, although this is accompanied by more small false alarms. For eddy radius, GLORYS2V4 has more misses than GLORYS12V1 up to approximately 150 km, while its false alarms become more numerous above about 50 km. In the amplitude distributions, misses remain the main source of error for GLORYS2V4. Above radii of approximately 50 km and amplitudes of 8–10 cm, hits become the dominant category for both reanalyses and both observational references. The largest discrepancies are therefore concentrated among smaller and lower-amplitude eddies.

Appendix D

This appendix assesses the sensitivity of the verification results to two methodological choices: (i) the maximum centroid distance allowed in the eddy-matching procedure and (ii) the cutoff wavelength of the Bessel high-pass filter applied before eddy detection. The reference values are 125 km for the matching distance and 500 km for the filter cutoff wavelength. The alternative values tested here are 62.5 km for matching and 200 and 800 km for filtering. All other methodological settings, including the 4 days minimum lifetime threshold, remain unchanged.

Sensitivity to a different cost function distance threshold for the eddy matching process: 62.5 km.

The matching analysis was repeated using a maximum centroid distance of 62.5 km, corresponding to half of the reference value of 125 km. This more restrictive criterion allows only eddies with closer centroid positions to be matched. When a

pair accepted with the 125 km threshold no longer satisfies the 62.5 km criterion, the reference eddy is classified as a miss and the corresponding reanalysis eddy as a false alarm.

Figure D1 shows an example for anticyclonic eddies on 1 January 2025. The stricter threshold reduces the number of hits and increases both misses and false alarms, demonstrating the sensitivity of the matching statistics to the maximum distance allowed between eddy centroids.

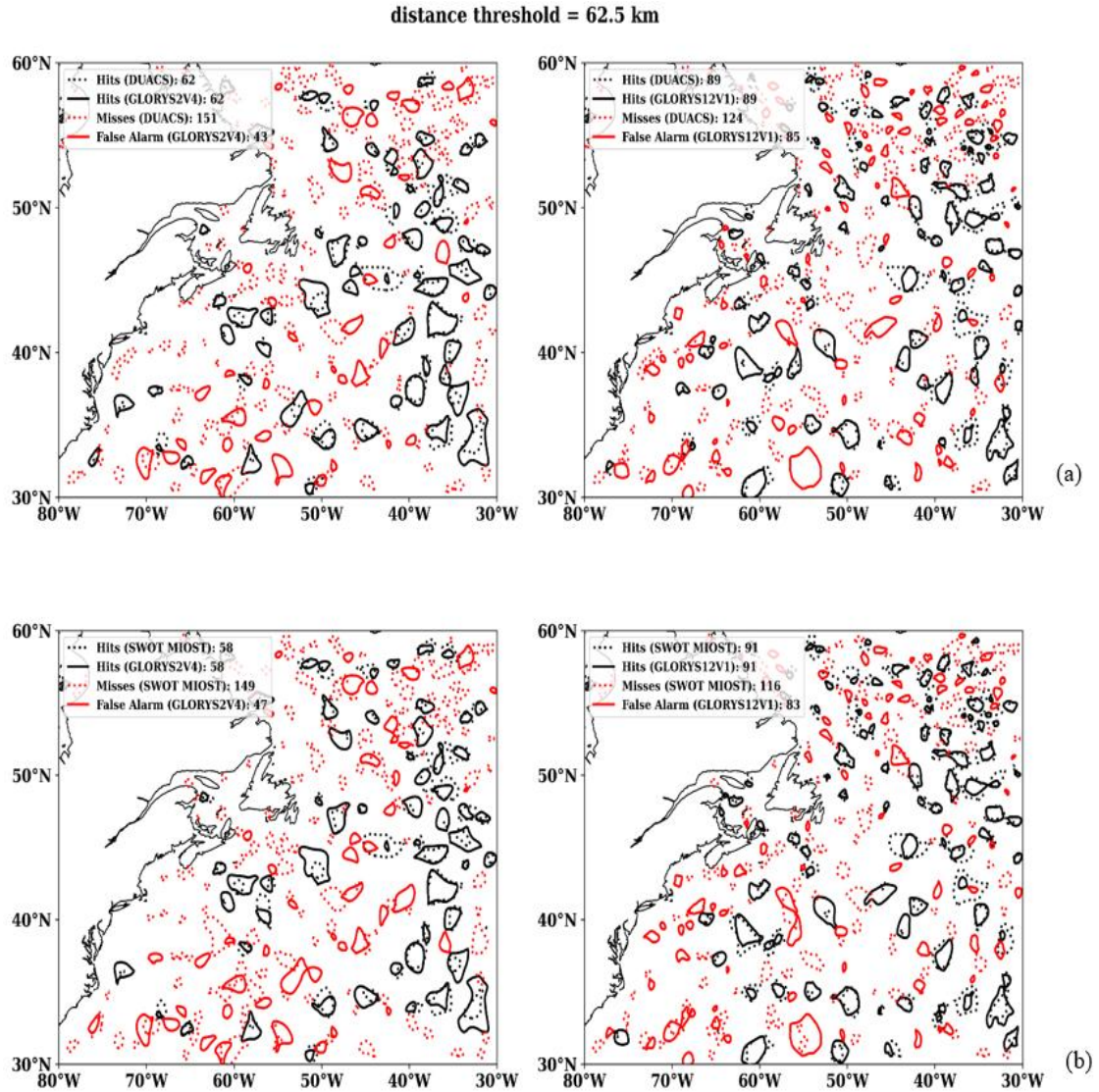


Fig. D1. Anticyclonic-eddy matching on 1 January 2025 using a maximum centroid distance of 62.5 km. GLORYS2V4 is shown on the left and GLORYS12V1 on the right, for the comparisons with (a) DUACS 1/8° and (b) SWOT MIOST Science.

The verification metrics are computed on a daily basis, following the same procedure described in Section 2.4 and Sect. 3.2, in order to evaluate the temporal evolution of the matching performance. Table D1 and Table D2 report the absolute and relative changes in the verification metrics for the two comparisons, respectively.

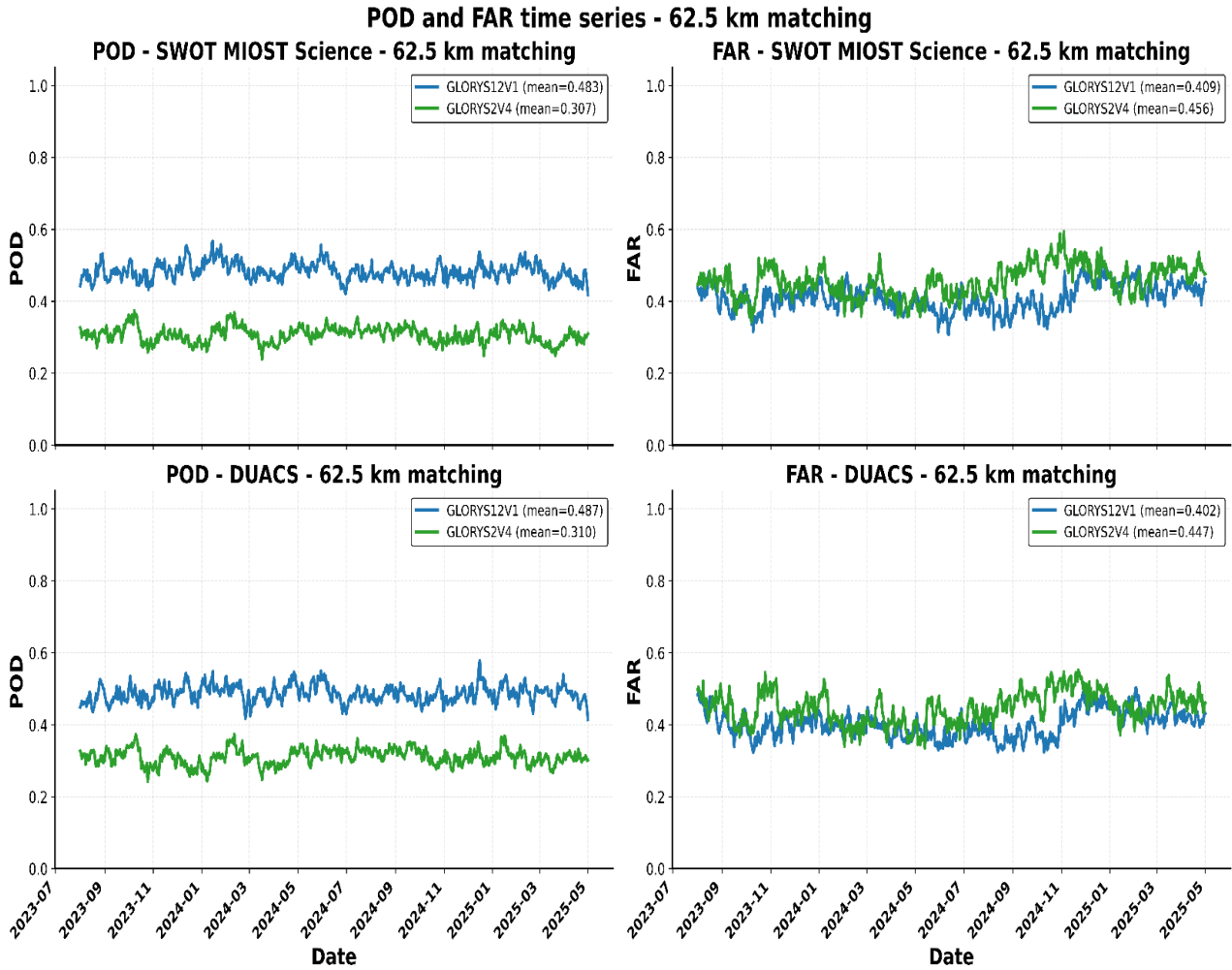
Verification metric	125 km threshold	62.5 km threshold	Difference	Difference %
POD 12V1	0.624	0.483	-0.141	-22.59%
FAR 12V1	0.237	0.409	0.172	72.57%

<i>POD 2V4</i>	<i>0.455</i>	<i>0.307</i>	<i>-0.148</i>	<i>-32.52%</i>
<i>FAR 2V4</i>	<i>0.194</i>	<i>0.456</i>	<i>0.262</i>	<i>135.05%</i>

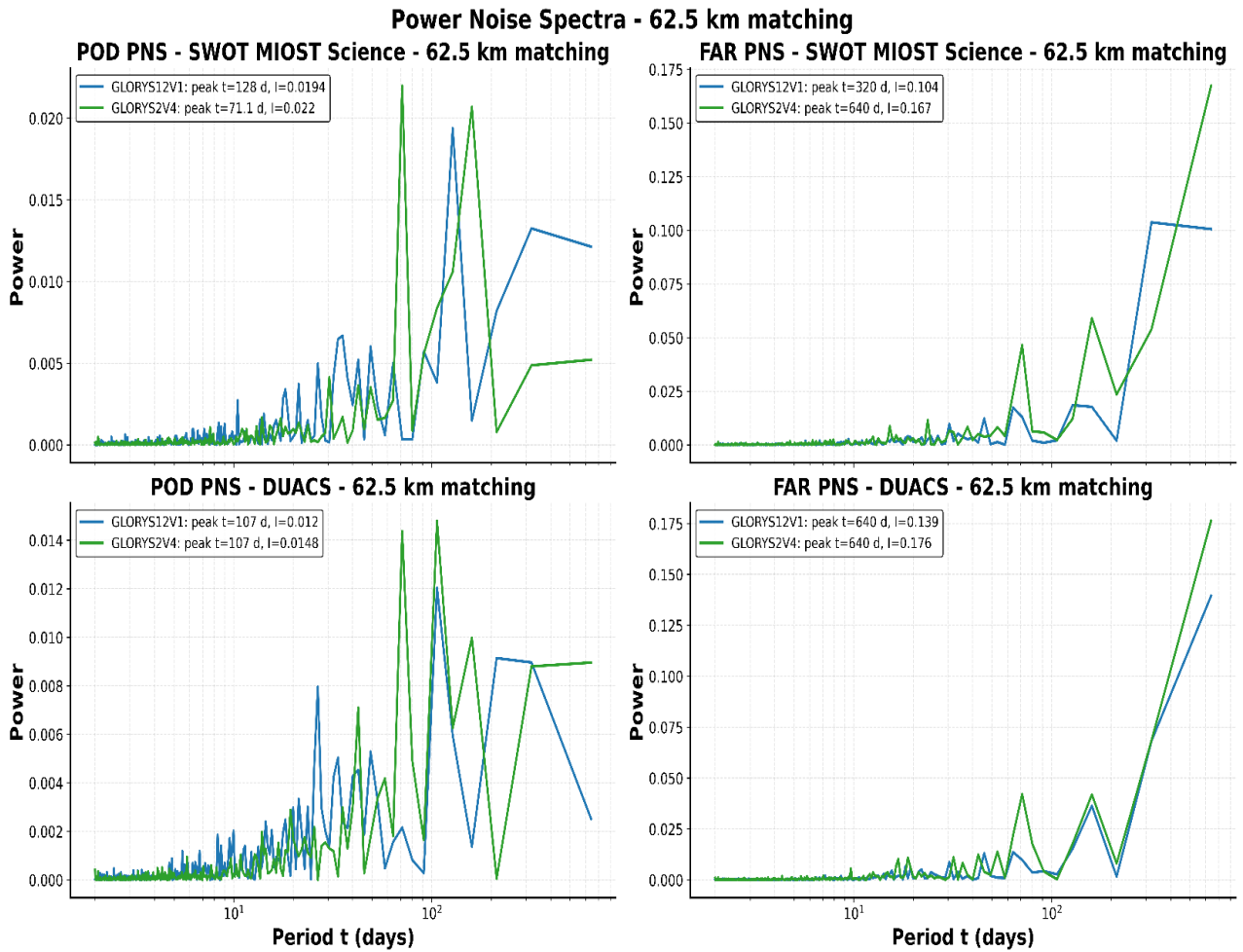
Table D1. Aggregated eddy verification statistics for the Probability of Detection (POD) and False Alarm Ratio (FAR) for the comparison with SWOT using maximum matching distances of 125 and 62.5 km. Absolute and percentage differences are calculated as the 62.5 km result minus the 125 km result.

<i>Verification metric</i>	<i>125 km threshold</i>	<i>62.5 km threshold</i>	<i>Difference</i>	<i>Difference %</i>
<i>POD 12V1</i>	<i>0.628</i>	<i>0.487</i>	<i>-0.141</i>	<i>-22.45%</i>
<i>FAR 12V1</i>	<i>0.230</i>	<i>0.402</i>	<i>0.172</i>	<i>74.78%</i>
<i>POD 2V4</i>	<i>0.460</i>	<i>0.310</i>	<i>-0.150</i>	<i>-32.60%</i>
<i>FAR 2V4</i>	<i>0.182</i>	<i>0.447</i>	<i>0.265</i>	<i>145.60%</i>

Table D2. Aggregated eddy verification statistics for the Probability of Detection (POD) and False Alarm Ratio (FAR) for the comparison with DUACS using maximum matching distances of 125 and 62.5 km. Absolute and percentage differences are calculated as the 62.5 km result minus the 125 km result.



(a)



(b)

Fig. D2. (a) Time series for the verification metrics, for the comparison with SWOT MIOST Science and DUACS $1/8^\circ$ comparison, with a minimum lifetime threshold of 4 days. In (b) power spectra for all the time series presented in (a).

Overall, when the maximum matching distance is reduced from 125 km to 62.5 km, FAR becomes higher for GLORYS2V4 than for GLORYS12V1, as shown in Table D1, Table D2, Fig. D1 and Fig. D2 (a). This does not necessarily mean that GLORYS2V4 produces a larger absolute number of false alarms. Rather, it reflects the relative nature of FAR, which depends on the proportion of unmatched reanalysis eddies with respect to the total number of reanalysis detections considered in the verification.

In contrast, POD remains higher for GLORYS12V1, indicating that the higher-resolution reanalysis is still able to match a larger fraction of the observed eddies, even under the more restrictive distance criterion. The most evident effect of reducing the matching distance is the strong increase in FAR, especially for GLORYS2V4, where FAR more than doubles in both satellite-based comparisons. A substantial increase is also found for GLORYS12V1, with FAR increasing by more than 70%. POD also decreases when the reduced distance threshold is applied, but the relative reduction is smaller than for FAR. The decrease is of the order of 20% for GLORYS12V1 and about 30% for GLORYS2V4. This indicates that the reduced distance threshold may affect the classification of unmatched reanalysis detections, leading to a large increase in false alarms, while the reduction in correctly matched observed eddies is less pronounced.

The power spectra in Fig. 2D (b) shows slightly larger main POD peaks for GLORYS2V4 than for GLORYS12V1. GLORYS2V4 also exhibits the larger FAR peak in both observational comparisons. In the SWOT comparison, the main POD peaks occur at approximately 71 days for GLORYS2V4 and 128 days for GLORYS12V1, whereas both are located

near 107 days in the DUACS comparison. The strongest FAR peaks occur at longer resolved periods. Because the evaluation period is only approximately 640 days long, peaks at or near 640 days should be interpreted as low-frequency variability rather than as evidence of a robust periodic cycle.

Sensitivity to different Bessel filters for eddy detection process

The SSH or ADT fields of all products were high-pass filtered using a Bessel filter before eddy detection. The reference cutoff wavelength is 500 km. To assess the sensitivity of the detection and verification results to the retained spatial scales, the complete detection, tracking, and matching procedure was repeated using cutoff wavelengths of 200 and 800 km. The 200 km cutoff applies more restrictive spatial filtering, whereas the 800 km cutoff retains a broader range of large-scale variability.

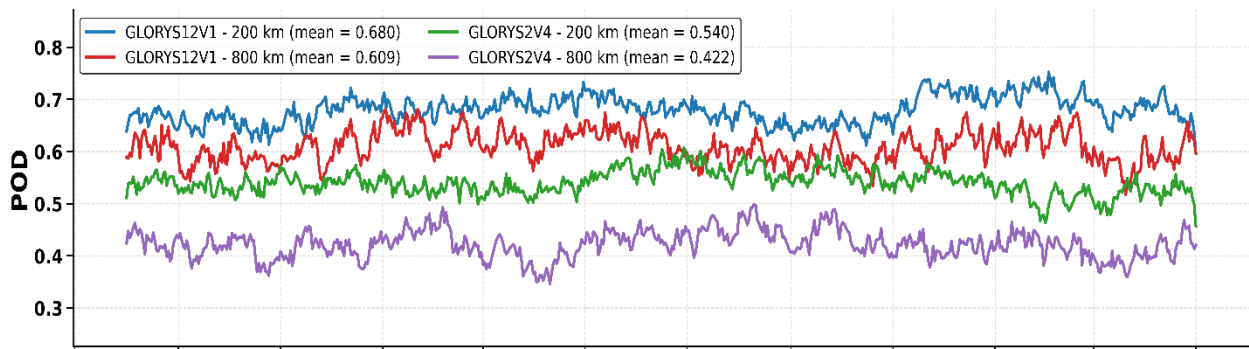
<i>Verification metric</i>	<i>500 km</i>	<i>200 km</i>	<i>Difference</i>	<i>Difference (%)</i>
<i>POD 12V1 (SWOT)</i>	<i>0.624</i>	<i>0.680</i>	<i>0.056</i>	<i>8.97%</i>
<i>POD 12V1 (DUACS)</i>	<i>0.628</i>	<i>0.679</i>	<i>0.051</i>	<i>8.12%</i>
<i>FAR 12V1 (SWOT)</i>	<i>0.237</i>	<i>0.197</i>	<i>-0.04</i>	<i>-16.88%</i>
<i>FAR 12V1 (DUACS)</i>	<i>0.230</i>	<i>0.189</i>	<i>-0.041</i>	<i>-17.83%</i>
<i>POD 2V4 (SWOT)</i>	<i>0.455</i>	<i>0.540</i>	<i>0.085</i>	<i>18.68%</i>
<i>POD 2V4 (DUACS)</i>	<i>0.460</i>	<i>0.541</i>	<i>0.081</i>	<i>17.61%</i>
<i>FAR 2V4 (SWOT)</i>	<i>0.194</i>	<i>0.129</i>	<i>-0.065</i>	<i>-33.51%</i>
<i>FAR 2V4 (DUACS)</i>	<i>0.182</i>	<i>0.118</i>	<i>-0.064</i>	<i>-35.16%</i>

Table D3. Mean eddy verification metrics, difference and percentage difference, crossing from a cutoff wavelength for Bessel filter of 500 km to 200 km for both comparisons.

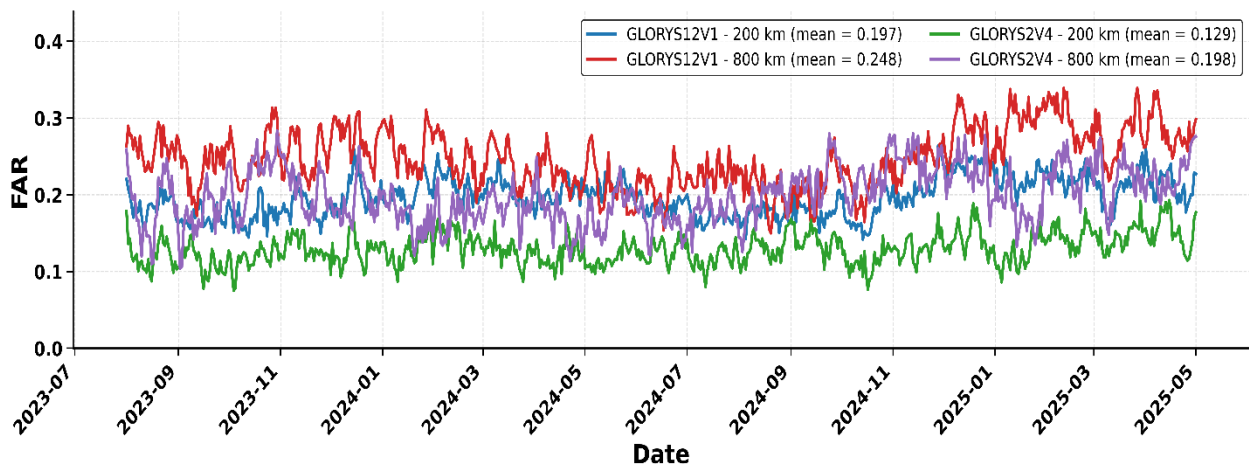
<i>Verification metric</i>	<i>500 km</i>	<i>800 km</i>	<i>Difference</i>	<i>Difference (%)</i>
<i>POD 12V1 (SWOT)</i>	<i>0.624</i>	<i>0.609</i>	<i>-0.015</i>	<i>-2.40%</i>
<i>POD 12V1 (DUACS)</i>	<i>0.628</i>	<i>0.614</i>	<i>-0.014</i>	<i>-2.23%</i>
<i>FAR 12V1 (SWOT)</i>	<i>0.237</i>	<i>0.248</i>	<i>0.011</i>	<i>4.64%</i>
<i>FAR 12V1 (DUACS)</i>	<i>0.230</i>	<i>0.243</i>	<i>0.013</i>	<i>5.65%</i>
<i>POD 2V4 (SWOT)</i>	<i>0.455</i>	<i>0.422</i>	<i>-0.033</i>	<i>-7.25%</i>
<i>POD 2V4 (DUACS)</i>	<i>0.460</i>	<i>0.428</i>	<i>-0.032</i>	<i>-6.96%</i>
<i>FAR 2V4 (SWOT)</i>	<i>0.194</i>	<i>0.198</i>	<i>0.004</i>	<i>2.06%</i>
<i>FAR 2V4 (DUACS)</i>	<i>0.182</i>	<i>0.190</i>	<i>0.008</i>	<i>4.40%</i>

Table D4. Mean eddy verification metrics, difference and percentage difference, crossing from a cutoff wavelength for Bessel filter of 500 km to 800 km for both comparisons.

Probability of Detection - SWOT MIOST Science: 200 vs 800 km Bessel filter

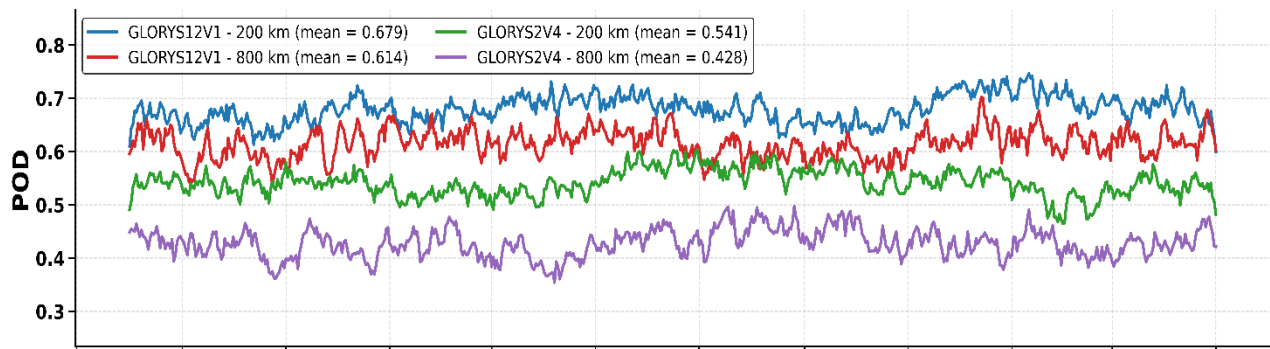


False Alarm Ratio - SWOT MIOST Science: 200 vs 800 km Bessel filter

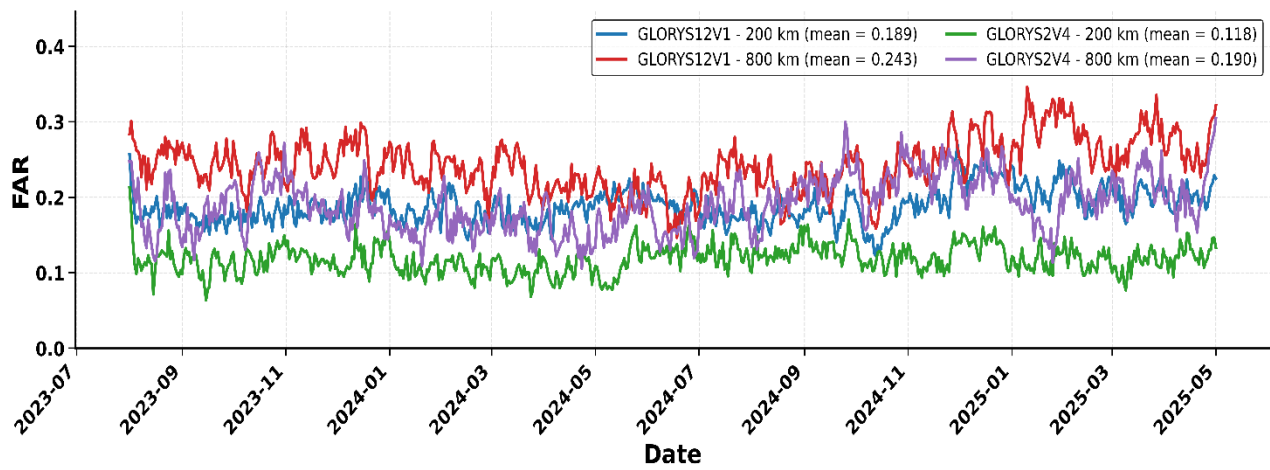


(a)

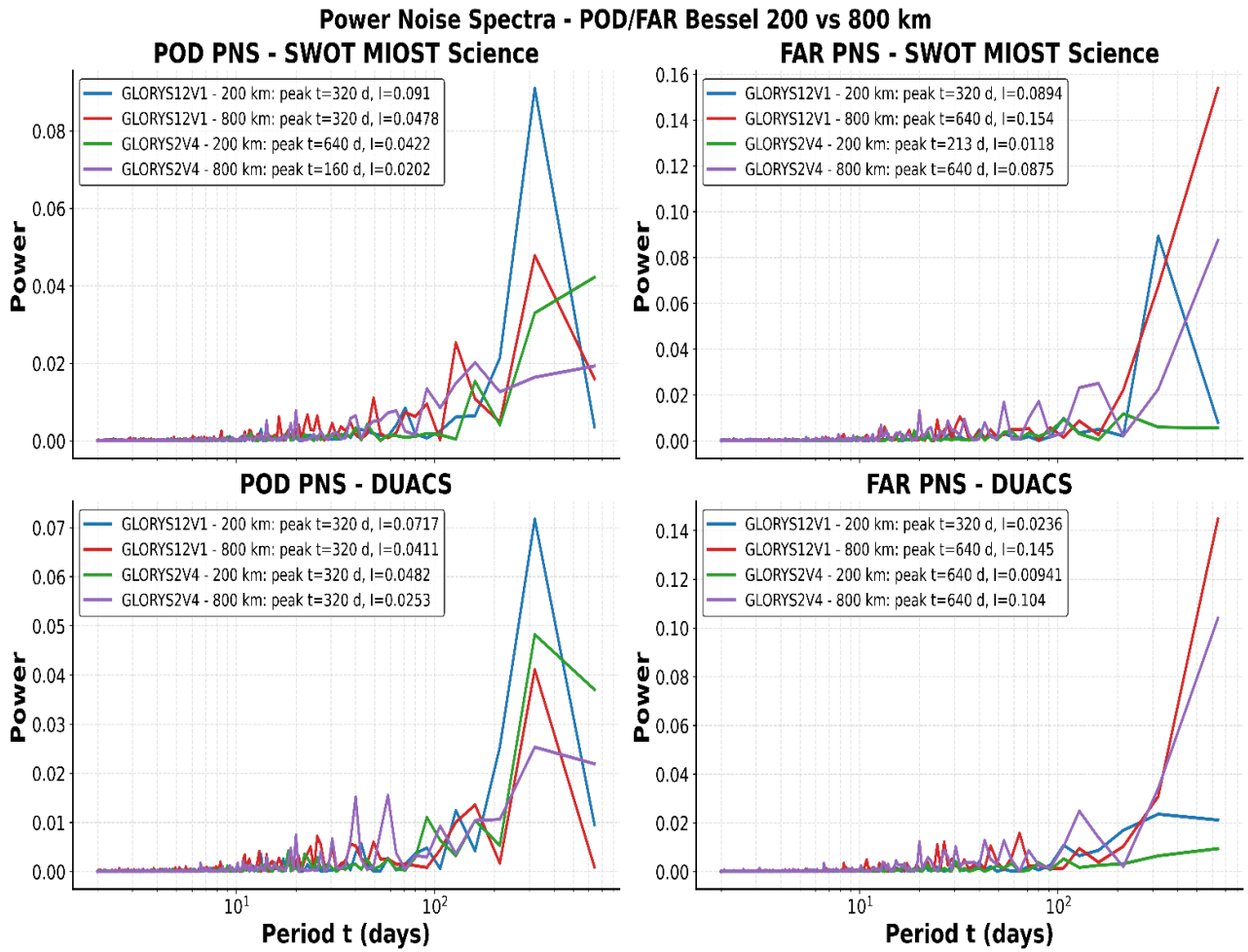
Probability of Detection - DUACS: 200 vs 800 km Bessel filter



False Alarm Ratio - DUACS: 200 vs 800 km Bessel filter



(b)



(c)

Fig. D3. POD and FAR time series obtained using Bessel-filter cutoff wavelengths of 200 and 800 km for the comparisons with (a) SWOT MIOST Science and (b) DUACS $1/8^\circ$. Panel (c) shows the corresponding power spectra. A 4-day minimum-lifetime threshold is used.

