July 8, 2026

Dear Editor:

Thank you very much for obtaining referee reports on 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 that referees evaluated our work positively and provided insightful comments to guide us to improve the paper. We have revised the manuscript to fully address all the referee comments. A detailed, point-by-point response to the referee's comments/suggestions is enclosed. The changes in the text are marked blue.

We would like to take this opportunity to thank the referees for their time and effort in evaluating our work and for the valuable comments that have resulted in an improved manuscript. We would also like to thank you for your professional handling of our paper. We hope the revised manuscript can be judged to have met the high standard of *GMD*.

With kind regards,

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

Point-by-point response to 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 lan-

guage 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.

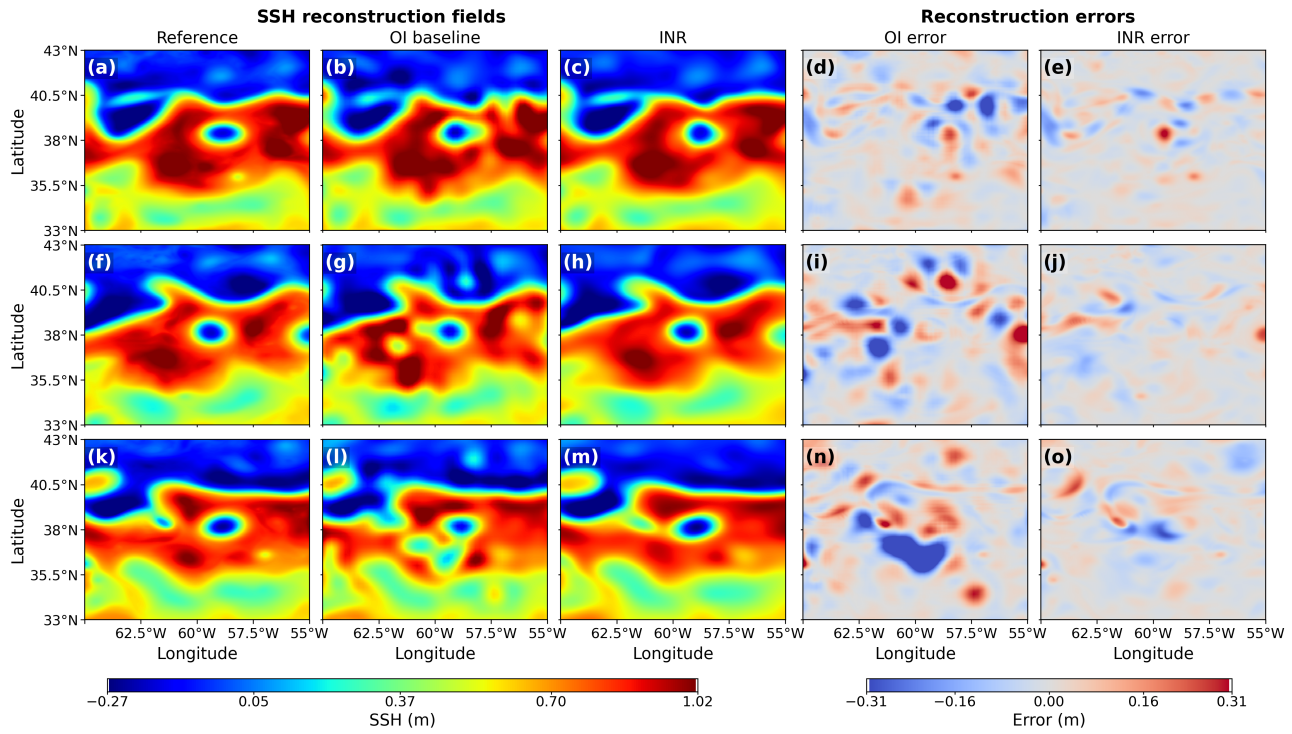

Figure 2: Qualitative comparison between the OI baseline and INR under the four-nadir observation setting. From top to bottom, the rows correspond to three representative evaluation times selected according to the 0.1, 0.5, and 0.9 quantiles of the OI SSH reconstruction error, respectively. The first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

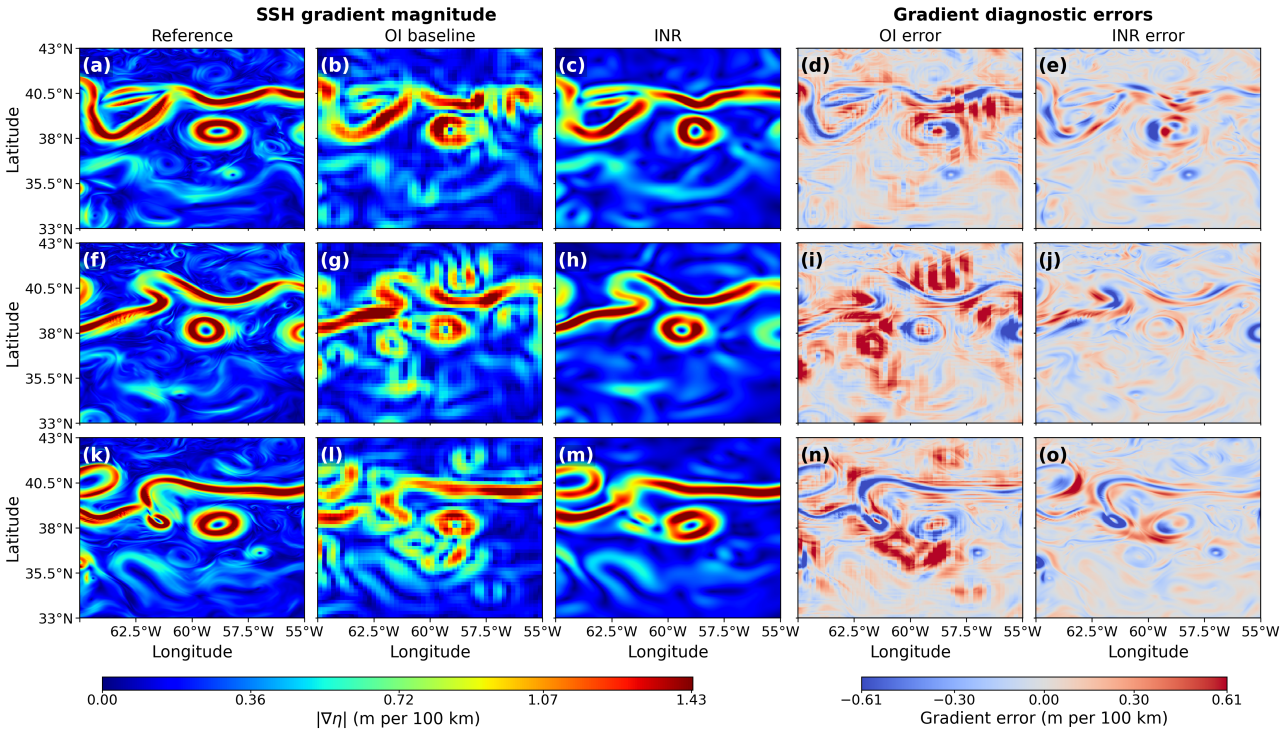

Figure 3: Derivative-based qualitative comparison between the OI baseline and INR under the four-nadir observation setting. From top to bottom, the rows correspond to three representative evaluation times selected according to the 0.1, 0.5, and 0.9 quantiles of the OI SSH reconstruction error, respectively. 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.

