July 8, 2026

Dear Referee:

Thank you very much for your careful evaluation of our manuscript **egusphere-2025-6389** entitled "A Continuous Implicit Neural Representation Framework with Gradient Regularization for Sea Surface Height Reconstruction From Satellite Altimetry". We are grateful for your constructive comments and suggestions, which helped us improve the clarity and balance of the manuscript. We have revised the manuscript accordingly, and provide below a point-by-point response to your comments. The corresponding changes in the manuscript are marked blue.

With kind regards,

Liming Pan
(on behalf of co-authors: Dongshuang Li, Liming Pan, Zhaoyuan Yu, and Linwang Yuan)

Point-by-point response to the First Referee

Report of the First Referee

General comment: "*The authors investigate the potential of an Implicit Neural Representation (INR) framework for reconstructing Sea Surface Height (SSH) fields from sparse altimetry observations. The approach is evaluated in two regional configurations: a realistic setting over the western Mediterranean basin and a simulation-based experiment in the Gulf Stream region. The study uses existing data challenge frameworks to benchmark the proposed method against established approaches (e.g. DUACS OI, BFN, 4DVARNET, etc.). The results suggest improved reconstruction performance with INR in both Observing System Experiments (OSE) and Observing System Simulation Experiments (OSSE), highlighting the potential of the method for recovering SSH fields from incomplete satellite measurements.*"

Overall, the manuscript is relatively well written, and the results are clearly presented. The study is relevant and the proposed approach appears promising. That said, several aspects could benefit from additional clarification and further discussion to strengthen the manuscript prior to publication in "Geoscientific Model Development". For this reason, I recommend the manuscript should be reconsidered after major revision. I outline below several points that the authors may wish to consider.

Response: We thank the referee for the positive assessment and for the constructive comments. We have revised the manuscript with two main aims: to make the experimental design easier to follow, and to make the interpretation of the INR results more balanced.

- *Experimental design and validation.* We now describe the OSE and OSSE settings separately and state the data roles directly in the text. This clarifies that the OSE experiment uses non-CryoSat-2 observations for training and withheld CryoSat-2 along-track observations for validation, whereas the OSSE experiment uses sparse pseudo-observations as training input and the full NATL60 field as the evaluation reference.

- *Method description and regularization.* We revised the description of the SIREN-based INR, the coordinate normalization, the data-fidelity term, and the TV regularization term. We also clarified that the regularization coefficient a is chosen once from preliminary diagnostic checks based on the available benchmark error diagnostics and inspection of the reconstructed SSH and gradient fields, and then kept fixed for all locations and times in a given experiment. The notation inconsistency involving λ has been corrected.

- *Additional diagnostics.* We added an OSSE spatio-temporal error analysis, including instantaneous error maps, a spatially averaged RMSE time series, and a time-mean RMSE map. We also added a no-TV ablation experiment, a PSD-score figure used to interpret the effective-scale diagnostics, and qualitative comparisons with the benchmark OI reconstruction.

- *Figure readability and interpretation.* We revised the error-diagnostic figures and captions, added coastlines to the spatial RMSE panels, and made the PSD-score and effective-scale language more cautious. In particular, the reported effective scales are now described as PSD-score-based diagnostics, not as nominal grid spacings or proof that all fine-scale SSH structures are resolved.

- *Limitations.* We expanded the discussion to state more clearly the limits of the present framework, including its dependence on observation coverage, the trade-off introduced by regularization, the different validation meanings of OSE and OSSE, and the absence of uncertainty quantification in the current implementation.

The detailed point-by-point responses are provided below.

Comment 1: "*The experimental design section would benefit from additional detail. In particular, it would help to more clearly distinguish between the OSE and OSSE frameworks and to explicitly describe the associated validation protocols. For instance, in the OSE configuration, the reconstructed fields are evaluated against independent along-track CryoSat-2 data, whereas in the OSSE case, validation relies on the full 4D SSH field from the numerical simulation used as reference.*"

Response: We thank the referee for pointing out this ambiguity. We have revised the experimental-design subsection to separate the two benchmark settings more clearly.

For the OSE experiment in the western Mediterranean Sea, the revised text now states that the model is trained with real along-track observations from operational nadir altimeters, while CryoSat-2 new-orbit tracks are kept out of the training data and used only for validation. Because no complete true SSH field is available in this real-ocean case, the evaluation is carried out only at the locations and times of the withheld CryoSat-2 observations.

For the OSSE experiment in the Gulf Stream region, we now describe the validation setting separately. The complete NATL60 SSH field is used as a known four-dimensional reference. For the reported INR reconstruction, sparse pseudo-observations over the target period are used as training input, and the reconstruction is evaluated against the full gridded NATL60 field over the same period. We also revised the text to state the data roles directly, rather than presenting them as a conventional training–test split.

These changes make clear that the OSE provides withheld along-track validation with real observations, whereas the OSSE provides a controlled full-field validation against a known numerical reference. The corresponding revisions in Section 5.1 (Experimental configurations and evaluation protocols) are as follows.

- The experiments are conducted in two study domains: the Western Mediterranean Sea (MEDIT) and the Gulf Stream region (GF), as shown in Figure 1. The two domains correspond to different Ocean Data Challenges benchmarks: the SSH MapMed OSE benchmark for real observations (Ocean Data Challenges, 2023) and the NATL60 OSSE benchmark for controlled evaluation with a known reference field (Ocean Data Challenges, 2020). The datasets used here have been

archived at Zenodo to ensure accessibility and reproducibility (Li, 2025a), and the experiments were implemented using the publicly available code (Li, 2025b).

Figure 1: Geographical domains used in the experiments: Western Mediterranean (MEDIT) and Gulf Stream (GF) regions.

In the OSE configuration, the experiment is conducted over the Western Mediterranean Sea within the domain $[1°\mathrm{E}, 20°\mathrm{E}] \times [30°\mathrm{N}, 45°\mathrm{N}]$. This experiment is designed to assess the reconstruction method under realistic satellite sampling and measurement conditions. The INR model is trained with sparse real along-track observations from the nadir missions available for mapping, excluding CryoSat-2 new-orbit observations, from Jan. 1 to Mar. 15, 2021. CryoSat-2 new-orbit observations from Jan. 15 to Mar. 15, 2021 are withheld from training and used only for evaluation. Thus, the OSE split is mission-based rather than a fully time-disjoint split. Because no complete true SSH field is available in this real-ocean setting, the evaluation is performed only at the locations and times of the withheld CryoSat-2 observations.

