***Response to RC1***

***l41-l42: Provide references for "recent studies" or rephrase. Aha: The next 3 sentences provide these references.***

Response: Thank you for noting this. As noted, the studies are cited in the immediately following sentences. We have nevertheless re-read the passage to ensure the flow is clear and find the current sequence to be natural.

**Changes in manuscript:** No changes.

***l119: Not sure what an "event-based perspective" could refer to.***

We acknowledge the ambiguity. By “event-based perspective” we meant the qualitative examination of individual eddy case studies presented later in Section 4.

**Changes in manuscript:** We have changed the wording to “case-study perspective” and clarified the sentence accordingly.

***l406: The word "captures" implies that these strain structures are realistic. I think a better word is "contains" since the realism of the structures is unknown.***

We agree with the referee's suggestion and have replaced “captures” with “contains” in the description of the FSLE filamentary patterns. The revised wording avoids implying that the strain structures are fully realistic. At the same time, we wish to clarify that the original use of “captures” was intended to convey a narrower, observation-based point: for the SME case presented (Case 3), 4DvarNet was the only product that reproduced a closed SSH contour and an identifiable eddy core consistent with SWOT L3, thereby “capturing” the existence of the eddy where the other products did not. We have therefore retained “captures” in the part of the discussion that refers directly to the SSH-based eddy identification, while adopting “contains” for the unvalidated FSLE diagnostics.

**Changes in manuscript:** We have replaced “captures” with “contains” in l419.

***l421: 4DvarNET "better preserves fine-scale spatial organization" is totally speculative. A more justified or balanced statement could be, "4DvarNET generates spurious small-scale flow features which look realistic, but which are not supported by any validation data. Hence, dynamical inferences from this product should be avoided."***

We accept the referee’s assessment and agree that the suggested statement more objectively describes 4DvarNet’s performance and its potential issues. We have therefore adopted the referee’s wording in the revised manuscript.

**Changes in manuscript:** The sentence “The nonlinear 4DvarNET framework exhibits enhanced sensitivity to submesoscale structures and better preserves fine-scale spatial organization, whereas QG-constrained and reduced-space linear approaches increasingly smooth or filter these signals.” has been revised to “The nonlinear 4DvarNET framework exhibits enhanced sensitivity to submesoscale structures, whereas QG-constrained and reduced-space linear approaches increasingly smooth or filter these signals. However, this enhanced sensitivity can also generate spurious small-scale flow features that look realistic but are not supported by any validation data. Hence, dynamical inferences from this product should be avoided.”

***l438: How do you know that 4DvarNET exhibits "enhanced sensitivity to fine-scale variability"? Could this not simply be due to spurious structures in the training data or mis-tuning of the assumed noise level?***

We appreciate the referee’s caution. In the original text we used “enhanced sensitivity” merely to describe the larger Rossby number magnitudes and sharper gradients produced by 4DvarNet, not as a claim of physical superiority. However, the wording was easily misinterpreted. We have now clarified that this property could equally reflect noise or artifacts, and that the origin of the variance amplification remains unresolved.

**Changes in manuscript:** The sentence “Complementary strengths emerge in dynamical diagnostics. 4DvarNET exhibits systematically sharper SSHA gradients, larger |Ro| magnitudes, and more intricate strain structures in the FSLE fields, reflecting enhanced sensitivity to fine-scale variability.” has been revised to “Complementary strengths emerge in dynamical diagnostics. 4DvarNET exhibits systematically sharper SSHA gradients, larger |Ro| magnitudes, and more intricate strain structures in the FSLE fields. These characteristics may indicate enhanced sensitivity to fine-scale variability, but they could equally reflect noise amplification or artifacts related to the training data or the assumed noise level; with the available SWOT L3 data alone, their physical realism cannot be determined.”

***l495-l496: omit "as it better preserves ..."***

We agree with the referee that the original phrasing was not sufficiently cautious, particularly regarding strain organization, whose physical realism cannot be assessed with the available validation data. However, the sharpness of SSH gradients was directly compared with SWOT L3 in the case studies (Case 3), where 4DvarNet more closely reproduced the observed gradient magnitudes than the other products. We have therefore retained a qualified reference to SSH gradient sharpness while removing the unsupported claim about strain organization, and have added a note urging caution in the interpretation of these features.