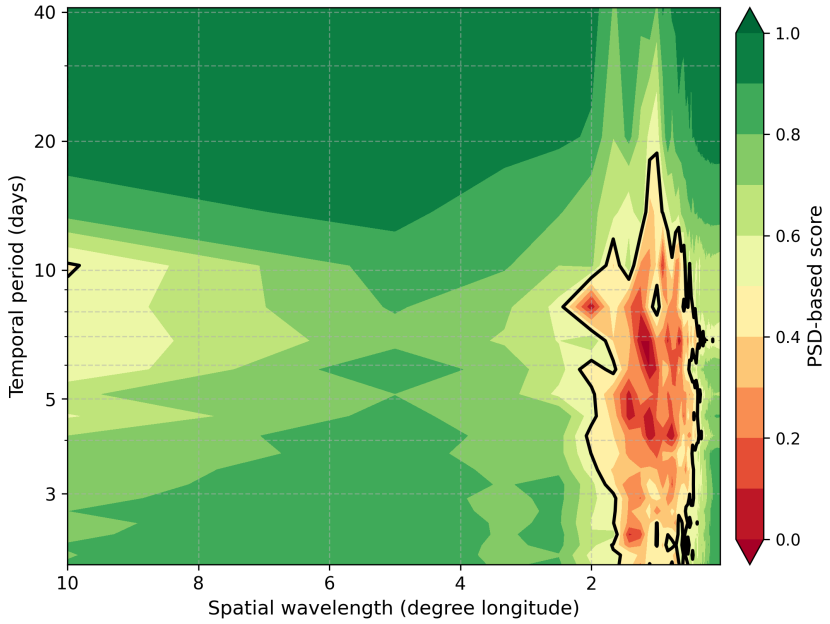

Figure 4: PSD-based skill score of the INR reconstruction in the wavelength–period domain. The black contour denotes the 0.5 level used to derive the PSD-score-based effective-scale diagnostics following the Ocean Data Challenges protocol. The diagnosed values $\lambda_x = 0.03°$ and $\lambda_t = 2.05$ 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

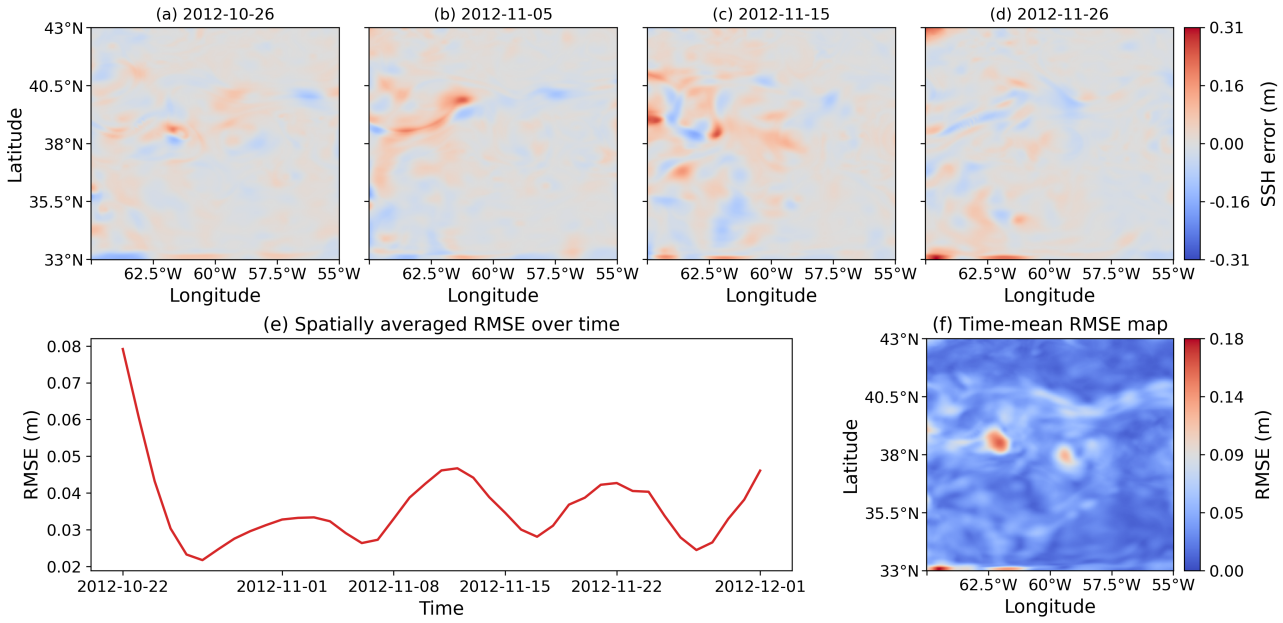

Figure 5: Spatio-temporal reconstruction error of the proposed method in the OSSE experiment. Panels (a)–(d) show the instantaneous error field e at the same four representative dates as in the preceding qualitative INR reconstruction figure. Panel (e) shows $\mathrm{RMSE}_{\mathrm{space}}(t_j)$, and 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.

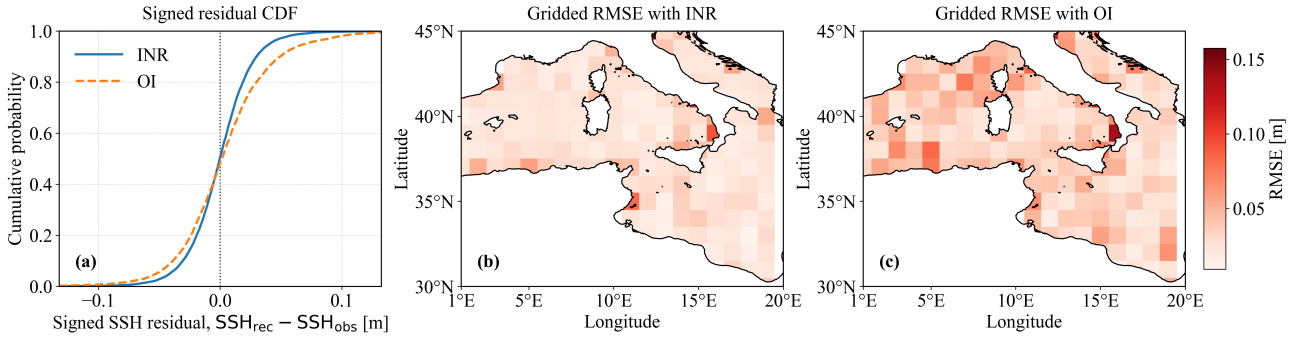

Figure 6: Combined error diagnostics for the INR-based reconstruction and OI. (a) Cumulative distribution function (CDF) of signed SSH residuals, defined as $\mathrm{SSH}_{\mathrm{rec}} - \mathrm{SSH}_{\mathrm{obs}}$, at withheld along-track observation locations. (b, c) Gridded RMSE maps for INR and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

Point-by-point response to the Second Referee

General comment: "*Reviewer 1 gave a good summary of the paper and has excellent comments. I am just going to supplement the discussion, starting from that review.*"

Strengths: The smoothness and differentiability of the method are pluses for ocean representation. Their examples show improved performance over most of the tested methods, and they evaluate an impressive array of methods.

Weaknesses: Some people may find it hard to learn about the method from just these few examples. These methods are changing rapidly, and one wonders how long it will be until this method is itself superseded.