Reducing the Bessel-filter cutoff wavelength from 500 to 200 km increases POD and decreases FAR for both reanalyses and both observational comparisons. The changes are particularly pronounced for GLORYS2V4, for which POD increases by approximately 18% and FAR decreases by approximately 34–35%. For GLORYS12V1, POD increases by about 8–9% and FAR decreases by approximately 17–18%. Increasing the cutoff wavelength from 500 to 800 km produces smaller and opposite changes: POD decreases and FAR increases for both reanalyses. The POD reduction is larger for GLORYS2V4, whereas the FAR changes remain relatively small. For both cutoff tests, POD is slightly more sensitive in the SWOT comparison, while FAR is generally more sensitive in the DUACS comparison. These changes show that the verification metrics depend on the spatial scales retained before eddy detection. However, the better verification scores obtained with the 200 km cutoff wavelength do not necessarily demonstrate that this value is physically more appropriate, because stronger filtering may remove structures that would otherwise remain unmatched. Regarding the power spectra, the largest peaks for POD are observed for GLORYS12V1 using the 200 km filter cutoff, at temporal scales of the order of several hundred days, approximately 320 days. For FAR, the highest values are instead found for GLORYS12V1 with the 800 km filter cutoff, with enhanced power around 640 days. At shorter temporal scales, the variability is nearly similar across all datasets and filter configurations. Overall, in terms of noise power, the strongest peak is found for the POD of GLORYS12V1 with the 200 km filter cutoff in the comparison with SWOT MIOST.”

Second reviewer's comments

This revision includes all the comments from Reviewer 2.

We have included several figures to provide a clearer explanation of the revisions and to better support the changes made in the manuscript. In addition, the four appendices, from Appendix A to Appendix D, have been included at the end of this revision document and the revised manuscript (after the Summary and Discussion section). Several additional analyses are reported in these appendices to support the statements made in the main text.

Please note that, following the changes made during the revision process, the line numbers in the manuscript have changed.

We would like to sincerely thank Reviewer 2 for the careful evaluation of our manuscript and for the detailed, constructive comments and very helpful suggestions. These comments helped us to improve the clarity, consistency, and scientific robustness of the manuscript.

Legend:

Black = reviewer's comments.

Blue = corrections (quoted text).

Green = response

In some cases, the text quoted in blue includes passages that are not limited strictly to the specific section being revised, but have been included to provide the necessary context and to better explain the concept requiring clarification or modification.

Some comments, especially for the minor issues, have been combined and addressed with a single response, because they refer to the same part of the text or to closely related aspects of the manuscript.

Major points

1) The use of two altimetry products is interesting since it shows that there is not a single truth. However, the North Atlantic has been used as a playground for several SWOT-enriched experimental products. Apart from MIOST, there are experimental maps produced with other software: <https://www.aviso.altimetry.fr/en/data/products/sea-surface-height-products/global/experimental-multimission-gridded-l4-sea-level-heights-and-velocities-with-swot.html>

The Authors should expand the literature review.

Response: We have expanded the literature review to include other SWOT-enriched experimental Level-4 products available over the North Atlantic, in particular 4DVarNET and 4DVarQG, and briefly described their methodological differences from MIOST. We have revised the manuscript, now at lines: 77– 118.

“The availability of observations from the SWOT (Surface Water and Ocean Topography) wide-swath satellite altimetry mission provides another possible reference dataset for evaluating the capability of higher-resolution ocean reanalyses to identify and to track mesoscale eddies over the time in a regional area. SWOT was launched in December 2022 and offers wide-swath measurements of sea surface height, which allows observations of ocean features at much finer spatial scales than traditional altimetry. One of the objectives of SWOT is the observation of two-dimensional ocean surface topography and the characterization of mesoscale and submesoscale ocean circulation. Nevertheless, one important application of SWOT observations is their use in ocean analysis and forecasting systems through data assimilation, since wide-swath SSH observations can help better constrain mesoscale features and smaller-scale variability (Morrow et al., 2019; Fouchet et al., 2025). In the present study, however, we do not use the native SWOT swath observations directly. We use the experimental SWOT MIOST Science product, a global daily Level-4 gridded reconstruction of sea-surface height and derived geostrophic velocities produced by SSALTO/DUACS. MIOST stands for Multiscale Interpolation Ocean Science Topography. It merges irregularly sampled altimeter observations into fields that are continuous in space and time using a linear optimal-interpolation framework in which different covariance functions can be used to represent variability at different spatial and temporal scales. In the configuration considered here, the covariance functions are designed primarily to represent mesoscale geostrophic variability (Ubelmann et al., 2021; CLS and CNES, 2025). SWOT Level-4 sea-level products are not limited to the MIOST reconstruction used in this study. AVISO/DUACS also distributes experimental regional products generated using 4DVarNET and 4DVarQG. The 4DVarNET approach combines a variational data-assimilation formulation with a deep-learning framework to learn both the representation of ocean dynamics and the associated solver (Fablet et al., 2021; Beauchamp et al., 2023). In contrast, 4DVarQG uses a four-dimensional variational scheme constrained by a reduced-order quasi-geostrophic model, introducing an explicit dynamic prior into the reconstruction (Le Guillou et al., 2021). Comparisons over the North Atlantic indicate that 4DVarNET and 4DVarQG are capable of providing finer effective resolution and improving the reconstruction of the energetic and nonlinear structures of the Gulf Stream compared to MIOST, although their performance may vary spatially and neither method currently provides a global mapping solution (Ballarotta et al., 2025). MIOST is selected because it provides a spatially consistent global reconstruction (such as AVISO DUACS 1/8°) covering the intended validation domain, the North Atlantic, up to the upper mid-latitudes (60°N). The 4DVarNET and 4DVarQG experimental maps, as of the study’s final implementation (October and November 2025), were currently regional products covering approximately 25–50° N and 80° W–10° W and therefore do not include the northern part of our study region. However, there are some important limitations that must also be considered. First, because the mission is very recent, the time coverage of SWOT data is limited. Therefore, SWOT data cannot yet be used to study decadal or long-term variability of mesoscale activity. In this study, SWOT is used only for a short-period evaluation of the representation of mesoscale and smaller-scale eddy features in high-resolution reanalyses. Second, since similar nadir altimetry observations are assimilated in some ocean reanalysis products considered in this study, neither AVISO DUACS 1/8° nor SWOT MIOST Science can be regarded as a fully independent verification dataset. DUACS 1/8° is based on multi-mission nadir altimetry, while SWOT MIOST Science combines SWOT wide-swath observations with a multi-mission nadir-altimetry background used to constrain the larger-scale signal. Therefore, both satellite products share, at least partly, observational information with the reanalyses. However, SWOT MIOST Science also introduces additional fine-scale information from SWOT wide-swath measurements, which is not directly assimilated in the reanalysis products

considered here. For these reasons, AVISO DUACS 1/8° and SWOT MIOST Science are used as complementary evaluation references rather than as absolute observational truth. The comparison with SWOT provides insight into how well the reanalyses represent fine-scale variability revealed by wide-swath altimetry, whereas the DUACS comparison provides continuity with the longer pre-SWOT altimetry record. This complementary approach allows us to assess the representation of mesoscale eddies in the reanalyses while explicitly acknowledging the limitations associated with the partial dependence of the reference datasets.”

2) The analysis is sometimes redundant, the text can be reshaped in a better way : -) According to the Authors' idea, the use of two similar lifetimes can be seen as a sensitivity analysis. I can agree on this point, but results should be presented accordingly. Rather than discussing values that are similar, priority can be given to one lifetime and introducing the second lifetime only when it shows results significantly different. The rest of the plots can be put in the supplementary material.

Response: We have revised the manuscript following the reviewer's suggestion. The 4-day minimum lifetime threshold is now used as the primary threshold throughout the main analysis. The results obtained using the 7-day threshold, together with an additional sensitivity analysis using a 28-day threshold, have been moved to Appendix A. In the main text, these alternative thresholds are discussed only where they provide relevant information on the sensitivity of the results. This restructuring reduces redundancy and gives priority to the primary 4-day analysis.

We have revised the manuscript, now at lines: 238 - 247.

“A minimum lifetime threshold is applied to reduce the identification of transient or spurious features, like in Mason et al. (2014). In this study, we adopt a 4-day minimum lifetime threshold, requiring each eddy to be detected on consecutive days without observational gaps. This relatively short threshold is used to verify the ability of the reanalyses to represent the full spectrum of mesoscale features detected in the satellite products, including short-lived eddies that may still affect the matching statistics. In addition, we have also introduced a calculation with a minimum lifetime threshold of 7 days and 28 days (see Appendix A, for more details). A lifetime threshold of 7 days is also used for certain evaluations to demonstrate the sensitivity to this choice (e.g 4 days). The 28-day threshold is useful for isolating persistent structures, but it is not our primary choice because the aim is also to evaluate the entire population, including short-lived eddies. Eddy properties such as effective radius, amplitude, trajectory, and lifetime are computed for each track. The effective radius is defined as the radius of a circle with the same area enclosed by the eddy contour.”

-) Similar comment can be done for the analysis of bootstrap resampling that largely confirms the standard analysis and can be partially put in a supplementary.

Response: We have revised the manuscript following the reviewer's suggestion. The main text now presents only the 4-day block-bootstrap analysis, which is consistent with the primary 4-day minimum lifetime threshold. Additional block-bootstrap sensitivity tests using block lengths of 7, 20, and 30 days, as well as block lengths corresponding to the mean lifetimes of the satellite eddies, have been moved to Appendix B. These additional analyses are referred to in the main

text only to assess the robustness of the results to the choice of block length. We have revised the manuscript, now at lines: 444 - 462.

“To further assess the robustness of these findings, a block bootstrap resampling (95% confidence intervals) was performed using 4-day blocks, consistent with the imposed minimum lifetime threshold. Results for both SWOT and DUACS (Tables 6 and 7) confirm that GLORYS12V1 systematically outperforms GLORYS2V4 in terms of POD (0.618 - 0.620 vs 0.448 - 0.452), with a statistically significant difference (~ 0.17) that is insensitive to block length. This gain is consistently associated with a moderate increase in FAR (~ 0.045 - 0.049), confirming a robust trade-off between detection sensitivity and reliability. The agreement between the two observational references strengthens confidence in the statistical significance of the results, although it does not fully remove uncertainties related to the definition and detectability of small-scale or short-lived eddy structures.

Verification Metric	GLORYS12V1	GLORYS2V4	<i>Difference (12V1 – 2V4)</i>
POD (Mean)	0.618	0.448	0.170
POD (95% CI)	[0.609, 0.627]	[0.441, 0.454]	[0.160, 0.182]
FAR (Mean)	0.235	0.190	0.045
FAR (95% CI)	[0.229, 0.242]	[0.183, 0.198]	[0.038, 0.052]

Table 6. Bootstrap resampling for SWOT MIOST Science comparison, for a minimum lifetime threshold of 4 days, and a block length of 4 days.

Verification Metric	GLORYS12V1	GLORYS2V4	<i>Difference (12V1 – 2V4)</i>
POD (Mean)	0.620	0.452	0.168
POD (95% CI)	[0.610, 0.629]	[0.442, 0.463]	[0.154, 0.180]
FAR (Mean)	0.235	0.187	0.049
FAR (95% CI)	[0.230, 0.237]	[0.181, 0.192]	[0.040, 0.054]

Table 7. Bootstrap resampling for DUACS 1/8° comparison, for a minimum lifetime threshold of 4 days, and a block length of 4 days.

As additional sensitivity tests of the coupled block bootstrap, we also performed the bootstrap resample using 7-day, 20-day and 30-day blocks, and the mean lifetimes of both satellite eddies. The results are reported in Appendix B. The relative ranking of GLORYS12V1 and GLORYS2V4 and the main conclusions of the 4-day analysis remain consistent across the different block lengths, although the estimated confidence intervals vary with the choice of block length. Therefore, the 30-day test is included only as an additional robustness check. Very large block lengths may reduce the variability of the resampling procedure and may partly weaken the main purpose of the block bootstrap, which is to account for temporal dependence in the data when the underlying statistical distribution is unknown.”

-) About the reanalysis comparison, the Authors described the hit/misses/false alarm skills and then the POD/FAR metrics moving back and forth multiple times. The two types of metrics can be described together for one reanalysis and then for the second.

Response: We have revised the manuscript at lines 463 – 515 (including the whole part about eddy matching statistics histograms) to reduce the repeated back-and-forth discussion between the eddy matching statistics (hits, misses, and false alarms) and the corresponding verification metrics (POD and FAR). The revised section now presents these quantities in a more integrated way, allowing the behaviour of the two reanalyses to be compared more directly across the same eddy-property ranges. We have also added Appendix C, which contains additional sensitivity analyses based on different amplitude thresholds and further representations of the matching statistics, allowing the main text to remain more focused.

“In general, as the minimum eddy lifetime threshold increases, the FAR and POD values of each reanalysis products, for both satellite comparisons, tend to converge toward similar levels as described in Appendix A for longer-lived eddies (especially for 28-day minimum eddy lifetime threshold). Figures 7 and 9 show the dependence of the eddy matching statistics (hits, misses and false alarms) on the physical properties of the eddies, namely the effective radius and amplitude, respectively. These statistics are calculated to show how eddies are distributed across different ranges of radius and amplitude, and thus to identify the ranges with the highest clustering. For hits and misses, the radius and amplitude were taken from the satellite (the reference) product, because these eddies refer to observed features that are either detected or not detected by the reanalysis. For false alarms, the radius and amplitude are taken from the corresponding reanalysis product. Each bin of the histogram has a width of 10 km and 2 cm, respectively for eddy radius and eddy amplitude. The highest counts of misses and false alarms are associated with small eddies, specifically for those with radii below 50 km and amplitudes below 5 cm. Figures 8 and 10 show the dependence of the verification metrics in terms of the same physical eddy features.

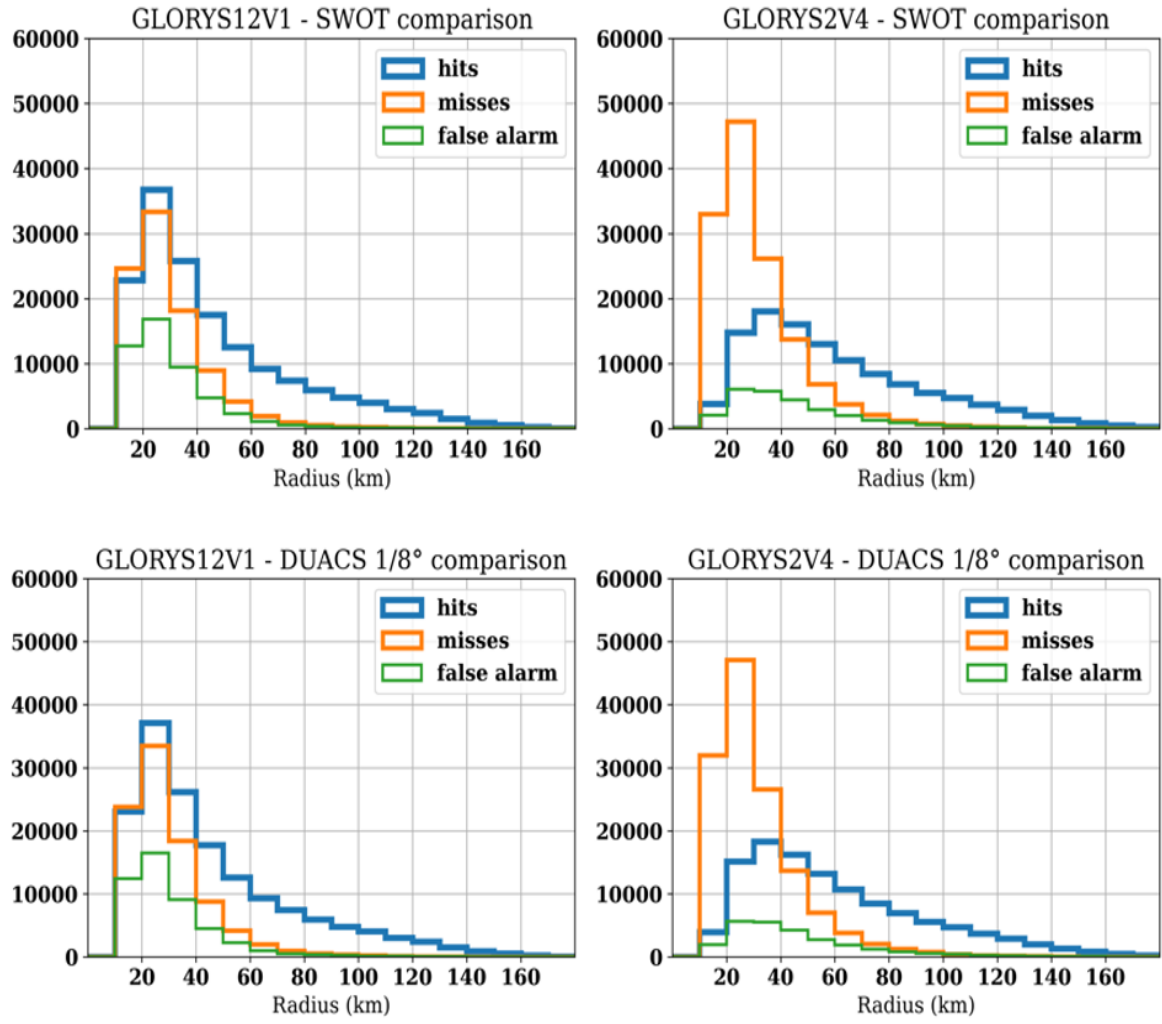


Fig. 7. On the top: eddy matching histograms for the total number of hits (blue), misses (orange), false alarms (green) for GLORYS12V1 (left) and GLORYS2V4 (right), compiled as a function of effective radius (km) over the evaluation period, for SWOT comparison, with a lifetime threshold of 4 days. In the bottom: eddy matching histograms compiled as a function of effective radius (km) over the same evaluation period, for comparison with DUACS, with a lifetime threshold of 4 days. The radius of hits and misses refers to the satellite (reference) eddies, while the radius of false alarms refers to those of unmatched eddies detected in the reanalysis.

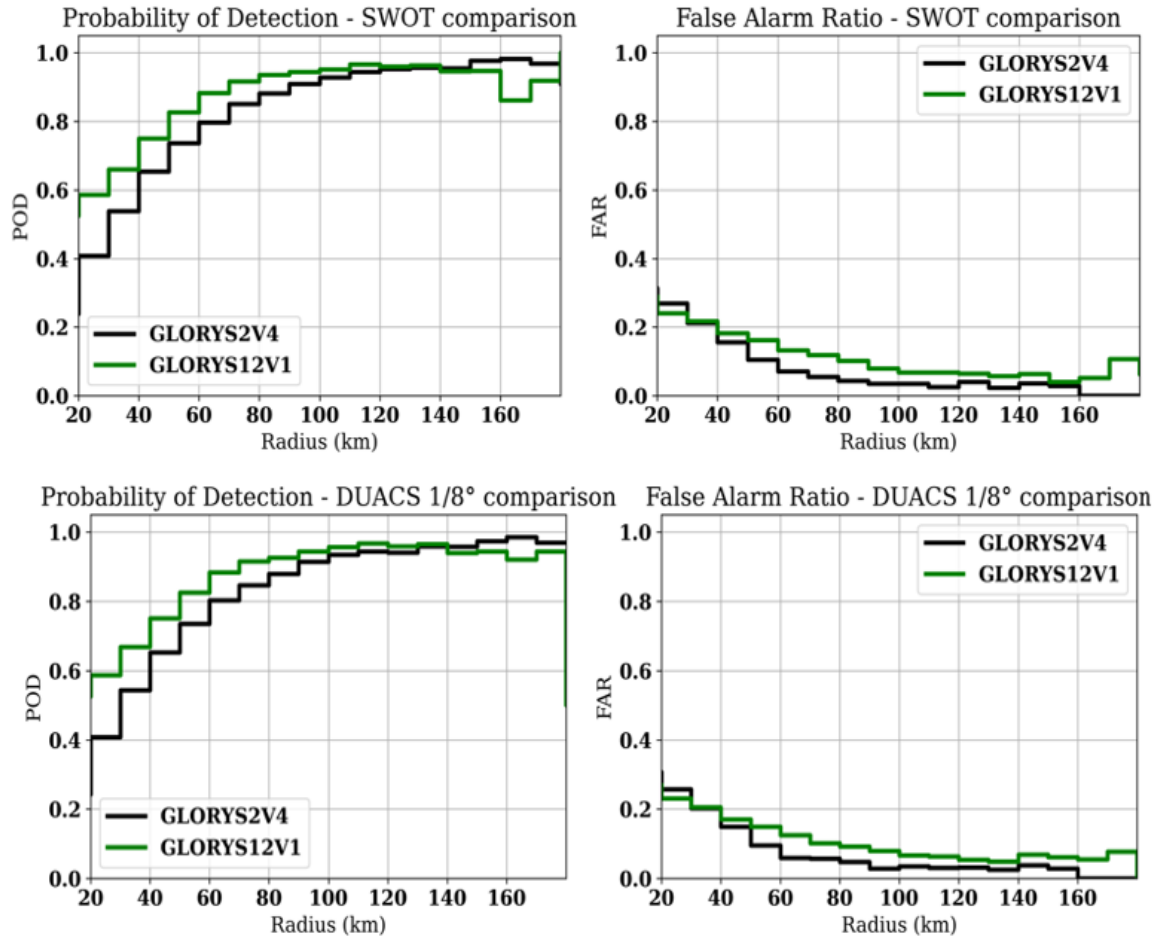


Fig.8. On the top: histograms of the probability of detection (POD; left) and the false alarm ratio (FAR; right), over the evaluation period, evaluated as a function of radius (km) for comparison with SWOT, with a lifetime threshold of 4 days. In the bottom: histograms of the POD and FAR as a function of radius (km), for the comparison with DUACS, with a lifetime threshold of 4 days. The radius of hits and misses refers to the satellite (reference) eddies, while the radius of false alarms refers to those of unmatched eddies detected in the reanalysis.

For the DUACS comparison, Figures 8 and 10 show that FAR is generally higher in the smallest effective-radius and amplitude bins, exceeding 30% in some of these classes and highlighting the greater difficulty in verifying the smallest features in this gridded product. The FAR remains close or higher than 20% until 50 km in terms of effective radius. At the same time, the POD of GLORYS12V1 is close to 60% even for small radius and amplitude values, showing a relevant fraction of small observed eddies is still detected by the reanalyses. The largest proportion of misses is consistently associated with small eddies, specifically those with radii below 50 km and amplitudes under 4 cm. The comparison with SWOT shows a similar dependence on eddy size. The largest differences in the verification metrics are again found for small-radius and small-amplitude eddies. The POD of GLORYS12V1 remains close to 60%, while the GLORYS2V4 FAR is generally higher in the smallest bins. This indicates that the main difficulty is not only the detection of small observed eddies, but also the presence of unmatched (especially for the small eddies) produced by the reanalyses. However, for both satellite comparisons, the verification skills generally improve as eddy radius and amplitude increase. This suggests a better matching capability for larger and stronger mesoscale eddies, which are less affected by small-scale noise or detection uncertainty. The differences between the DUACS and SWOT comparison are most evident for small eddies, while they become smaller for larger eddies. For this reason, Appendix C includes additional analyses of the verification metrics using different amplitude-based subsets, including eddies with amplitudes greater than 5, 10, 15, and 20 cm, as well as eddies with amplitudes below 10 cm. Appendix C also presents the eddy matching histograms using a semi-

logarithmic scale for the y-axis, which represents the number of eddies. These additional analyses allow us to assess more clearly how the verification results change when low-amplitude eddies are progressively excluded, when the analysis focuses on stronger mesoscale structures, and when the contribution of the most numerous small-amplitude eddies is examined in greater detail.

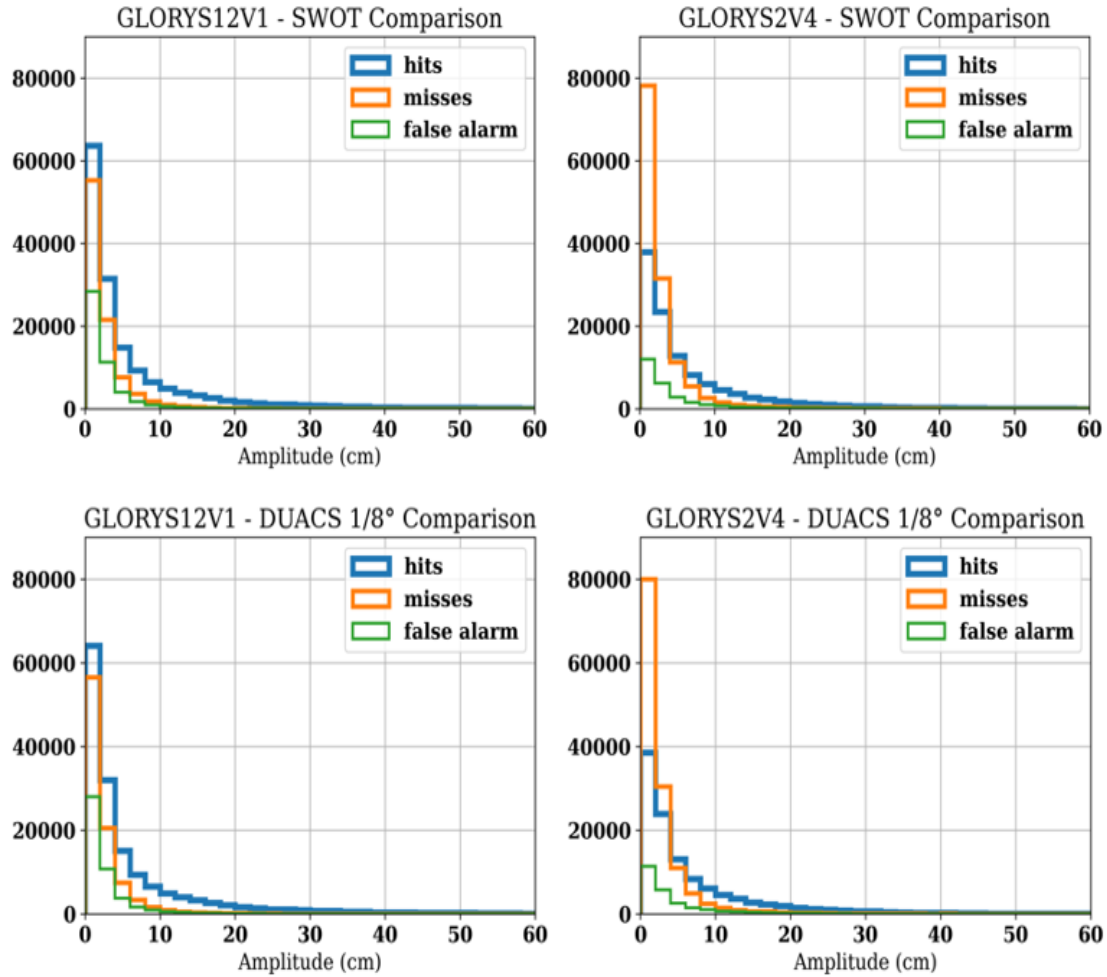


Fig.9. On the top: eddy matching histograms for the total number of hits (blue), misses (orange), false alarms (green) for GLORYS12V1 (left) and GLORYS2V4 (right), compiled as a function of amplitude (cm) over the evaluation period, for SWOT comparison. In the bottom: eddy matching histograms compiled as a function of amplitude (cm) over the evaluation period, for comparison with DUACS 1/8°. The amplitude of hits and misses refers to the satellite (reference) eddies, while the amplitude of false alarms refers to those of unmatched eddies, detected in the reanalysis.

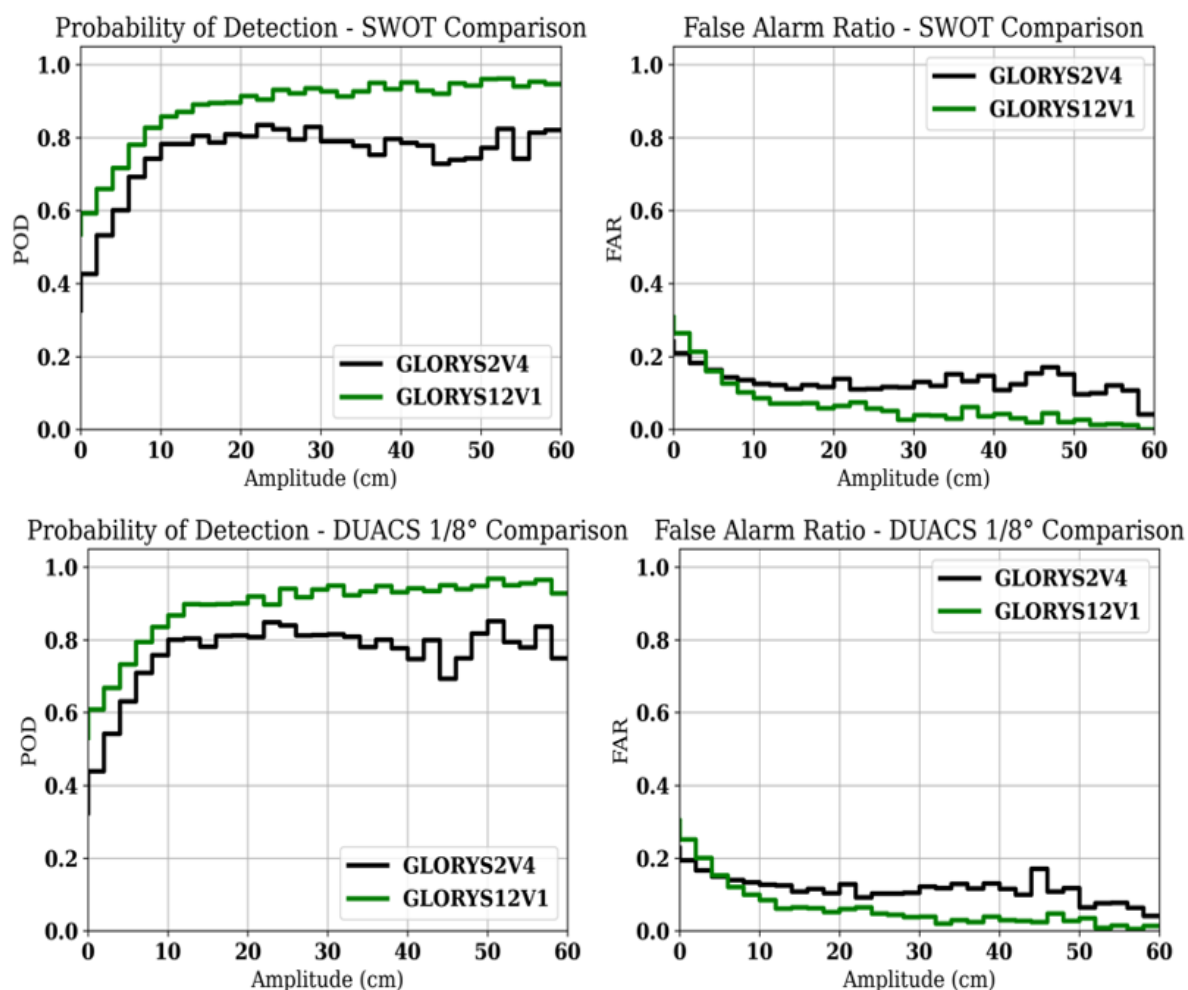


Fig. 10. On the top: histograms of the probability of detection (POD; left) and the false alarm ratio (FAR; right), over the evaluation period, evaluated as a function of amplitude (cm) for comparison with SWOT. In the bottom: histograms of the POD and FAR as a function of amplitude (cm), over the evaluation period, for the comparison with DUACS. The amplitude of hits and misses refers to the satellite (reference) eddies, while the amplitude of false alarms refers to those of unmatched eddies detected in the reanalysis.”

-) There are some mismatches between the text described and the corresponding figures and tables that do not facilitate the readability. I can suggest the use of acronyms to not repeat AVISO DUACS 1/8, AVISO DUAC, MIOST SWOT, SWOT MIOST etc. Some sentences tend to assess the quality of altimetry products based on the skill-scores of the Reanalyses. This mutual assessment should be avoided; the quality of observation-only data can be quantified by comparing the two datasets among themselves.

Response: we tried to completely get rid of all typos related to name mismatches and errors in the mentioned tables.

Further comments on the text will be provided in the minor comments.