**Changes in manuscript:** The sentence “Conversely, investigations targeting fine-scale frontal structures, filaments, or submesoscale eddy detection may find 4DvarNET more informative, as it better preserves sharp SSH gradients and strain organization.” has been revised to “Conversely, investigations targeting fine-scale frontal structures, filaments, or submesoscale eddy detection may find 4DvarNET more informative, as it more closely reproduced the observed SSH gradient sharpness in Case 3. However, the physical realism of the enhanced strain organization produced by 4DvarNET cannot be assessed with the available validation data and should be interpreted with caution.”

***l502: Why does inclusion of the neural networks "appear particularly promising"? Your results appear to show the opposite is true.***

We thank the referee for this valid criticism. We have adjusted the original text to better separate the current performance of the 4DvarNet product from its conditional future potential, rather than implying that the method is inherently superior based on the present results. In particular, we have clarified that the “promising” aspect refers to a longer-term methodological perspective, not to the overall skill of the current experimental product.

We would like to explain our cautiously positive view of 4DvarNet based on the following considerations. First, observational data for submesoscale processes remain relatively scarce, which means that the training and validation of neural-network-based methods still have considerable room for improvement. For instance, 4DvarNet was the only product able to detect a coherent ~10 km eddy clearly visible in SWOT L3, suggesting that data-driven variational methods can extract fine-scale information that other approaches currently miss; with more high-resolution observations and high-quality model outputs, this capability could be further strengthened. Second, the current 4DvarNet implementation does not incorporate explicit physical constraints, and its spatial discontinuity is amplified in Lagrangian diagnostics, which are particularly sensitive to small-scale velocity inconsistencies. This does not negate the potential of the methodology; rather, it highlights that the present configuration is not yet fully optimized. If such limitations are addressed—for example, by integrating dynamical regularization or improving spatial coherence—data-driven variational methods could complement physics-based approaches in resolving fine-scale dynamics. We have therefore adjusted the original passage to reflect this more balanced view, while emphasizing that any advancement of such methods must be accompanied by rigorous validation.

**Changes in manuscript:** The sentence “The other emphasizes targeted reconstruction of specific dynamical processes, for which hybrid methodologies integrating neural networks with physically based constraints appear particularly promising.” has been revised to “The other emphasizes targeted reconstruction of specific dynamical processes, for which hybrid methodologies integrating neural networks with physically based constraints may hold longer-term methodological potential, rather than a demonstrated advantage based on the present results.”

***Please proofread the references list more carefully. I note some problems in Le Guillou 2025 with a quick spot-check, for example.***

We thank the referee for this observation. We have carefully proofread the entire reference list and corrected several formatting inconsistencies. In particular, we have fixed the title, spacing, and journal name entry to accurately reflect the published version. All other references have also been checked for correctness and consistency with the journal style.

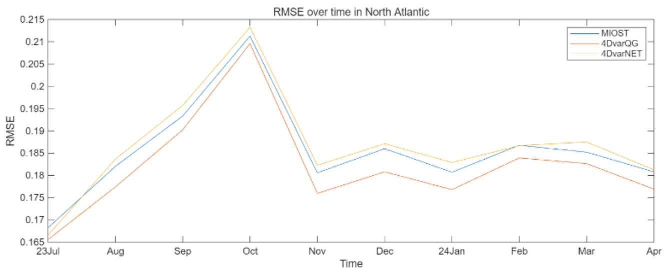
**Changes in manuscript:** Fixed the formatting problems in the reference section.

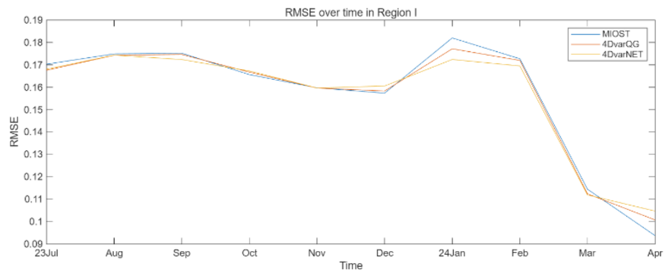
***Response to RC2***

***I do think it could be useful to include even a limited discussion of seasonality: this could entail recreating Figures 3/4 for a couple of different seasons and adding the plots to a supplement or, if seasonal changes in the statistics are negligible, saying that somewhere in the text.***