In the OSSE configuration, the experiment is conducted over a dynamically active portion of the Gulf Stream region within the domain $[65°\mathrm{W}, 55°\mathrm{W}] \times [33°\mathrm{N}, 43°\mathrm{N}]$. This experiment provides a controlled setting in which the reconstructed SSH fields can be evaluated against a known reference field. The complete four-dimensional SSH field from the NATL60 numerical simulation is used as the reference field. For the reported INR reconstruction, pseudo-altimetric nadir and SWOT observations over Oct. 22–Dec. 2, 2012 are generated by subsampling this reference field and are used as sparse input data. The reconstruction is then evaluated against the full gridded NATL60 SSH field over the same target period.

These data roles should not be interpreted as a conventional supervised training–test split with fully independent ocean states. They follow the corresponding Ocean Data Challenges benchmark rules. In the OSE case, the split is mission-based: non-CryoSat-2 nadir observations are used for training, and CryoSat-2 new-orbit observations are withheld for along-track validation. In the OSSE case, sparse pseudo-observations over the target reconstruction period are used

as training input, while the full NATL60 field over the same period is used only as the known gridded reference for evaluation.

Comment 2: "*Including a qualitative assessment in the main text could further support the conclusions. As the manuscript highlights the ability of the method to resolve finer-scale structures, illustrative examples (e.g., snapshots showing sharp gradients or mesoscale features) together with comparisons to other approaches would be valuable. This could be particularly informative in the OSSE framework, potentially complemented by diagnostics based on spatial derivatives.*"

Response: We thank the reviewer for this constructive suggestion. We agree that qualitative examples are useful here, especially in the OSSE framework where the NATL60 reference field is available. We therefore added two qualitative comparisons with the reproducible OI baseline. The broader comparison with multiple methods is kept in the quantitative benchmark table, while the field-by-field figures are shown for the four-nadir setting because the benchmark provides reproducible gridded OI outputs for that case.

The first comparison shows the NATL60 reference SSH field, the OI reconstruction, the proposed INR reconstruction, and the corresponding SSH reconstruction error maps. This provides a direct visual assessment of how the two methods represent mesoscale SSH structures and regions with strong spatial variability.

In addition, we added a derivative-based qualitative comparison using the SSH gradient magnitude ($|\nabla\eta|$). We compare the gradient magnitude computed from the NATL60 reference field, the OI reconstruction, and the INR reconstruction, and further show the corresponding gradient diagnostic errors relative to the NATL60 reference. This diagnostic complements the SSH snapshot comparison by showing whether rapid spatial transitions, such as fronts and eddy boundaries, are represented coherently.

We have also revised the related interpretation to avoid over-emphasizing the recovery of unconstrained fine-scale structures from sparse observations. In the revised manuscript, the gradient-based diagnostics are used to assess whether the dominant high-gradient structures are represented coherently, and where the main gradient discrepancies occur relative to the NATL60 reference.

The corresponding revisions in Section 5.5.5 (Comparison of Multiple Methods for SSH Reconstruction) are as follows.

- In addition to the quantitative comparison above, we provide a qualitative comparison with the benchmark OI baseline. The comparison is shown for the four-nadir setting, for which reproducible gridded OI outputs are provided by the benchmark. This allows the spatial reconstruction patterns, reconstruction errors, and derivative-based diagnostics of OI and INR to be compared on the same input observations and evaluation grid.

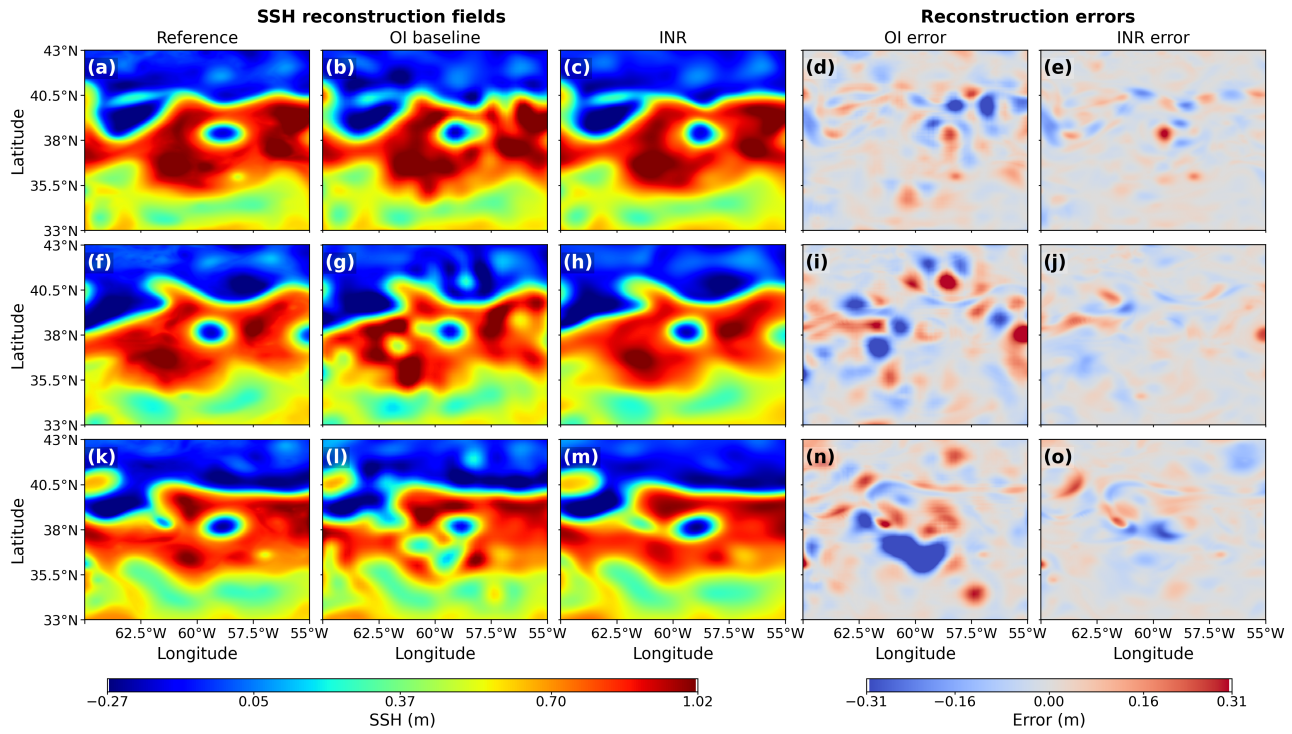

Figure 2: Qualitative comparison between the OI baseline and INR under the four-nadir observation setting. From top to bottom, the rows correspond to three representative evaluation times selected according to the 0.1, 0.5, and 0.9 quantiles of the OI SSH reconstruction error, respectively. The first three columns show the NATL60 reference SSH field, the OI baseline reconstruction, and the INR reconstruction, respectively. The last two columns show the corresponding SSH reconstruction errors of the OI baseline and INR relative to the NATL60 reference.