The black box nature of the method makes it difficult to build understanding. Anything the authors can do to counteract this point of view would increase the impact of their work.

Response: We thank the reviewer for the positive assessment of the smoothness, differentiability, and reconstruction performance of the proposed method. We also take the reviewer's concern about interpretability seriously. Accordingly, we revised the manuscript to make the INR formulation less opaque to readers who may not be familiar with this class of methods, and to clarify what the two benchmark examples do and do not demonstrate.

With this in mind, we revised the manuscript in three directions. We rewrote parts of Sections 2–4 in more geophysical language, clarified the coordinate normalization, defined the layer and gradient notation more explicitly, and described the TV term as a spatial-gradient regularization rather than as a physical dynamical constraint. We also revised Section 5 to state the OSE and OSSE protocols more clearly, added the requested signed-residual CDF and temporal paired-difference diagnostics, and made fuller use of the complete NATL60 reference in the OSSE analysis. Finally, we expanded Section 6 to discuss the dependence on observation coverage, the deterministic nature of the reconstruction, the absence of OI-like uncertainty maps, and the fact that the learned weights should not be interpreted as physical parameters or regional dynamical time scales.

The detailed changes are given point by point below, with the corresponding revised manuscript text shown in blue.

Comment 1: "*What are the lessons a reader will take away from it for other applications? When do we expect this method to work well, and what are the limitations? E.g., in the first experiment, how variable could the results be if it were tried on other time ranges, and how much time coverage is needed to have a good training data set? Is their technique of training on the first 15 days and the last 15 days something that they learned was good, or is it the first thing they tried? They should also make the language clearer about whether the training and test datasets overlap on Jan 15 and Mar 15, which is currently somewhat ambiguous, although I assume they do not overlap.* "

Response: We thank the reviewer for raising these useful points. We revised the manuscript in two places: first, the Discussion now states more clearly what the method can and cannot support; second, Section 5.1 now spells out the OSE data roles under the benchmark protocol.

On the lessons for other applications, we clarified the conditions under which the method is expected to work well and where it is likely to be limited. The revised Section 6 (Discussion) states that the method depends on the spatial–temporal coverage of the observations and should be interpreted as a reconstruction method constrained by data and regularization, not as a fully general dynamical model. The corresponding revision in Section 6.2 (Applicability and limitations) is as follows:

- 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. The PSD-score-based effective scale should also be interpreted as a diagnostic of reconstruction skill under a given sampling and evaluation protocol, rather than as the nominal grid spacing of the output field or the NATL60 reference.

For the OSE time split, we clarified that the experiment uses a leave-one-mission-out validation design rather than two non-overlapping training windows. The model is trained with non-CryoSat-2 observations over the benchmark input-observation period, and CryoSat-2 new-orbit observations are kept out of the training data and used for evaluation. Thus the split is by mission, not by two separated time blocks. The setting was not chosen after testing different window lengths, and we do not present it as an optimized or generally sufficient amount of training coverage. The main clarification was added in Section 5.1 (Experimental configurations and evaluation protocols), and the later OSE data-preparation subsection now refers back to this protocol.

- 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.

To address the question of how much temporal coverage is needed, we added a limitation and future-work statement in Section 6.3 (Training strategy, scalability, and operational applicability). It makes clear that the OSE split is a prescribed benchmark setting; the effect of using other windows or observation densities remains an application-specific validation issue rather than a conclusion drawn from this experiment:

- The present training strategy follows the target reconstruction setting described in Section 5.1: for each domain and target period, the INR model is optimized using the available input observations and a gradient-regularized objective. In the OSE experiment, the input observations are non-CryoSat-2 along-track measurements and the evaluation uses withheld CryoSat-2 new-orbit observations. In the OSSE experiment, the input observations are sparse pseudo-altimetric samples and the full NATL60 field is used only as the gridded reference. The experiments use preset optimization settings chosen from preliminary convergence checks; the exact settings are provided in the released code. These choices follow the benchmark setting and our implementation, rather than a sensitivity search over input periods. We therefore interpret the results as benchmark evaluations under the corresponding sampling scenarios, not as a general estimate of the minimum observation coverage needed for other regions or seasons. When the method is extended to other settings, the target period and observation density should be examined as part of the application-specific validation, for example through withheld-data tests or sensitivity experiments. A systematic assessment across alternative periods, observation densities, and seasons is left for future work. 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. The main value of this study is to show that a continuous and differentiable INR can complement existing SSH mapping approaches, especially for regional reconstruction and derivative-based diagnostics under sparse sampling.

Comment 2: "*A classic reviewer question is to ask about the sensitivities of the method, and one lesson we might take from these results is how long it takes for this method to accurately learn the dynamics of each region.* "

Response: We thank the reviewer for raising this point. The comment highlights a practical issue: the present experiments should make clear what they do, and do not, show about sensitivity to temporal coverage and the amount of data needed for a given region.

We have revised Section 6 (Discussion) to make two points explicit. The INR is not used here as a model that learns reusable ocean dynamics or regional time scales. Its parameters are optimized from the available observations for the target reconstruction problem, so its skill depends on whether the available observations sufficiently constrain the SSH variability in that setting. The corresponding revision in Section 6.1 (Methodological characteristics and relation to existing approaches) is as follows:

- 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. 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.

We also clarified that the OSE training and validation data roles follow the benchmark protocol and were not selected after a sensitivity search. We therefore treat the present result as evidence for this prescribed benchmark configuration, while leaving a systematic test of other target periods, observation densities, and seasons as future work. The corresponding revision in Section 6.3 (Training strategy, scalability, and operational applicability) is as follows:

- The present training strategy follows the target reconstruction setting described in Section 5.1: for each domain and target period, the INR model is optimized using the available input observations and a gradient-regularized objective. In the OSE experiment, the input observations are non-CryoSat-2 along-track measurements and the evaluation uses withheld CryoSat-2 new-orbit observations. In the OSSE experiment, the input observations are sparse pseudo-altimetric samples and the full NATL60 field is used only as the gridded reference. The experiments use preset optimization settings chosen from preliminary convergence checks; the exact settings are provided in the released code. These choices follow the benchmark setting and our implementation, rather than a sensitivity search over input periods. We therefore interpret the results as benchmark evaluations under the corresponding sampling scenarios, not as a general estimate of the minimum observation coverage needed for other regions or seasons. When the method is extended to other settings, the target period and observation density should be examined as part of the application-specific validation, for example through withheld-data tests or sensitivity experiments. A systematic assessment across alternative periods, observation densities, and seasons is left for future work. 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. The main value of this study is to show that a continuous and differentiable INR can complement existing SSH mapping approaches, especially for regional reconstruction and derivative-based diagnostics under sparse sampling.

Comment 3: "*The presentation should be made more accessible to earth scientists, who may be less familiar with formal mathematical descriptions. In particular, the formal math exposition, it leaves out small details that would have helped me to understand. As an example, in line 104, they describe a general mapping from spatial coordinates to physical quantities. I wish they could have*

noted that there were m spatial coordinates and n corresponding physical quantities. This only adds two characters to the exposition, but it would have made it much clearer. Along the same lines, when they introduce the constraints on line 109, they could have given an example, as they do later with the altimeter, which would have helped. "

Response: We thank the reviewer for this helpful suggestion. The original text introduced the INR idea in a rather abstract form, so we revised it to define the reconstruction problem in terms that are easier for Earth-science readers to follow. We made three specific changes.