We thank the referee for this suggestion. In the course of this study, we did explore the possibility of analyzing the temporal or seasonal dependence of the reconstruction errors. However, the error distributions computed over sub-periods or individual seasons exhibited considerable variability that could not be robustly interpreted. This is illustrated in the figure below, which shows the RMSE error distributions for different sub-periods within the study window. The temporal fluctuations may be partly related to changes in drifter availability and spatial sampling: at any given time, the number of drifters within a specific dynamical regime is limited, and their spatial sampling is sparse and uneven (cf. Figure R1). However, we have not conclusively identified the source of this variability, which could also reflect genuine temporal changes in ocean dynamics. Given this uncertainty, we considered that a dedicated seasonal analysis would not provide robust conclusions and chose not to include it in the revised manuscript.

**Changes in manuscript:** No Changes.



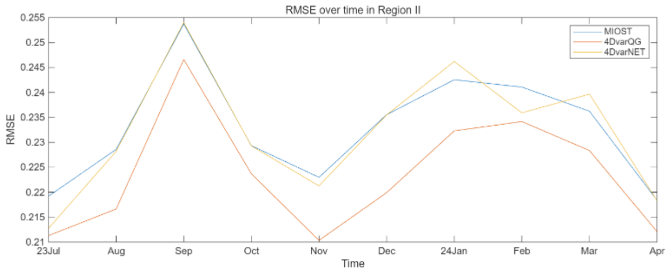



Figure R1

***I also think it could be worth it to tweak the raw SWOT L3 evaluation: are the gradients that the authors use for Ro and FSLE calculated directly on the SWOT L3 SSHA and hence contaminated by waves? It may be useful to apply a filter here to account for that in the SWOT - L4 product comparisons. One could fall back to the spline-filtered product in the v3.0 release of the L3 product, rather than go down a filtering rabbit hole.***

We thank the referee for this helpful suggestion. We have examined the impact of the SWOT L3 version choice on the diagnostics presented in Section 4. For the Rossby number, using the v3.0 spline-filtered fields indeed brings the SWOT L3 reference much closer to the L4 products (Figure R2), whereas with the v2.0.1 fields the L3-derived |Ro| values are nearly an order of magnitude larger. This difference is expected: the v3.0 release computes geostrophic currents and related gradients with the 2D spline-fit methodology of Tranchant et al. (2025), replacing the stencil derivatives used in earlier versions, which effectively imposes additional spatial smoothing on the SSHA field.

Regarding the FSLE, we clarify that it was not computed from SWOT L3 swath data in the original analysis, because the swath geometry does not allow a complete FSLE field to be constructed (values are missing at swath edges). The FSLE fields shown in the case studies are therefore derived from the L4 products only.

We nevertheless chose to keep v2.0.1 as the reference in the main analysis for two reasons. First, the L4 products evaluated here were generated from SWOT L3 v1.0 input data; v1.0 and v2.0.1 belong to the same processing chain in that neither applies the spline-based filtering introduced in v3.0. Using v2.0.1 therefore provides a reference that is consistent with the unfiltered processing paradigm of the L4 inputs, whereas adopting the v3.0 filtered fields as the reference would pre-suppose that the small-scale and potentially unbalanced signals have to be smoothed out before comparison. Second, whether the additional smoothing in v3.0 removes wave-contaminated or true oceanic signals is not a criterion we wish to impose a priori; using the filtered v3.0 as the reference would systematically favour mapping approaches that suppress fine-scale variability (MIOST, 4DvarQG) and penalize methods that attempt to retain such signals (4DvarNet), thereby biasing the intercomparison.