Figure 2 compares the NATL60 reference SSH field, the OI baseline reconstruction, and the INR reconstruction at three representative evaluation times. The rows correspond to cases selected according to the 0.1, 0.5, and 0.9 quantiles of the OI SSH reconstruction error, respectively. Both methods reproduce the broad SSH distribution from sparse along-track observations, while noticeable differences appear in dynamically active regions with strong spatial variability. The error maps show that the largest residuals tend to occur near strong-gradient frontal and eddy-like features visible in the reference, indicating that these regions remain challenging under sparse observational constraints. Compared with the OI baseline, the INR error maps in these selected cases contain fewer broad high-amplitude residual patches and show a more spatially continuous SSH pattern. This qualitative comparison complements the quantitative metrics by showing where the main residual patterns occur in the selected OSSE examples.

To further examine the behavior of the reconstructed derivative fields, we compare the SSH gradient magnitude derived from the NATL60 reference field, the OI baseline reconstruction, and the INR reconstruction. This derivative-based diagnostic complements the SSH snapshot comparison by focusing on fronts, eddy boundaries, and other high-gradient regions. The same

three evaluation times as in Figure 2 are used for consistency.

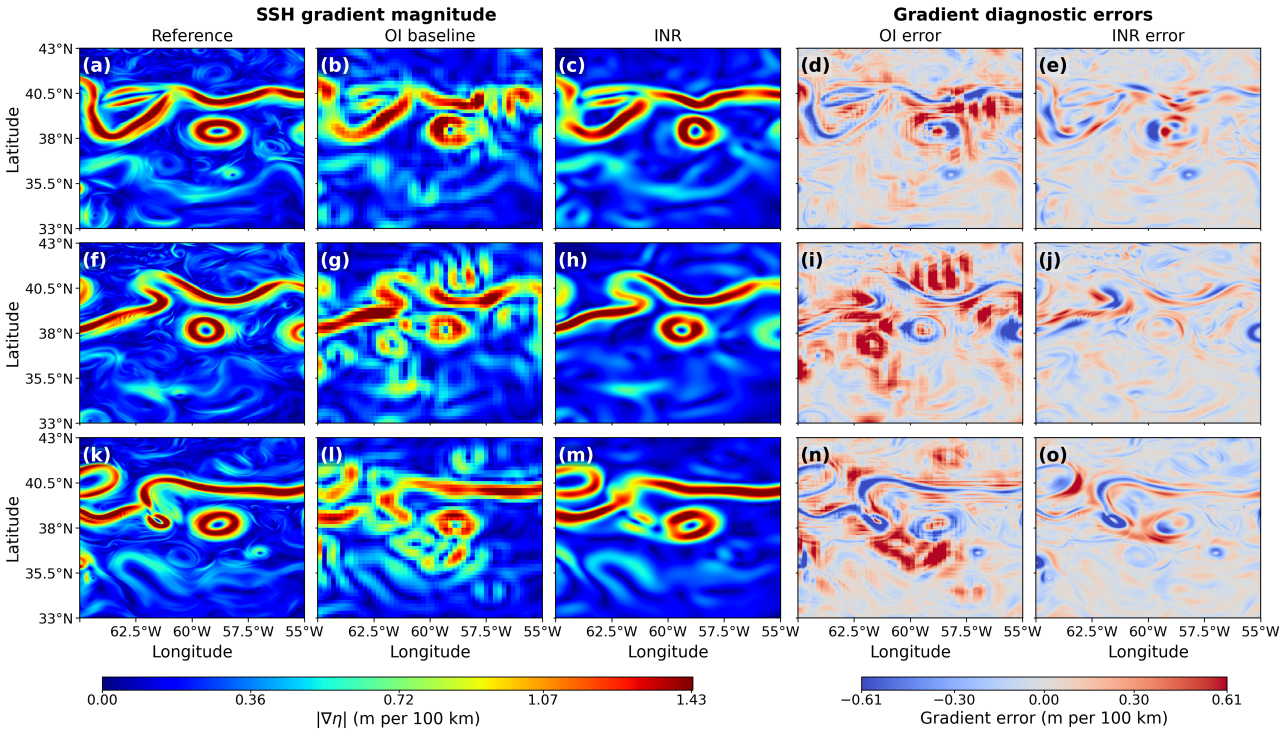

Figure 3: Derivative-based qualitative comparison between the OI baseline and INR under the four-nadir observation setting. From top to bottom, the rows correspond to three representative evaluation times selected according to the 0.1, 0.5, and 0.9 quantiles of the OI SSH reconstruction error, respectively. The first three columns show the SSH gradient magnitude $|\nabla\eta|$ computed from the NATL60 reference field, the OI baseline reconstruction, and the INR reconstruction, respectively. The last two columns show the corresponding gradient diagnostic errors of the OI baseline and INR relative to the NATL60 reference.

Figure 3 shows the corresponding gradient-magnitude fields and gradient diagnostic errors. The OI-derived gradients are comparatively diffuse and show some small-scale artifacts in these examples, whereas the INR-derived gradients follow the main frontal and eddy-like bands more continuously. The diagnostic error maps also show that sharp fronts and eddy boundaries remain the main sources of gradient discrepancies for both methods, partly because small spatial displacements can lead to pronounced derivative errors. Some very narrow gradient filaments in the NATL60 reference are absent from both reconstructions, as expected under sparse satellite-like sampling. We therefore use this comparison mainly to assess whether the dominant high-gradient structures in the reference are represented coherently in the reconstructed fields.

Comment 3: "*The discussion section could be expanded to provide additional perspective on the method and results. In particular, it may be useful to elaborate on: (i) the main limitations of the approach, including the method, training, and testing; (ii) how it differs from other existing approaches;*

(iii) its potential applicability at the global scale; (iv) the associated training strategy and computational requirements; (v) the feasibility of its use in an operational context; and (vi) explicit references to the data challenge frameworks used in this study."

Response: We agree that the Discussion needed to do more than restate the scores. We reorganized Section 6 (Discussion) around the points raised by the reviewer: what the method is, how it differs from existing mapping approaches, when it is expected to be useful, and what is still missing for global or operational use.