To make the INR background less abstract, we rewrote it so that the reader first sees the idea as a continuous coordinate-to-field representation, rather than as a generic formal mapping. The corresponding revision in Section 2 (Background: Implicit Neural Representations) is as follows:

- Implicit neural representations (INRs) provide a way to describe a continuous field as a function of its coordinates. Instead of assigning values only to a finite set of locations, an INR learns a coordinate-to-value mapping that can be queried at arbitrary positions within the domain. This idea is useful for geophysical reconstruction problems, where observations are often sparse or irregular, but the target variable is expected to vary continuously in space and time.

  Let $\mathbf{s}$ denote a general input coordinate. The neural representation is written as $\Phi_\theta(\mathbf{s})$, where θ denotes the trainable network parameters and Φ_θ gives the field value at the queried coordinate.

We then made the SSH-specific input and output notation explicit. In this application, the input coordinate is longitude, latitude, and time, and the target value is SSH. The corresponding revision in Section 3 (Problem Formulation) is as follows:

- Satellite radar altimeters measure SSH only along their ground tracks, resulting in sparse and irregular spatio-temporal observations. Let $\mathbf{s} = (\mathrm{lon}, \mathrm{lat}, \mathrm{time})$ denote a general spatio-temporal coordinate, where lon, lat, and time represent longitude, latitude, and time, respectively. The coordinate of the i-th satellite observation is denoted by $\mathbf{s}_i = (\mathrm{lon}_i, \mathrm{lat}_i, \mathrm{time}_i)$, and the corresponding observed SSH value is denoted by $\mathrm{SSH}_{\mathrm{true},i}$.

We also added a concrete altimetry example to explain why additional constraints are needed. The corresponding revision in Section 3 (Problem Formulation) is as follows:

- At the observed coordinates, the reconstructed values should be consistent with the available satellite measurements. For locations away from the observed tracks, however, the SSH field must be determined by the learned coordinate-to-SSH relationship rather than by direct observations.

  Because satellite altimetry provides sparse and irregular along-track measurements, the SSH field away from the observed tracks must be inferred from the learned coordinate-to-SSH relationship rather than directly constrained by observations. Such a reconstruction may contain unsupported local fluctuations, which motivates the use of spatial regularization in the proposed framework.

Comment 4: "*OI can analytically produce gradient estimates and has the additional benefit of being easily understandable, transparent, and generating uncertainty maps. For example: An Improved Mapping Method of Multisatellite Altimeter Data, P. Y. Le Traon, F. Nadal, and N. Ducet (1998, DOI); Maps from the Mid-Ocean Dynamics Experiment: Part I. Geostrophic Streamfunction, James C. McWilliams (1976, DOI); and An Objective Analysis of the POLYMODE Local Dynamics Experiment. Part II: Streamfunction and Potential Vorticity Fields during the Intensive Period, Bach Lien Hua, James C. McWilliams, and W. Brechner Owens (1986, DOI). This extends to the other analytical methods discussed, which generally bring in trusted dynamical frameworks to do the time evolution. It is unfortunate that we currently have to give up explainability and uncertainty quantification to get improved performance.*"

Response: We thank the reviewer for this comment. This is an important distinction. OI and related analytical or dynamically constrained methods are not only accuracy benchmarks; they also have clear assumptions, useful derivative diagnostics, and, in the case of OI, uncertainty estimates. We have revised the discussion to make the comparison less one-sided.

For this comparison, we now make the contrast more concrete. The INR model learns a continuous map from longitude, latitude, and time to SSH. Once optimized for the target reconstruction problem, it can be evaluated within the reconstruction domain and period, including between satellite tracks, and differentiated for gradient diagnostics. We present this flexibility as a complement to OI, not as a replacement for OI's explicit assumptions or uncertainty estimates. The corresponding revision in Section 6.1 (Methodological characteristics and relation to existing approaches) is as follows.

- 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 uncertainty quantification, we clarified what is still missing in the present INR implementation. Although the reconstructed SSH field is differentiable and can be used for gradient-based diagnostics,

the method remains a deterministic reconstruction and does not provide an OI-like uncertainty map or posterior error estimate. The corresponding revision in Section 6.2 (Applicability and limitations) is as follows.

- 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: "*When they introduce the SIREN framework, again they use a very formal presentation without a translation into more geophysical language, which doesn't appear until equation 4, 2 pages later. Perhaps an expanded exposition could be in an appendix if they don't want to expand the main body.*"

Response: We thank the reviewer for this suggestion. We agree that the original presentation moved too quickly into network notation. We therefore kept the explanation in the main text and added a short geophysical description before the formal equations, because this context is needed for the loss function and regularization that follow.

At the beginning of Section 4 (Proposed Framework), we revised the text so that the two components of the method are described in SSH-reconstruction terms before the mathematical details are introduced. The corresponding revision is as follows.

- This section presents the proposed INR-based framework for reconstructing a continuous SSH field from sparse satellite altimetry observations. The approach combines a SIREN-based implicit neural representation with TV-based spatial-gradient regularization. The following subsections describe the coordinate normalization, network architecture, spatial-gradient regularization, and training objective.

In Section 4.1 (SIREN-based Implicit Neural Representation), we now define the SIREN model directly as a map from longitude, latitude, and time to SSH before presenting the layer equations. The corresponding revision is as follows.

- The first component of the proposed framework is a SIREN-based INR for representing SSH as a continuous function of longitude, latitude, and time. For a given spatio-temporal coordinate $\mathbf{s} = (\mathrm{lon}, \mathrm{lat}, \mathrm{time})$, the network is designed to output the corresponding reconstructed SSH value. In geophysical terms, the network represents a coordinate-to-SSH mapping learned from the available satellite-track samples. After training, this mapping is evaluated over the prescribed reconstruction domain and time window. In the present experiments, it is used to estimate SSH at target space–time coordinates in this domain, including locations and times not directly sampled by the satellite tracks. Thus, the continuous coordinate-based formulation provides interpolation within the sampled reconstruction problem, rather than implying unconstrained extrapolation to unrelated regions or periods. Before being passed to the SIREN network, the coordinate $\mathbf{s}$ is converted into a normalized coordinate, as described below.

Comment 6: "*On line 164, they could have noted that k is the layer number. It's obvious eventually, but please say it here and help the reader.*"

Response: We thank the reviewer for pointing this out. This was a notation issue in the SIREN description. We now define k at the point where the layer equation is introduced, and also state the range of the layer index.

The corresponding revisions in Section 4.1 (SIREN-based Implicit Neural Representation) are as follows.

- For the k-th hidden layer, the layer output is defined as

$$\mathbf{z}_{k+1} = \sin(\mathbf{W}_k \mathbf{z}_k + \mathbf{b}_k), \qquad k = 0, \ldots, L-1, \tag{4}$$

  where k denotes the hidden-layer index, L is the number of hidden layers, $\mathbf{W}_k$ and $\mathbf{b}_k$ are the trainable weights and biases of the k-th hidden layer, and the sine function is applied element-wise. Here, $k = 0$ corresponds to the first hidden layer, and $k = L-1$ corresponds to the last hidden layer.

Comment 7: "*The normalization of the coordinates is important in regression. This is a good sign of the care they took in their analysis.*"

Response: We appreciate this point. We kept the coordinate normalization step and made its role more explicit, because longitude, latitude, and time have different units and numerical ranges. The revised text also defines the normalized coordinate used as the SIREN input.