Minor Points

-) Line 20: “[..] small and weak eddies” Please add the definition of weak eddy

Response: we have revised the abstract's manuscript, deleting “weak” eddies. It was unnecessary for the abstract description. We have revised the manuscript, now at lines: 18 – 20.

“For both AVISO DUACS 1/8° and SWOT MIOST Science, the POD increases by more than 30% when moving from GLORYS2V4 to GLORYS12V1, while the FAR also rises by about 20%, mainly due to the detection of more small eddies.”

-) Line 24: “[..] eddies in ocean reanalyses and highlight the value of wide-swath.” Such behaviour is proven only in energetic regions.

Response: we have revised the abstract's manuscript, now at lines: 24 – 25.

“The results underline the added value of high model resolution for the representation mesoscale eddies in ocean reanalyses and highlight the value of wide-swath altimetry for model verification in regions as North Atlantic.”

) Line 45: “The detection method used [...]” The line should be rephrased, some of the methods listed in the sentence are similar to Chelton et al 2011.

Response: We revised the manuscript, now the lines are: 45 – 59.

“Several studies have investigated mesoscale eddy properties, such as population, duration, lifetime, and trajectories, in numerical models and reanalyses (de Vos et al., 2018; Mason et al., 2019; Xie et al., 2020; Bonaduce et al., 2021; Smith and Fortin, 2022). Eddy detection methods commonly used in these studies include approaches based on the Okubo-Weiss parameter, wavelet analysis of the relative vorticity field, and the geometry of streamlines (Souza et al., 2011). Another widely used approach, introduced by Chelton et al. (2011), relies on satellite-altimetry fields to identify coherent eddy structures. The py-eddy-tracker package developed by Mason et al. (2014) is also widely used to detect and track eddies at regional and global scales, as in de Vos et al. (2018) and López-Álzate et al. (2022). In the present study, we use py-eddy-tracker to perform the eddy detection and tracking process. This step is applied independently to each dataset and provides the eddy catalogues, including eddy position, amplitude, radius, polarity, and trajectories. However, evaluating the skill of an ocean reanalysis requires an additional step beyond detecting and tracking eddies independently in each dataset. In particular, it is necessary to determine whether an eddy detected in a reanalysis corresponds to the same eddy observed in satellite altimetry at a given time and location (a feature-matching and verification problem). To fill the gap, Smith and Fortin (2022) developed a verification method to evaluate the capability of ocean reanalyses in representing the eddy features at exact locations and time detected in satellite altimetry data. In this study we will adopt this method to evaluate the representation of eddies in two ocean reanalyses, thereby opening the door to a greater exploitation of ocean reanalyses.”

-) Line 59: “For this reason [...]” The Authors suggest that similar problems affect also the product at 1/8 at lines 100-109. Is there a scientific/technical reference supporting the use of the product at 1/8 rather than 1/4?

-) Line 61: “The AVISO DUACS 1/8 satellite [...]” I do not understand the comment. The AVISO at 1/4 is also available since 1993, moreover the manuscript works on a limited period.

Response: We have revised the manuscript to clarify the rationale for using the AVISO DUACS 1/8° product and to avoid implying that its finer nominal grid spacing necessarily corresponds to a proportionally finer effective resolution. We agree that the 1/4° AVISO DUACS product is also available over the satellite-altimetry era and that the temporal availability of the 1/8° product is not the reason for its selection. The 1/8° product was selected because its finer nominal grid spacing reduces the grid-spacing mismatch with GLORYS12V1 and is consistent with the 1/8° grid of the SWOT MIOST Science product, allowing the same verification framework to be applied to both satellite references. We have revised the manuscript, now at lines: 60 – 76.

“A reference dataset is needed to validate and evaluate the performance of ocean numerical models in representing mesoscale eddies, also as done in previous feature-based eddy verification studies (Smith and Fortin, 2022). The finer spatial resolution of ocean reanalyses creates a problem when using the AVISO DUACS ¼° dataset as a reference, because this dataset cannot capture many small-scale features. Amores et al. (2018) show that signals smaller than about 50 km can be aliased into longer wavelengths in satellite altimetry. Ballarotta et al. (2019) also showed that the effective resolution of AVISO DUACS gridded altimetry maps is lower than their nominal grid spacing. This means that some of the small-scale variability found in high-resolution reanalyses may not be shown correctly in the AVISO DUACS product. For this reason, we did not use the ¼° AVISO DUACS product in this study. Instead, we used the AVISO DUACS dataset at 1/8°, which provides a finer nominal grid spacing and is consistent with the 1/8° grid of the SWOT MIOST product. Its use reduces the nominal grid-spacing mismatch with GLORYS12V1, but it does not overcome the effective-resolution limits of mapped altimetry, which depends on satellite sampling and the mapping procedure rather than on grid spacing alone. Small eddies close to the resolving capability of AVISO DUACS 1/8° are therefore interpreted with caution (Amores et al., 2018; Ballarotta et al., 2019).

The AVISO DUACS 1/8° satellite altimetry dataset, available from 1993 to the present at the global scale, has been widely used as a benchmark to validate and assess numerical models, ocean reanalyses, mesoscale eddy properties in many studies (Taburet et al., 2019; Mason et al., 2019; Xie et al., 2020). In this study, however, we use AVISO DUACS 1/8° only over the period overlapping with SWOT MIOST Science, from 1 August 2023 to 1 May 2025, in order to apply the same verification framework to both satellite reference products.”

Line 68: “Second, the primary purpose of SWOT [...]” I find hard to believe that SWOT was design and put in orbit for the data assimilation. Please add a citation or rephrase the sentence.

-) Line 84: “This comparison also provides [...] pre SWOT”. The uncertainties, especially at the beginning of the altimetry era, depends on the heavy usage of interpolation methodologies and the error associated to that satellite retrivials.

Response: We have clarified that SWOT was designed to observe ocean surface topography and mesoscale–submesoscale circulation, while data assimilation is an important application rather than its primary purpose. Relevant references have been added. We also clarified that the longer pre-SWOT record is affected by satellite sampling, retrieval, mapping, and interpolation uncertainties, particularly during the early altimetry era. Finally, we specify that our comparison is limited to the common 2023–2025 period. We have revised the manuscript, now at lines: 77 – 118.

“The availability of observations from the SWOT (Surface Water and Ocean Topography) wide-swath satellite altimetry mission provides another possible reference dataset for evaluating the capability of higher-resolution ocean reanalyses to identify and to track mesoscale eddies over the time in a regional area. SWOT was launched in December 2022 and offers wide-swath measurements of sea surface height, which allows observations of ocean features at much finer spatial scales than traditional altimetry. One of the objectives of SWOT is the observation of two-dimensional ocean surface

topography and the characterization of mesoscale and submesoscale ocean circulation. Nevertheless, one important application of SWOT observations is their use in ocean analysis and forecasting systems through data assimilation, since wide-swath SSH observations can help better constrain mesoscale features and smaller-scale variability (Morrow et al., 2019; Fouchet et al., 2025). In the present study, however, we do not use the native SWOT swath observations directly. We use the experimental SWOT MIOST Science product, a global daily Level-4 gridded reconstruction of sea-surface height and derived geostrophic velocities produced by SSALTO/DUACS. MIOST stands for Multiscale Interpolation Ocean Science Topography. It merges irregularly sampled altimeter observations into fields that are continuous in space and time using a linear optimal-interpolation framework in which different covariance functions can be used to represent variability at different spatial and temporal scales. In the configuration considered here, the covariance functions are designed primarily to represent mesoscale geostrophic variability (Ubelmann et al., 2021; CLS and CNES, 2025). SWOT Level-4 sea-level products are not limited to the MIOST reconstruction used in this study. AVISO/DUACS also distributes experimental regional products generated using 4DVarNET and 4DVarQG. The 4DVarNET approach combines a variational data-assimilation formulation with a deep-learning framework to learn both the representation of ocean dynamics and the associated solver (Fablet et al., 2021; Beauchamp et al., 2023). In contrast, 4DVarQG uses a four-dimensional variational scheme constrained by a reduced-order quasi-geostrophic model, introducing an explicit dynamic prior into the reconstruction (Le Guillou et al., 2021). Comparisons over the North Atlantic indicate that 4DVarNET and 4DVarQG are capable of providing finer effective resolution and improving the reconstruction of the energetic and nonlinear structures of the Gulf Stream compared to MIOST, although their performance may vary spatially and neither method currently provides a global mapping solution (Ballarotta et al., 2025). MIOST is selected because it provides a spatially consistent global reconstruction (such as AVISO DUACS $1/8^\circ$) covering the intended validation domain, the North Atlantic, up to the upper mid-latitudes (60°N). The 4DVarNET and 4DVarQG experimental maps, as of the study's final implementation (October and November 2025), were currently regional products covering approximately $25\text{--}50^\circ\text{N}$ and $80^\circ\text{W}\text{--}10^\circ\text{W}$ and therefore do not include the northern part of our study region. However, there are some important limitations that must also be considered. First, because the mission is very recent, the time coverage of SWOT data is limited. Therefore, SWOT data cannot yet be used to study decadal or long-term variability of mesoscale activity. In this study, SWOT is used only for a short-period evaluation of the representation of mesoscale and smaller-scale eddy features in high-resolution reanalyses. Second, since similar nadir altimetry observations are assimilated in some ocean reanalysis products considered in this study, neither AVISO DUACS $1/8^\circ$ nor SWOT MIOST Science can be regarded as a fully independent verification dataset. DUACS $1/8^\circ$ is based on multi-mission nadir altimetry, while SWOT MIOST Science combines SWOT wide-swath observations with a multi-mission nadir-altimetry background used to constrain the larger-scale signal. Therefore, both satellite products share, at least partly, observational information with the reanalyses. However, SWOT MIOST Science also introduces additional fine-scale information from SWOT wide-swath measurements, which is not directly assimilated in the reanalysis products considered here. For these reasons, AVISO DUACS $1/8^\circ$ and SWOT MIOST Science are used as complementary evaluation references rather than as absolute observational truth. The comparison with SWOT provides insight into how well the reanalyses represent fine-scale variability revealed by wide-swath altimetry, whereas the DUACS comparison provides continuity with the longer pre-SWOT altimetry record. This complementary approach allows us to assess the representation of mesoscale eddies in the reanalyses while explicitly acknowledging the limitations associated with the partial dependence of the reference datasets.”

: -) Line 135: “ Finally there are [...]”. There are different SWOT products over North Atlantic that are not based MIOST (see mayor point).

Response: we have revised the manuscript, now at lines: 128 – 155.

“2.1 Satellite Altimetry Products

The AVISO DUACS global 1/8° daily gridded product provides sea level anomaly (SLA), absolute dynamic topography (ADT), and geostrophic velocities derived from multi-mission satellite altimetry. The product is generated by the CNES/CLS DUACS system and distributed by the Copernicus Marine Service (Taburet et al., 2019; Ballarotta et al., 2019). Although its nominal grid spacing is 1/8° (approximately 12.5 km), its effective resolution is lower because the optimal-interpolation procedure smooths part of the small-scale variability. Its effective resolving capability is typically between 50 and 100 km, depending on latitude and the observing constellation, which limits the representation of the smallest mesoscale and submesoscale features (Chelton et al., 2011b; Ballarotta et al., 2019; CMEMS Sea Level Product User Manual, 2025).

To complement the standard DUACS product, we also use the experimental global 1/8° SWOT MIOST Science gridded sea-level dataset distributed by AVISO/DUACS (CLS and CNES, 2025). This product combines multi-mission nadir altimetry with SWOT observations and uses the MIOST framework to reconstruct fields that are continuous in space and time through a multiscale optimal-interpolation approach designed to represent mesoscale geostrophic variability (Ubelmann et al., 2015; Rogé et al., 2017; CLS and CNES, 2025). However, SWOT MIOST remains a daily Level-4 mapped product and should not be interpreted as a direct representation of the native KaRIn observations. Its smallest-scale variability is still affected by observational noise, filtering, interpolation, and the spatial and temporal sampling of the observing system. In addition, its relatively short temporal coverage limits its use for investigating decadal or longer-term variability.

The SWOT Level-4 products over the North Atlantic are not limited to MIOST. There are, as noted in Introduction, other experimental regional products, such as 4DVarNET and 4DVarQG, are also available and use different data-driven or dynamically constrained mapping approaches (Le Guillou et al., 2021; Beauchamp et al., 2023; Ballarotta et al., 2025). As explained in Introduction, we select MIOST because it provides a spatially consistent global reconstruction covering the complete verification domain, differently from these other products that don't include the whole study area.

The AVISO DUACS 1/8° product used in this study corresponds to the Copernicus Marine Service product SEALEVEL_GLO_PHY_L4_MY_008_047, dataset cmems_obs-sl_glo_phy-ssh_my_allsat-l4-duacs-0.125deg_P1D, DOI: 10.48670/moi-00148. This product was accessed and downloaded on 31 October 2025.

The SWOT MIOST Science product used in this study corresponds to the AVISO/DUACS Experimental Multimission Gridded Level-4 Sea Surface Heights and Velocities product including SWOT observations, generated using the MIOST mapping method, version 2.0.1, DOI: 10.24400/527896/A01-2025.001. This product was accessed and downloaded from the v_2_0_1/miost/science directory on 1 November 2025.”

-) Line 186: “A closed contour [...]” Equation 1 is also used at detection level, which are the values of the parameters in that case?

Response: To avoid adding excessive technical detail to the manuscript (because the main goal is the eddy matching and not eddy detection), we preferred not to include a separate complete list of all detection inputs and settings. For clarity, the detection uses SSH, u_0 , v_0 (reanalyses), and ADT, u_{gos} , v_{gos} (satellite products). The main settings are a 500 km Bessel-

filter cutoff frequency, a 0.2 cm contour interval, a minimum algorithmic amplitude of 0.4 cm, a maximum shape error of 55%, and a pixel count between 5 and 2000. The Bessel-filter cutoff scale is reported in Section 2.3, while the matching-distance sensitivity and other cutoff frequencies of Bessel filter are discussed in Appendix D. We revised the manuscript, at lines 219 – 237.

“ A closed contour is “accepted” as an eddy, in other words, it is considered a “detected” eddy for py-eddy-tracker routine, if it satisfies the following criteria:

- *It must contain a single maximum (for Anticyclonic) or minimum (for Cyclonic).*
- *The shape must be roughly circular, with a contour area not deviating more than 55% from a perfect circle.*
- *Pixel count 5 pixels = $I_{min} \leq I \leq I_{max} = 2000$ pixels: have between 5 and 2000 pixels.*
- *The amplitude, computed as the difference between the contour extremum and the outermost valid contour, must be at least twice the contour interval. With the 0.2 cm contour interval used here, this gives a minimum algorithmic threshold of 0.4 cm. This low value is used only as a permissive detection criterion and not as a physical threshold for eddies; weak features are assessed later through lifetime filtering and amplitude-dependent verification statistics (see Sect. 3.2 and Appendix C for more details).*

After the detection of eddies for each date, eddies are tracked in time using cost function C (Eq. 1), which links eddies on consecutive days based on similarity in amplitude, radius, and centroid position. For a particular eddy, all eddies within a radius of 125 km are identified. The cost function, adopted for each pair of eddies, is based on the difference in amplitude A, in effective radius R and distance D, and it is defined as:

$$C = \sqrt{\frac{(A_1 - A_2)^2}{A_1^2} + \frac{(R_1 - R_2)^2}{R_1^2} + \frac{(D)^2}{(125 \text{ km})^2}} \quad (1)$$

where the subscripts 1 and 2 denote, respectively, the satellite eddy and the candidate reanalysis eddy in the matching procedure. In the tracking procedure, they denote a detected eddy at time t and a candidate continuation at time t+1, respectively (see Appendix D for more details about the use of a different distance threshold of the matching cost function).
”

-) Fig. 2: “At large scales, the four [...]” This comment should go in the text rather than in the caption.

Response: We have moved the interpretation of the spectra from the figure caption to the main text. The revised caption now contains only the methodological information needed to describe the figure, while the discussion of the spectral differences is provided at lines 276 – 305.

“In analogy to the approach adopted in Droghei et al. (2018) and Ciani et al. (2019), we selected the North-West Atlantic region and computed one-dimensional wavenumber spectra using a fast Fourier transform (FFT) to quantify the spatial scales effectively represented by each dataset. To avoid the latitude-dependent distortion associated with zonal wavenumbers expressed in degrees, all SSH fields were first projected onto a Lambert Azimuthal Equal-Area coordinate system centred on the study domain. The fields were then interpolated onto a regular kilometre-scale grid with $dx = dy = 12.5$ km. The one-dimensional zonal and meridional spectra were computed on this projected grid using FFTs. The resulting wavenumbers are therefore expressed in cycles / km, with the corresponding wavelengths shown in kilometres.

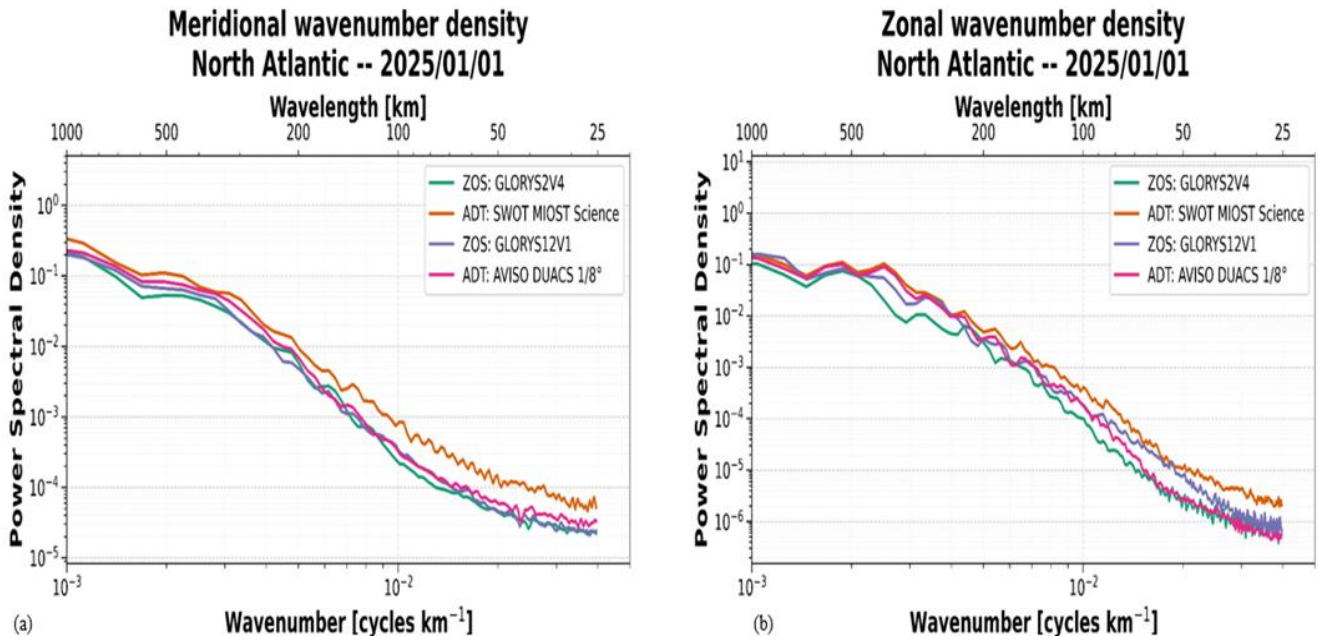


Fig.2. Power spectral density of SSH, corresponding to ADT for the satellite altimetry products, for the four datasets on 1 January 2025 over the selected North Atlantic region. All fields were projected onto a Lambert Azimuthal Equal-Area coordinate system and interpolated onto a regular kilometre-scale grid with $dx = dy = 12.5$ km, before computing the spectra. Panel (a) shows the meridional spectrum and panel (b) the zonal spectrum. Wavenumbers are expressed in cycles / km, with the corresponding wavelengths shown in kilometres on the upper x-axis.

Fig. 2 reports the comparative spectral analysis of the four SSH fields along the zonal and meridional directions on the projected grid. At large spatial scales, approximately larger than 200 km, all spectra exhibit comparable energy levels, indicating a consistent representation of the dominant mesoscale structures. At shorter wavelengths, differences become more evident. In the meridional direction, SWOT systematically exhibits the highest spectral energy over most of the resolved range, with a clearer separation from the other products below about 100 km. DUACS and GLORYS12V1 show intermediate and broadly similar energy levels, whereas GLORYS2V4 generally displays the weakest small-scale energy. In the zonal direction, the four spectra are also similar at large scales, but a progressive separation emerges toward smaller wavelengths, with SWOT again retaining the highest spectral energy. Over much of the mesoscale range, GLORYS12V1 and DUACS remain relatively close, while GLORYS2V4 decreases more rapidly. At wavelengths shorter than about 50 km, all spectra show a marked decay, but SWOT continues to preserve the largest amount of variance, and GLORYS12V1 generally retains more small-scale zonal energy than GLORYS2V4. Overall, these spectra indicate that the SWOT product better preserves fine-scale variability, while the higher-resolution GLORYS12V1 reanalysis captures small-scale mesoscale energy more effectively than GLORYS2V4.”

-) Fig. 3 The description of dashed lines is missing.

Response: we have deleted Fig.3 and we have put the eddy detection on SSH filtered fields in Fig.1. At this point, we have included the eddies detected in all the datasets with SSH fields, we deleted Figure 3, which shows the detected eddies as well. Furthermore, Figure 1 also shows the path of the Gulf Stream (as the first reviewer suggested to show). We have revised the manuscript, now at lines: 248 - 275.

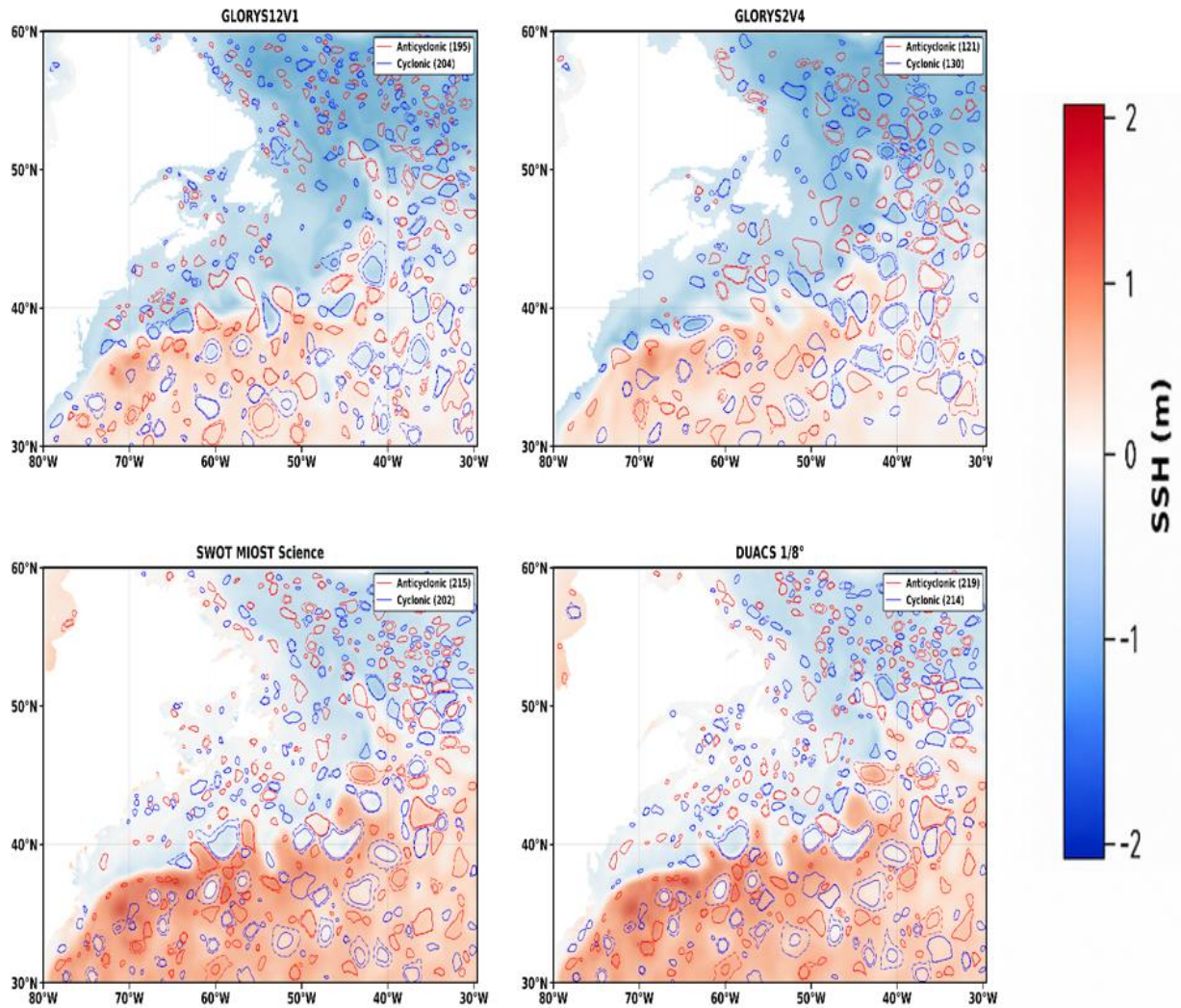


Fig.1. Daily sea-level fields on 1 January 2025 for GLORYS12V1, GLORYS2V4, SWOT MIOST Science, and AVISO DUACS 1/8°, after application of a high-pass Bessel filter with a cutoff wavelength of 500 km. The common colour bar is labelled SSH (m), with SSH used here as a general label for the sea-surface height field. Specifically, the colour shading represents modelled SSH for GLORYS12V1 and GLORYS2V4 and satellite-derived Absolute Dynamic Topography (ADT) for SWOT MIOST Science and AVISO DUACS 1/8°. Although SSH and ADT are not formally identical quantities, they are dynamically comparable and are treated consistently for contour-based eddy detection. Detected anticyclonic and cyclonic eddies are overlaid in red and blue, respectively. Solid contours indicate the contour of maximum circum-averaged geostrophic speed, whereas dashed contours indicate the effective contour used to calculate eddy area and effective radius. All fields were remapped onto the same regular 1/8° grid to allow direct comparison.

This consistent methodology for all datasets ensures that the detected eddies are comparable and that any differences in eddy representation between reanalyses or between them and observations are due to the intrinsic quality of the reanalysis systems and not to processing artifacts (Droghei et al., 2018; Ciani et al., 2019). First, an analysis of the power spectral analysis for the four products considered in this study, using 1 January 2025 as a representative test case. All datasets are remapped onto the same regular 1/8° grid, which is the native grid of both SWOT and DUACS. Figure 1 shows the SSH fields used in this study from GLORYS12V1, GLORYS2V4, SWOT MIOST Science and AVISO DUACS 1/8°. For the two reanalysis products, GLORYS12V1 and GLORYS2V4, SSH represents the modelled sea-surface height above the model geoid or reference equipotential surface, as described in Lellouche et al. (2021). For the satellite altimetry products used in this study, the corresponding SSH field is represented by Absolute Dynamic Topography (ADT), which is the sea-surface height above the geoid and is obtained by adding the Mean Dynamic Topography (MDT) to the Sea Level Anomaly

(SLA) (i.e. $ADT = SLA + MDT$, as explained in Mulet et al. (2021)). Although the reanalysis SSH and satellite ADT are not identical data products, since the former is produced by ocean models constrained by data assimilation and the latter is derived from satellite altimetry and an estimated MDT, they can be considered dynamically comparable, because they provide dynamically comparable representations of sea-surface dynamic topography, from which surface geostrophic currents are derived. As described in Mason et al. (2014), Mason et al. (2019) and Pegliasco et al. (2021), contour-based eddy detection and tracking methods, such as those implemented in py-eddy-tracker, can be consistently applied to both fields because they mainly rely on local extrema, closed contours, spatial gradients, and centre-to-edge height differences. The superimposition of the detected eddies on the SSH/ADT background also allows their sizes and positions to be evaluated in relation to the Gulf Stream pathway.”

-) Fig. 5 “Most eddies have an amplitude of less than 10 cm. However, their mean lifetime is lower than 50 [...]” This comment should go in the text rather than in the caption.

-) Line 252: “ The minimum radius and amplitude are 25km and 0.4cm” The threshold of 0.4cm can be used at detection level as a technical parameter (see line 194) but the actual eddy population should not consider eddies with an amplitude that is much smaller than the error on SLA (1 or 2 cm).

Response: We moved the discussion of eddy amplitude and lifetime from the caption of the former Fig. 5, now Fig. 4, to the main text. We also clarified that 0.4 cm is only a permissive technical threshold used for detection and not a physical threshold defining a robust eddy population. Features with amplitudes of only a few millimetres are therefore interpreted cautiously. We have also added here, for completeness, the Fig.3 (former Fig.4). The manuscript has been revised at lines 306 – 329.

“In Fig.3, we see the dependence between radius and amplitude, in which we note that eddies display similar properties in terms of effective radius and amplitude. Many small eddies are detected with a radius below 50 km and amplitude less than 5 cm. In Fig. 4, most eddies have amplitudes lower than 10 cm. However, their mean lifetime is lower than 50 days (see Appendix C for more details about mean lifetime and eddy amplitude thresholds). However, features with amplitudes of only a few millimetres, should not be interpreted as physically robust eddies. The 0.4 cm value corresponds only to the permissive algorithmic threshold used during the py-eddy-tracker detection step, and is not used here as a physical threshold to define the robust eddy population. It can be noted, generally, that larger amplitude eddies are also found to have larger radius (Fig. 3). The eddy identification and tracking procedure are independently applied for each product and produce plots about physical features and time series of anticyclonic and cyclonic eddies.

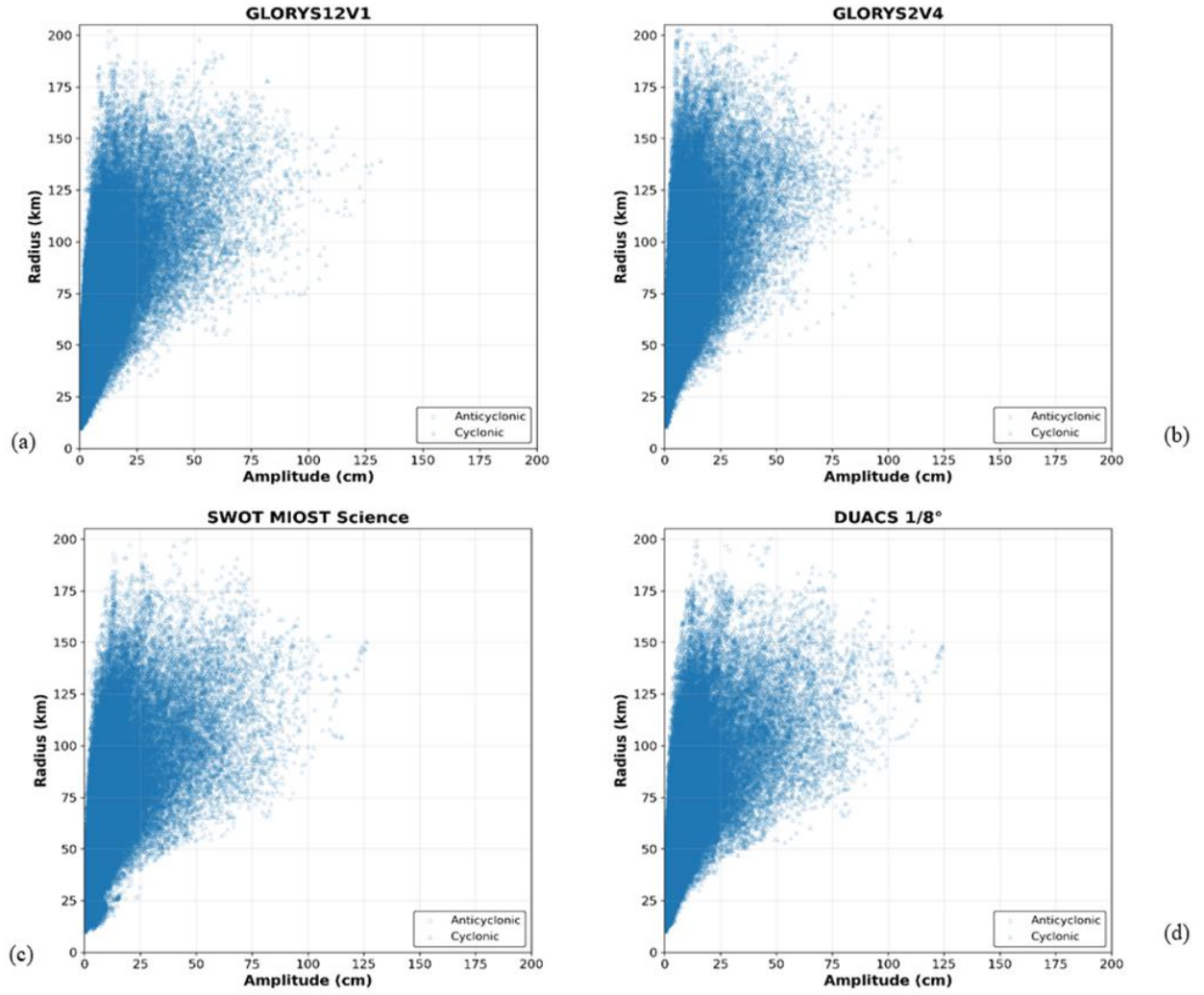


Fig. 3. Scatter plots of detected eddies, with a minimum lifetime threshold of 4 days, between Radius and Amplitude, from 2023-08-01 to 2025-05-01, for (a) GLORYS12V1 (bilinear remapped onto DUACS and SWOT grid), (b) GLORYS2V4 (bilinear remapped onto DUACS and SWOT grid), (c) SWOT MIOST Science, and (d) AVISO DUACS 1/8°. The eddy amplitude and the eddy radius are calculated at the track level, by averaging the amplitude and radius over the entire lifetime of the eddy.