For the relation to existing approaches, the revised Section 6.1 (Methodological characteristics and relation to existing approaches) now makes the comparison with OI, model-based assimilation, and supervised deep-learning methods explicit:

- The proposed framework differs from conventional SSH mapping approaches mainly in how the field is represented. OI-based products estimate values on a predefined grid using prescribed covariance models, while model-based assimilation methods introduce explicit dynamical constraints. By contrast, the INR model learns a continuous coordinate-to-SSH mapping from the available observations. Once optimized for a target reconstruction problem, this mapping can be evaluated at requested coordinates within the reconstruction domain and period, and its differentiability supports gradient diagnostics. Compared with supervised deep-learning approaches such as 4DVarNet-type methods, it does not require a large set of paired sparse-observation and full-field training examples. Instead, its parameters are optimized from the available observations for the target reconstruction problem. This formulation is flexible, but it remains a reconstruction model constrained by observations and regularization, not a replacement for full dynamical data assimilation. OI and related analytical approaches remain useful because their assumptions are explicit and, in the case of OI, can provide uncertainty estimates associated with the prescribed covariance model. The optimized INR should therefore be viewed as a reconstruction constrained by the available observations over the prescribed reconstruction window. The length of this window describes the temporal coverage of the input observations, not a physical time scale learned by the network for the regional dynamics.

For limitations, training, and testing, Section 6.2 (Applicability and limitations) now states that the result is constrained by observation coverage and regularization choices:

- The main limitation is that reconstruction quality remains constrained by the available observations. The INR learns a coordinate-to-SSH relation from sparse samples and can be queried at unobserved space–time points within the chosen reconstruction domain and time window. This interpolation capability should not be interpreted as unconstrained extrapolation: poorly sampled SSH structures or rapidly evolving events remain difficult to recover.

For global-scale and operational applicability, Section 6.3 (Training strategy, scalability, and operational applicability) now summarizes the additional work needed before broader or operational use:

- In principle, the coordinate-based formulation can be applied beyond the two regional domains considered here. In practice, global-scale SSH reconstruction would require additional design choices because sampling density, dynamical regimes, coastlines, and dominant spatial scales vary strongly across the ocean. Domain decomposition, regional tiling, multi-resolution representations, or hierarchical training may be needed for larger domains.

- The experiments were carried out with GPU acceleration, and the released code documents the optimization settings used here. A systematic benchmark of training time, memory use, and update frequency across different domains and data-availability scenarios would require a dedicated operational assessment. For operational use, the method would also need testing under variable data availability, observation errors, hyperparameter uncertainty, and latency constraints, as well as integration with quality control and uncertainty-estimation procedures.

Comment 4: "*In the Observed SSH Experiment section, the relatively short temporal gap between training and testing datasets may introduce some degree of correlation. A brief discussion of this aspect, and its possible implications for the results, would strengthen the manuscript. Exploring additional data challenges available on the same platform could also provide further support for the robustness of the approach.*."

Response: We thank the reviewer for pointing this out. We have revised this part because the previous wording could make the OSE validation sound like a time-disjoint training–testing experiment. The actual benchmark split is by mission: CryoSat-2 new-orbit observations are withheld from training, while the training and validation samples still belong to the same evolving ocean period. Some dynamical correlation may therefore remain, and the OSE evaluation should be interpreted as mapping skill against withheld observations, not as prediction skill for fully independent ocean states.

Specifically, we now clarify in the Discussion that the OSE validation follows the benchmark mission split and should not be read as a test on a fully independent ocean state. We also added the reviewer's suggestion about broader testing across additional challenges. The corresponding revision in Section 6.2 (Applicability and limitations) is as follows:

- The OSE and OSSE results should also be interpreted according to their different validation settings. The OSSE experiment allows full-field evaluation against NATL60, whereas the OSE experiment uses withheld-mission along-track validation because no complete true SSH field is available. The OSE statistics are therefore best read as reconstruction or mapping skill under the benchmark sampling configuration, not as a strict test of long-term temporal extrapolation. In the OSE case, some dynamical correlation may remain between training and validation samples because they are drawn from the same evolving ocean period. Future evaluations over additional Ocean Data Challenges, regions, seasons, and sampling configurations, together with estimates of SSH autocorrelation scales, would help assess robustness and the effective independence of validation samples. The learned network weights should therefore not be interpreted as physical parameters or dynamical time scales. The present study reports deterministic reconstructions,

and it does not provide an OI-like uncertainty map or posterior error estimate. Uncertainty quantification remains an important direction for future work.

Comment 5: "*For the simulated SSH experiment (OSSE), the inclusion of wavenumber–frequency spectral diagnostics, e.g., PSD*(ω, k) *and associated spectral scores, could provide a more detailed characterization of the spatio-temporal scales effectively resolved by the method. The 0.03° spatial resolution seems extremely small.*"

Response: We thank the reviewer for raising this point. In response, we added a wavenumber–frequency PSD-score diagnostic for the OSSE experiment and revised the wording around the reported 0.03° value.

The main clarification is that $\lambda_x = 0.03°$ is not a nominal grid spacing or a direct physical resolution. It is the shortest spatial wavelength reached by the 0.5 contour of the PSD skill score under the adopted Ocean Data Challenges diagnostic. We therefore changed the wording from "spatial resolution" to "PSD-score-based effective scale" and added a caution that this value should not be read as evidence that all SSH structures at 0.03° are physically resolved.

We first clarified the definition of the PSD-based wavenumber–frequency skill score in Section 5.2 (Evaluation Indexes), under *(2) Wavenumber–frequency PSD skill score*, as follows:

- To assess the scale-dependent consistency of the reconstructed SSH fields, we compute a power spectral density skill score in the wavenumber–frequency domain. To make the notation explicit, we write the PSD as a function of spatial wavenumber k and temporal frequency f. The wavenumber–frequency PSD skill score is defined as

$$\mathrm{PSD}_{\mathrm{SS}}(k, f) = 1 - \frac{\mathrm{PSD}[\mathrm{SSH}_{\mathrm{rec}} - \mathrm{SSH}_{\mathrm{true}}]\,(k, f)}{\mathrm{PSD}[\mathrm{SSH}_{\mathrm{true}}]\,(k, f)}. \tag{1}$$

  Here, $\mathrm{PSD}[\cdot]\,(k, f)$ denotes the wavenumber–frequency power spectral density of a given field evaluated at spatial wavenumber k and temporal frequency f. The numerator represents the spectral energy of the reconstruction error, and the denominator represents the spectral energy of the reference SSH field. This score therefore measures the relative reconstruction skill at each wavenumber–frequency bin. A value close to 1 indicates that the reconstruction error is small relative to the reference signal at the corresponding spatio-temporal scale, whereas a value close to 0 indicates that the error energy is comparable to the signal energy.