The corresponding revisions in Section 4.1.1 (Coordinate Normalization) are as follows.

- The input variables in SSH reconstruction have different physical units and numerical ranges. Longitude and latitude are measured in degrees, whereas time is measured in days. If these raw coordinates are directly used as network inputs, the difference in numerical scale may make the training less stable. Therefore, each coordinate is shifted and rescaled before being passed to the network.

  The normalized coordinate is denoted by $\tilde{\mathbf{s}} = (\widetilde{\mathrm{lon}}, \widetilde{\mathrm{lat}}, \widetilde{\mathrm{time}})$. This normalization is only a preprocessing step for the network input and does not change the physical meaning of longitude, latitude, or time. This preprocessing improves numerical conditioning and makes SIREN training more stable for the sparse multi-mission altimetry data used here.

Comment 8: "*Equations 4 and 7 show no weighting in the norm. Normally, there would be a metric tensor as part of the quadratic product, or at least a diagonal matrix that scales by local uncertainty. Their regions may be small enough that they can get away with this, but it should be acknowledged. Reviewer one may hint at this by asking whether variable a is changed regionally and how large the total variation constraint is.*"

Response: We thank the reviewer for raising this point. The present implementation does use unweighted norms, so we revised the text to state this modeling choice directly where the objective is defined, and to discuss its implication later in the limitations. We address three related issues below: the data-fidelity term, the spatial-gradient norm, and the regularization coefficient a.

For the data-fidelity term, the present loss uses an unweighted mean squared error over the available along-track SSH observations. This means that all observed SSH samples are assigned equal weight in the training objective. We keep this point in Section 4.2 (TV-based Spatial-Gradient Regularization), because it defines the objective actually used in this study:

- The unweighted mean squared error can also be related to a standard least-squares interpretation, in which the available along-track observations are treated with a common error variance. Thus, the present experiments use an ordinary least-squares data term in which the observed along-track samples contribute equally.

For the TV regularization term, we agree that the spatial-gradient norm should not be confused with a physical-distance gradient. In the present implementation, the TV penalty is computed from gradients with respect to the normalized longitude and latitude inputs and is used as a practical regularization measure. We have clarified this in Section 4.2:

- These gradients describe how rapidly the reconstructed SSH changes with respect to the normalized horizontal input coordinates. In the TV term, they are used as a practical regularization measure rather than as physical SSH gradients per unit distance.

For the regularization coefficient a, we have clarified that it is not changed regionally in the current implementation. Instead, a is treated as a global hyperparameter for each experimental configuration. It controls the overall balance between data fidelity and TV regularization, and should not be interpreted as a local uncertainty parameter. The corresponding revision in Section 4.2 (TV-based Spatial-Gradient Regularization) is as follows:

- where $a > 0$ controls the relative contribution of the TV regularization term compared with the data-fidelity loss. A smaller a gives greater relative emphasis to matching the observed along-track SSH values, whereas a larger a imposes stronger control on local spatial gradients. 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. Several trial values of a are tested in preliminary runs by examining the benchmark error diagnostics available for the experiment and inspecting the reconstructed SSH and gradient fields for isolated noisy gradients or excessive smoothing.

We also added the limitation to Section 6.2 (Applicability and limitations), together with possible extensions involving observation-dependent error variances, physically scaled spatial-gradient penalties, and spatially varying regularization coefficients:

- The balance between data fitting and regularization is another practical issue. The objective function should be understood as a regularized least-squares formulation rather than a full

Bayesian model. The data-fidelity loss treats the along-track observations as having a common error variance, so the observed samples contribute equally in the current implementation. The TV term acts as a regularization preference on spatial gradients, and the coefficient a is a global hyperparameter rather than a local uncertainty- or region-dependent parameter. Strong regularization can improve spatial coherence but may smooth physically meaningful variability, whereas weak regularization can leave unstable gradients in sparsely observed areas. A more systematic strategy for choosing the regularization strength, and possible extensions using observation-dependent error variances, physically scaled spatial-gradient penalties, spatially varying regularization coefficients, or Bayesian uncertainty quantification, would be useful in future work.

Comment 9: "*Also, for comparison to Bayesian methods, they can characterize this as maximizing a likelihood with Gaussian distributions for the observation uncertainty and the total variation metric. This would also give guidance for how to weight the trade-off between data fit and structure, which is currently left to the parameter a. Regarding the parameter a: on line 203, it seems to be renamed as λ, as flagged by reviewer 1.*"

Response: We thank the reviewer for this suggestion. We have revised the manuscript to make the statistical interpretation of the loss more explicit, to clarify what the trade-off parameter a does and does not represent, and to correct the inconsistent notation for the regularization coefficient.

On the probabilistic interpretation, we agree that the data-fidelity part of the objective can be discussed from this perspective. In the revised manuscript, we clarify that the data-fidelity loss is an unweighted mean squared error. Under a standard least-squares interpretation, this corresponds to treating the available along-track SSH observations with a common error variance. We keep the TV term as a spatial-gradient regularization term, rather than recasting it as a full probabilistic prior. The corresponding revision in Section 4.2 (TV-based Spatial-Gradient Regularization) is as follows:

- This term constrains the network output $\mathrm{SSH}_{\mathrm{rec},i}$ to match the observed value $\mathrm{SSH}_{\mathrm{true},i}$ at the available satellite-track locations. The unweighted mean squared error can also be related to a standard least-squares interpretation, in which the available along-track observations are treated with a common error variance. Thus, the present experiments use an ordinary least-squares data term in which the observed along-track samples contribute equally.

For the trade-off parameter, we now state more clearly that a controls the balance between the data-fidelity loss and the TV regularization term. It is treated as a hyperparameter for each experimental configuration and is kept fixed over the reconstruction domain. The corresponding revision in Section 4.2 (TV-based Spatial-Gradient Regularization) is as follows:

- where $a > 0$ controls the relative contribution of the TV regularization term compared with the data-fidelity loss. A smaller a gives greater relative emphasis to matching the observed along-track SSH values, whereas a larger a imposes stronger control on local spatial gradients. 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. Several trial values of a are tested in preliminary runs by examining the benchmark error diagnostics available for the experiment and inspecting the reconstructed SSH and gradient fields for isolated noisy gradients or excessive smoothing.

We also clarified the limitation of this formulation in Section 6.2 (Applicability and limitations). The current objective is a regularized least-squares objective, not a calibrated Bayesian posterior, so it does not provide posterior uncertainty estimates. The corresponding revision is as follows:

- The balance between data fitting and regularization is another practical issue. The objective function should be understood as a regularized least-squares formulation rather than a full Bayesian model. The data-fidelity loss treats the along-track observations as having a common error variance, so the observed samples contribute equally in the current implementation. The TV term acts as a regularization preference on spatial gradients, and the coefficient a is a global hyperparameter rather than a local uncertainty- or region-dependent parameter. Strong regularization can improve spatial coherence but may smooth physically meaningful variability, whereas weak regularization can leave unstable gradients in sparsely observed areas. A more systematic strategy for choosing the regularization strength, and possible extensions using observation-dependent error variances, physically scaled spatial-gradient penalties, spatially varying regularization coefficients, or Bayesian uncertainty quantification, would be useful in future work.