**Changes in manuscript:** We have revised the figure captions and the text to state this explicitly.

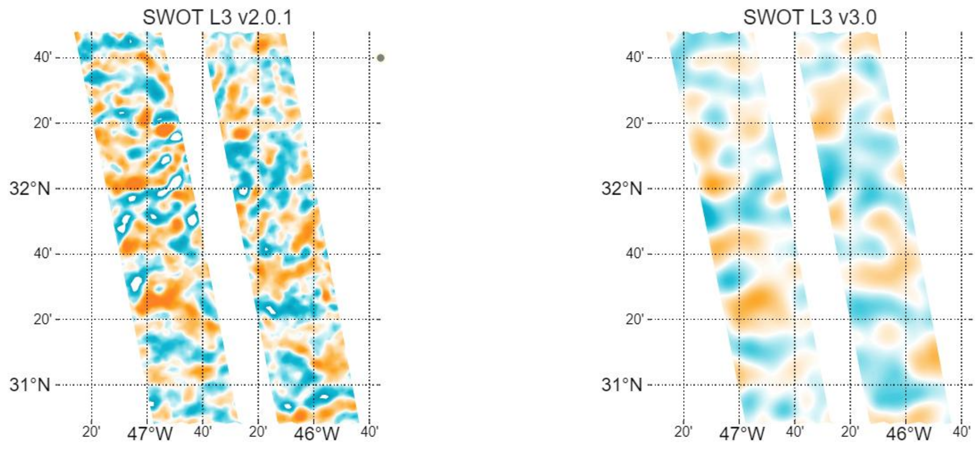



Figure R2

***(l. 15) typo: mprediction -> prediction. I did not find any other obvious typos in the main text but maybe good to double check.***

Thank you for catching this. The typo has been corrected, and we have re-checked the whole manuscript for similar errors.

**Changes in manuscript:** Corrected the spelling error in that location.

***(l. 49-59): there are a couple of other studies the authors can cite wrt the ongoing efforts to understand the SWOT noise / effective resolution and extract velocities from the SWOT data. These include: Archer et al., 2025, Carli et. al., 2025, Wang et al., 2025, and Tranchant et al., 2025.***

We thank the referee for pointing out these relevant studies. Archer et al., 2025 and Wang et al., 2025 were already cited in the original manuscript; we have now added Carli et al., 2025 and Tranchant et al., 2025 to the Introduction, where they support the discussion of SWOT effective resolution and velocity extraction methods. These additions strengthen the context for our L4 product evaluation.

**Changes in manuscript:** We have inserted the following sentences in the second paragraph of the Introduction: "Direct comparisons of SWOT-derived velocities with surface drifters in the Southern Ocean further show that balanced motions dominate SSH down to ~15 km, supporting the derivation of geostrophic currents from SWOT swaths (Tranchant et al., 2025). These fine-scale signals, however, are often attenuated when the swath observations are merged into gridded products (Carli et al., 2025)." Corresponding entries for Carli et al. (2025) and Tranchant et al. (2025) have been added to the reference list.

***(l. 95-109): Can the author provide some context about how the various schemes treat small-scale SSH variability? Are unbalanced motions implicitly filtered out of MIOST and 4DvarQG? A short discussion here could help provide context for the later results.***

We agree that this context is important for interpreting the later dynamical diagnostics, and we have added a short paragraph in Section 2.1 to clarify how each scheme handles small-scale and unbalanced variability. The three approaches differ fundamentally in this respect. MIOST does not filter unbalanced motions through an explicit dynamical constraint; instead, it represents a prescribed set of variability modes—geostrophic mesoscale, barotropic, and equatorial wave components—in a reduced wavelet basis with mode-specific covariances. Variability that falls outside these modeled modes is not represented in the reconstruction, since it lies outside the modeled covariance subspaces. 4DvarQG applies quasi-geostrophic dynamics as a weak dynamical constraint within a reduced-order variational scheme; because the control vector is confined to the QG state space, ageostrophic motions are strongly suppressed rather than explicitly resolved, although the weak constraint allows the solution to deviate from exact QG balance to fit the observations. 4DvarNet is trained end-to-end on 1/60° NATL60 simulations, and its learned prior is not restricted to geostrophic balance; it therefore has, in principle, the capacity to retain small-scale SSH variability of the kind present in the training data. However, since the learned prior does not physically separate balanced from unbalanced components, we cannot establish whether the additional small-scale features produced by 4DvarNet correspond to true ageostrophic motions or to amplified noise; this caution is now stated in the revised text.

**Changes in manuscript:** The paragraph in Section 2.1 was changed to "MIOST is based on a multiscale statistical interpolation framework, in which SSH variability is decomposed into a set of prescribed dynamical components—geostrophic mesoscale, barotropic, and equatorial wave modes—represented in a reduced wavelet space with mode-specific covariances. This formulation allows for scale-dependent covariance modeling and provides a robust reconstruction under heterogeneous sampling conditions. Because no explicit dynamical constraint is imposed, variability that falls outside these modeled modes is not reconstructed and is effectively damped by the scale-dependent covariance regularization. The MIOST product provides near-global coverage (80°S–90°N, 0°–360°E) at a spatial resolution of 1/10° with daily temporal sampling (Ballarotta et al., 2025).