We then revised the explanation of λ_x and λ_t. Since the PSD score is a relative spectral-error measure, these values are now described as contour-derived effective-scale diagnostics under the benchmark protocol, rather than as proof that all SSH structures at the corresponding scales are fully resolved. The corresponding revisions in Section 5.2 (Evaluation Indexes) are as follows:

- Following the Ocean Data Challenges protocol, two PSD-score-based effective-scale diagnostics, λ_x and λ_t, are derived from the 0.5 contour of the normalized wavenumber–frequency PSD skill score $\mathrm{PSD}_{\mathrm{SS}}(k, f)$. This threshold is used to summarize the shortest spatial and temporal

scales that satisfy the adopted spectral-skill criterion. Specifically, λ_x is defined as the shortest spatial wavelength along this contour, while λ_t is defined as the shortest temporal period along the same contour. Smaller values of λ_x and λ_t indicate shorter effective scales under this diagnostic.

These quantities should be interpreted as threshold-based effective-scale diagnostics, rather than as nominal grid spacings or direct observational resolutions. In particular, when λ_x approaches the smallest wavelength represented by the evaluation grid and spectral discretization, it indicates that the adopted spectral-score criterion remains satisfied down to the smallest scales accessible in the diagnostic, rather than guaranteeing that all physical SSH structures at that scale are fully resolved.

To show how the reported values are obtained, we added Figure 4 in Section 5.5.5 and reproduce it below. The figure shows the two-dimensional PSD-score field in the wavelength–period domain, together with the 0.5 contour used in the Ocean Data Challenges protocol. In this diagnostic, $\lambda_x = 0.03^\circ$ is the shortest spatial wavelength reached by the 0.5 contour, and $\lambda_t = 2.05$ days is the shortest temporal period reached by the same contour.

We retained this diagnostic because it is useful for relative comparison: all methods in Table 1 are evaluated with the same OSSE configuration, spectral discretization, and Ocean Data Challenges PSD-score protocol. The corresponding revisions in Section 5.5.5, *Comparison of Multiple Methods for SSH Reconstruction*, are as follows.

- As reported in Table 1, INR gives the smallest PSD-score-based effective spatial and temporal scales under the adopted evaluation criterion. To clarify how these effective-scale values are obtained and how they should be interpreted, Figure 4 shows the PSD-based score in the wavelength–period domain. The score is generally high for long temporal periods, especially above approximately 15–20 days, over a broad range of spatial wavelengths, indicating that slowly varying SSH components are well reconstructed. Relatively high scores are also observed at broad spatial wavelengths, roughly larger than 3°, suggesting that large-scale SSH variability is captured with small relative spectral error. In contrast, the main low-score region is concentrated at shorter temporal periods and intermediate spatial wavelengths, approximately within 3–10 days and 0.5–2°. This indicates that rapidly evolving and spatially complex SSH variability remains more challenging to reconstruct.

  Some local high-score patches also appear near the shortest-wavelength edge of the diagnostic. We do not interpret these patches as evidence that the smallest SSH structures are easier to reconstruct. The PSD-based score is a relative spectral-error measure and can be sensitive where the reference spectral energy is weak and where the spectral estimate is affected by discretization near the shortest wavelengths represented in the diagnostic. Therefore, these local high-score regions should not be directly interpreted as evidence that small-scale SSH structures are fully resolved. Following the Ocean Data Challenges protocol, the PSD-score-based effective scales are summarized by the contour where the PSD-based score equals 0.5. Therefore, the

reported λ_x and λ_t values in Table 1 should be regarded as contour-derived summary diagnostics of the two-dimensional PSD-based score. In particular, $\lambda_x = 0.03°$ represents the shortest spatial wavelength satisfying the adopted PSD-based score criterion, rather than a nominal grid spacing or a guarantee that all SSH structures at this scale are fully resolved. Nevertheless, these diagnostics remain meaningful for relative comparison because all methods in Table 1 are evaluated under the same Ocean Data Challenges protocol. Under this common evaluation framework, smaller values of λ_x and λ_t indicate that the reconstruction maintains a PSD-based score above the 0.5 threshold down to shorter spatial wavelengths and shorter temporal periods, respectively. For SSH reconstruction, this indicates that, under the adopted PSD-score criterion, the INR reconstruction maintains lower relative spectral errors over a broader range of spatial wavelengths and temporal periods.

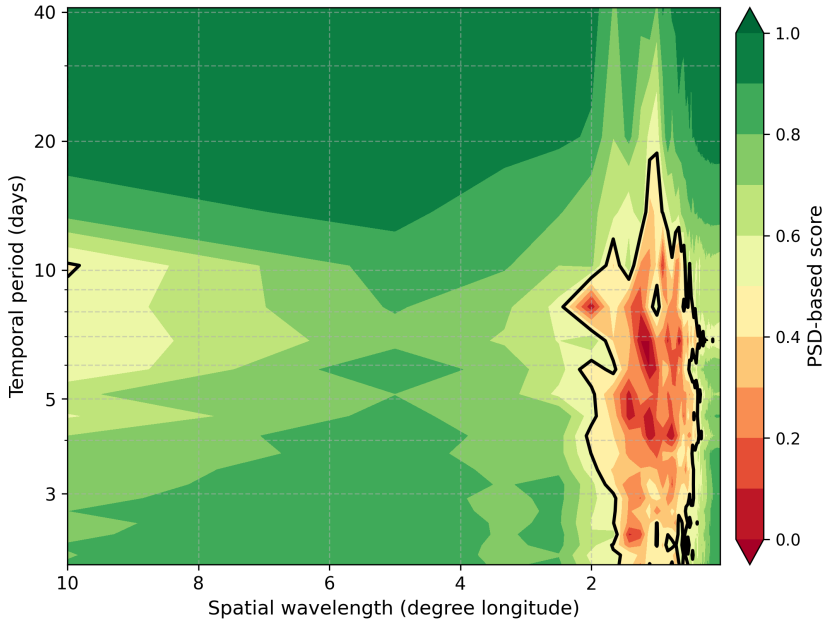

Figure 4: PSD-based skill score of the INR reconstruction in the wavelength–period domain. The black contour denotes the 0.5 level used to derive the PSD-score-based effective-scale diagnostics following the Ocean Data Challenges protocol. The diagnosed values $\lambda_x = 0.03°$ and $\lambda_t = 2.05$ days correspond to the shortest spatial wavelength and shortest temporal period reached by this contour, respectively. These values should be interpreted as PSD-score-based diagnostics rather than exact resolved scales.