We also thank the reviewer for pointing out the notation inconsistency. The regularization coefficient was incorrectly denoted as λ in one place in the previous version. We have corrected this and now use a consistently throughout the manuscript.

Comment 10: "*Reviewer one also asked about the correlation between training and test data, which is an important point. The autocorrelation decay times of the mapped SSH fields and the model SSH fields could be used as a time scale to see how independent the values are.*"

Response: We thank the reviewer for this comment. This is a useful qualification of the OSE validation: even when training and validation samples are separated by mission, they are still samples from the same evolving SSH field and may remain correlated over finite spatial and temporal scales.

We therefore revised the Discussion to state that the OSE metrics should be read as reconstruction skill under sparse sampling, not as performance on fully independent dynamical states. We also added the reviewer's suggested autocorrelation-scale diagnostic as a useful future check on the effective independence of validation samples.

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 11: "*My reading of the text is that, in addition to withholding the Cryosat-2 altimeter, they train on the first 15 days of January and the last 15 days of March and then test on the withheld data in between. In this case, why withhold any altimeter data? Should we expect that Cryosat would be that different from the other altimeters, so why not compare to all of them? This would also allow us to look at the presumably increasing and decreasing area-averaged RMS error day by day during the period between the two training data sets.*"

Response: We thank the reviewer for this comment. The previous text did not explain the OSE validation design clearly enough, and we can see how it could be read as two separated training windows with a test period in between. After checking the data-preparation and training scripts, we confirmed that this two-window interpretation was not how the experiment was implemented. The confusion arose because the earlier text used a compact training/testing wording for the benchmark periods, instead of separating the observations used to train the INR from the observations withheld for validation. In the revised manuscript, we state explicitly that the OSE experiment follows the leave-one-mission-out setting of the Ocean Data Challenges benchmark: non-CryoSat-2 observations are used for training, and CryoSat-2 new-orbit observations are used as the withheld validation data. This separation is part of the benchmark design; it is not based on an assumption that CryoSat-2 samples dynamically different SSH variability from the other missions. The OSE results should therefore be interpreted as a benchmark consistency test against withheld along-track observations over the same region and period.

The main clarification was added in Section 5.1 (Experimental configurations and evaluation protocols), and the OSE data-preparation subsection was shortened to refer back to that protocol:

- 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.

- Following the protocol defined in Section 5.1, the OSE input data consist of real along-track SSH observations from SARAL/Altika (alg), Haiyang-2B (h2b), Jason-3 (j3), Sentinel-3A (s3a), and Sentinel-3B (s3b), while CryoSat-2 new-orbit observations (c2n) are retained for withheld along-track validation. This leave-one-mission-out setting follows the predefined benchmark protocol. It allows reconstruction skill to be assessed against real satellite observations that are not used during model training, while still reflecting the same evolving ocean period.

On the day-by-day error evolution, we agree that temporal variability of the evaluation error is useful. In the OSE experiment, however, a true full-domain area-averaged RMS error is not available because there is no complete gridded SSH truth. Also, because the experiment is not based on two separated training windows, the daily curve should not be interpreted as error growth or decay across a gap between training periods. We therefore use the withheld CryoSat-2 tracks to show daily along-track validation behaviour. The revised temporal-performance figure shows the daily number of validation samples, the daily RMSE skill scores for INR and OI, and their paired daily difference.

The corresponding revision in Section 5.4.3 (Detailed Comparison with Optimal Interpolation) is as follows:

- **(2) Temporal performance: RMSE skill score.** The temporal behaviour of the reconstruction is assessed using the RMSE skill score, denoted as $\mathrm{RMSE}_{\mathrm{SS}}$. Figure 10 summarizes the temporal variability of the evaluation data and the corresponding reconstruction skill. Figure 10a shows the daily number of available along-track observations used for the evaluation. The observation count varies noticeably over time, reflecting the non-uniform temporal sampling of the along-track data.

  Figure 10b shows the daily RMSE skill scores for the INR-based reconstruction and OI. The INR method yields consistently higher $\mathrm{RMSE}_{\mathrm{SS}}$ values than OI over most of the valid paired evaluation days, indicating better temporal reconstruction skill. Across 49 paired evaluation days, the daily RMSE skill score is 0.748 ± 0.120 for INR and 0.631 ± 0.151 for OI, where the uncertainty denotes the standard deviation over days.

  Figure 10c further reports the paired daily difference between INR and OI. We define this difference as $\Delta = \mathrm{RMSE}_{\mathrm{SS}}^{\mathrm{INR}} - \mathrm{RMSE}_{\mathrm{SS}}^{\mathrm{OI}}$. Positive values indicate days when INR outperforms OI. The paired daily difference is 0.117 ± 0.095, with an approximate 95 % confidence interval of $[0.090, 0.143]$, and INR outperforms OI on 95.9 % of the evaluated days. The positive confidence interval and the large fraction of positive daily differences indicate that the INR improvement is not driven by a few isolated dates, but is maintained over most of the OSE evaluation period. The remaining day-to-day variability indicates that the improvement is not identical on every date. Some dates are more difficult because the withheld tracks provide fewer samples or sample more challenging parts of the domain.

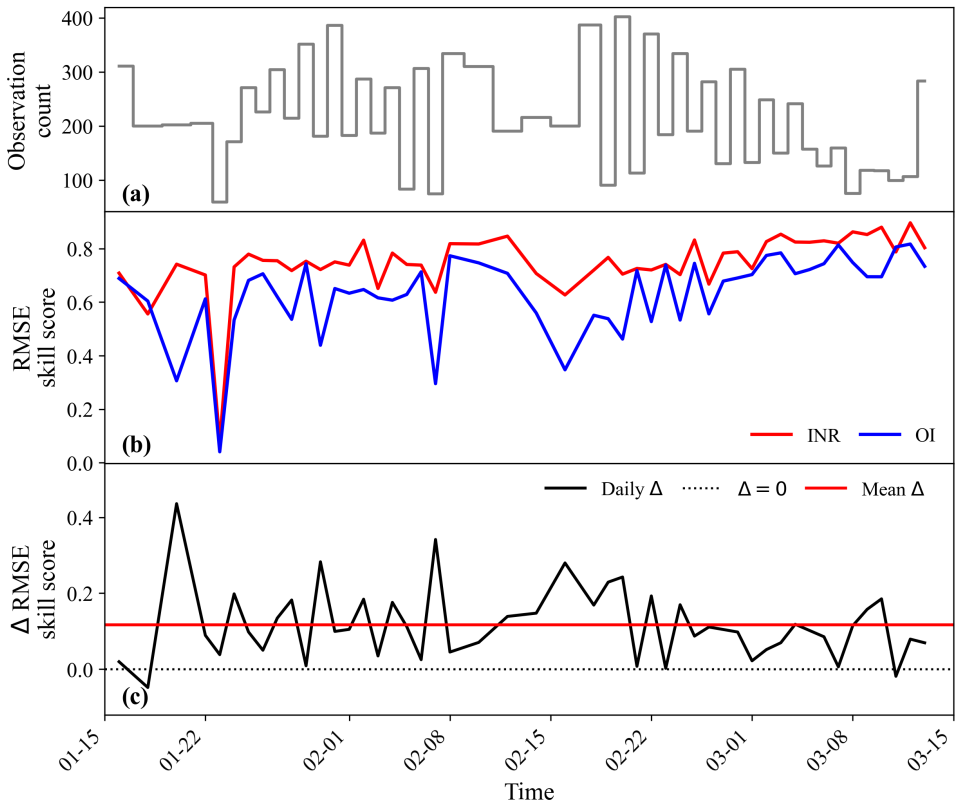

Figure 7: Temporal variability of the reconstruction skill. (a) Daily number of available along-track observations used for the evaluation. (b) Daily RMSE skill scores for the INR-based reconstruction and OI. (c) Daily paired difference in RMSE skill score, $\Delta = \mathrm{RMSE}_{\mathrm{SS}}^{\mathrm{INR}} - \mathrm{RMSE}_{\mathrm{SS}}^{\mathrm{OI}}$; the dotted line denotes $\Delta = 0$, and the red line denotes the mean difference. Only valid paired evaluation days are shown.