The same cost function (1) used for the tracking procedure is also used to match the eddies between the reanalysis and the observed eddy (satellite). This function allows us to see how the eddies can be matched even if they have different values for radius, amplitude, mean centroid position, and shape. Although one or more of these parameters can vary, for an eddy to be considered matched to the corresponding satellite dataset, the cost function must be minimized, provided that the distance between the eddy centroids does not exceed 125 km (is set to 125 km, i.e., that's slightly greater than the average diameter of eddies in the North Atlantic region; see Ballarota et al., 2019).

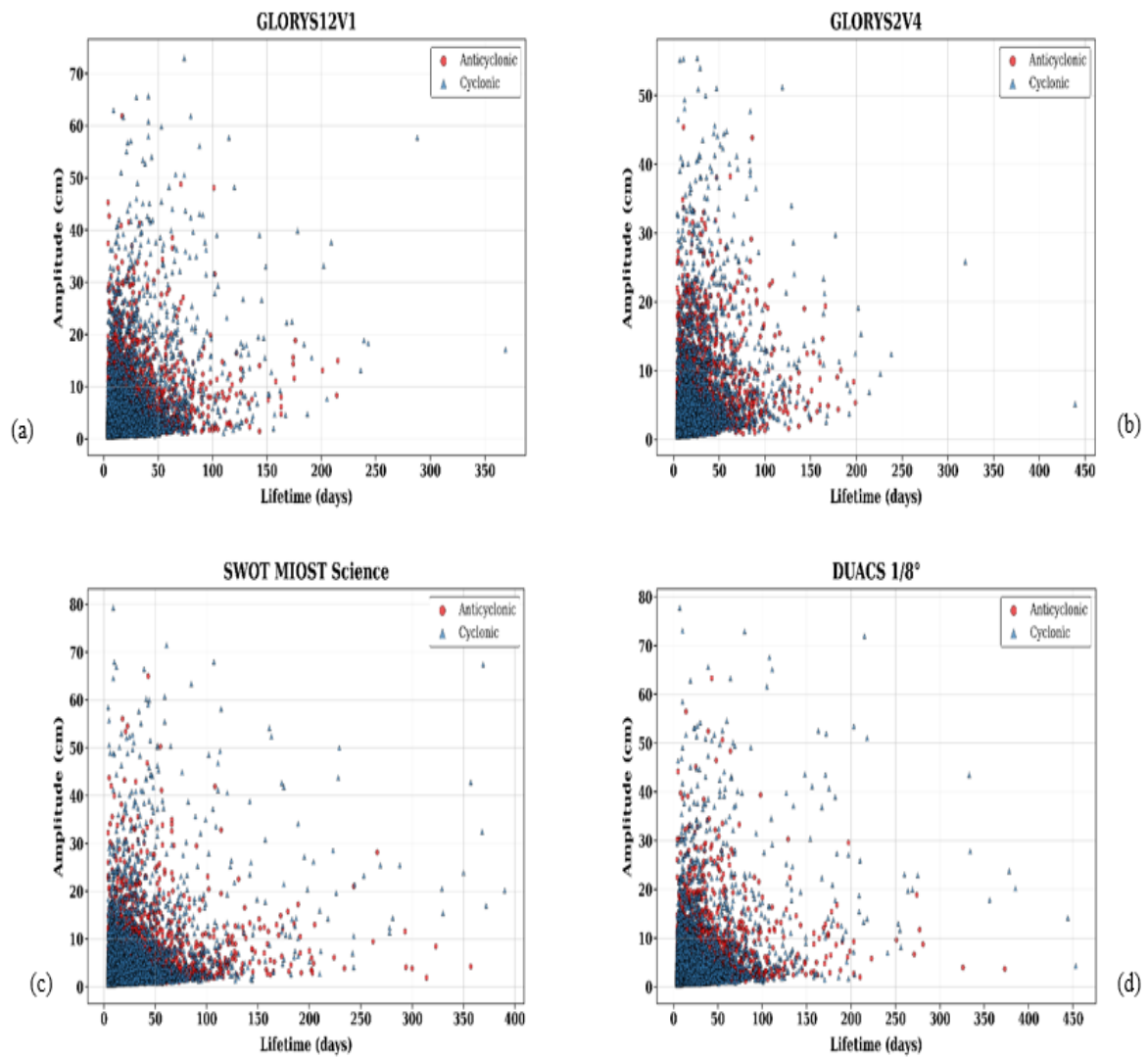


Fig.4. Scatter plots of identified eddies, with a lifetime threshold of 4 days, between Amplitude vs Lifetime of eddies for (a) GLORYS12V1, (b) GLORYS2V4, (c) SWOT, (d) DUACS. Each eddy is counted only once. The amplitude is calculated at the trace level, by averaging the amplitude over the entire lifetime of the eddy. “

-) Fig. 6 : “[...](satellite data: source for the comparison) [...]” The source is missing

Response: We revised the caption of the former Fig. 6, now Fig. 5, to explicitly identify AVISO DUACS 1/8° and SWOT MIOST Science as the satellite reference datasets used to define hits, misses, and false alarms. The manuscript has been revised at lines 346 – 353.

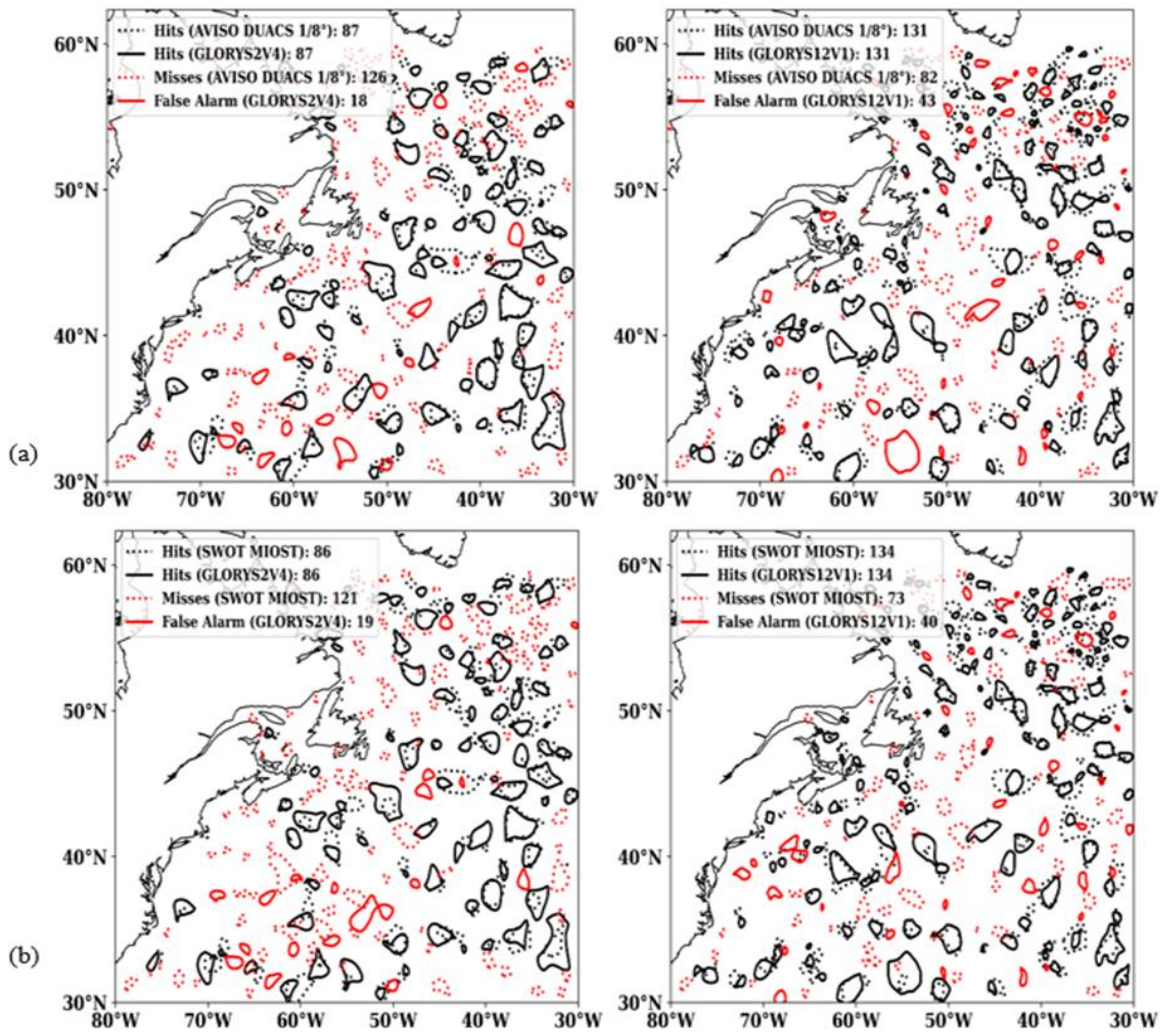


Fig. 5. The first one (a) is an example showing Anticyclonic eddy matching statistics for 1st January 2025 for GLORYS2V4 (left), and GLORYS12V1 (right), compared with DUACS. Eddies for which a corresponding match is found (hits) between DUACS and GLORYS2V4/GLORYS12V1 are shown as dotted and solid black contours. Eddies identified in DUACS (satellite data) with no corresponding eddy in GLORYS2V4/GLORYS12V1 are deemed misses (red dotted contours). Eddies identified in GLORYS2V4/GLORYS12V1 but without a corresponding eddy to DUACS are deemed false alarms. The second one (b) is an example showing Anticyclonic eddy matching statistics for 1st January 2025 for GLORYS2V4 (left), and GLORYS12V1 (right), compared with SWOT (satellite data)..

Line 304 :"))"

Typo

Response: corrected.

-) Line 305-309, "Many small eddies..." The Authors described the correlation between amplitude and lifetime for two populations: eddies with amplitude below and above 10 cm. I do not see particular differences in the results, why the description of the population is divided in two?

Response: Following both reviewers's suggestions, we agree that the 10 cm threshold does not identify two distinct eddy populations. We use "10 cm" only as a descriptive threshold. The revised text describes a continuous moderate positive relationship between eddy amplitude and lifetime. The manuscript has been revised at lines 361 – 369.

"The properties of identified eddies for the four datasets are quite similar (see Fig. 3) over the entire period. In Fig. 3, we see similar trends between radius and amplitude. Many small eddies, with a radius lower than 60 km (see Fig.3 and

Fig.4) not far from the threshold of detection, and an amplitude lower than 5 cm, have been identified. We do not clearly see two distinct populations based on an amplitude threshold of 10 cm, compared to the corresponding average lifetime. This is merely a descriptive threshold, intended to show that most of the detected eddies are clustered within a 10-cm amplitude range. For more details, see Appendix C, which provides a more detailed description of the datasets based on specific eddy amplitude thresholds. In Fig. 4 we see that most have amplitudes below 10 cm, but at the same time there is a moderate positive correlation between amplitude and lifetime, whereby as amplitude increases, so does the average survival time of the eddies.”

-) Line 313-316: “Following the tabulation of hits, misses, and false alarms [...]” The sentence is not clear.

-) Table 2: There is no reference in the text to Table 2. Why the number for GLORYS2V4 and GLORYS12V1 differ passing from Table 2 and 3?

-) Line 338: “An exception is observed for GLORYS2V4[...]” This should be referred to GLORYS12, right?

-) Line 342: “In this case, we demonstrate [...]” It seems a final statement and should go after the description of Table 4 and 5.

Response: We rewrote the passage to clarify that hits, misses, and false alarms are aggregated over the common evaluation period and across both eddy polarities before computing POD and FAR. Table 2 is now cited in the text. The reanalysis counts in Tables 2 and 3 are identical because detection and tracking use the same period and configuration; only the percentages differ because DUACS and SWOT provide different reference totals. We also corrected GLORYS2V4 to GLORYS12V1 and moved the concluding statement after Tables 4 and 5. The manuscript has been revised at lines 386 – 415.

“For each comparison, hits, misses, and false alarms are tabulated over the full evaluation period (1 year and 9 months for both comparisons). These counts tabulated, meaning that they are evaluated by adding up all the counts of hits, misses, and false alarms for the entire verification period, for both cyclonic and anticyclonic eddies are used to compute the Probability of Detection (POD) and the False Alarm Ratio (FAR) (following the Eqs. (2) and (3)). For the total counts of eddies, the results are reported separately for anticyclonic and cyclonic. The results for DUACS-based comparison are shown in Fig. 6(a), Table 2 and Table 4, while in Fig. 6(b), Table 3 and Table 5 for SWOT comparison results.

Dataset	Anticyclonic	% w.r.t. DUACS	Cyclonic	% w.r.t. DUACS
DUACS	125580 (4 days)	-	125229 (4 days)	-
GLORYS2V4	67961 (4 days)	54.12%	72953 (4 days)	58.26%
GLORYS12V1	97788 (4 days)	77.87%	106861 (4 days)	85.33%

Table 2: total number of anticyclonic and cyclonic eddies identified in the region of study over the evaluation period for DUACS, GLORYS2V4, and GLORYS12V1, for a minimum lifetime threshold of 4 days. The percentage of eddies identified in GLORYS2V4 and GLORYS12V1 with respect to AVISO is also provided.

Dataset	Anticyclonic	% w.r.t. SWOT	Cyclonic	% w.r.t. SWOT
SWOT	124154 (4 days)	-	125676 (4 days)	-
GLORYS2V4	67961 (4 days)	54.74%	72953 (4 days)	58.05%
GLORYS12V1	97788 (4 days)	78.76%	106861 (4 days)	85.03%

Table 3: total number of anticyclonic and cyclonic eddies identified in the region of study over the evaluation period for SWOT MIOST, GLORYS2V4, GLORYS12V1, for a minimum lifetime threshold of 4 days. The percentage of eddies identified in GLORYS2V4 and GLORYS12V1 with respect to SWOT MIOST is also provided.

Verification metric	GLORYS12V1	GLORYS2V4	Diff.	Diff. %
POD	0.628	0.460	0.168	36.52%
FAR	0.230	0.182	0.048	26.37%

Table 4: aggregated eddy verification statistics for the Probability of Detection (POD) and False Alarm Ratio (FAR) evaluated over the period, from 1st August 2023 to 1st May 2025, for the comparison with AVISO DUACS 1/8°, with a minimum lifetime threshold of 4 days.

Verification metric	GLORYS12V1	GLORYS2V4	Diff.	Diff. %
POD	0.624	0.455	0.169	37.14%
FAR	0.237	0.194	0.043	22.16%

Table 5: aggregated eddy verification statistics for the Probability of Detection (POD) and False Alarm Ratio (FAR) evaluated over the period 1st August 2023 to 1st May 2025, for the comparison with SWOT MIOST, with a 4-day minimum lifetime threshold.

There is certainly a slight decrease, when comparing the total number of anticyclonic eddies to that of cyclonic eddies, which is particularly evident for GLORYS12V1. Based on the observed data, the two satellite products show nearly balanced polarity counts, with differences not exceeding 1.3 %. In contrast, cyclonic daily detections exceed anticyclonic detections by approximately 7.3 % in GLORYS2V4 and 9.3 % in GLORYS12V1. As shown in the eddy matching example in Fig. 5, there is a significant difference in terms of hits, misses, and false alarms between the two reanalyses. While the number of hits increases by about 37% when switching from GLORYS2V4 to GLORYS12V1, the number of misses decreases significantly, by about 31%, whereas the number of false alarms increases dramatically, around 77%. This demonstrates how GLORYS12V1 shows greater agreement with both observational datasets, but at the same time generates a significantly higher number of false alarms, which perfectly explains the increase in the FAR, as shown in Fig. 6. Furthermore, the daily POD and FAR fluctuate but remain stationary over time, with no apparent seasonal cycle as shown in Fig. 6(a) and Fig. 6(b). As noted above, the calculation is also applied to the minimum threshold of 7 and 28 days without any observational gaps for each eddy within this time interval, in Appendix A. This sensitivity test shows that short-lived eddies contribute to part of the variability in the verification metrics. “

-) Line 345-356: “GLORYS12V1 achieves a POD exceeding 50% (higher than 60% for SWOT MIOST comparison) [...]” The description jumps from eddies longer than 4 days to 7 days, from one reanalysis to another, percentages have sometimes two decimals sometimes zero. Table 3 is described after Table 5.

Response: We reorganized the discussion to present the 4-day results consistently and in the same order as Tables 2–5, using uniform numerical precision. The detailed results for the 7- and 28-day thresholds are now discussed separately in Appendix A. The manuscript has been revised at lines 416 – 441.

“On average values, GLORYS12V1 achieves a POD exceeding 60% and a FAR lower than 24%. For the comparison with GLORYS2V4 FAR is lower than 20%. GLORYS12V1 shows an improvement over GLORYS2V4, that’s greater than 36% for POD, but with high values for FAR increasing, that is 26.37% for DUACS comparison. The same verification approach was applied to the SWOT, yielding a slightly different balance of hits, misses, and false alarms. Mean eddy lifetimes are slightly lower for SWOT than for DUACS for both minimum-lifetime thresholds considered (Appendix B). GLORYS12V1 tends to align more closely with SWOT observations than GLORYS2V4, with the highest values, with a greater relative gain in POD (+37.14%) and a lower FAR increasing (22.16%) than DUACS comparison, for the minimum lifetime of 4 days.

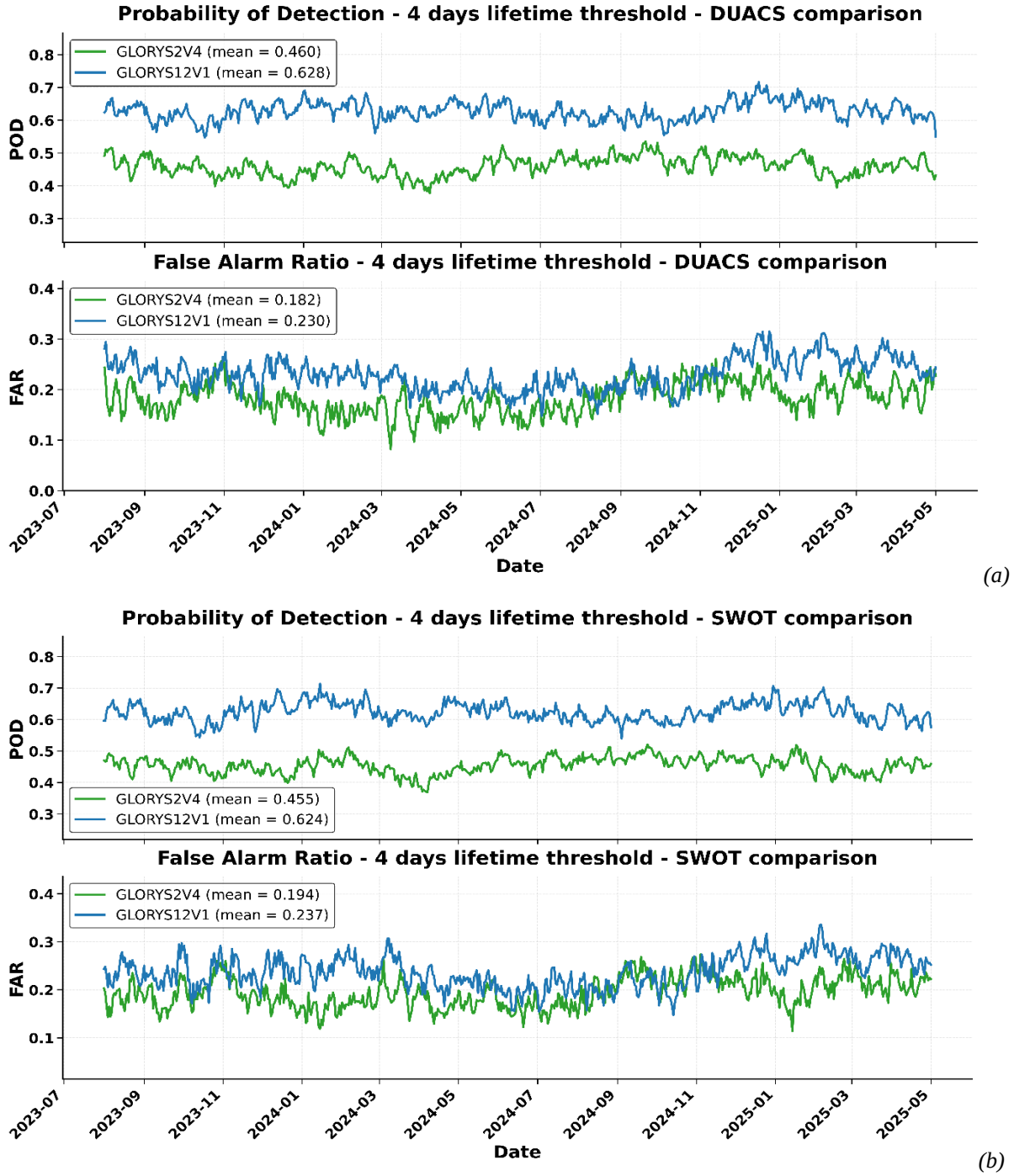


Fig.6. Timeseries of matching statistics for GLORYS2V4 and GLORYS12V1, compared with DUACS 1/8°, over the period from 1st August 2023 to 1st May 2025, for a 4-day minimum lifetime threshold (a). and in the bottom, the timeseries with SWOT MIOST Science, over the period from 1st August 2023 to 1 May 2025, for a 4-day minimum lifetime threshold (b).

This indicates that, specifically regarding GLORYS2V4, the POD values are very similar in the DUACS and SWOT comparisons. There is a slight decrease when SWOT is used as reference. The decrease is about 0.005 for the 4-day threshold. GLORYS2V4 can match part of the eddies detected by SWOT, but with slightly lower efficiency than for DUACS. The FAR also slightly increases in the SWOT comparison, by about 0.012, showing a small increase in terms of false alarms. Overall, these differences are small, suggesting that the verification results for GLORYS2V4 are relatively insensitive to the choice between the two satellite reference products. However, the overall POD values, that are around 62%, for both comparisons with GLORYS12V1, show that a significant proportion of observed eddies (around 38%) remains unmatched. It highlights the presence of limitations in their high-resolution applications, such as difficulties in

representing small or short-lived features and maintaining eddy-track continuity. As the minimum lifetime threshold increases, the POD improvement of GLORYS12V1 relative to GLORYS2V4 becomes smaller, whereas the reduction in FAR becomes more pronounced. GLORYS12V1 has a higher FAR than GLORYS2V4 for the 4-day and 7-day lifetime thresholds, likely because it detects a larger number of eddies, some of which do not find a corresponding match in the satellite datasets. At the 28-day threshold, this relationship reverses, with GLORYS12V1 showing a lower FAR than GLORYS2V4 (see Appendix A for further details)."

-) Line 365-375: "This indicates [...]" The word "This" is used multiple times in the next sentences.

-) Line 366-371 "This confirms [...]" It seems that GLORYS2V4 is used to assess the quality of the two altimetry datasets, but the mutual assessment is something to avoid.

-) Line 369: "[...] and that, therefore, it would be necessary to introduce an effective methodology capable of eliminating these ones." Too long a sentence please rephrase or remove.

-) Line 373-374: "[...] such a scale lower than mesoscale [...]" I do not understand why the trajectory prediction is affected, can the Authors clarify the reason?

Response: We revised the paragraph to avoid repeated use of "This", shortened the unclear sentence, and removed any interpretation implying a mutual assessment of the satellite products. The results are now described in terms of reanalysis–satellite consistency. We also clarified that unmatched eddies may reflect difficulties in representing small or short-lived features and maintaining track continuity. We have revised the manuscript, combining these comments, now at lines: 429 – 441.

"This indicates that, specifically regarding GLORYS2V4, the POD values are very similar in the DUACS and SWOT comparisons. There is a slight decrease when SWOT is used as reference. The decrease is about 0.005 for the 4-day threshold. GLORYS2V4 can match part of the eddies detected by SWOT, but with slightly lower efficiency than for DUACS. The FAR also slightly increases in the SWOT comparison, by about 0.012, showing a small increase in terms of false alarms. Overall, these differences are small, suggesting that the verification results for GLORYS2V4 are relatively insensitive to the choice between the two satellite reference products. However, the overall POD values, that are around 62%, for both comparisons with GLORYS12V1, show that a significant proportion of observed eddies (around 38%) remains unmatched. It highlights the presence of limitations in their high-resolution applications, such as difficulties in representing small or short-lived features and maintaining eddy-track continuity. As the minimum lifetime threshold increases, the POD improvement of GLORYS12V1 relative to GLORYS2V4 becomes smaller, whereas the reduction in FAR becomes more pronounced. GLORYS12V1 has a higher FAR than GLORYS2V4 for the 4-day and 7-day lifetime thresholds, likely because it detects a larger number of eddies, some of which do not find a corresponding match in the satellite datasets. At the 28-day threshold, this relationship reverses, with GLORYS12V1 showing a lower FAR than GLORYS2V4 (see Appendix A for further details)."

-) Table 7: "days days" Typo

-) Line 401-405: A final [...] 7 days. If I understood well, results from 20-day block show similar values in term of POD/FAR w.r.t. 4 day and 7-day block, but at same time they are different from POD in table 4 (table 3 does not show POD). Since the Authors are discussing the 20-day block for "sake of completeness" than such values should be added. Otherwise rephrase the paragraph.

Response: We also added Appendix B, which reports the block-bootstrap results using 7-day, 20-day, and 30-day blocks and block lengths based on the mean lifetimes of the satellite-observed eddies. The main text now summarizes these sensitivity tests and refers to Appendix B for the complete values and confidence intervals. We have revised the manuscript, now at lines: 442 – 460.

“To further assess the robustness of these findings, a block bootstrap resampling (95% confidence intervals) was performed using 4-day blocks, consistent with the imposed minimum lifetime threshold. Results for both SWOT and DUACS (Tables 6 and 7) confirm that GLORYS12V1 systematically outperforms GLORYS2V4 in terms of POD (0.618 - 0.620 vs 0.448 - 0.452), with a statistically significant difference (~ 0.17) that is insensitive to block length. This gain is consistently associated with a moderate increase in FAR ($\sim 0.045 - 0.049$), confirming a robust trade-off between detection sensitivity and reliability. The agreement between the two observational references strengthens confidence in the statistical significance of the results, although it does not fully remove uncertainties related to the definition and detectability of small-scale or short-lived eddy structures.

Verification Metric	GLORYS12V1	GLORYS2V4	<i>Difference (12V1 – 2V4)</i>
POD (Mean)	0.618	0.448	0.170
POD (95% CI)	[0.609, 0.627]	[0.441, 0.454]	[0.160, 0.182]
FAR (Mean)	0.235	0.190	0.045
FAR (95% CI)	[0.229, 0.242]	[0.183, 0.198]	[0.038, 0.052]

Table 6. Bootstrap resampling for SWOT MIOST Science comparison, for a minimum lifetime threshold of 4 days, and a block length of 4 days.

Verification Metric	GLORYS12V1	GLORYS2V4	<i>Difference (12V1 – 2V4)</i>
POD (Mean)	0.620	0.452	0.168
POD (95% CI)	[0.610, 0.629]	[0.442, 0.463]	[0.154, 0.180]
FAR (Mean)	0.235	0.187	0.049
FAR (95% CI)	[0.230, 0.237]	[0.181, 0.192]	[0.040, 0.054]

Table 7. Bootstrap resampling for DUACS 1/8° comparison, for a minimum lifetime threshold of 4 days, and a block length of 4 days.

As additional sensitivity tests of the coupled block bootstrap, we also performed the bootstrap resample using 7-day, 20-day and 30-day blocks, and the mean lifetimes of both satellite eddies. The results are reported in Appendix B. The relative ranking of GLORYS12V1 and GLORYS2V4 and the main conclusions of the 4-day analysis remain consistent across the different block lengths, although the estimated confidence intervals vary with the choice of block length. Therefore, the 30-day test is included only as an additional robustness check. Very large block lengths may reduce the variability of the

resampling procedure and may partly weaken the main purpose of the block bootstrap, which is to account for temporal dependence in the data when the underlying statistical distribution is unknown.”

-) Line 406:” In general, it can be observed that, as the minimum lifetime threshold increases [...]” The Authors discussed long-lived eddies, but I did not understand whether it refers to 7-day or 20-day. Whether it refers to 20-day that the corresponding values should be added. If it refers to 7-day than I would avoid calling them long-lived eddies.

-) Line 407: “This is shown [...]” The connection between long-term eddies and small eddies is not clear, the discussion started from Figures 8 and 10 then they move to Figures 9 and 11 and then back to 8 and 10.

Response: We clarified that 7 and 28 days refer to minimum eddy-lifetime thresholds, whereas 20 and 30 days refer only to bootstrap block lengths, putting these ones in Appendix A and Appendix B, respectively. We also reorganized the discussion to present Figs. 7–10 sequentially and separated the lifetime-sensitivity analysis from the radius- and amplitude-dependent results. Additional amplitude analyses are reported in Appendix C.

We have revised the manuscript, now at lines: 461 - 483.

“In general, as the minimum eddy lifetime threshold increases, the FAR and POD values of each reanalysis products, for both satellite comparisons, tend to converge toward similar levels as described in Appendix A for longer-lived eddies (especially for 28-day minimum eddy lifetime threshold). Figures 7 and 9 show the dependence of the eddy matching statistics (hits, misses and false alarms) on the physical properties of the eddies, namely the effective radius and amplitude, respectively. These statistics are calculated to show how eddies are distributed across different ranges of radius and amplitude, and thus to identify the ranges with the highest clustering. For hits and misses, the radius and amplitude were taken from the satellite (the reference) product, because these eddies refer to observed features that are either detected or not detected by the reanalysis. For false alarms, the radius and amplitude are taken from the corresponding reanalysis product. Each bin of the histogram has a width of 10 km and 2 cm, respectively for eddy radius and eddy amplitude. The highest counts of misses and false alarms are associated with small eddies, specifically for those with radii below 50 km and amplitudes below 5 cm.

Figures 8 and 10 show the dependence of the verification metrics in terms of the same physical eddy features.