Comment 6: "*Regarding the methodology, it could be insightful to assess the role of the Total Variation (TV) regularization term by performing complementary experiments without this constraint, in order to better isolate its impact.*"

Response: We thank the reviewer for this helpful suggestion. We have added a no-TV ablation experiment in the OSSE setting. In this test, the TV term is removed and all other settings are kept un-

changed. We use the OSSE case for this ablation because the NATL60 reference provides a complete space–time SSH field. This allows the TV and no-TV reconstructions to be compared with the same full-domain RMSE and PSD-score diagnostics, including both the spatial and temporal effective-scale summaries. In the OSE case, the withheld observations are sparse along-track samples; the same locations are not generally observed repeatedly in time, so the same full space–time diagnostics cannot be computed.

The ablation shows that TV regularization has little effect on the domain-averaged RMSE score or on the diagnosed temporal scale, but it does affect the PSD-score-based spatial-scale diagnostic. We have therefore clarified that, unless otherwise specified, INR refers to the full proposed model with TV regularization, while "INR without TV" is only used as an ablation variant. The corresponding revisions in Section 5.5.5 (Comparison of Multiple Methods for SSH Reconstruction) are as follows.

- Table 1 reports the performance of the proposed INR method in comparison with several representative state-of-the-art approaches. Unless otherwise specified, INR denotes the full proposed model with TV regularization, while "INR without TV" is used only as an ablation variant. In terms of the OSSE $\mathrm{RMSE}_{\mathrm{SS}}$ computed over all valid grid points and evaluation times, INR performs on par with the top-performing schemes. Although its $\mathrm{RMSE}_{\mathrm{SS}}$ value is slightly lower than that of 4DVarNet v2022, INR attains the smallest PSD-score-based effective spatial and temporal scales among all evaluated methods. In particular, INR with TV regularization gives the smallest diagnosed spatial scale λ_x and temporal scale λ_t under the adopted Ocean Data Challenges PSD-score protocol.

  The effect of the TV regularization term is isolated by comparing the full INR model with an ablation variant in which the TV term is removed while all other settings are kept unchanged. The ablation result shows that removing the TV term has little influence on the normalized RMSE score, which remains approximately 0.94, and on the diagnosed temporal scale λ_t, which remains 2.05 days. In contrast, the diagnosed spatial scale λ_x, derived from the 0.5 contour of the PSD-score field, changes from 0.12° without TV regularization to 0.03° with TV regularization. This indicates that the TV term mainly affects the spatial PSD-score diagnostic, by reducing isolated large gradients and part of the high-wavenumber error energy, while the domain-averaged pointwise error changes little.

Table 1: Performance comparison of SSH reconstruction methods on the simulated altimetry dataset in the Gulf Stream domain. Here, $\mathrm{RMSE_{SS}}$ is the normalized RMSE skill score, computed once over all valid grid points and evaluation times in the OSSE experiment. Unless otherwise specified, INR denotes the full proposed model with TV regularization, while "INR without TV" denotes the ablation variant used to assess the effect of the TV term. Bold values denote the best performance for each metric.

Method	$\mathrm{RMSE_{SS}}$	λ_x (degree)	λ_t (days)
DUACS	0.92	1.22	11.37
BFN	0.93	1.00	10.24
DYMOST	0.93	1.19	10.04
MIOST	0.94	1.18	10.33
4DVarNet	0.95	0.70	6.48
4DVarNet v2022	**0.96**	0.62	4.35
INR without TV	0.94	0.12	2.05
INR	0.94	**0.03**	**2.05**

Comment 7: "*Since the OSSE framework provides access to the full spatio-temporal reconstruction error, presenting and discussing this information would further enrich the analysis.*"

Response: We thank the reviewer for this helpful suggestion. We agree that the OSSE setting should be used more fully because the NATL60 SSH field is available as a complete reference. We therefore added a new subsubsection entitled "Spatio-temporal reconstruction error".

In this subsection, we added Figure 5 to examine the reconstruction error from both spatial and temporal perspectives. Specifically, the figure includes instantaneous SSH reconstruction error maps at the same four selected dates as those used in the preceding INR reconstruction figure, the time-mean RMSE map over the full evaluation period, and the spatially averaged RMSE time series. The four dates were selected by dividing the evaluation period into four approximately equal sub-periods and choosing the middle valid snapshot from each sub-period.

This additional analysis summarizes the spatial and temporal variability of the reconstruction errors. In the selected instantaneous maps and in the time-mean RMSE map, larger errors are mainly found in dynamically active regions with stronger SSH variability and sharper gradients. The spatially averaged RMSE curve further shows how the domain-mean error level changes among the evaluated time slices. The detailed revisions in Section 5.5.4 (Spatio-temporal reconstruction error) are as follows.

- Since the OSSE configuration provides the complete NATL60 SSH field as the reference, the full spatio-temporal reconstruction error can be directly evaluated. The error field and the two