Comment 12: "*In their exposition of the spatial gradients, they don't mention the metric terms from lat and lon to distance, the scaling from SSH to geostrophic velocity, or the concerns about cyclogeostrophy and other ageostrophic features at small scales. If they mean gradients instead of velocities, then why call them* v_lat *and* v_lon*? However, later they describe the ability to estimate velocity as a benefit of the method, so these concerns are valid.*"

Response: We thank the reviewer for pointing this out. The previous notation could indeed blur the distinction between SSH gradients and velocity-like quantities. We have removed the misleading velocity-style notation and now describe the quantities as SSH-gradient components.

In the method section, the spatial-gradient components are now defined with respect to the normalized longitude and latitude coordinates. We also state that this gradient is used for regularization, while the physical-distance gradient shown in the qualitative OSSE reconstruction figure (Figure 8) is computed separately. The corresponding revision in Section 4.2 (TV-based Spatial-Gradient Regularization) is as follows:

- Spatial gradients of the reconstructed SSH field. Let $\Phi_\theta(\tilde{\mathrm{s}})$ denote the reconstructed SSH field, where $\tilde{\mathrm{s}} = (\widetilde{\mathrm{lon}}, \widetilde{\mathrm{lat}}, \widetilde{\mathrm{time}})$ is the normalized spatio-temporal coordinate. The spatial gradient is computed with respect to the normalized longitude and latitude coordinates:

$$\nabla_{\mathrm{sp}}\Phi_\theta(\tilde{\mathrm{s}}) = \left(\frac{\partial\Phi_\theta}{\partial\widetilde{\mathrm{lon}}}, \frac{\partial\Phi_\theta}{\partial\widetilde{\mathrm{lat}}}\right) = (g_{\mathrm{lon}}, g_{\mathrm{lat}}), \tag{5}$$

  where g_{lon} and g_{lat} denote the two spatial-gradient components. These gradients describe how rapidly the reconstructed SSH changes with respect to the normalized horizontal input coordinates. In the TV term, they are used as a practical regularization measure rather than as physical SSH gradients per unit distance.

For Figure 8, we clarified how the bottom-row SSH-gradient magnitude is obtained from the differentiable SIREN field. The derivatives first returned by automatic differentiation are with respect to the normalized input coordinates; they are then rescaled to longitude and latitude derivatives and converted to derivatives per physical distance on the sphere. We also state that this plotted quantity is a physical SSH-gradient magnitude, not a direct velocity or vorticity estimate. The corresponding revision in Section 5.5.3 (Evaluation of reconstructed sea surface height) is as follows.

- The physical SSH-gradient magnitude shown in the bottom row is derived from the automatic differentiation of the SIREN representation. This diagnostic directly uses one of the key properties of the INR framework, namely that the reconstructed SSH field is represented as a continuous and differentiable function of the spatio-temporal coordinates. Let $\eta(\mathrm{lon}, \mathrm{lat}, t) = \Phi_\theta(\tilde{\mathrm{s}})$ denote the reconstructed SSH field, where $\tilde{\mathrm{s}} = (\widetilde{\mathrm{lon}}, \widetilde{\mathrm{lat}}, \widetilde{\mathrm{time}})$ is the normalized spatio-temporal coordinate, and lon and lat denote longitude and latitude in radians. Automatic differentiation first gives the derivatives of Φ_θ with respect to the normalized input coordinates. These normalized-coordinate derivatives are then rescaled to derivatives with respect to longitude and latitude:

$$\frac{\partial\eta}{\partial\mathrm{lon}} = \frac{\partial\Phi_\theta}{\partial\widetilde{\mathrm{lon}}}\frac{\partial\widetilde{\mathrm{lon}}}{\partial\mathrm{lon}}, \qquad \frac{\partial\eta}{\partial\mathrm{lat}} = \frac{\partial\Phi_\theta}{\partial\widetilde{\mathrm{lat}}}\frac{\partial\widetilde{\mathrm{lat}}}{\partial\mathrm{lat}}. \tag{6}$$

  These longitude and latitude derivatives are then converted into physical-distance derivatives using the Earth-radius scale factors:

$$\frac{\partial\eta}{\partial x} = \frac{1}{R\cos(\mathrm{lat})}\frac{\partial\eta}{\partial\mathrm{lon}}, \qquad \frac{\partial\eta}{\partial y} = \frac{1}{R}\frac{\partial\eta}{\partial\mathrm{lat}}, \tag{7}$$

  where R is the Earth radius. The physical SSH-gradient magnitude is defined as

$$|\nabla\eta| = \left[\left(\frac{\partial\eta}{\partial x}\right)^2 + \left(\frac{\partial\eta}{\partial y}\right)^2\right]^{1/2}. \tag{8}$$

  For visualization, we plot $10^5|\nabla\eta|$, corresponding to units of m per 100 km. Therefore, the bottom-row panels should be interpreted as physical SSH-gradient magnitudes derived from the differentiable SSH representation, rather than as direct velocity or vorticity fields. Their further conversion into geostrophic velocity would require additional dynamical assumptions.