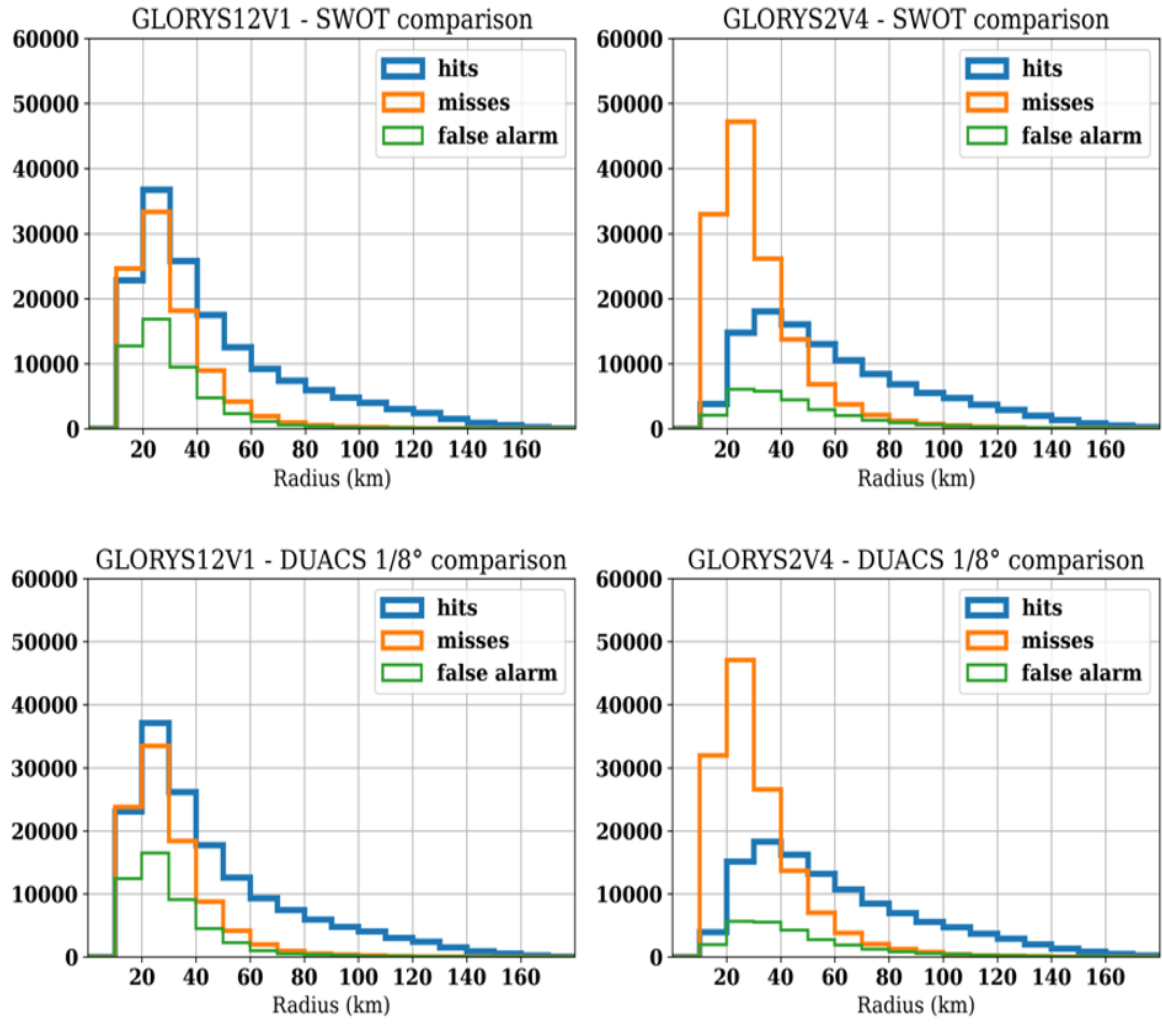


Fig. 7. On the top: eddy matching histograms for the total number of hits (blue), misses (orange), false alarms (green) for GLORYS12V1 (left) and GLORYS2V4 (right), compiled as a function of effective radius (km) over the evaluation period, for SWOT comparison, with a lifetime threshold of 4 days. In the bottom: eddy matching histograms compiled as a function of effective radius (km) over the same evaluation period, for comparison with DUACS, with a lifetime threshold of 4 days. The radius of hits and misses refers to the satellite (reference) eddies, while the radius of false alarms refers to those of unmatched eddies detected in the reanalysis.

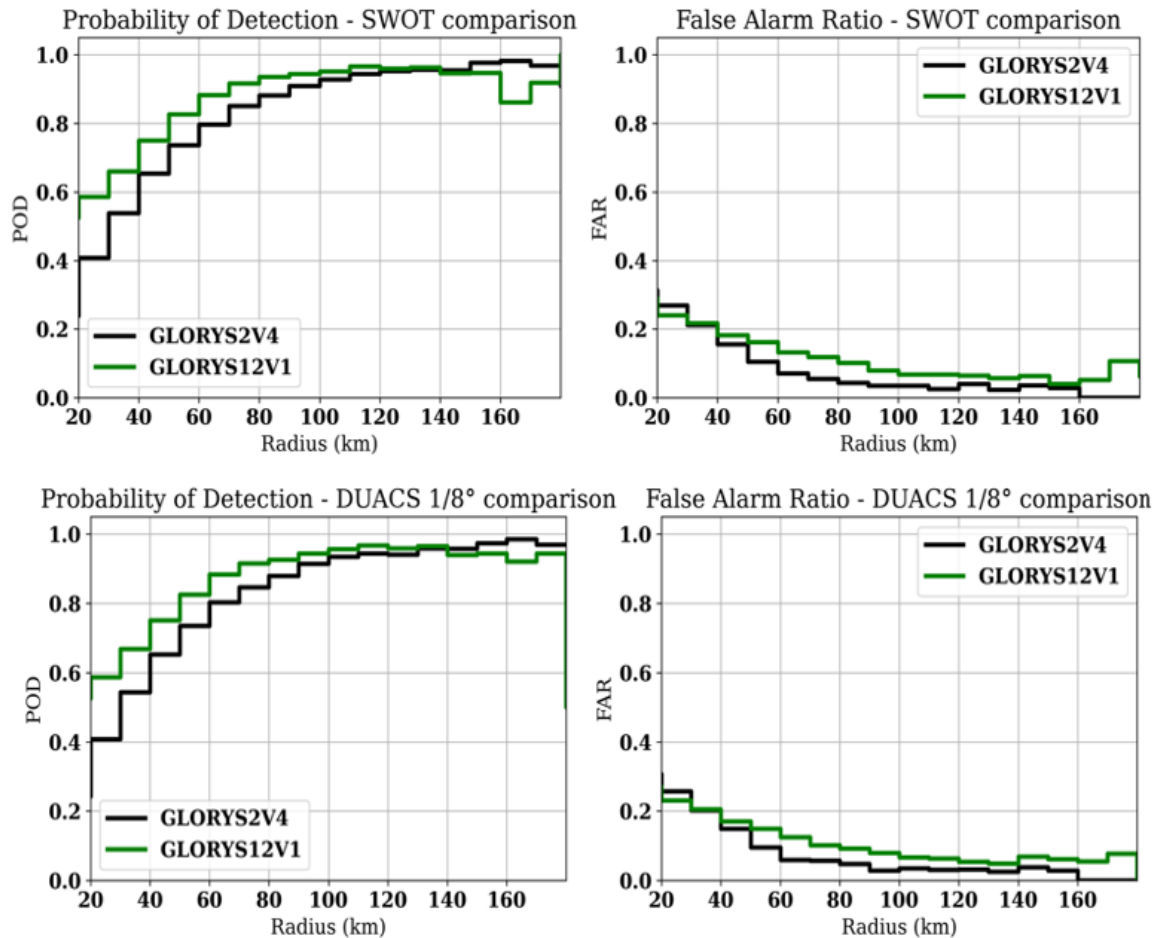


Fig.8. On the top: histograms of the probability of detection (POD; left) and the false alarm ratio (FAR; right), over the evaluation period, evaluated as a function of radius (km) for comparison with SWOT, with a lifetime threshold of 4 days. In the bottom: histograms of the POD and FAR as a function of radius (km), for the comparison with DUACS, with a lifetime threshold of 4 days. The radius of hits and misses refers to the satellite (reference) eddies, while the radius of false alarms refers to those of unmatched eddies detected in the reanalysis.”

-) Fig 9 The plots on AVISO DUACS are placed twice (instead of SWOT ones), please check

-) Line 431: “[...] they accurately represent eddies [...]” It seems to me it should be the opposite. Moreover, if the minimum threshold in radius is 25 km, what is the purpose to discuss the statistics at 30km?

-) Line 441 “This suggests [...]” The assessment of the quality of conventional and SWOT altimetry cannot be based on reanalysis skill but only by comparing the two datasets among themselves. I would drop the sentence

Response: We replaced the duplicated DUACS panels in Fig. 9 with the correct SWOT panels. We also corrected the interpretation of the smallest-radius bins. We removed any assessment of satellite-product quality and now discuss only reanalysis–satellite consistency. The manuscript has been revised at lines 484 – 513.

“ For the DUACS comparison, Figures 8 and 10 show that FAR is generally higher in the smallest effective-radius and amplitude bins, exceeding 30% in some of these classes and highlighting the greater difficulty in verifying the smallest features in this gridded product. The FAR remains close or higher than 20% until 50 km in terms of effective radius. At the same time, the POD of GLORYS12V1 is close to 60% even for small radius and amplitude values, showing a relevant fraction of small observed eddies is still detected by the reanalyses. The largest proportion of misses is consistently

associated with small eddies, specifically those with radii below 50 km and amplitudes under 4 cm. The comparison with SWOT shows a similar dependence on eddy size. The largest differences in the verification metrics are again found for small-radius and small-amplitude eddies. The POD of GLORYS12V1 remains close to 60%, while the GLORYS2V4 FAR is generally higher in the smallest bins. This indicates that the main difficulty is not only the detection of small observed eddies, but also the presence of unmatched (especially for the small eddies) produced by the reanalyses. However, for both satellite comparisons, the verification skills generally improve as eddy radius and amplitude increase. This suggests a better matching capability for larger and stronger mesoscale eddies, which are less affected by small-scale noise or detection uncertainty. The differences between the DUACS and SWOT comparison are most evident for small eddies, while they become smaller for larger eddies. For this reason, Appendix C includes additional analyses of the verification metrics using different amplitude-based subsets, including eddies with amplitudes greater than 5, 10, 15, and 20 cm, as well as eddies with amplitudes below 10 cm. Appendix C also presents the eddy matching histograms using a semi-logarithmic scale for the y-axis, which represents the number of eddies. These additional analyses allow us to assess more clearly how the verification results change when low-amplitude eddies are progressively excluded, when the analysis focuses on stronger mesoscale structures, and when the contribution of the most numerous small-amplitude eddies is examined in greater detail.

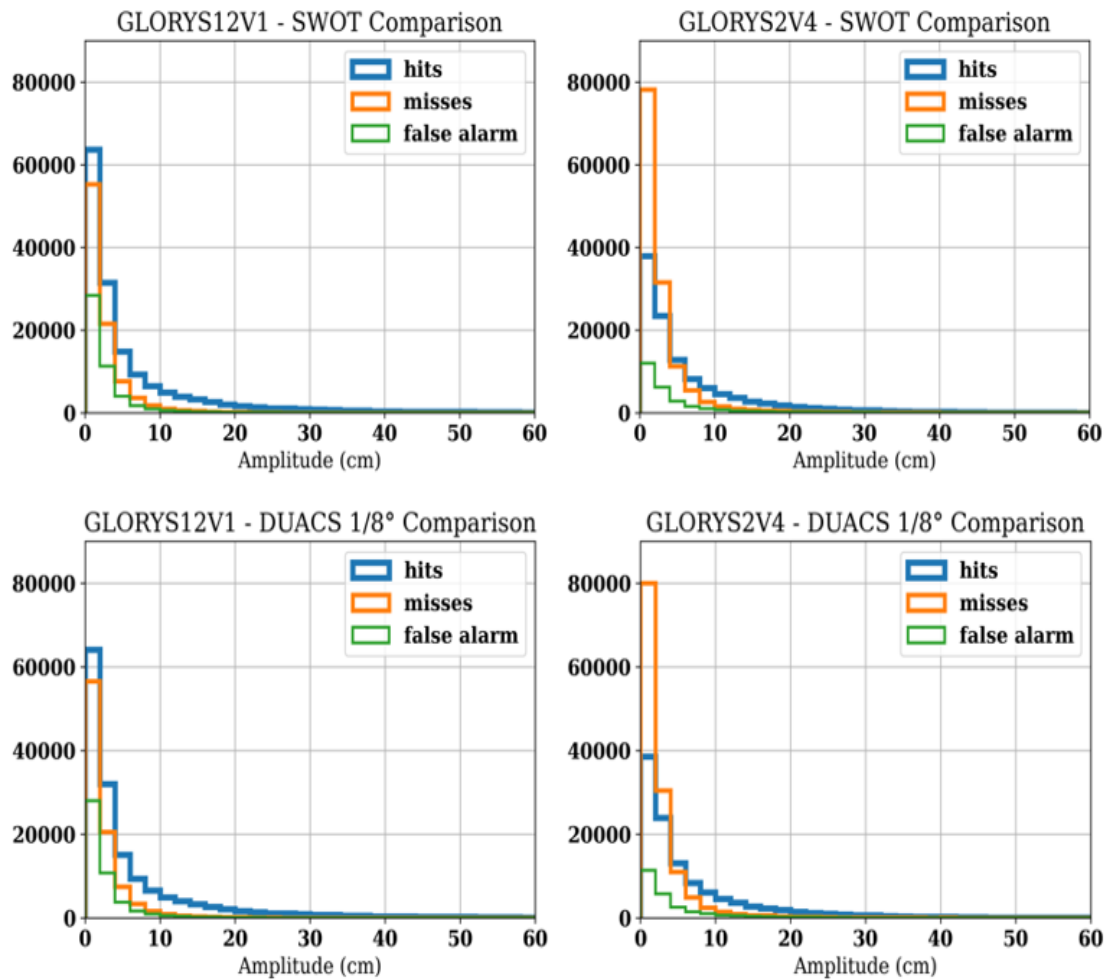


Fig.9. On the top: eddy matching histograms for the total number of hits (blue), misses (orange), false alarms (green) for GLORYS12V1 (left) and GLORYS2V4 (right), compiled as a function of amplitude (cm) over the evaluation period, for SWOT comparison. In the bottom: eddy matching histograms compiled as a function of amplitude (cm) over the evaluation period, for comparison with DUACS 1/8°. The amplitude of hits and misses refers to the satellite (reference) eddies, while the amplitude of false alarms refers to those of unmatched eddies, detected in the reanalysis.

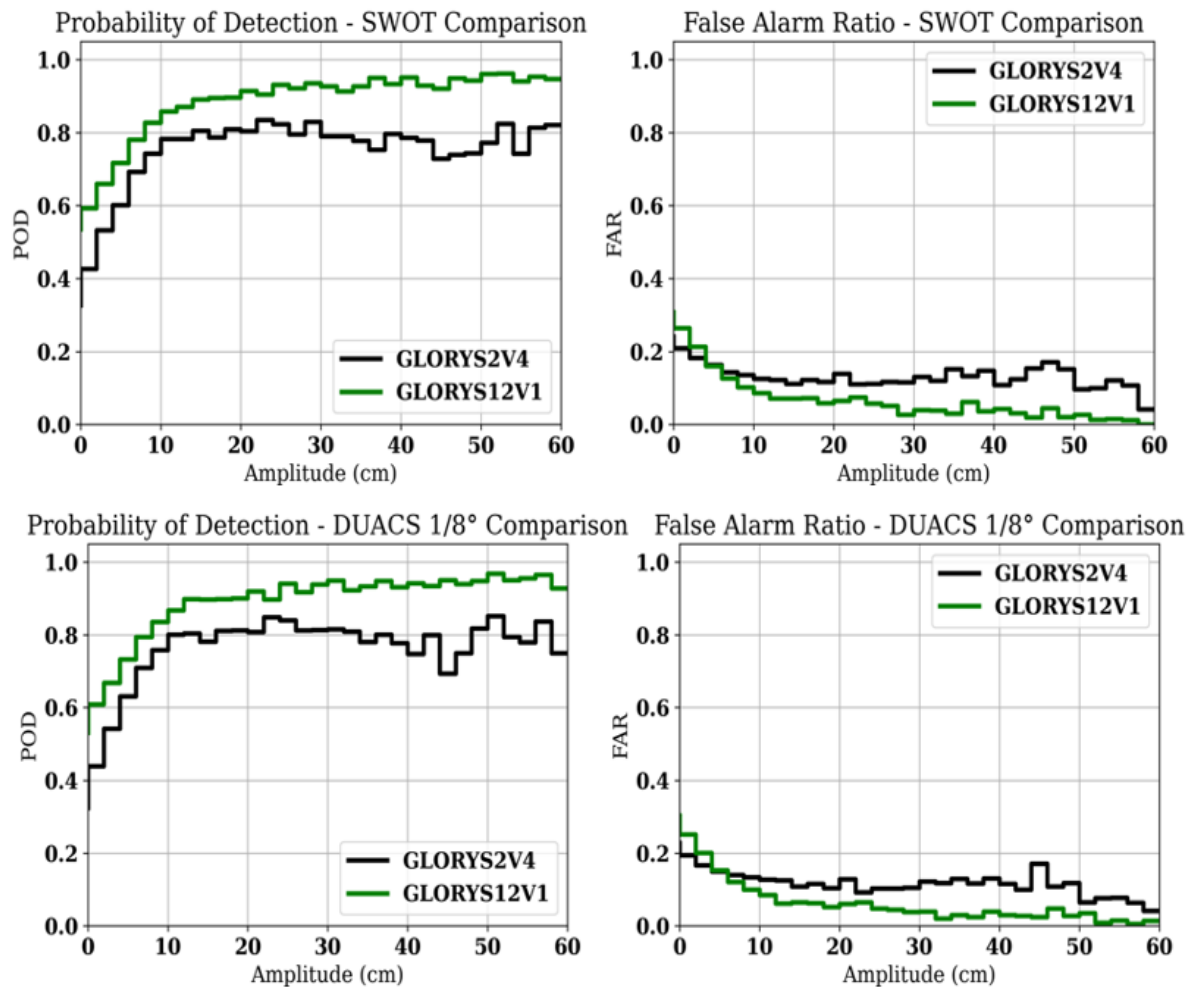


Fig. 10. On the top: histograms of the probability of detection (POD; left) and the false alarm ratio (FAR; right), over the evaluation period, evaluated as a function of amplitude (cm) for comparison with SWOT. In the bottom: histograms of the POD and FAR as a function of amplitude (cm), over the evaluation period, for the comparison with DUACS. The amplitude of hits and misses refers to the satellite (reference) eddies, while the amplitude of false alarms refers to those of unmatched eddies detected in the reanalysis.”

-) Line 479 “GLORYS12V1 shows a 15–20% [...]” The description of eddies with amplitude higher than 20cm does not seem to include more insights.

Response: We removed the detailed discussion of eddies with amplitudes greater than 20 cm because it did not provide additional insight. We retained only a concise description of the general improvement in POD and FAR for larger-radius and higher-amplitude eddies. We have revised the manuscript, now at lines: 523 – 525.

“For eddies larger than these limits, both comparisons show strong agreement. In DUACS comparison, for radii > 100 km, the POD reaches ~90% and FAR falls to around 10–15%. In the SWOT-based comparison, the performance is quite similar: FAR drops below 10%, while POD rises above 80–95% for amplitudes greater than 10 cm.”

-) Line 486: “Below what radius do 90% of misses?” The question is not clear to me, please rephrased.

-) Line 491: “[..]Fig 11 [..]” The reference should be Figure 12

-) Line 492: “90% of false [...]” There is only one point listed plus it seems there are too many decimals.

-) Line 499 : “[..]Fig 12[..]” The reference should be Figure 13

-) Line 515: “This shows that SW [...]” The sentence seems to quantify the better quality of SWOT-based data based on reanalysis skills. See mayor point.

Response: we have revised the manuscript, combining all the comments above, now at lines: 526 – 564.

“An additional perspective on the differences between GLORYS12V1 and GLORYS2V4 can be obtained by examining the integrated matching statistics as a function of eddy radius and amplitude (Fig. 11 and Fig. 12). We focus on two descriptive diagnostics: (i) the radius below which 90% of the false alarms occur and (ii) the percentage of matched eddies with amplitudes greater than 10 cm. The first identifies the size range in which unmatched reanalysis detections are concentrated, whereas the second describes the contribution of high-amplitude eddies to successful matches. These values are used for descriptive comparison and should not be interpreted as physical thresholds defining separate eddy populations or as criteria for filtering detections.

For DUACS comparison (Fig. 11), we can see that:

90% of false alarms are concentrated below the 49.9 km radius for GLORYS12V1 and 79.7 km for GLORYS2V4, corresponding to a difference of 29.8 km between the two cumulative distributions. Regarding Amplitude, 95.63% of GLORYS12V1 hits and 86.93% of GLORYS2V4 hits occur below 10 cm.

We can estimate that the 4.27% of GLORYS12V1 hits are above 10 cm versus 13.07% for GLORYS2V4, with a difference of 8.8 in terms of percentage.

For the SWOT comparison, using the results of Fig.12, we can estimate:

90% of false alarms fall below the 52.70 km radius for GLORYS12V1 and 80.30 km for GLORYS2V4, corresponding to a difference of 27.60 km. In terms of amplitude, 87.1% of GLORYS2V4 hits and 95.55% of GLORYS12V1 hits lie below 10 cm.

For SWOT, the corresponding proportions are 4.45% and 12.9%, respectively, with a difference of –8.45 percentage points.

Thus, the proportion of matched eddies exceeding 10 cm is smaller for GLORYS12V1 in both comparisons. Similar statements could be done for eddy radius.

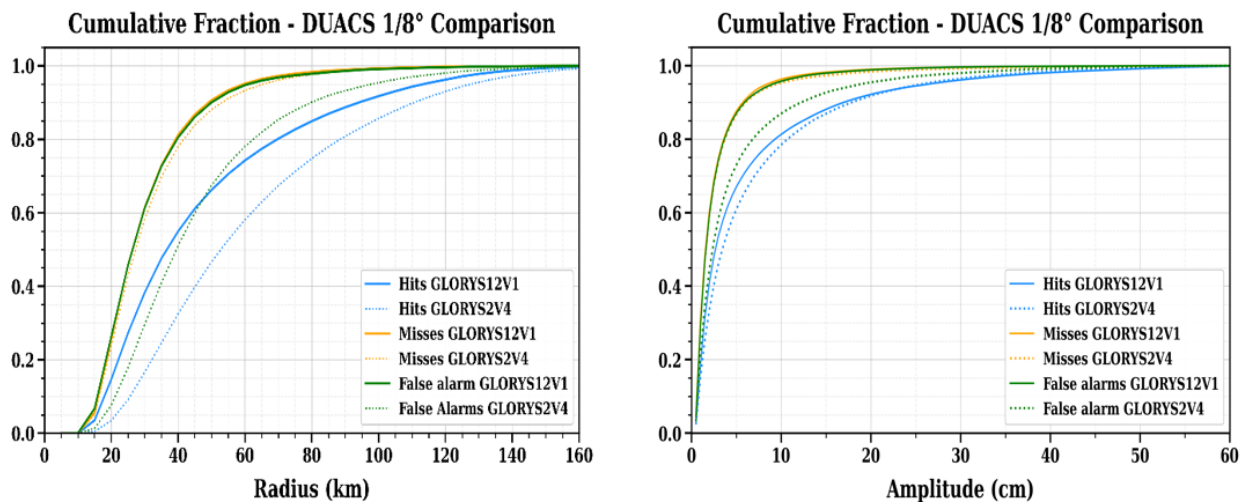


Fig. 11. Cumulative distribution of hits (blue), misses (orange) and false alarms (green) for GLORYS2V4 (dotted) and for GLORYS12V1 (solid) as a function of radius (left) and amplitude (right), for the comparison with DUACS, with a lifetime threshold of 4 days. The radius and the amplitude of hits and misses refers to the satellite (reference) eddies, while the radius and the amplitude of false alarms refers to the unmatched eddies detected in the reanalysis.

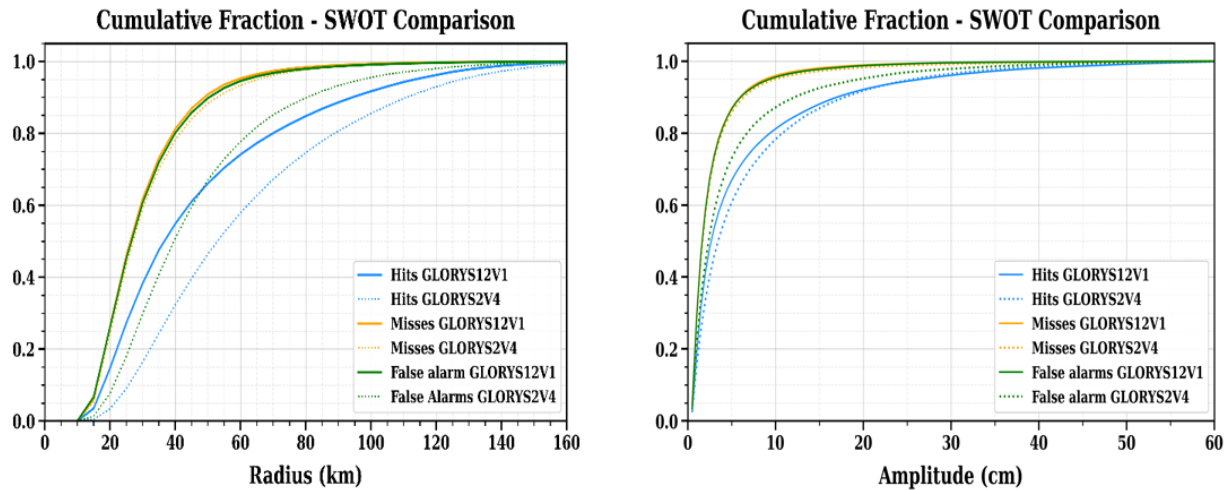


Fig.12. Cumulative distribution of hits (blue), misses (orange) and false alarms (green) for GLORYS2V4 (dotted) and for GLORYS12V1 (solid) as a function of radius (left) and amplitude (right), for the comparison with SWOT, with a lifetime threshold of 4 days. The radius and the amplitude of hits and misses refers to the satellite (reference) eddies, while the radius and the amplitude of false alarms refers to the unmatched eddies detected in the reanalysis.

Overall, these cumulative diagnostics show that most matching errors are associated with small eddies, those with radii below 50 km and amplitude below 4-5 cm. We see differences between DUACS and SWOT comparisons. They should be interpreted only as differences in reanalysis–satellite consistency over the common evaluation period, and not as a ranking of the quality of the two satellite products. However it also confirms a key weakness of GLORYS12V1: although it detects more eddies overall, it also produces more unmatched eddies (small eddies), which increases FAR. These cumulative plots show that both reanalysis systems still struggle to match very small eddies (radius < 50 km and amplitude < 4 – 5 cm), and that many of these features are likely not true eddies but artefacts in either the satellite products or the model. This suggests that future verification studies should apply filters to remove or down-weight small eddies, so that evaluation focuses on dynamically meaningful ocean structures.”

-) Table 9: the caption misses the reference to SWOT

Response: we have revised manuscript, now at lines: 594 – 596.

“

Mean error	Cost	Radius (Km)	Amplitude (cm)	Distance (Km)
GLORYS2V4	1.79	24.92	5.53	57.86
GLORYS12V1	1.33	16.79	3.59	50.29
% diff.	-25.70%	-32.62%	-35.08%	-13.06%

Table 9. Error contributions to the cost function for all hits evaluated over the period 1st August 2023 to 1st May 2025, for SWOT comparison Mean values for the cost function as well as contributions from radius (km), amplitude (cm), and distance errors (km). The percentage difference crossing from GLORYS2V4 to GLORYS12V1 is provided in the bottom row.”

-) Line 573: “[...] AVISO DUACS product (Table 7)” Why is the Table 7 cited here?

Response: we have revised the manuscript, now at lines: 609 – 611.

“These results indicate that, for all terms of the cost function, GLORYS12V1 achieves nearly identical results in both comparisons (Table 8 and Table 9).”

-) Line 605: “It should be noted...” Products at $\frac{1}{4}$ is not analyzed in the manuscript. What is the purpose of discussing it in the conclusion?

Response: deleted.

Line 617 ” [...] the Merged Interpolation of Observations with Spatial and Temporal constraints (MIOST) [...] ” MIOST acronym is previously used in the conclusion, please put the definition the first time it is used.

Response: We have deleted here. We have put the definition of SWOT MIOST in the Introduction section.

Line 628: “For amplitudes above 10 cm [...]” The values are given with too many decimals. Please discuss the limitation coming from the fact that the period analyzed covers less than two years.

Response: we have revised the manuscript, now at lines: 658 – 689.

“Both satellite reference products are distributed on a nominal $1/8^\circ$ grid, while SWOT MIOST Science additionally incorporates wide-swath SWOT observations through the MIOST mapping methodology. However, the Copernicus Marine Service Synthesis Quality Overview Document for Sea Level Products (CMEMS-SL-SQO-008-033-068) notes that the effective resolution of gridded sea-level products remains constrained by altimeter sampling and mapping procedures. Therefore, caution is required when interpreting results close to the effective resolution limit, because very small and short-lived eddies may still be influenced by noise and mapping artefacts. The remapping of all datasets onto the same regular $1/8^\circ$ grid allows a consistent comparison but does not make their native or effective resolutions equivalent. Despite these limitations, the results support the added value of the eddy-resolving GLORYS12V1 reanalysis compared with the eddy-permitting GLORYS2V4. The amplitude-based analyses show that GLORYS12V1 generally maintains a higher POD than GLORYS2V4. However, progressively excluding low-amplitude eddies does not produce a monotonic improvement in POD and FAR, as shown in Appendix C. False alarms are mainly concentrated among small eddies. In the DUACS comparison, 90% of the false alarms occur below a radius of approximately 50 km for GLORYS12V1 and 80 km for GLORYS2V4. The corresponding thresholds in the SWOT comparison are approximately 53 and 80 km. These results indicate that the false alarms associated with GLORYS12V1 are concentrated at smaller radii; they do not imply a lower total number or ratio of false alarms. In the DUACS comparison, approximately 4.3% of the GLORYS12V1 hits have amplitudes above 10 cm, compared with 13.1% for GLORYS2V4. In the SWOT comparison, the corresponding proportions are approximately 4.5% and 12.9%, respectively. These values describe the amplitude distribution of the matched eddies and should not be interpreted as direct measures of the overall performance of the two reanalyses. The common evaluation period covers less than two years and may therefore not fully sample the interannual variability of the North Atlantic eddy field. This limitation is particularly relevant for large and energetic eddies, which are less frequent and more sensitive to the specific dynamical conditions of the analysed period. Moreover, the results should not be interpreted as a ranking of the quality of the two satellite products. DUACS and SWOT MIOST Science provide complementary observational references with different mapping characteristics and levels of small-scale variability. SWOT MIOST Science provides additional information from wide-swath observations, but it remains a Level-4 mapped product. Furthermore, neither satellite product can be considered a fully independent reference because they share nadir-altimetry information with each other and with observations assimilated into the reanalyses.

Future work should develop objective and scale-aware strategies to filter or down-weight uncertain small features without automatically excluding dynamically relevant eddies. The present matching and verification framework could also be extended to include the “coincidence status” classification described by Ma et al. (2020), distinguishing matched, split, merged, missed, and artefactual eddies. This would provide a more detailed assessment and improve the diagnostic

capability of the verification framework. Finally, extending the analysis to longer periods, using satellite products with sufficient temporal coverage, and applying the methodology to other ocean regions would make it possible to better account for interannual variability and different dynamical regimes.”

Here, we have inserted the four Appendices (in the revised manuscript, after the Summary and Discussion section: A, B, C, and D).