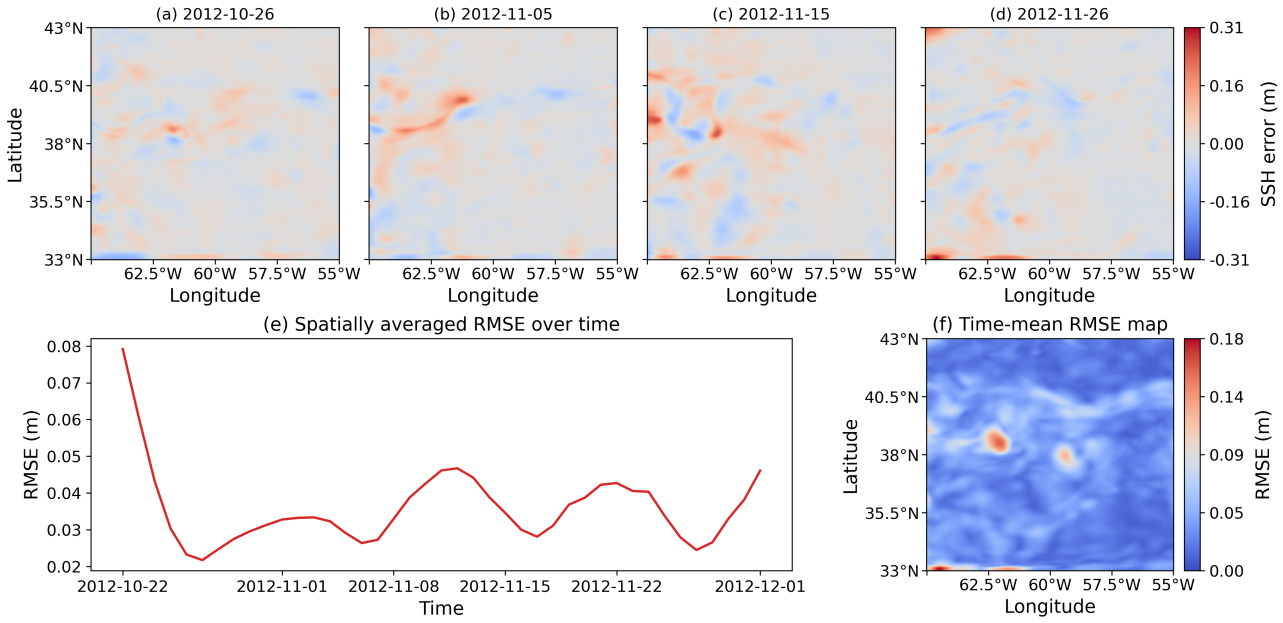

Figure 5: Spatio-temporal reconstruction error of the proposed method in the OSSE experiment. Panels (a)–(d) show the instantaneous error field e at the same four representative dates as in the preceding qualitative INR reconstruction figure. Panel (e) shows $\mathrm{RMSE}_{\mathrm{space}}(t_j)$, and panel (f) shows $\mathrm{RMSE}_{\mathrm{time}}(x_m, y_n)$, as defined in Eq. (2). All errors are computed in the original SSH physical space, with units of metres.

RMSE summaries used below are defined together as

$$
\begin{aligned}
e(x_m, y_n, t_j) &= \mathrm{SSH}_{\mathrm{rec}}(x_m, y_n, t_j) - \mathrm{SSH}_{\mathrm{true}}(x_m, y_n, t_j), \\
\mathrm{RMSE}_{\mathrm{space}}(t_j) &= \left[\frac{1}{N_x N_y} \sum_{m=1}^{N_x} \sum_{n=1}^{N_y} e(x_m, y_n, t_j)^2 \right]^{1/2}, \\
\mathrm{RMSE}_{\mathrm{time}}(x_m, y_n) &= \left[\frac{1}{N_t} \sum_{j=1}^{N_t} e(x_m, y_n, t_j)^2 \right]^{1/2}.
\end{aligned}
\tag{2}
$$

Here, $\mathrm{SSH}_{\mathrm{rec}}$ denotes the reconstructed SSH field and $\mathrm{SSH}_{\mathrm{true}}$ denotes the NATL60 reference field. All errors in this analysis are computed in the original SSH physical space, with units of metres.

Figure 5 presents the spatio-temporal reconstruction error of the proposed method in the OSSE experiment. Figures 5a–d show the instantaneous SSH reconstruction error at the same four selected dates as those used in the preceding INR reconstruction figure. These dates were obtained by dividing the evaluation period into four approximately equal sub-periods and selecting the middle valid snapshot from each sub-period. The instantaneous error maps show that errors are relatively small over much of the domain. Larger errors are mainly localized in dynamically active regions with stronger SSH variability and sharper gradients. This indicates that

energetic mesoscale structures remain more challenging to reconstruct accurately under sparse satellite-like sampling.

Figure 5e shows the spatially averaged RMSE over time. This time series summarizes how the domain-mean reconstruction error varies among the evaluated time slices. The RMSE is relatively larger near the beginning and then fluctuates during the evaluation period. Larger RMSE values indicate time slices in which the SSH structures and sampling constraints are more challenging for the reconstruction.

To summarize the spatial distribution of the reconstruction error over the full evaluation period, Figure 5f shows the time-mean RMSE map. The time-mean RMSE remains low over most of the domain, while several localized regions exhibit relatively larger values. These localized high-error regions are mainly associated with areas where the SSH field has stronger spatial variability and sharper gradients. This result further shows that the proposed method can reconstruct the broad SSH pattern and maintain spatial coherence over most of the domain, while dynamically active mesoscale structures remain more difficult to reproduce accurately.

In addition, because the OSSE experiment provides a complete reference field, we further added a qualitative comparison with the benchmark OI baseline under the four-nadir setting in Section 5.5.5 (Comparison of Multiple Methods for SSH Reconstruction). This additional comparison is not discussed in detail here, as it has been addressed in our response to Comment 2.

Comment 8: "*Clarify how the parameter a is selected to balance the loss terms in Equation (9). Does it vary regionally? please define the parameter λ.*"

Response: We thank the reviewer for pointing out this ambiguity. We have clarified that a is the trade-off coefficient that controls the relative contribution of the TV regularization term in the total loss. It balances the data-fidelity loss and the spatial-gradient regularization term.

We have also clarified that a is treated as a global hyperparameter for each experimental configuration and is kept fixed over the entire reconstruction domain. Therefore, a is not a proxy for local observation uncertainty and is not varied by region. In practice, a was not optimized by an automated rule or adjusted locally. Several trial values were tested in preliminary runs. For each value, we examined the benchmark error diagnostics available for the experiment and inspected the reconstructed SSH and gradient fields for isolated noisy gradients or excessive smoothing of mesoscale structures. We then selected the value separately for each experiment and used it for all locations and times within that experiment.

In addition, we corrected the notation inconsistency in the previous manuscript. The parameter previously denoted by λ was not an additional parameter, but referred to the same trade-off coefficient as a. We now use a consistently throughout the manuscript. The corresponding revisions in Section 4.2 (TV-based Spatial-Gradient Regularization) are as follows.

- The final training objective combines the data-fidelity loss and the TV regularization term:

$$\mathcal{L}_{\text{total}}(\theta) = \mathcal{L}_{\text{data}}(\theta) + a\mathcal{L}_{\text{TV}}(\theta), \tag{3}$$

where $a > 0$ controls the relative contribution of the TV regularization term compared with the data-fidelity loss. In this study, a is treated as a global hyperparameter for each experimental configuration and is kept fixed over the reconstruction domain. The coefficient a is therefore not a proxy for local observation uncertainty and is not varied by region. In practice, a is not optimized by an automated rule or adjusted locally. Several trial values are tested in preliminary runs. For each value, we examine the benchmark error diagnostics available for the experiment and inspect the reconstructed SSH and gradient fields for isolated noisy gradients or excessive smoothing of mesoscale structures. One value is then fixed for each experiment and used for all locations and times. A very small a may provide insufficient regularization and lead to noisy local variations, whereas an overly large a may over-smooth the SSH field and weaken dynamically meaningful sharp structures.

Comment 9: "*a schematic illustrating the training and testing periods could improve clarity.*"

Response: We thank the reviewer for this helpful suggestion. We agree that the previous wording did not make the data roles clear enough. A time-period schematic alone would not fully resolve this point, because the relevant distinction is not only chronological: in the OSE case it is a leave-one-mission-out validation, while in the OSSE case sparse pseudo-observations and the full NATL60 field play different roles over the same target period. We therefore revised the manuscript text to state the training input and evaluation reference directly.