4DvarQG relies on a four-dimensional variational assimilation scheme constrained by a reduced-order quasi-geostrophic (QG) dynamical model. By introducing physically grounded dynamical constraints, this approach aims to reconstruct dynamically consistent SSH and surface current fields while mitigating limitations associated with SWOT's sparse temporal sampling. The 4DvarQG product is currently available over the North Atlantic domain (25°N–50°N, 80°W–10°W), with a spatial resolution of approximately 1/10° and a temporal resolution of 6 hours (Le Guillou et al., 2025).

4DvarNet adopts a data-driven variational framework in which both the dynamical prior and the assimilation operator are learned using deep neural networks. The model is trained within an observing system simulation experiment (OSSE) framework based on high-resolution NATL60 (1/60°) simulations, with the objective of capturing nonlinear and non-Gaussian characteristics of fine-scale ocean variability. Because its learned prior is not restricted to geostrophic balance, 4DvarNet can in principle retain small-scale unbalanced variability; however, it does not physically separate balanced from unbalanced components, and the realism of such features cannot be established from the available validation data. The current implementation focuses on the same North Atlantic domain as 4DvarQG and provides daily fields on a 1/8° grid (Beauchamp et al., 2023)."

***(l. 115-120): Small quibble with the SWOT L3 version - could the author specify whether they are using the "Smoothed" or "Unsmoothed" version of the v2.0.1 SWOT SSH data?***

We have clarified this in the revised manuscript. In the original analysis we used the Expert sub-product of the SWOT L3_LR_SSH v2.0.1 dataset (SSHA filtered), which is provided on the 2-km grid rather than the 250-m native-grid Unsmoothed sub-product. We note that "Expert" refers to the extended content of the product (additional layers such as sigma0, MSS, absolute geostrophic currents, quality flags, and corrections) and is not itself a filtering designation; the SSH field was used at its delivered 2-km resolution without any additional spatial smoothing.

**Changes in manuscript:** The sentence "To facilitate a qualitative assessment of fine-scale dynamics, individual surface features are identified using SWOT Level-3 (L3) Expert SSH data (https://doi.org/10.24400/527896/A01-2023.018), which provide two-dimensional SSH fields at 2 km × 2 km resolution." has been revised to "To facilitate a qualitative assessment of fine-scale dynamics, individual surface features are identified using the SWOT Level-3 (L3) KaRIn low-rate SSH Expert product, version 2.0.1 (https://doi.org/10.24400/527896/A01-2023.018), which provides two-dimensional SSH fields on the 2 km × 2 km delivered grid."

***(l. 115-120): Release v3.0 of the L3 data includes SSH gradients calculated on spline-filtered fields ... Presumably the authors used v2.0.1 for consistency with the various L4 products, but I'm curious if the using the v3.0 filtered fields improves the agreement in Section 4. Could be worth it to take a look at this (for example Figures 9-12) and make a note of any difference or lack thereof in the text.***

We thank the referee for this suggestion. For consistency with the L4 product evaluation, we chose v2.0.1 as the L3 reference, as the L4 inputs were generated from an earlier processing chain without the spline-based gradient filtering introduced in v3.0. We have nonetheless examined the v3.0 filtered fields: they do bring the L3-derived gradients and Rossby numbers closer to the L4 products, as expected from the additional smoothing. However, using them as the reference would systematically favour products that suppress fine-scale variability (MIOST, 4DvarQG) while penalizing those that retain it (4DvarNet).

***(l. 125-128): Have the authors experimented with correcting for the Ekman component as in Tranchant (2025)?***

We thank the referee for raising this point. We agree that removing the Ekman component is meaningful for reducing representativeness errors in velocity validation. In the present study, however, the influence of the Ekman contribution on the relative comparison between products is largely eliminated by the way the metric is constructed: the gain/loss ratio is based on the difference between two RMSE values, and the Ekman-related error, which is common to all L4 products under the same drifter observations, is substantially removed during this subtraction step. The relative inter-product comparison is therefore expected to be far less sensitive to the Ekman signal than the absolute RMSE values themselves. We nonetheless acknowledge that the absolute RMSE estimates would benefit from an explicit Ekman correction, and we plan to explore this in future work.