As already mentioned at the beginning of this response document, we have added some calculations that had already been carried out and were previously included in the main text. In addition, we have included new calculations to further support and complete the statements made in the main text.

“Appendix A

In this appendix we aim to assess how the verification metrics, used to evaluate the ability of global ocean reanalyses to detect eddies observed in satellite datasets over a highly eddy-populated region, change under these different methodological choices.

Eddy detection-tracking and matching results for 7 and 28 minimum lifetime thresholds.

To assess the sensitivity of the verification results to the lifetime criterion, the detection, tracking, and matching analyses were repeated using minimum-lifetime thresholds of 7 and 28 days. Tables A1, A2, A5, and A6 report the numbers of anticyclonic and cyclonic eddies retained in each catalogue, whereas Tables A3, A4, A7, and A8 report POD, FAR, and the differences between GLORYS12V1 and GLORYS2V4. Fig. A1 shows the corresponding time series and power spectra. Increasing the minimum lifetime substantially reduces the number of retained eddies, especially at 28 days. This reduction does not, however, lead to lower FAR values: relative to the 4-day analysis, POD decreases and FAR increases for both reanalyses and both observational references. In Fig. A1, panel (a) shows the time series of the verification metrics over the full study period for both comparisons, while panel (b) shows the power spectra associated with each time series. The aim for the calculation of the power spectra is assessing how the variability of the verification metrics changes as the minimum eddy lifetime threshold is progressively increased.

Dataset	Anticyclonic	% w.r.t. DUACS	Cyclonic	% w.r.t. DUACS
DUACS	116644	-	116295	-
GLORYS2V4	62479	53.56%	67486	58.03%
GLORYS12V1	86976	74.57%	95836	82.40%

Table A1. Total number of anticyclonic and cyclonic eddies identified in the region of study over the evaluation period for DUACS, GLORYS2V4, and GLORYS12V1, for eddies with a minimum lifetime threshold of 7 days. The percentage of eddies identified in GLORYS2V4 and GLORYS12V1 with respect to DUACS is also provided.

Dataset	Anticyclonic	% w.r.t. SWOT	Cyclonic	% w.r.t. SWOT
SWOT	114015	-	116036	-
GLORYS2V4	62479	54.79%	67486	58.15%
GLORYS12V1	86976	76.28%	95836	82.59%

Table A2. Total number of anticyclonic and cyclonic eddies identified in the region of study over the evaluation period for SWOT, GLORYS2V4, and GLORYS12V1, for eddies with a minimum lifetime threshold of 7 days. The percentage of eddies identified in GLORYS2V4 and GLORYS12V1 with respect to SWOT is also provided.

Verification metric	GLORYS12V1	GLORYS2V4	Diff.	Diff. %
POD	0.608	0.449	0.159	35.41%
FAR	0.225	0.196	0.029	14.80%

Table A3. aggregated eddy verification statistics for the Probability of Detection (POD) and False Alarm Ratio (FAR) evaluated over the period, from 1st August 2023 to 1st May 2025, for the comparison with DUACS, with a minimum lifetime threshold of 7 days.

Verification metric	GLORYS12V1	GLORYS2V4	Diff.	Diff. %
POD	0.606	0.447	0.159	35.57%
FAR	0.237	0.209	0.028	13.4%

Table A4. Aggregated eddy verification statistics for the Probability of Detection (POD) and False Alarm Ratio (FAR) evaluated over the period, from 1st August 2023 to 1st May 2025, for the comparison with SWOT, with a minimum lifetime threshold of 7 days.

Dataset	Anticyclonic	% w.r.t. DUACS	Cyclonic	% w.r.t. DUACS
DUACS	67686	-	69657	-
GLORYS2V4	39218	57.94%	43705	62.74%
GLORYS12V1	41768	61.70%	49935	71.69%

Table A5. Total number of anticyclonic and cyclonic eddies identified in the region of study over the evaluation period for DUACS, GLORYS2V4, and GLORYS12V1. The percentage of eddies identified in GLORYS2V4 and GLORYS12V1 with respect to DUACS is also provided, with a lifetime threshold of 28 days.

Dataset	Anticyclonic	% w.r.t. SWOT	Cyclonic	% w.r.t. SWOT
SWOT	67089	-	68842	-
GLORYS2V4	39218	58.45%	43705	63.49%
GLORYS12V1	41768	62.26%	49935	72.54%

Table A6. Total number of anticyclonic and cyclonic eddies identified in the region of study over the evaluation period for SWOT, GLORYS2V4, and GLORYS12V1. The percentage of eddies identified in GLORYS2V4 and GLORYS12V1 with respect to SWOT is also provided, with a lifetime threshold of 28 days.

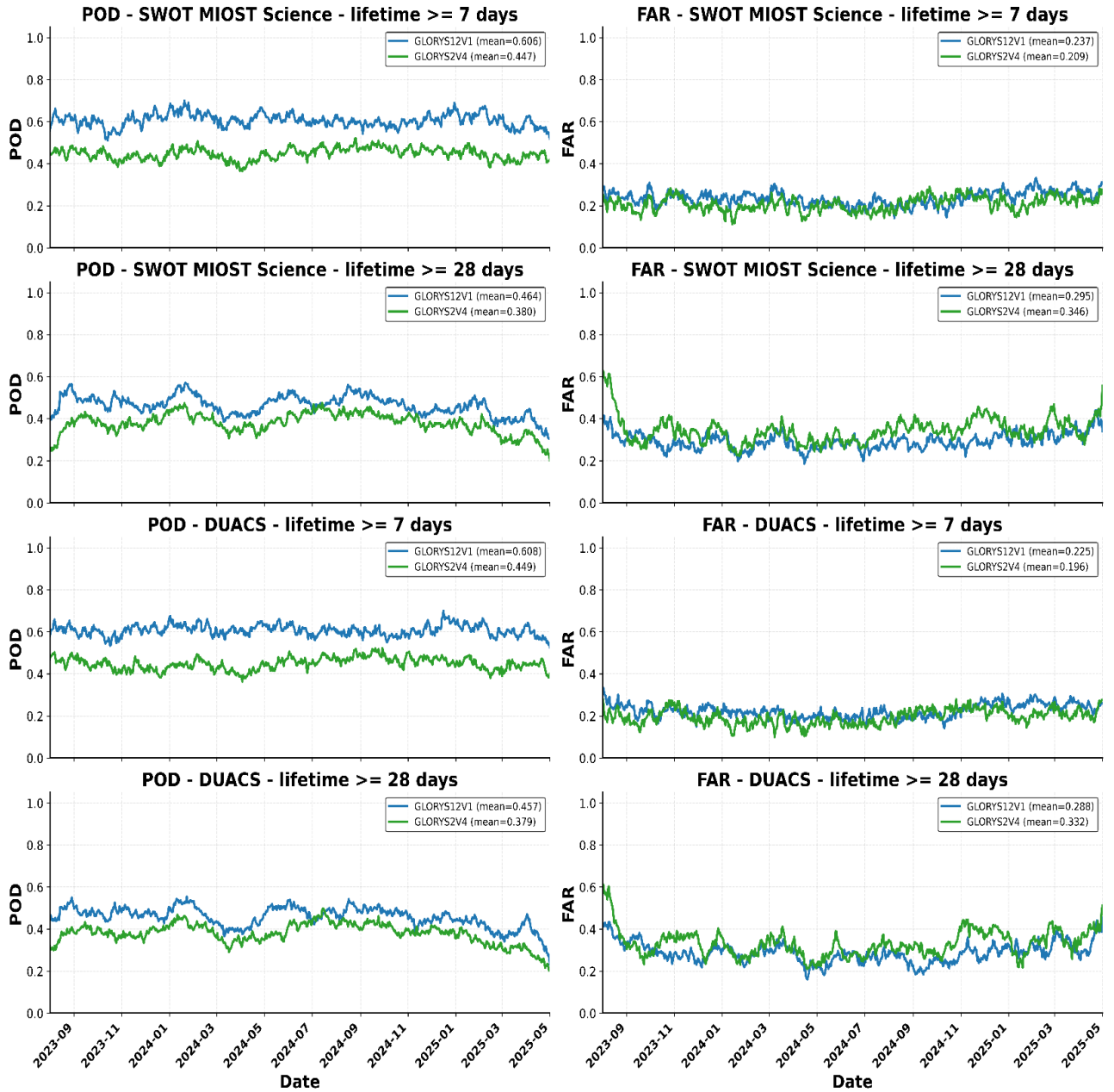
Statistics	GLORYS12V1	GLORYS2V4	Diff.	Diff. %
POD	0.464	0.380	0.084	22.1%
FAR	0.295	0.346	-0.051	-14.7%

Table A7. aggregated eddy verification statistics for the Probability of Detection (POD) and False Alarm Ratio (FAR) evaluated over the period, from 1st August 2023 to 1st May 2025, for the comparison with DUACS, with a minimum lifetime threshold of 28 days.

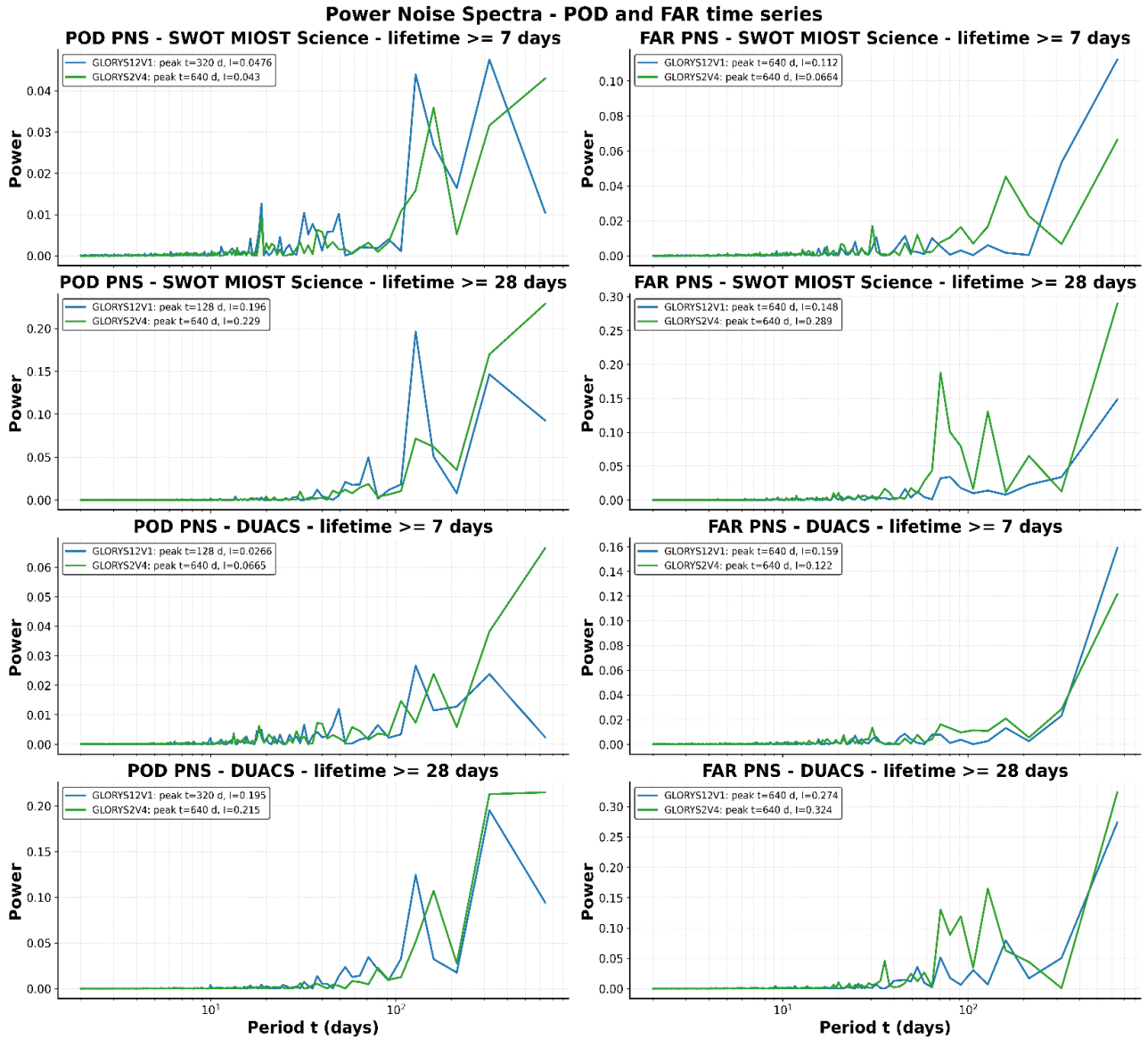
Statistics	GLORYS12V1	GLORYS2V4	Diff.	Diff. %
POD	0.457	0.379	0.078	20.5%
FAR	0.288	0.322	-0.034	-10.5%

Table A8. aggregated eddy verification statistics for the Probability of Detection (POD) and False Alarm Ratio (FAR) evaluated over the period, from 1st August 2023 to 1st May 2025, for the comparison with SWOT, with a minimum lifetime threshold of 28 days.

POD and FAR time series - SWOT MIOST Science and DUACS



(a)



(b)

Fig. A1. Time series of matching statistics for GLORYS2V4 and GLORYS12V1, compared with SWOT and DUACS (a), over the period from 1st August 2023 to 1st May 2025, with a minimum lifetime of 7 and 28 days. In (b), the power spectra of each time series, for both products and comparisons.

At the 7-day threshold, POD decreases for both reanalyses. FAR increases for GLORYS2V4, whereas for GLORYS12V1 it decreases slightly against DUACS and remains essentially unchanged against SWOT. At the 28-day threshold, POD decreases and FAR increases relative to the 4-day configuration, while GLORYS12V1 retains higher POD and attains lower FAR than GLORYS2V4 for both references. This condition highlights not only the greater ability of GLORYS12V1 to capture observational eddies compared to GLORYS2V4, for both satellite datasets, but also its lower error rate in detecting and tracking false alarms. The GLORYS12V1 FAR becomes lower than that of GLORYS2V4 when only long-lived eddies are considered. However, the application of a longer lifetime threshold also leads to a substantial reduction in the total number of detected and tracked eddies compared to the lower lifetime cases, as shown in Tables A1, A2, A5, and A6. This reduction is stronger for the reanalyses than for the satellite datasets, producing an overall decrease in POD and an overall increase in FAR for both comparisons and for both reanalysis products with respect to the results discussed in Sect. 3.2 and reported in Tables 4 and 5. This behaviour indicates that, although the 28-day threshold isolates more stable eddy structures, it also exposes a significant discrepancy between the reanalyses and the observational datasets. In particular, some of the eddies matched in the reanalyses at lower lifetime thresholds appear to be less

temporally stable than those observed in the satellite products. As a result, the number of hits clearly tends to decrease as the minimum lifetime threshold increases.

This sensitivity test therefore suggests that temporally more stable eddies tend to reduce the resolution-related gap between the two reanalysis products. Nevertheless, this does not lead to an increase in POD or to a systematic decrease in FAR. This confirms that both ocean reanalyses still show limitations in capturing observed eddies and in avoiding the generation of unmatched eddy-like structures. More generally, these results show that one of the main challenges is to identify a methodology able to remove spurious and unstable structures, for which it is not always clear whether they can be considered robust mesoscale eddies.

In panel (b), the power spectra show that, at short temporal scales, the variability of the verification-metric time series is almost similar across the different configurations. The spectral differences are mainly concentrated at the longest resolved periods. For FAR, GLORYS12V1 has the larger peak with the 7-day threshold, whereas GLORYS2V4 has the larger peak with the 28-day threshold. Because the analysed record is only 640 days long, peaks at periods of 320–640 days should be interpreted as low-frequency variability rather than as evidence of robust periodic or seasonal signals. It should also be noted that the axes do not have the same numerical range across all panels. In general, the numerical range of the FAR power spectra is larger than that of POD, indicating that the strongest spectral peaks are generally associated with FAR.

Appendix B

Sensitivity of bootstrap resampling results for verification metrics to the block length

To assess the sensitivity of the confidence intervals to temporal dependence, the coupled block-bootstrap analysis was repeated using block lengths of 7, 20, and 30 days, as well as block lengths derived from the mean lifetimes of the satellite-observed eddies. Results are reported for the 4-day and 7-day minimum-lifetime thresholds. Across the tested block lengths, GLORYS12V1 maintains higher POD than GLORYS2V4 for both observational references. The FAR difference also remains positive, although it becomes smaller for the 7-day minimum-lifetime threshold.

Verification Metric	GLORYS12V1	GLORYS2V4	<i>Difference (12V1 – 2V4)</i>
POD (Mean)	0.603	0.442	0.161
POD (95% CI)	[0.594, 0.611]	[0.431, 0.453]	[0.148, 0.173]
FAR (Mean)	0.234	0.200	0.034
FAR (95% CI)	[0.224, 0.245]	[0.189, 0.213]	[0.020, 0.048]

Table B1. Bootstrap resampling for SWOT, for a minimum lifetime threshold of 7 days and a block length of 7 days.

Verification Metric	GLORYS12V1	GLORYS2V4	<i>Difference (12V1 – 2V4)</i>
---------------------	------------	-----------	--------------------------------

POD (Mean)	0.599	0.438	0.161
POD (95% CI)	[0.586, 0.610]	[0.430, 0.446]	[0.147, 0.174]
FAR (Mean)	0.241	0.208	0.034
FAR (95% CI)	[0.233, 0.250]	[0.197, 0.219]	[0.024, 0.043]

Table B2. Bootstrap resampling for DUACS comparison, for a minimum lifetime threshold of 7 days and a block length of 7 days.

Verification Metric	GLORYS12V1	GLORYS2V4	Difference (12V1 – 2V4)
POD (Mean)	0.620 0.603	0.455 0.444	0.165 0.159
POD (95% CI)	[0.608 , 0.632] [0.591 , 0.614]	[0.442 , 0.468] [0.429 , 0.460]	[0.149 , 0.183] [0.142 , 0.177]
FAR (Mean)	0.236 0.235	0.187 0.200	0.049 0.034
FAR (95% CI)	[0.228 , 0.244] [0.222 , 0.250]	[0.175 , 0.202] [0.182, 0.221]	[0.031 , 0.065] [0.011 , 0.054]

Table B3. Block-bootstrap estimates of POD and FAR using 20-day blocks for the comparison with DUACS. Values are reported for GLORYS12V1, GLORYS2V4, and their difference, together with 95% confidence intervals. In the first row, there is the calculation for the 4 days minimum lifetime threshold. In the second row, there is the calculation for 7 days minimum lifetime threshold.

Verification Metric	GLORYS12V1	GLORYS2V4	Difference (12V1 – 2V4)
POD (Mean)	0.617 0.598	0.450 0.440	0.167 0.158
POD (95% CI)	[0.603 , 0.633] [0.582, 0.614]	[0.440 , 0.459] [0.429, 0.451]	[0.151 , 0.187] [0.140, 0.178]
FAR (Mean)	0.236 0.242	0.192 0.207	0.044 0.035
FAR (95% CI)	[0.230, 0.241] [0.233, 0.251]	[0.180 , 0.207] [0.193, 0.225]	[0.031 , 0.056] [0.021, 0.046]

Table B4. Block-bootstrap estimates of POD and FAR using 20-day blocks for the comparison with SWOT MIOST. Values are reported for GLORYS12V1, GLORYS2V4, and their difference, together with 95% confidence intervals. In the first row, there is the calculation for the 4 days minimum lifetime threshold. In the second row, there is the calculation for 7 days minimum lifetime threshold.

Verification Metric	GLORYS12V1	GLORYS2V4	Difference (12V1 – 2V4)
POD (Mean)	0.620	0.452	0.168

	0.603	0.441	0.162
POD (95% CI)	[0.604, 0.634] [0.589, 0.615]	[0.437, 0.466] [0.425, 0.459]	[0.148, 0.190] [0.142, 0.185]
FAR (Mean)	0.235 0.234	0.186 0.201	0.049 0.033
FAR (95% CI)	[0.227, 0.245] [0.223, 0.250]	[0.173, 0.200] [0.184, 0.217]	[0.031, 0.066] [0.011, 0.055]

Table B5. Block-bootstrap estimates of POD and FAR using 30-day blocks for the comparison with DUACS. Values are reported for GLORYS12V1, GLORYS2V4, and their difference, together with 95% confidence intervals. In the first row, there is the calculation for the 4 days minimum lifetime threshold. In the second row, there is the calculation for 7 days minimum lifetime threshold.

Verification Metric	GLORYS12V1	GLORYS2V4	<i>Difference (12V1 – 2V4)</i>
POD (Mean)	0.618 0.616	0.448 0.460	0.170 0.166
POD (95% CI)	[0.600, 0.636] [0.602, 0.626]	[0.438, 0.456] [0.448, 0.484]	[0.149, 0.194] [0.131, 0.173]
FAR (Mean)	0.236 0.240	0.191 0.206	0.045 0.035
FAR (95% CI)	[0.231, 0.240] [0.236, 0.245]	[0.180, 0.202] [0.194, 0.217]	[0.032, 0.056] [0.023, 0.046]

Table B6. Block-bootstrap estimates of POD and FAR using 30-day blocks for the comparison with SWOT. Values are reported for GLORYS12V1, GLORYS2V4, and their difference, together with 95% confidence intervals. In the first row, there is the calculation for the 4 days minimum lifetime threshold. In the second row, there is the calculation for 7 days minimum lifetime threshold.

For the sake of completeness, we also calculated the average lifetimes of the observed eddies for each satellite dataset using two different minimum lifetimes (4 and 7 days). The idea is to use these average lifetimes as the block lengths for the coupled bootstrap resampling. The mean eddy lifetime is calculated from the number of daily observations associated with each individual trajectory, using the “nb_obs_by_track” property provided by py-eddy-tracker. Tracks with zero observations are excluded.

For DUACS, we calculate the average lifetime of eddies:

- Anticyclonic eddies with a minimum lifetime threshold of 4 days: 18.03 days.
- Cyclonic eddies with a minimum lifetime threshold of 4 days: 18.23 days.
- Anticyclonic eddies with a minimum lifetime threshold of 7 days: 22.69 days.
- Cyclonic eddies with a minimum lifetime threshold of 7 days: 23.05 days.

Therefore, considering both anticyclonic and cyclonic eddies, we calculate an average eddy lifetime of: 18.13 days (minimum lifetime of 4 days) and 22.87 days (minimum lifetime of 7 days). Therefore, for the block bootstrap resample, we rounded up to the next whole day, setting a value of 19 days for a minimum lifetime of 4 days, and a value of 23 days for a minimum lifetime of 7 days.

For SWOT, we calculate the average lifetime of eddies:

- Anticyclonic eddies with a minimum lifetime threshold of 4 days: 17.51 days.
- Cyclonic eddies with a minimum lifetime threshold of 4 days: 17.58 days.
- Anticyclonic eddies with a minimum lifetime threshold of 7 days: 22.69 days.
- Cyclonic eddies with a minimum lifetime threshold of 7 days: 22.38 days.

Therefore, considering both anticyclonic and cyclonic eddies, we calculate an average eddy lifetime of: 17.54 days (minimum lifetime of 4 days) and 22.54 days (minimum lifetime of 7 days). Therefore, for the block bootstrap resample, we rounded up to the next whole day, setting a value of 18 days for a minimum lifetime of 4 days, and a value of 23 days for a minimum lifetime of 7 days.

Verification Metric	GLORYS12V1	GLORYS2V4	Difference (12V1 – 2V4)
POD (Mean)	0.620	0.452	0.168
	0.602	0.443	0.159
POD (95% CI)	[0.608, 0.632]	[0.438, 0.466]	[0.153, 0.186]
	[0.589, 0.614]	[0.426, 0.460]	[0.143, 0.180]
FAR (Mean)	0.235	0.186	0.049
	0.235	0.200	0.034
FAR (95% CI)	[0.228, 0.243]	[0.175, 0.199]	[0.035, 0.062]
	[0.223, 0.250]	[0.182, 0.221]	[0.011, 0.055]

Table B7. Block-bootstrap estimates of POD and FAR using 18 days and 23 days blocks for the comparison with DUACS. Values are reported for GLORYS12V1, GLORYS2V4, and their difference, together with 95% confidence intervals. In the first row, there is the calculation for the 4 days minimum lifetime threshold. In the second row, there is the calculation for 7 days minimum lifetime threshold.

Verification Metric	GLORYS12V1	GLORYS2V4	Difference (12V1 – 2V4)
POD (Mean)	0.618	0.448	0.170
	0.598	0.441	0.157
POD (95% CI)	[0.601, 0.632]	[0.440, 0.457]	[0.153, 0.188]
	[0.576, 0.613]	[0.430, 0.452]	[0.139, 0.176]
FAR (Mean)	0.236	0.192	0.044
	0.241	0.207	0.034
FAR (95% CI)	[0.229, 0.244]	[0.180, 0.204]	[0.032, 0.056]
	[0.234, 0.250]	[0.195, 0.220]	[0.019, 0.047]

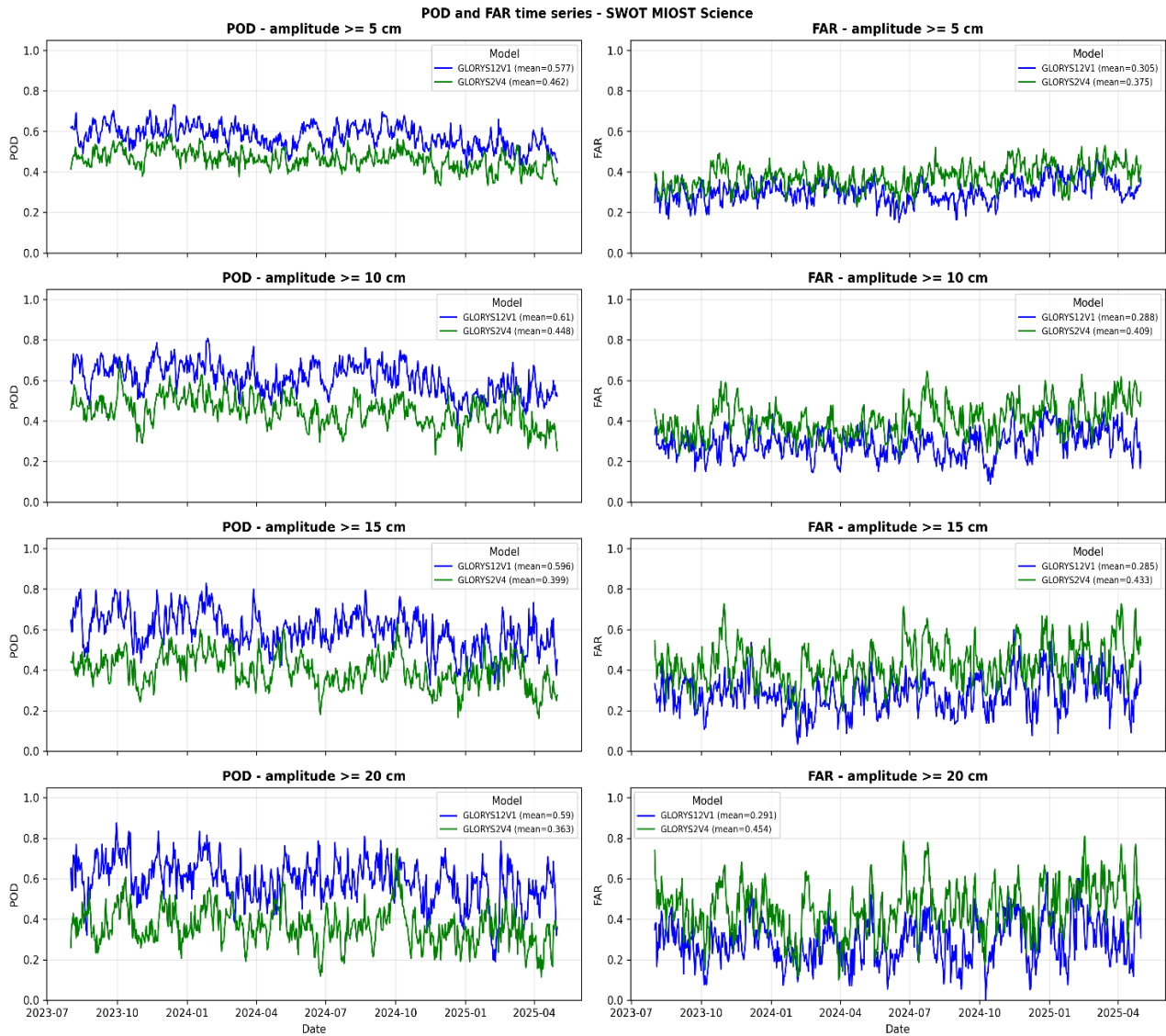
Table B8. Block-bootstrap estimates of POD and FAR using 19 days and 23 days blocks for the comparison with SWOT. Values are reported for GLORYS12V1, GLORYS2V4, and their difference, together with 95% confidence intervals. In the first row, there is the calculation for the 4 days minimum lifetime threshold. In the second row, there is the calculation for 7 days minimum lifetime threshold.

Appendix C

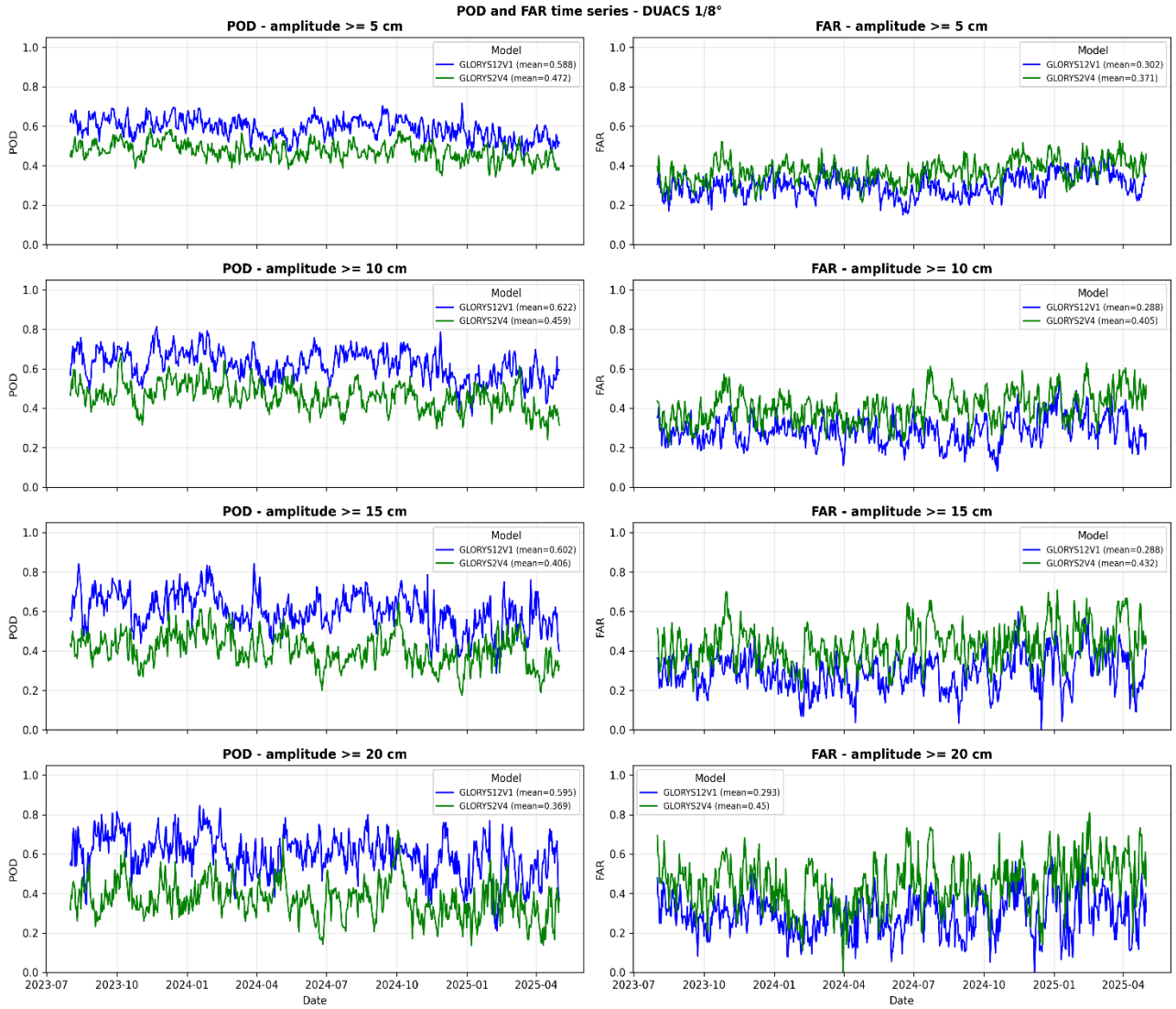
Sensitivity of verification metrics to eddy amplitude thresholds.