After re-checking both the Ocean Data Challenges documentation and our data-preparation/training scripts, we confirmed that the computations followed the benchmark data use. The ambiguity was introduced by our earlier manuscript wording, which compressed two different points into a conventional "training–test" description: the benchmark data-availability windows and the data actually used by the INR. In the OSE benchmark, the assessment period is Jan. 15–Mar. 15, 2021; CryoSat-2 new-orbit observations are reserved for independent evaluation, while the other nadir missions are available for reconstruction. The two 15-day windows mentioned by the benchmark describe additional observations that methods may use for spin-up or training around the assessment period; they were not the two isolated training windows used in our INR implementation. In the revised text, we state directly that the INR is trained with non-CryoSat-2 observations from Jan. 1 to Mar. 15, 2021 and evaluated against withheld CryoSat-2 observations from Jan. 15 to Mar. 15, 2021. For the OSSE case, the benchmark assessment period is Oct. 22–Dec. 2, 2012. Sparse pseudo-observations over this period are used as the INR training input, and the full NATL60 field over the same period is used only as the gridded reference for evaluation. The 2013 NATL60 reference-data period is part of the benchmark data availability for methods that require full-field learning data, but it is not the training period used for the reported INR reconstruction.

The corresponding revisions in Section 5.1 (Experimental configurations and evaluation protocols) are as follows.

- In the OSE configuration, the experiment is conducted over the Western Mediterranean Sea within the domain $[1°\mathrm{E}, 20°\mathrm{E}] \times [30°\mathrm{N}, 45°\mathrm{N}]$. This experiment is designed to assess the reconstruction method under realistic satellite sampling and measurement conditions. The INR

model is trained with sparse real along-track observations from the nadir missions available for mapping, excluding CryoSat-2 new-orbit observations, from Jan. 1 to Mar. 15, 2021. CryoSat-2 new-orbit observations from Jan. 15 to Mar. 15, 2021 are withheld from training and used only for evaluation. Thus, the OSE split is mission-based rather than a fully time-disjoint split. Because no complete true SSH field is available in this real-ocean setting, the evaluation is performed only at the locations and times of the withheld CryoSat-2 observations.

In the OSSE configuration, the experiment is conducted over a dynamically active portion of the Gulf Stream region within the domain $[65°\mathrm{W}, 55°\mathrm{W}] \times [33°\mathrm{N}, 43°\mathrm{N}]$. This experiment provides a controlled setting in which the reconstructed SSH fields can be evaluated against a known reference field. The complete four-dimensional SSH field from the NATL60 numerical simulation is used as the reference field. For the reported INR reconstruction, pseudo-altimetric nadir and SWOT observations over Oct. 22–Dec. 2, 2012 are generated by subsampling this reference field and are used as sparse input data. The reconstruction is then evaluated against the full gridded NATL60 SSH field over the same target period.

These data roles should not be interpreted as a conventional supervised training–test split with fully independent ocean states. They follow the corresponding Ocean Data Challenges benchmark rules. In the OSE case, the split is mission-based: non-CryoSat-2 nadir observations are used for training, and CryoSat-2 new-orbit observations are withheld for along-track validation. In the OSSE case, sparse pseudo-observations over the target reconstruction period are used as training input, while the full NATL60 field over the same period is used only as the known gridded reference for evaluation.

Comment 10: "*Line 339: minor typo: "interpolation". Line 342: consider adding the appropriate reference for MIOST, i.e., Ubelmann et al. (2021). Line 211: include a reference to the data challenge platform.*"

Response: We thank the reviewer for these helpful corrections. We have corrected the typo in "interpolation". We have also added the appropriate reference to Ubelmann et al. (2021) when MIOST is introduced. In addition, we have revised the description of the data source to cite the Ocean Data Challenges platform, while retaining the Zenodo archive citation for the datasets used in this study. These changes clarify both the methodological references and the provenance of the benchmark data and evaluation protocols.

Comment 11: "*For improved readability, adding coastlines to Figures 3 and 4 could be helpful.*"

Response: We thank the reviewer for this helpful suggestion. We have added coastlines to the spatial RMSE panels, which makes the error patterns easier to interpret relative to the domain boundaries and nearby land. While revising these panels, we also reorganized the accompanying error diagnostics by combining the signed-residual CDF with the gridded RMSE maps. This provides a more compact view of both the distribution of withheld along-track residuals and the spatial structure of the reconstruction

errors. The corresponding revised text, figure, and caption in Section 5.4.3 (Detailed Comparison with Optimal Interpolation) are as follows.

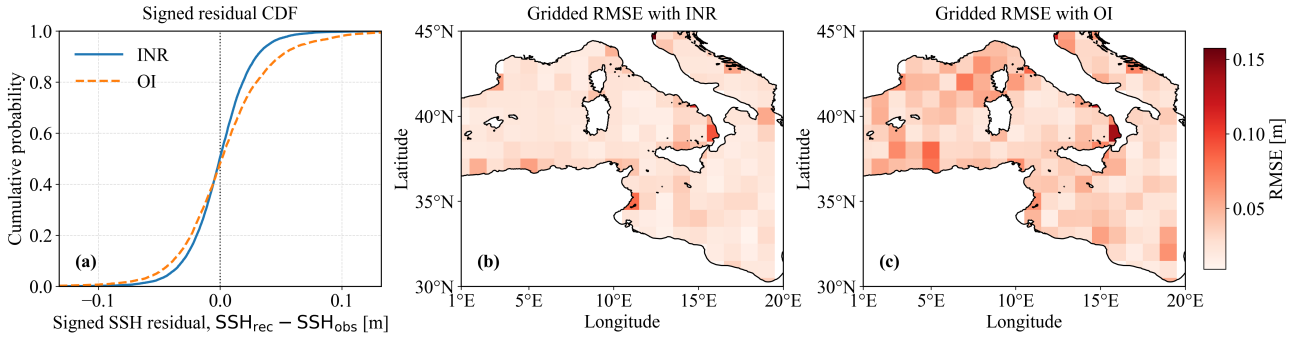

Figure 6: Combined error diagnostics for the INR-based reconstruction and OI. (a) Cumulative distribution function (CDF) of signed SSH residuals, defined as $\mathrm{SSH}_{\mathrm{rec}} - \mathrm{SSH}_{\mathrm{obs}}$, at withheld along-track observation locations. (b, c) Gridded RMSE maps for INR and OI, respectively, computed from withheld residuals within each spatial grid cell. Coastlines are shown in panels (b) and (c).

We sincerely thank the referee again for the valuable comments and constructive suggestions. We believe that the revisions made in response to these comments have improved the clarity, readability, and overall presentation of the manuscript. We have also carefully checked the entire manuscript and made our best effort to improve its accuracy, consistency, and readability.