**Changes in manuscript:** No changes.

***Figures 3-4: (very minor suggestion) it may be useful to include the region boundaries on one or more of the subplots in each figure, if there is a way to do that cleanly.***

A good suggestion. We have added the region boundaries to the RMSE and gain/loss maps in Figures 3 and 4 using thin black or white contours, which does not interfere with the color shading. This makes the spatial interpretation of the regional differences easier.

**Changes in manuscript:** Figures 3 and 4 have been redrawn, and the corresponding captions have been revised.

***Section 3.1: it would be helpful to see a plot or plots of the performance changes binned across the different regions that the author is describing verbally in the text. Scatter plots or histograms showing the spread in RMSE or gain/loss ratios over all of the pixels in each region would be interesting. It may be ok to exclude Region 1 here, or push these plots to an appendix or supplement.***

We thank the referee for this suggestion. Visualizing the spread of the performance metrics within each region will add useful quantitative support to the regional comparisons described in the text. We will add such plots to the appendix in the revised manuscript. In addition, we have identified some errors in the statistical information presented in Figure 3 of the original manuscript, which we will also correct in the revised version.

**Changes in manuscript:** Figure 3 has been redrawn, and histograms of the RMSE PDFs have been added to the figure.

***(l. 252-289, Figure 5, Figure 6): This is a really interesting section, I'm grateful for the authors contribution here. I wonder if it would be possible to include some discussion of the spread in prediction skill for the various methods across regions and/or seasons (maybe push to a supplement). It would be interesting to see the standard deviation in Gain/Loss ratios for each method overlaid on Figure 6 as shaded envelopes, for example.***

We thank the referee for this positive feedback and for the thoughtful suggestion to examine the spread in prediction skill across regions and seasons. We did explore such refinements during the course of the analysis. Spatially, individual drifter trajectories may cross from one dynamical region to another, which makes a clean region-based attribution of the prediction errors difficult. Temporally, the limited sample sizes in both the Eulerian and Lagrangian error estimates mean that the statistics are highly sensitive to random factors, such as the start or interruption of individual drifter records. Overall, a further breakdown in space and time did not lend itself to a clear causal interpretation, and we therefore chose not to pursue this decomposition of the total error.

**Changes in manuscript:** No changes.

***(l. 324-329): I think the author should expand on this discussion of the role of the potential role of the training dataset in 4DvarNET's accuracy either here or in the conclusions. If I recall correctly NATL60 does not include large-scale tidal signals, and seems to underrepresent SSHA signal from fast-moving intertia-gravity waves. Could this deficiency play a role in it's over-representation of Ro? By eye it seems like the highest gain regions for 4DvarNET are in the more wave-dominated region outside of the Gulf Stream.***

We thank the referee for this perceptive remark. The suggestion that the NATL60-based training data may influence 4DvarNet's tendency to over-represent Rossby number signals is well founded. Indeed, the standard NATL60 configuration does not include explicit tidal forcing, and fast-moving inertia-gravity waves are expected to be only partially represented at the model resolution and sampling timescales relevant here. These deficiencies could plausibly cause the learned mapping to misattribute SSH variance arising from unresolved wave-like variability to vortical structures, which would be consistent with the spatial pattern of the largest 4DvarNet Rossby number enhancements outside the energetic Gulf Stream. We have expanded the discussion in the revised manuscript to present this interpretation, while being careful to frame it as a plausible explanatory hypothesis rather than a demonstrated causal mechanism, because the available independent observations are insufficient to definitively separate balanced and unbalanced contributions at these scales. We have also added a cautionary note to alert potential users of 4DvarNet that Rossby-number-based diagnostics in weakly energetic, wave-influenced regions should be interpreted with particular care.

**Changes in manuscript:** We have inserted the following sentences in section 4:"Notably, the NATL60 configuration used for training does not include explicit tidal forcing and only partially represents fast-moving inertia-gravity waves, so unresolved wave-like SSH variance may be misattributed to vortical structures. This is consistent with the largest $|R_o|$ enhancements occurring in wave-influenced regions outside the energetic Gulf Stream. As such, $|R_o|$-based enhancements in dynamically quiescent regimes should be interpreted with caution, particularly in wave-dominated environments."