To assess the sensitivity of the verification results to eddy amplitude, POD and FAR were recomputed using minimum amplitude thresholds of 5, 10, 15, and 20 cm. We also considered the complementary subset of eddies with amplitudes

lower than or equal to 10 cm to isolate the contribution of low-amplitude features. The 10 cm value is used only as a diagnostic criterion and not as a physical boundary between different eddy regimes.



(a)

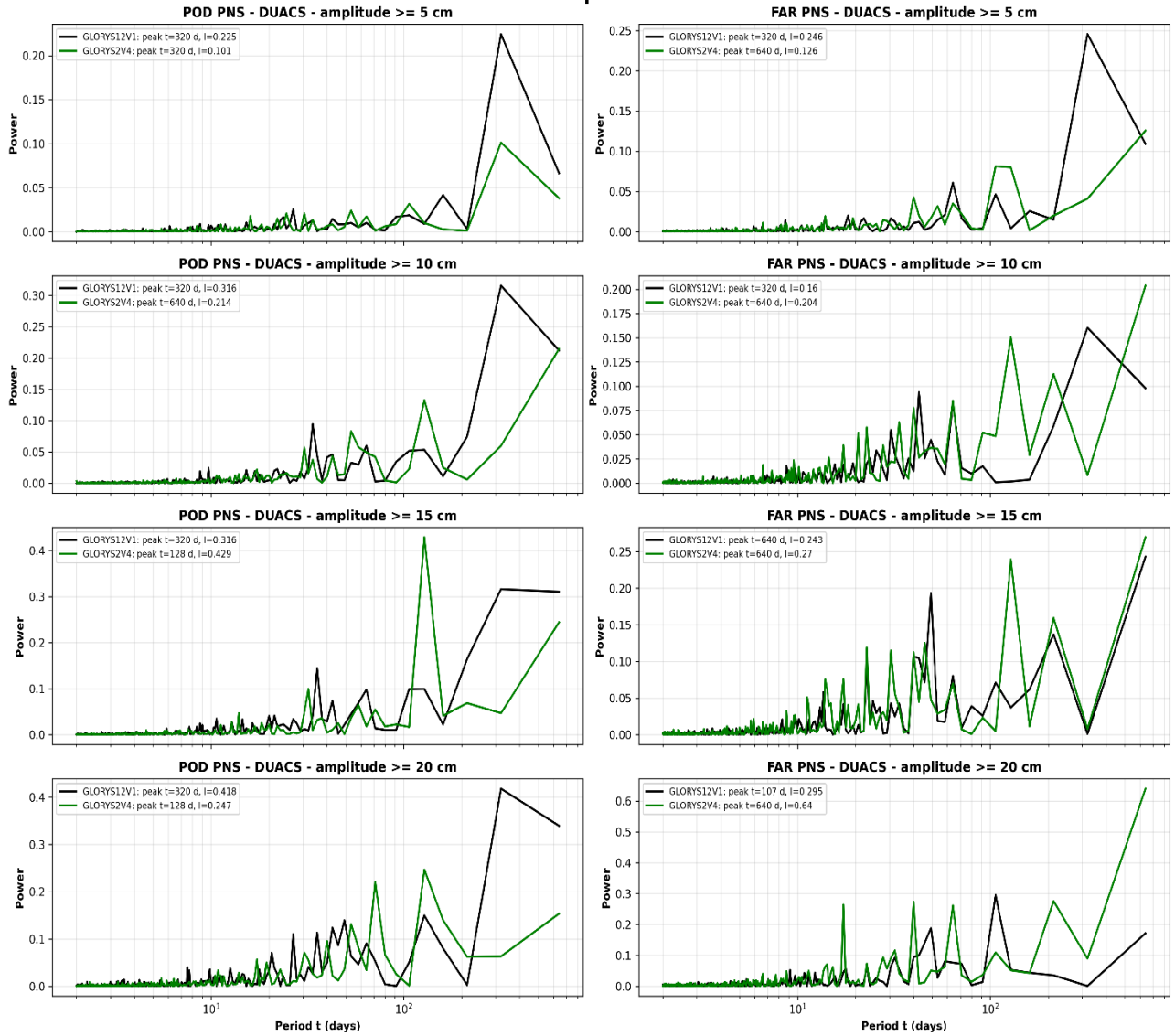


(b)

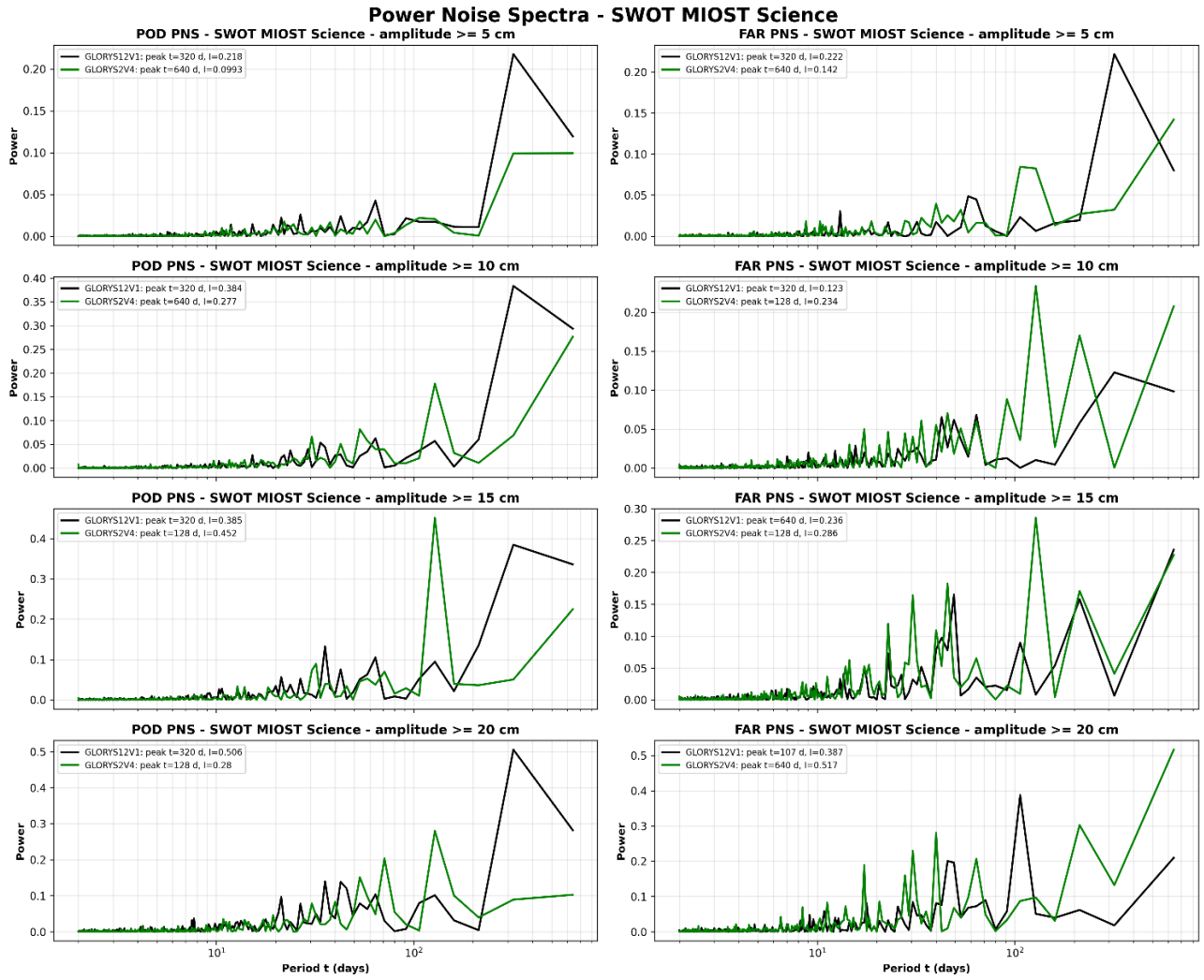
Fig. C1. Time series for POD & FAR by amplitude threshold: 5 cm, 10 cm, 15 cm, 20 cm, for SWOT comparison (a) and for DUACS 1/8° comparison (b), with a minimum lifetime threshold of 4 days.

The verification metrics do not improve monotonically as low-amplitude eddies are progressively excluded. For GLORYS12V1, POD remains approximately between 0.58 and 0.62, while FAR remains between about 0.27 and 0.31. For GLORYS2V4, POD decreases from approximately 0.46 to 0.47 at the 5 cm threshold to about 0.36 – 0.37 at 20 cm, while FAR increases from about 0.37–0.38 to approximately 0.45. The higher thresholds also retain fewer eddies, producing larger day-to-day variability. Therefore, these results do not indicate a uniform improvement in the representation of stronger eddies.

Power Noise Spectra - DUACS



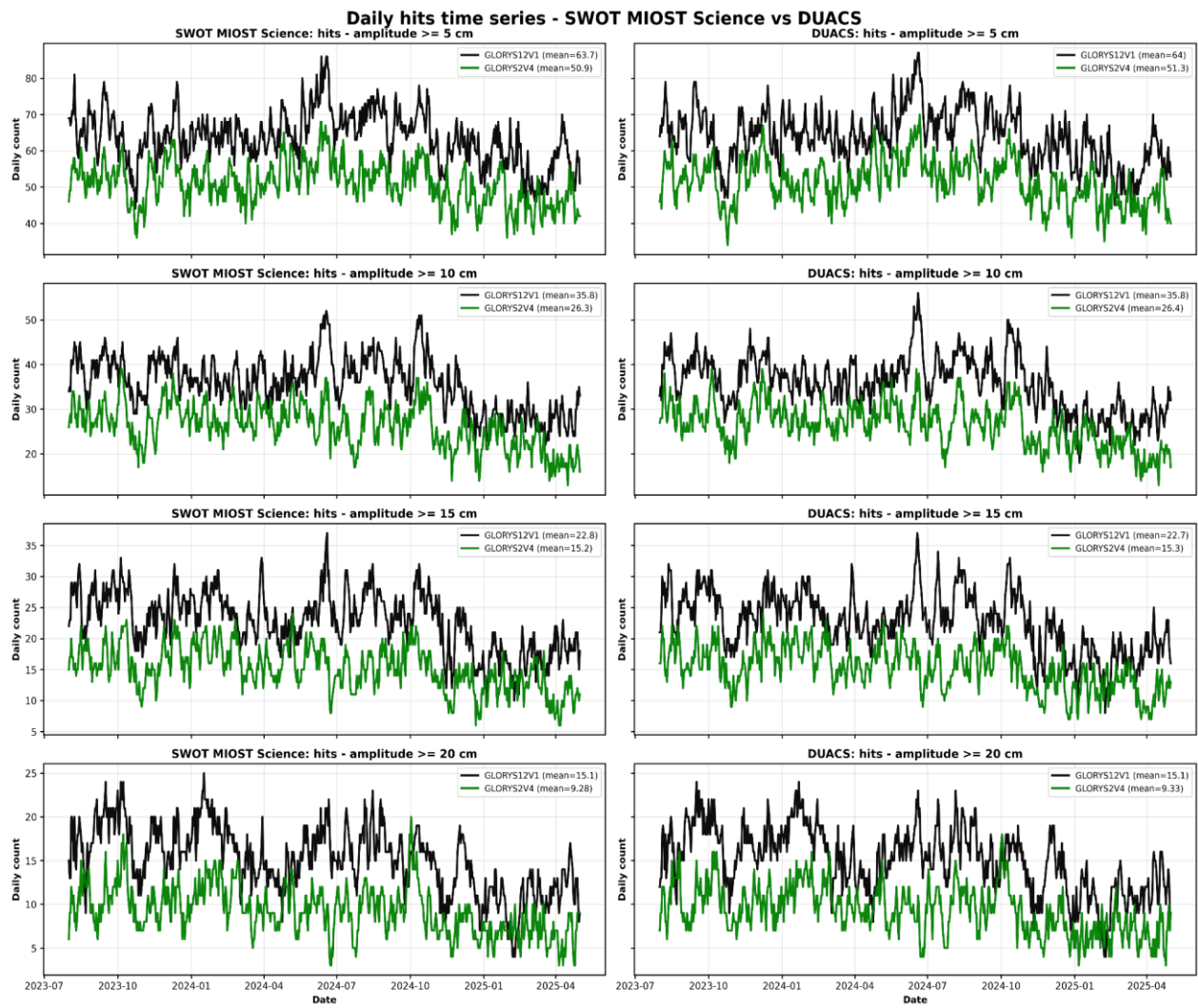
(a)



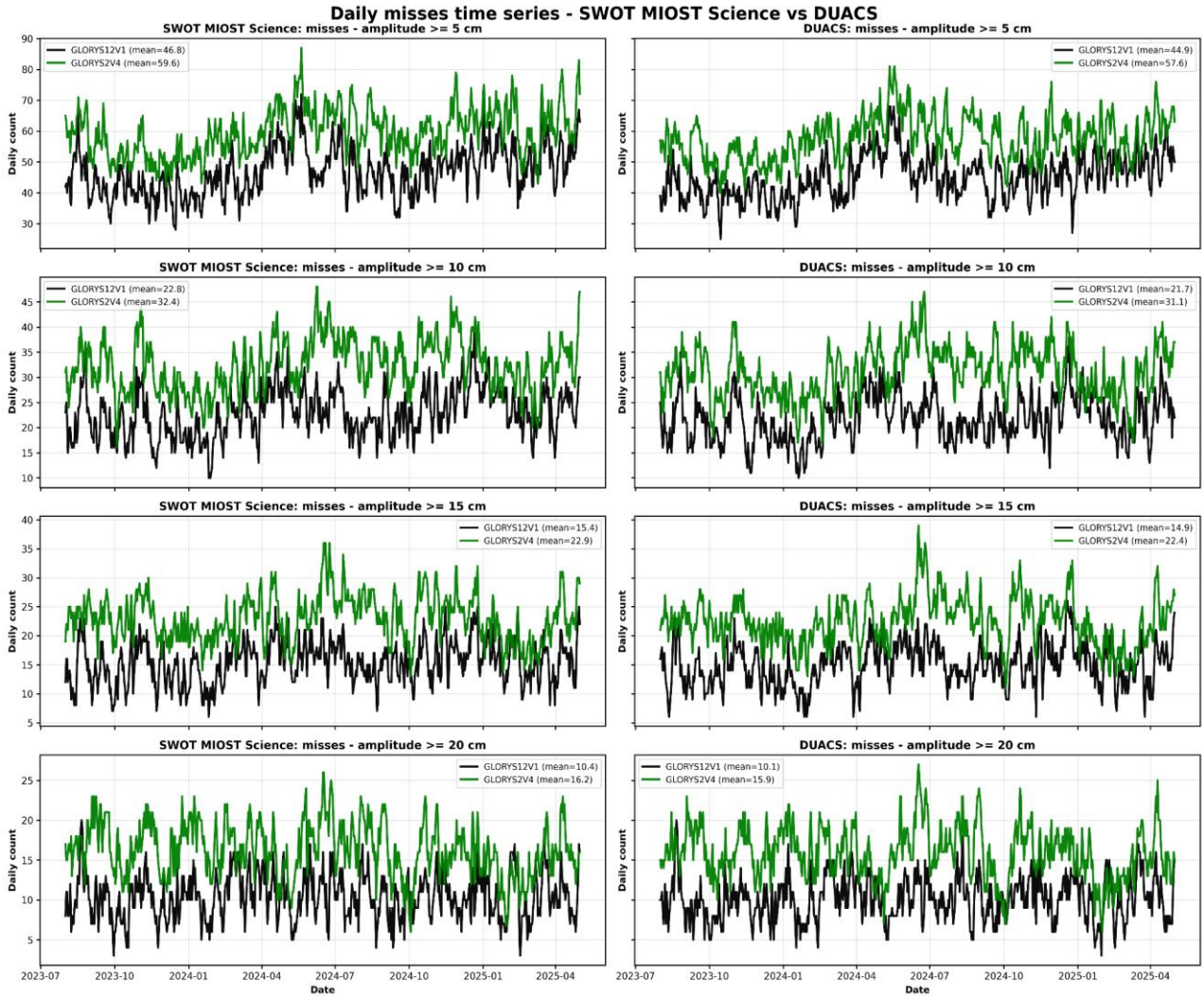
(b)

Fig. C2. Power Spectra of time series of both verification metrics for each reanalysis products, for DUACS (a) and SWOT MIOST Science comparison (b). It is shown that the agreement does not improve significantly compared with the more general case, namely that extended to all eddies. Generally, for both observational datasets, we observe that the agreement between the two reanalyses is never markedly better. Only in terms of FAR is it observed that GLORYS2V4 shows a net deterioration in agreement, as there is a clear increase in FAR, particularly at high amplitudes.

For higher amplitude thresholds, the metric values show much larger variations over time. On some days, the POD values are close to 1 and the FAR values are close to 0, while on other days they are much farther from these ideal values. This variability is mainly caused by the smaller number of eddies with amplitudes above 10 or 20 cm. For this reason, the POD and FAR time series can show large variations and can rapidly move towards values close to 0 or 1. These results suggest that the imperfect matching between the reanalyses and the observational datasets mainly concerns small ocean eddies. However, this improvement does not mean that the complete eddy field is better represented.



(a)

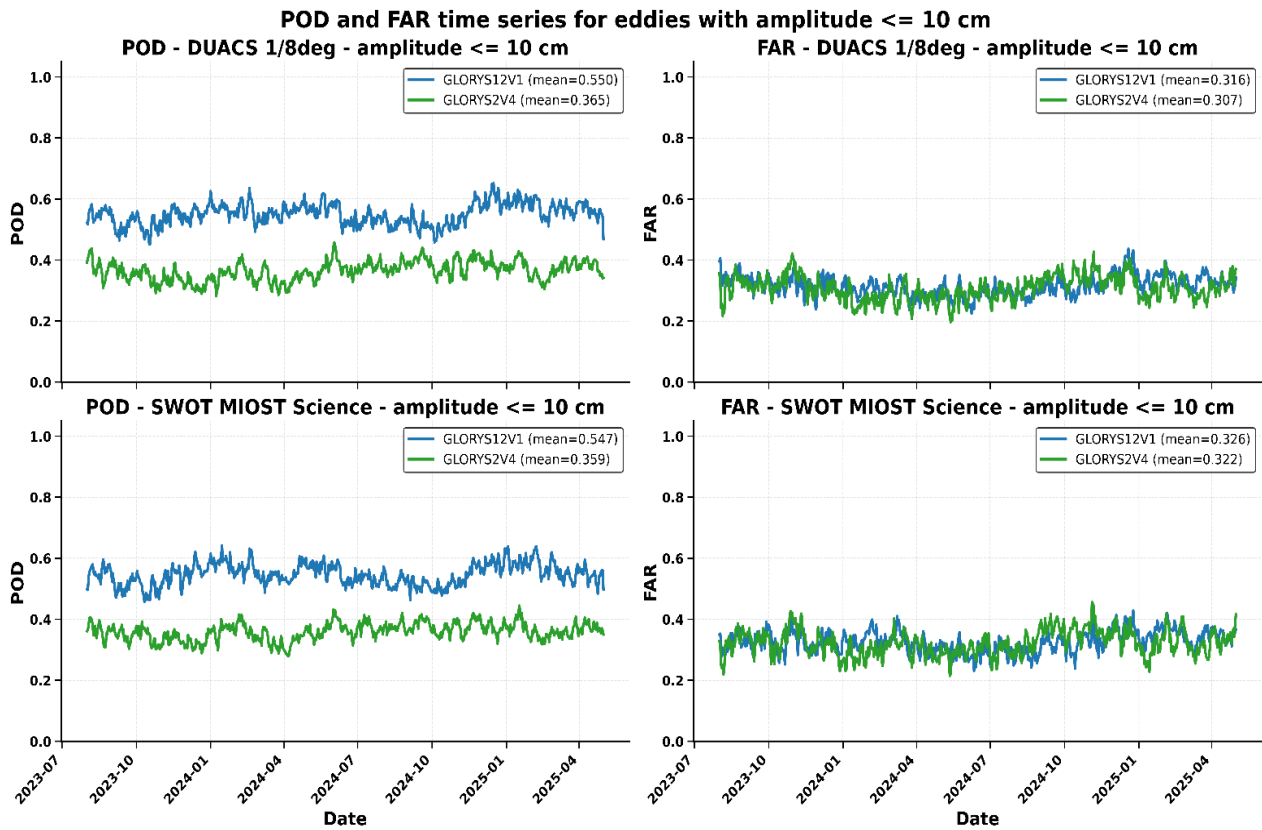


(b)

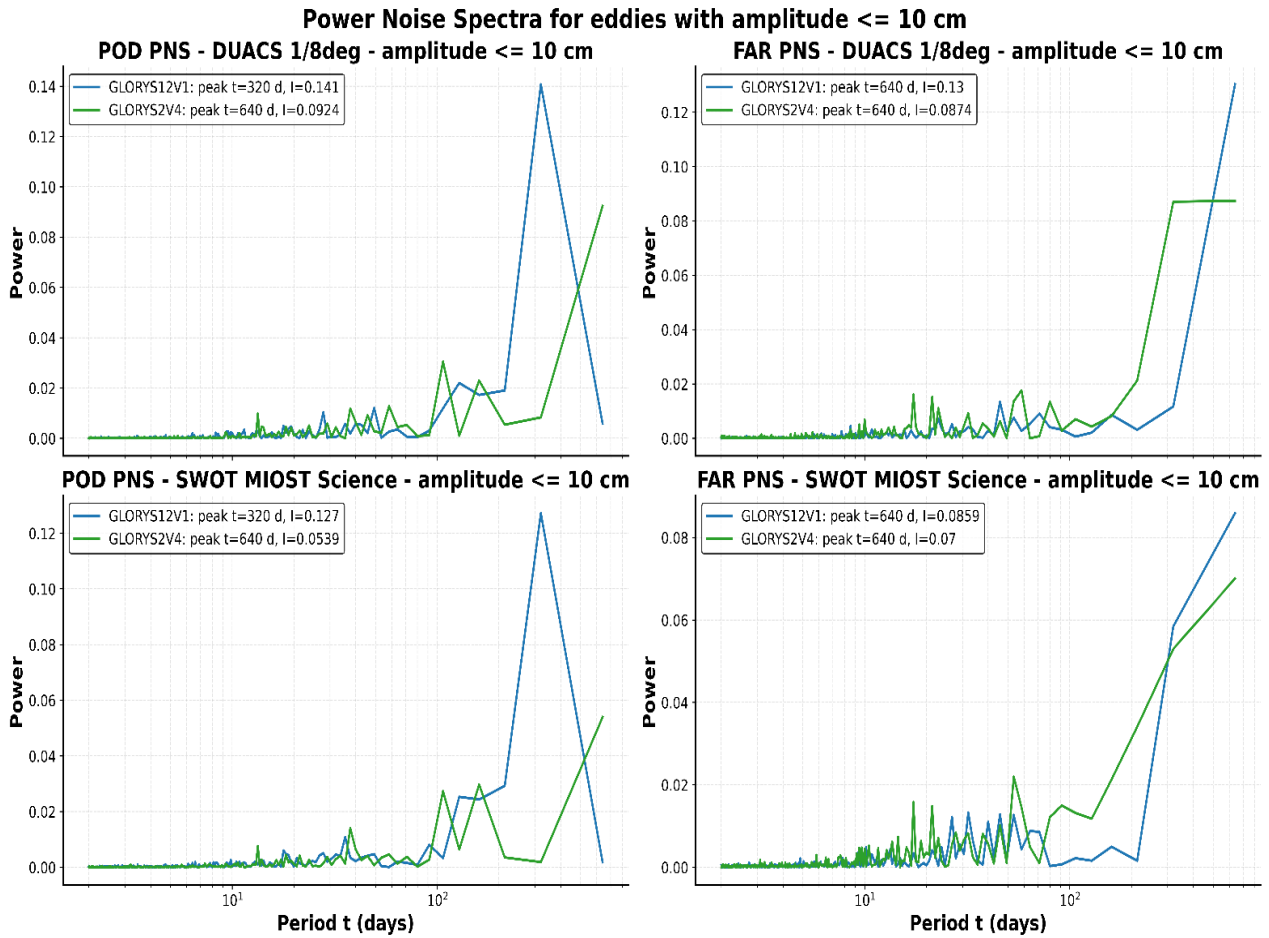
Fig. C3. Time series for daily number of hits (a), misses (b), for both satellite data compared to the reanalyses. On the left, the comparison with SWOT MIOST Science, while on the right, the comparison with DUACS.

As the minimum amplitude threshold increases, the absolute numbers of both hits and misses decrease because fewer eddies satisfy the criterion. For GLORYS12V1, the mean number of hits remains higher than the number of misses at all thresholds. For GLORYS2V4, misses remain more numerous than hits and their relative importance increases at higher thresholds. This behaviour is consistent with the reduction in POD observed for GLORYS2V4. The power spectra in Fig. C2 shows that temporal variability generally increases with the minimum amplitude threshold, mainly at longer periods, while differences remain limited at short periods. GLORYS12V1 generally shows larger POD spectral peaks, whereas GLORYS2V4 generally shows larger FAR peaks, with exceptions at 15 cm for POD and 5 cm for FAR. Fig. C3 shows that both hits and misses decrease in absolute number as the threshold increases because fewer eddies are retained. However, GLORYS12V1 maintains more hits than misses, whereas misses remain more numerous than hits for GLORYS2V4, indicating a greater reduction in matching performance for GLORYS2V4 at higher amplitude thresholds. We also evaluated POD and

FAR for low-amplitude eddies (≤ 10 cm), whose weak SSH signatures make their detection and matching more sensitive to mapping, filtering, and background variability. This subset is therefore expected to show lower POD and higher FAR because small differences between observational and reanalysis fields can alter eddy position and contours or generate unmatched features.



(a)



(b)

Fig. C4. (a) POD and FAR time series for eddies with amplitudes ≤ 10 cm for the comparisons with DUACS $1/8^\circ$ and SWOT MIOST Science. (b) Corresponding power spectra. A 4-day minimum-lifetime threshold is used.

Fig. C5 shows the distributions of hits, misses, and false alarms as functions of eddy radius and amplitude.

We further introduce the histograms of the eddy matching statistics, namely hits, misses, and false alarms, as a function of both eddy radius and eddy amplitude. The aim is to provide a more detailed view, similar to a zoomed-in analysis, of how the matching statistics vary with the two main physical eddy features. This additional representation is useful because, on a linear scale, the histograms are not always able to clearly show the distinct behaviour of each matching statistic as eddy amplitude and radius increase. A logarithmic y-axis is used to distinguish the matching categories in bins containing fewer eddies. This is shown in Fig. C5.

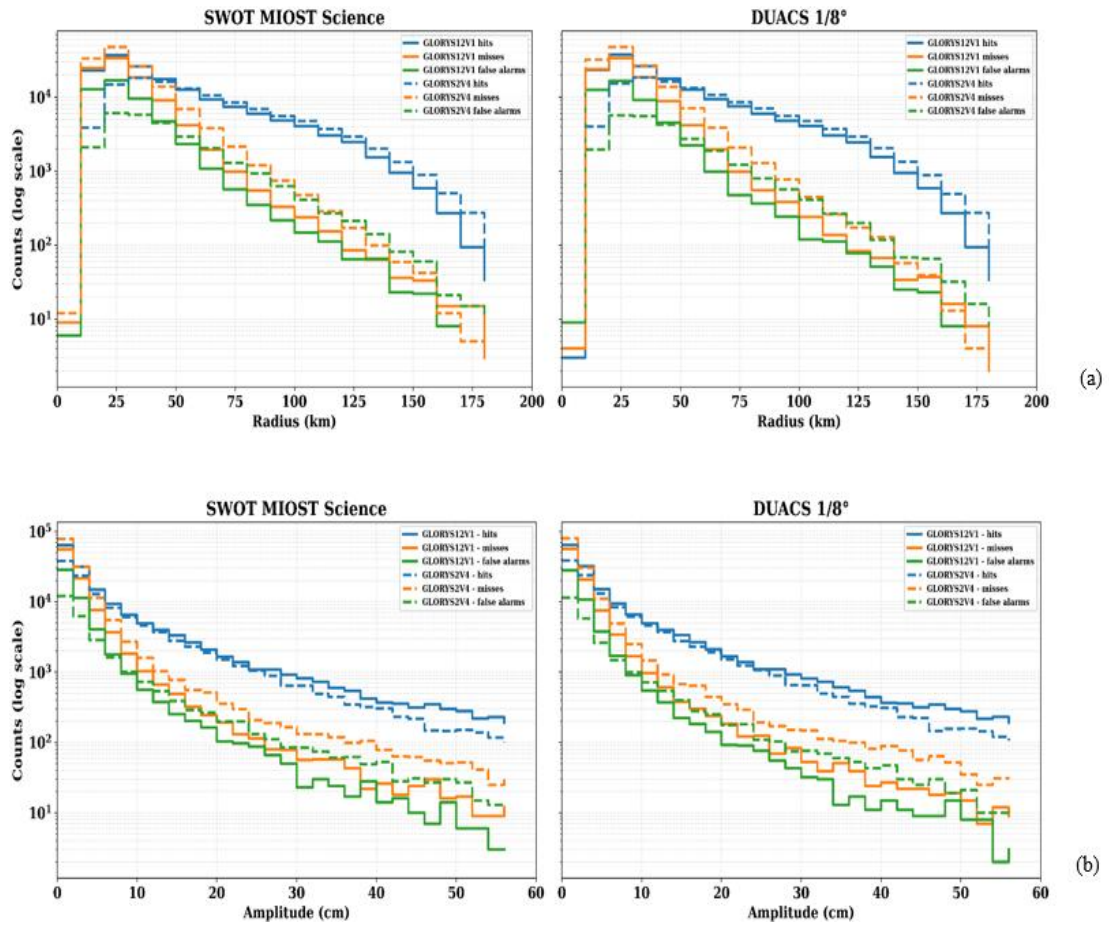


Fig. C5. Histograms for eddy matching statistics on a logarithmic scale, for both comparisons, for eddy radius (a) and eddy amplitude (b).

At the smallest radii and amplitudes, misses are particularly prominent, especially for GLORYS2V4, confirming its greater difficulty in representing small eddies. GLORYS12V1 produces more hits in these classes, although this is accompanied by more small false alarms. For eddy radius, GLORYS2V4 has more misses than GLORYS12V1 up to approximately 150 km, while its false alarms become more numerous above about 50 km. In the amplitude distributions, misses remain the main source of error for GLORYS2V4. Above radii of approximately 50 km and amplitudes of 8–10 cm, hits become the dominant category for both reanalyses and both observational references. The largest discrepancies are therefore concentrated among smaller and lower-amplitude eddies.

Appendix D

This appendix assesses the sensitivity of the verification results to two methodological choices: (i) the maximum centroid distance allowed in the eddy-matching procedure and (ii) the cutoff wavelength of the Bessel high-pass filter applied before eddy detection. The reference values are 125 km for the matching distance and 500 km for the filter cutoff wavelength. The alternative values tested here are 62.5 km for matching and 200 and 800 km for filtering. All other methodological settings, including the 4 days minimum lifetime threshold, remain unchanged.

Sensitivity to a different cost function distance threshold for the eddy matching process: 62.5 km.

The matching analysis was repeated using a maximum centroid distance of 62.5 km, corresponding to half of the reference value of 125 km. This more restrictive criterion allows only eddies with closer centroid positions to be matched. When a

pair accepted with the 125 km threshold no longer satisfies the 62.5 km criterion, the reference eddy is classified as a miss and the corresponding reanalysis eddy as a false alarm.

Figure D1 shows an example for anticyclonic eddies on 1 January 2025. The stricter threshold reduces the number of hits and increases both misses and false alarms, demonstrating the sensitivity of the matching statistics to the maximum distance allowed between eddy centroids.

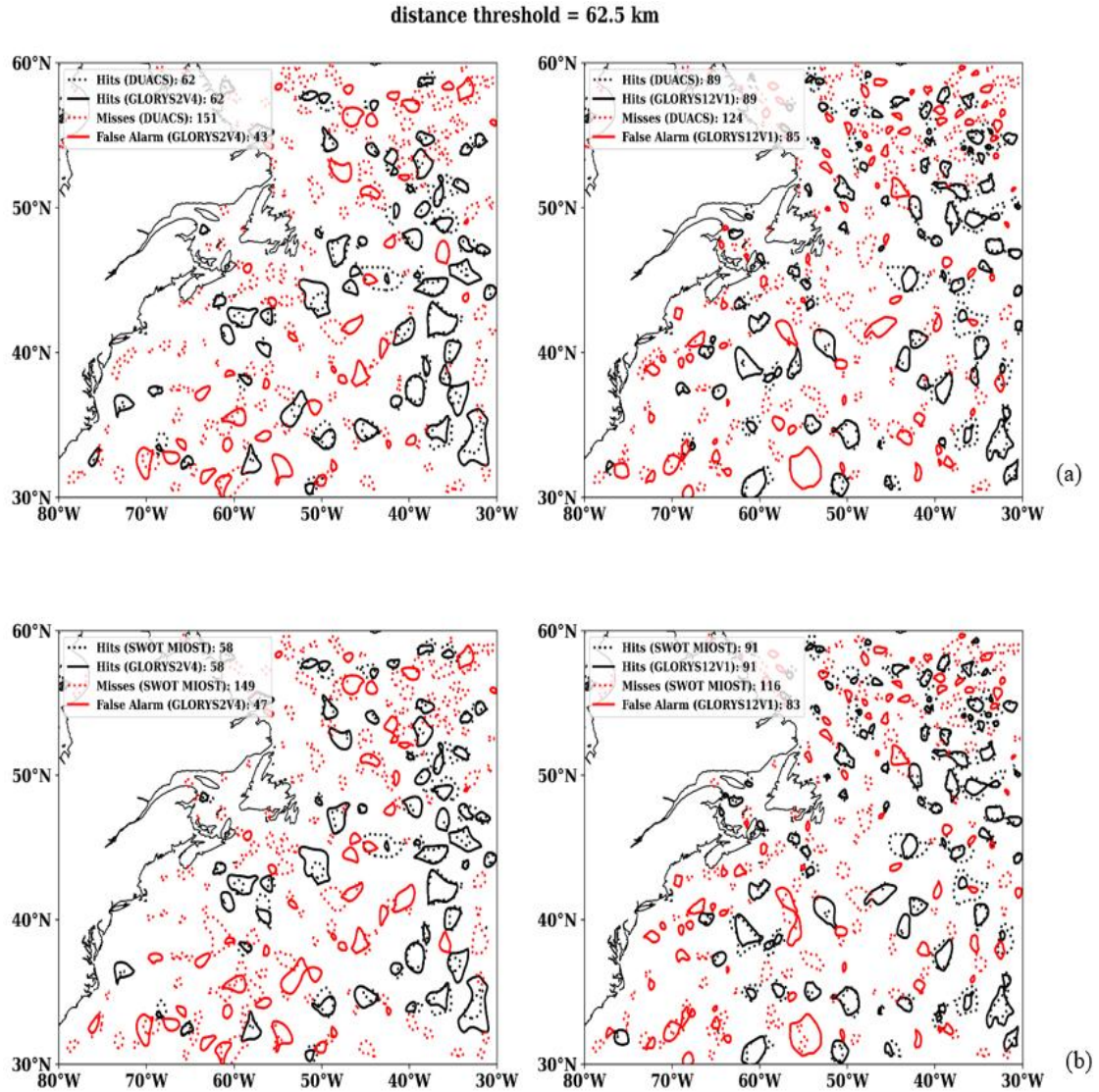


Fig. D1. Anticyclonic-eddy matching on 1 January 2025 using a maximum centroid distance of 62.5 km. GLORYS2V4 is shown on the left and GLORYS12V1 on the right, for the comparisons with (a) DUACS 1/8° and (b) SWOT MIOST Science.

The verification metrics are computed on a daily basis, following the same procedure described in Section 2.4 and Sect. 3.2, in order to evaluate the temporal evolution of the matching performance. Table D1 and Table D2 report the absolute and relative changes in the verification metrics for the two comparisons, respectively.

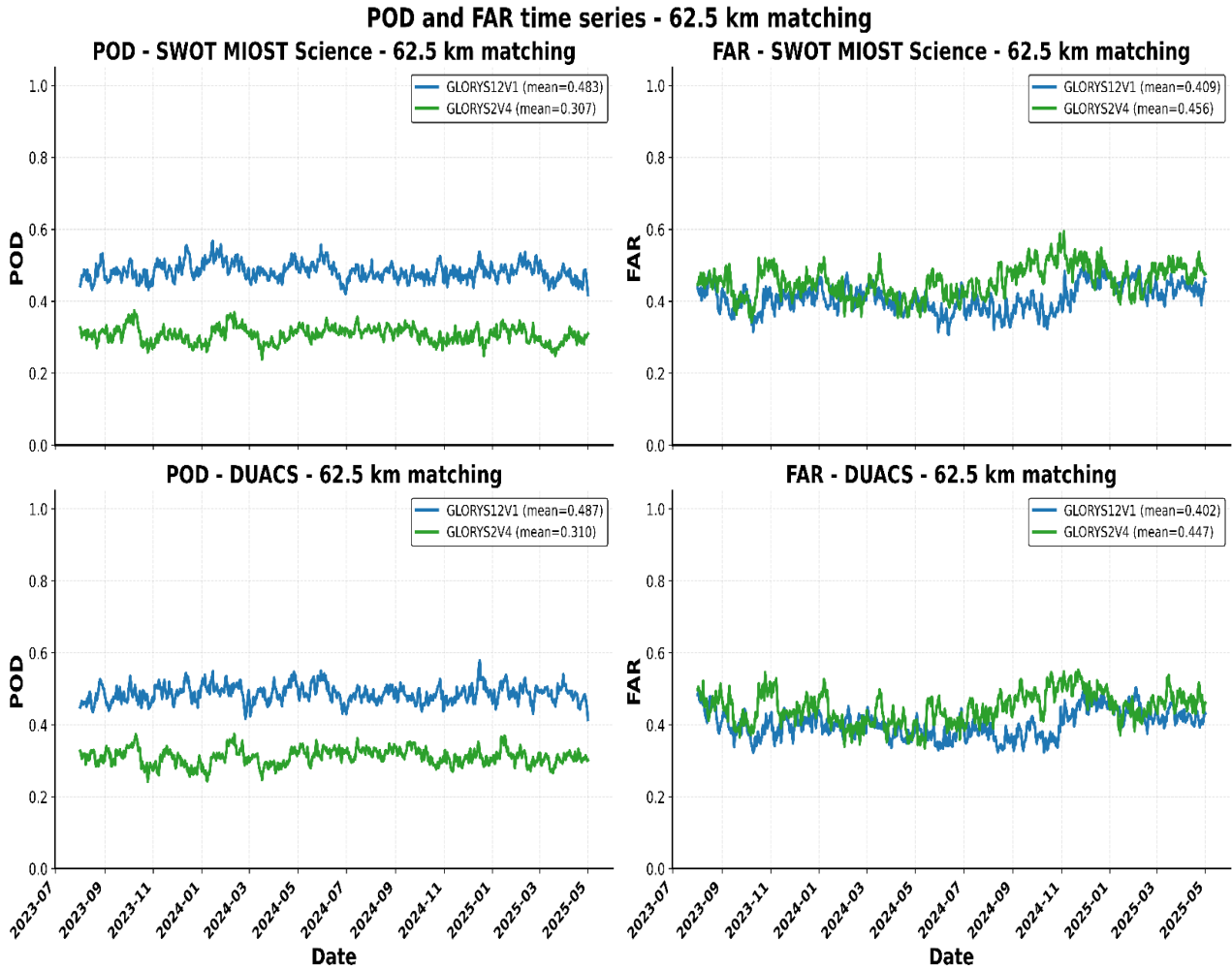
Verification metric	125 km threshold	62.5 km threshold	Difference	Difference %
POD 12V1	0.624	0.483	-0.141	-22.59%
FAR 12V1	0.237	0.409	0.172	72.57%

<i>POD 2V4</i>	<i>0.455</i>	<i>0.307</i>	<i>-0.148</i>	<i>-32.52%</i>
<i>FAR 2V4</i>	<i>0.194</i>	<i>0.456</i>	<i>0.262</i>	<i>135.05%</i>

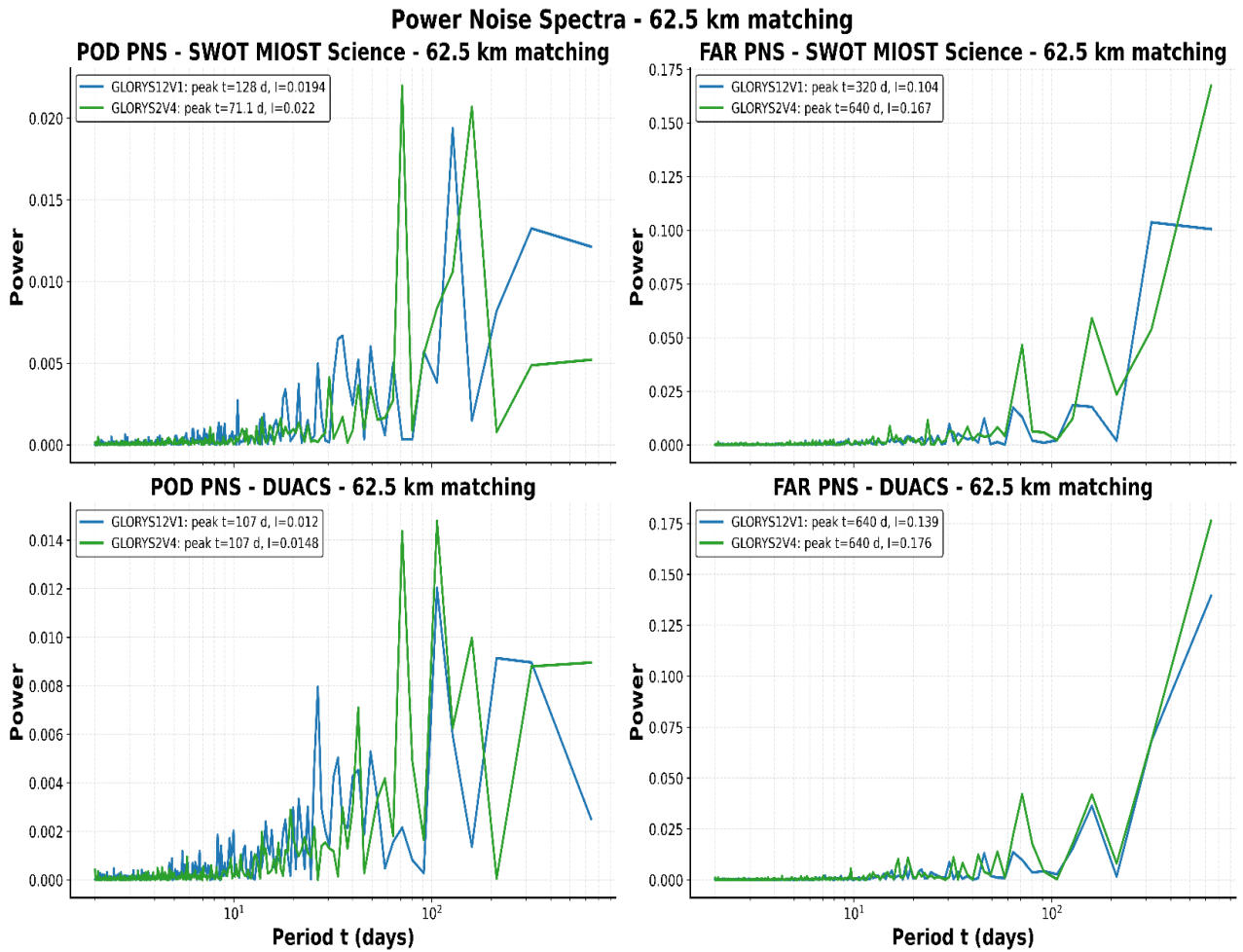
Table D1. Aggregated eddy verification statistics for the Probability of Detection (POD) and False Alarm Ratio (FAR) for the comparison with SWOT using maximum matching distances of 125 and 62.5 km. Absolute and percentage differences are calculated as the 62.5 km result minus the 125 km result.

<i>Verification metric</i>	<i>125 km threshold</i>	<i>62.5 km threshold</i>	<i>Difference</i>	<i>Difference %</i>
<i>POD 12V1</i>	<i>0.628</i>	<i>0.487</i>	<i>-0.141</i>	<i>-22.45%</i>
<i>FAR 12V1</i>	<i>0.230</i>	<i>0.402</i>	<i>0.172</i>	<i>74.78%</i>
<i>POD 2V4</i>	<i>0.460</i>	<i>0.310</i>	<i>-0.150</i>	<i>-32.60%</i>
<i>FAR 2V4</i>	<i>0.182</i>	<i>0.447</i>	<i>0.265</i>	<i>145.60%</i>

Table D2. Aggregated eddy verification statistics for the Probability of Detection (POD) and False Alarm Ratio (FAR) for the comparison with DUACS using maximum matching distances of 125 and 62.5 km. Absolute and percentage differences are calculated as the 62.5 km result minus the 125 km result.



(a)



(b)

Fig. D2. (a) Time series for the verification metrics, for the comparison with SWOT MIOST Science and DUACS 1/8° comparison, with a minimum lifetime threshold of 4 days. In (b) power spectra for all the time series presented in (a).

Overall, when the maximum matching distance is reduced from 125 km to 62.5 km, FAR becomes higher for GLORYS2V4 than for GLORYS12V1, as shown in Table D1, Table D2, Fig. D1 and Fig. D2 (a). This does not necessarily mean that GLORYS2V4 produces a larger absolute number of false alarms. Rather, it reflects the relative nature of FAR, which depends on the proportion of unmatched reanalysis eddies with respect to the total number of reanalysis detections considered in the verification.

In contrast, POD remains higher for GLORYS12V1, indicating that the higher-resolution reanalysis is still able to match a larger fraction of the observed eddies, even under the more restrictive distance criterion. The most evident effect of reducing the matching distance is the strong increase in FAR, especially for GLORYS2V4, where FAR more than doubles in both satellite-based comparisons. A substantial increase is also found for GLORYS12V1, with FAR increasing by more than 70%. POD also decreases when the reduced distance threshold is applied, but the relative reduction is smaller than for FAR. The decrease is of the order of 20% for GLORYS12V1 and about 30% for GLORYS2V4. This indicates that the reduced distance threshold may affect the classification of unmatched reanalysis detections, leading to a large increase in false alarms, while the reduction in correctly matched observed eddies is less pronounced.

The power spectra in Fig. 2D (b) shows slightly larger main POD peaks for GLORYS2V4 than for GLORYS12V1. GLORYS2V4 also exhibits the larger FAR peak in both observational comparisons. In the SWOT comparison, the main POD peaks occur at approximately 71 days for GLORYS2V4 and 128 days for GLORYS12V1, whereas both are located

near 107 days in the DUACS comparison. The strongest FAR peaks occur at longer resolved periods. Because the evaluation period is only approximately 640 days long, peaks at or near 640 days should be interpreted as low-frequency variability rather than as evidence of a robust periodic cycle.

Sensitivity to different Bessel filters for eddy detection process

The SSH or ADT fields of all products were high-pass filtered using a Bessel filter before eddy detection. The reference cutoff wavelength is 500 km. To assess the sensitivity of the detection and verification results to the retained spatial scales, the complete detection, tracking, and matching procedure was repeated using cutoff wavelengths of 200 and 800 km. The 200 km cutoff applies more restrictive spatial filtering, whereas the 800 km cutoff retains a broader range of large-scale variability.

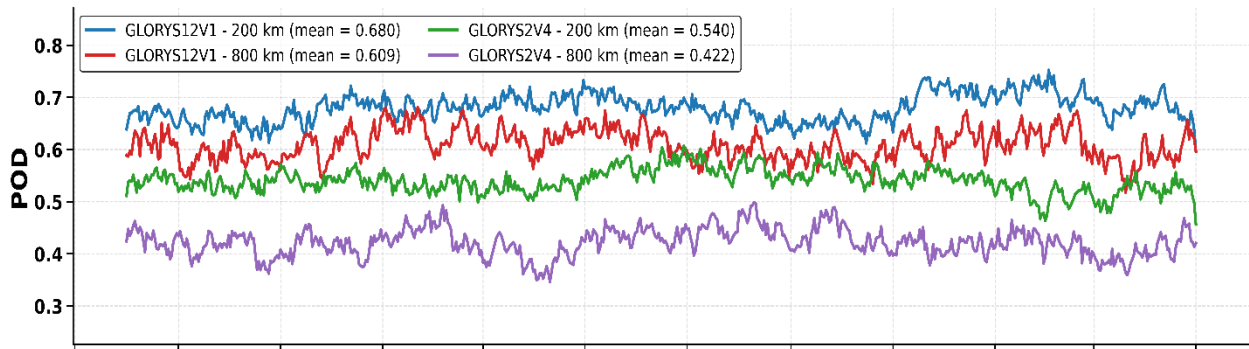
<i>Verification metric</i>	<i>500 km</i>	<i>200 km</i>	<i>Difference</i>	<i>Difference (%)</i>
<i>POD 12V1 (SWOT)</i>	<i>0.624</i>	<i>0.680</i>	<i>0.056</i>	<i>8.97%</i>
<i>POD 12V1 (DUACS)</i>	<i>0.628</i>	<i>0.679</i>	<i>0.051</i>	<i>8.12%</i>
<i>FAR 12V1 (SWOT)</i>	<i>0.237</i>	<i>0.197</i>	<i>-0.04</i>	<i>-16.88%</i>
<i>FAR 12V1 (DUACS)</i>	<i>0.230</i>	<i>0.189</i>	<i>-0.041</i>	<i>-17.83%</i>
<i>POD 2V4 (SWOT)</i>	<i>0.455</i>	<i>0.540</i>	<i>0.085</i>	<i>18.68%</i>
<i>POD 2V4 (DUACS)</i>	<i>0.460</i>	<i>0.541</i>	<i>0.081</i>	<i>17.61%</i>
<i>FAR 2V4 (SWOT)</i>	<i>0.194</i>	<i>0.129</i>	<i>-0.065</i>	<i>-33.51%</i>
<i>FAR 2V4 (DUACS)</i>	<i>0.182</i>	<i>0.118</i>	<i>-0.064</i>	<i>-35.16%</i>

Table D3. Mean eddy verification metrics, difference and percentage difference, crossing from a cutoff wavelength for Bessel filter of 500 km to 200 km for both comparisons.

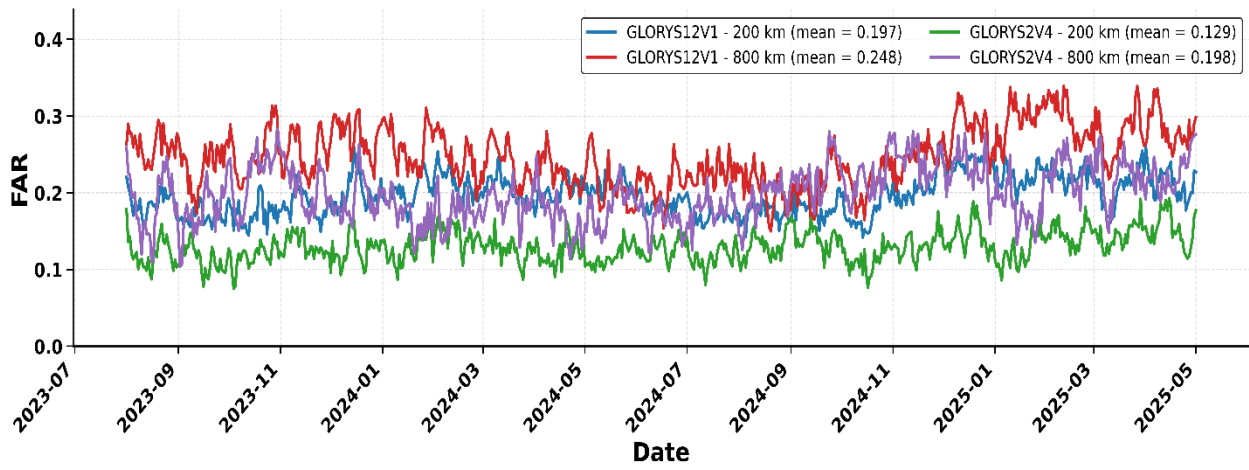
<i>Verification metric</i>	<i>500 km</i>	<i>800 km</i>	<i>Difference</i>	<i>Difference (%)</i>
<i>POD 12V1 (SWOT)</i>	<i>0.624</i>	<i>0.609</i>	<i>-0.015</i>	<i>-2.40%</i>
<i>POD 12V1 (DUACS)</i>	<i>0.628</i>	<i>0.614</i>	<i>-0.014</i>	<i>-2.23%</i>
<i>FAR 12V1 (SWOT)</i>	<i>0.237</i>	<i>0.248</i>	<i>0.011</i>	<i>4.64%</i>
<i>FAR 12V1 (DUACS)</i>	<i>0.230</i>	<i>0.243</i>	<i>0.013</i>	<i>5.65%</i>
<i>POD 2V4 (SWOT)</i>	<i>0.455</i>	<i>0.422</i>	<i>-0.033</i>	<i>-7.25%</i>
<i>POD 2V4 (DUACS)</i>	<i>0.460</i>	<i>0.428</i>	<i>-0.032</i>	<i>-6.96%</i>
<i>FAR 2V4 (SWOT)</i>	<i>0.194</i>	<i>0.198</i>	<i>0.004</i>	<i>2.06%</i>
<i>FAR 2V4 (DUACS)</i>	<i>0.182</i>	<i>0.190</i>	<i>0.008</i>	<i>4.40%</i>

Table D4. Mean eddy verification metrics, difference and percentage difference, crossing from a cutoff wavelength for Bessel filter of 500 km to 800 km for both comparisons.

Probability of Detection - SWOT MIOST Science: 200 vs 800 km Bessel filter

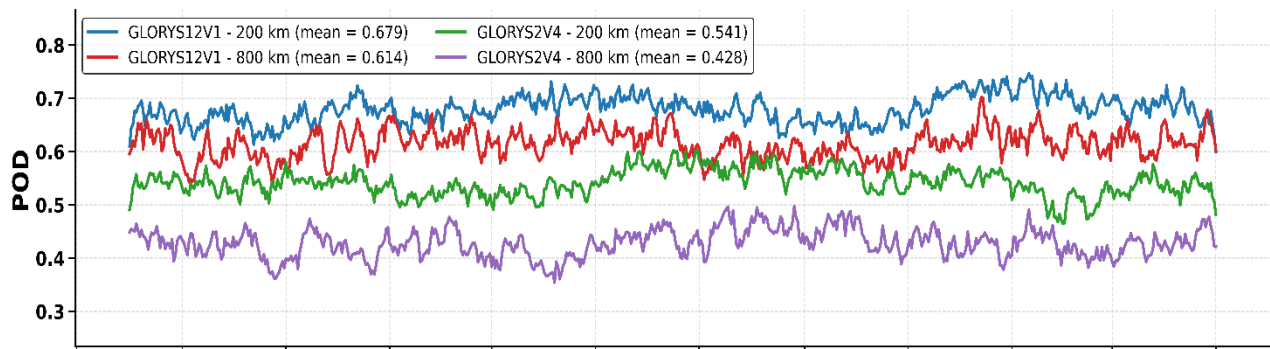


False Alarm Ratio - SWOT MIOST Science: 200 vs 800 km Bessel filter

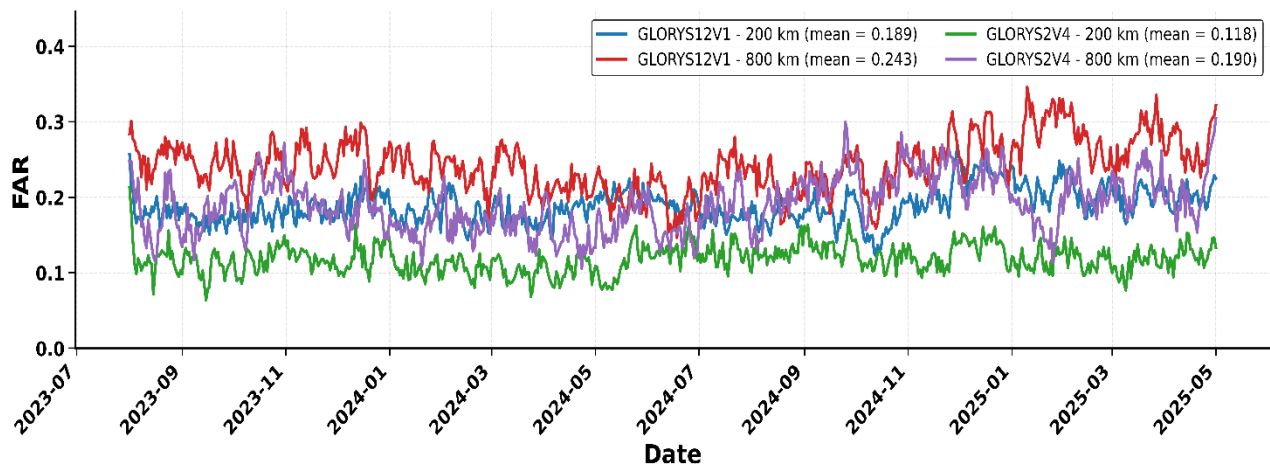


(a)

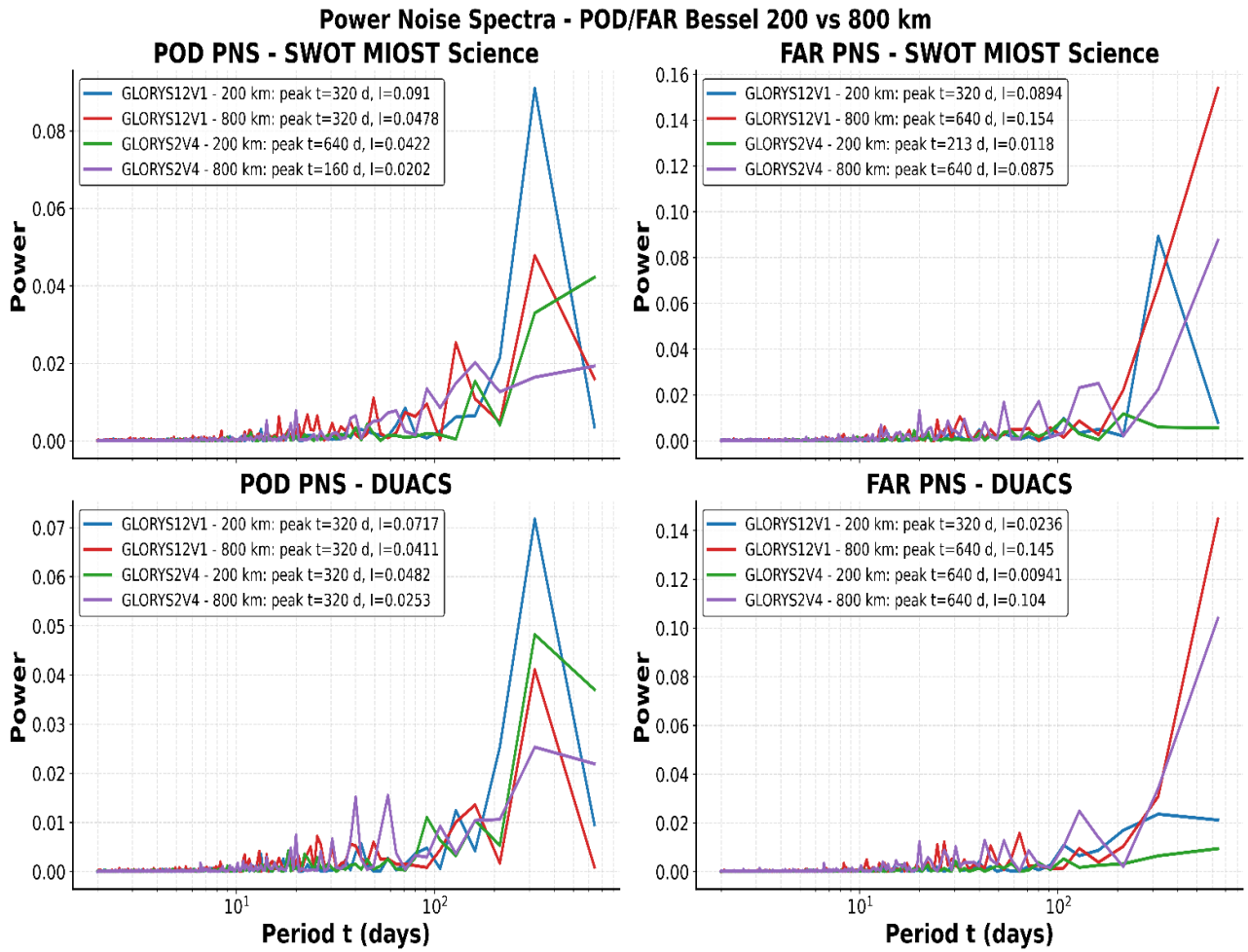
Probability of Detection - DUACS: 200 vs 800 km Bessel filter



False Alarm Ratio - DUACS: 200 vs 800 km Bessel filter



(b)



(c)

Fig. D3. POD and FAR time series obtained using Bessel-filter cutoff wavelengths of 200 and 800 km for the comparisons with (a) SWOT MIOST Science and (b) DUACS $1/8^\circ$. Panel (c) shows the corresponding power spectra. A 4-day minimum-lifetime threshold is used.

Reducing the Bessel-filter cutoff wavelength from 500 to 200 km increases POD and decreases FAR for both reanalyses and both observational comparisons. The changes are particularly pronounced for GLORYS2V4, for which POD increases by approximately 18% and FAR decreases by approximately 34–35%. For GLORYS12V1, POD increases by about 8–9% and FAR decreases by approximately 17–18%. Increasing the cutoff wavelength from 500 to 800 km produces smaller and opposite changes: POD decreases and FAR increases for both reanalyses. The POD reduction is larger for GLORYS2V4, whereas the FAR changes remain relatively small. For both cutoff tests, POD is slightly more sensitive in the SWOT comparison, while FAR is generally more sensitive in the DUACS comparison. These changes show that the verification metrics depend on the spatial scales retained before eddy detection. However, the better verification scores obtained with the 200 km cutoff wavelength do not necessarily demonstrate that this value is physically more appropriate, because stronger filtering may remove structures that would otherwise remain unmatched. Regarding the power spectra, the largest peaks for POD are observed for GLORYS12V1 using the 200 km filter cutoff, at temporal scales of the order of several hundred days, approximately 320 days. For FAR, the highest values are instead found for GLORYS12V1 with the 800 km filter cutoff, with enhanced power around 640 days. At shorter temporal scales, the variability is nearly similar across all datasets and filter configurations. Overall, in terms of noise power, the strongest peak is found for the POD of GLORYS12V1 with the 200 km filter cutoff in the comparison with SWOT MIOST.”