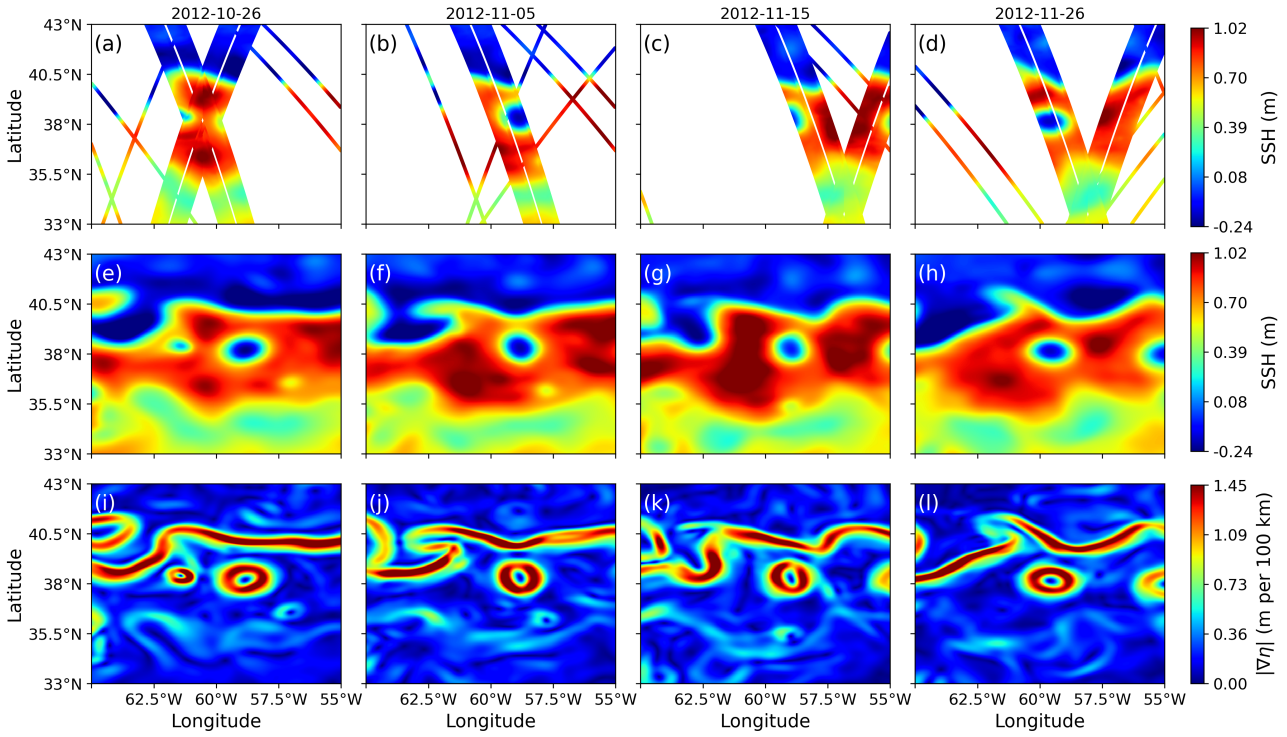

Figure 8: Qualitative OSSE snapshots of SSH reconstruction from sparse satellite-like observations. The four columns show selected dates from four approximately equal sub-periods of the evaluation period. Rows show sparse nadir/SWOT-like pseudo-observations, reconstructed SSH, and the physical SSH-gradient magnitude derived from the differentiable SIREN field. SSH is in metres, $|\nabla\eta|$ is in metres per 100 km, and the full temporal evolution is provided in Supplementary Video S1 (`https://zenodo.org/records/21231911`).

Comment 13: *"In section 5.1, line 231, they introduce the "RMSE-based score". It would be easier to understand as an RMSE skill score."*

Response: We agree that "RMSE skill score" is clearer. We have replaced "RMSE-based score" with this wording and use the notation $\mathrm{RMSE}_{\mathrm{SS}}$ consistently. We also consolidated the standalone RMSE definition into the same item, so that RMSE is introduced as the quantity used to form the skill score rather than as a separate evaluation metric.

The corresponding revisions in Section 5.2 (Evaluation Indexes) are as follows.

- **(1) RMSE skill score.** The RMSE between reconstructed and true SSH is first computed as

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left[\mathrm{SSH}_{\mathrm{rec}}(\mathrm{lon}_i, \mathrm{lat}_i, \mathrm{time}_i) - \mathrm{SSH}_{\mathrm{true}}(\mathrm{lon}_i, \mathrm{lat}_i, \mathrm{time}_i)\right]^2}, \tag{9}$$

  where N denotes the number of evaluation points. $\mathrm{SSH}_{\mathrm{rec}}(\mathrm{lon}_i, \mathrm{lat}_i, \mathrm{time}_i)$ is the reconstructed SSH at the i-th spatio-temporal location, and $\mathrm{SSH}_{\mathrm{true}}(\mathrm{lon}_i, \mathrm{lat}_i, \mathrm{time}_i)$ is the corresponding reference SSH value.

Following the common definition of a skill score, this RMSE is normalized by the root-mean-square amplitude of the reference SSH field:

$$\mathrm{RMSE}_{\mathrm{SS}} = 1 - \frac{\mathrm{RMSE}}{\mathrm{RMS}(\mathrm{SSH}_{\mathrm{true}})}, \tag{10}$$

where $\mathrm{RMS}(\cdot)$ denotes the root-mean-square of the reference SSH field. A score of 1 corresponds to a perfect reconstruction, while larger values indicate better reconstruction skill.

Comment 14: "*The notation in equation 12 is a little hard to follow. I could be helped by writing PSD as an explicit function of w and f, and perhaps changing to more traditional notation for frequency and wavenumber.*"

Response: We thank the reviewer for this suggestion. We have rewritten the PSD score explicitly as a function of spatial wavenumber k and temporal frequency f, so that the score is clearly evaluated at each wavenumber–frequency bin. The revised text in Section 5.2 (Evaluation Indexes) is as follows:

- **(2) Wavenumber–frequency PSD skill score.** 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

$$\mathrm{PSD}_{\mathrm{SS}}(k, f) = 1 - \frac{\mathrm{PSD}[\mathrm{SSH}_{\mathrm{rec}} - \mathrm{SSH}_{\mathrm{true}}]\,(k, f)}{\mathrm{PSD}[\mathrm{SSH}_{\mathrm{true}}]\,(k, f)}. \tag{11}$$

  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.

Comment 15: "*On line 239, they give minimum resolvable wavelengths in x and time, but don't include why. Their total variation metric is isotropic, but that doesn't mean that their learned weights will provide isotropic estimates. I suggest that they discuss this.*"

Response: We thank the reviewer for this comment. We revised the text to explain that λ_x and λ_t are taken from the 0.5 contour of the normalized wavenumber–frequency PSD skill score, following the Ocean Data Challenges protocol. We also clarified that these are effective-scale diagnostics under the adopted spectral-skill criterion, not grid spacings or direct resolutions. The corresponding revision in Section 5.2 (Evaluation Indexes) is 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.

Comment 16: "*Figures 3 and 4 seem to be somewhat redundant, since 3 is the absolute error at points and 4 is a pixel-averaged RMS error. They should address this question for other readers, and, if it is a repeat, what is the reason for plotting such similar measures? It might instead be useful to show a cumulative distribution function for the signed differences between the observations and the reconstruction methods.*"

Response: This comment helped us see that the roles of the previous two error figures were not sufficiently distinct, since both mainly emphasized error magnitude. We therefore reorganized the diagnostics into one combined figure.

For the distribution of errors, we replaced the pointwise absolute-error map with a cumulative distribution function of signed SSH residuals at the withheld along-track observation locations, as suggested by the reviewer. For the spatial structure of the remaining errors, we retained the gridded RMSE maps and present them as complementary diagnostics in the same figure. The corresponding text and caption were revised as follows.

The corresponding revisions in Section 5.4.3 (Detailed Comparison with Optimal Interpolation) are as follows.

- **(1) Error distribution and spatial RMSE.** Figure 9 provides a combined error diagnostic for the INR-based reconstruction and the OI baseline. Figure 9a shows the cumulative distribution function (CDF) of signed SSH residuals at the withheld along-track observation locations. The residual is defined as $\mathrm{SSH}_{\mathrm{rec}} - \mathrm{SSH}_{\mathrm{obs}}$, so positive values indicate overestimation and negative values indicate underestimation relative to the withheld observations. Compared with OI, the INR residuals are more tightly concentrated around zero, indicating a smaller residual spread and little systematic offset.

  Figures 9b and 9c show gridded RMSE maps for the INR-based reconstruction and OI, respectively. The maps are obtained by assigning the withheld along-track residuals to spatial grid cells and computing the RMSE within each cell, with the withheld observations used as the reference values. Overall, the INR-based reconstruction exhibits lower RMSE values than OI over much of the domain, while localized high-RMSE regions are weaker and less extensive in the INR map.

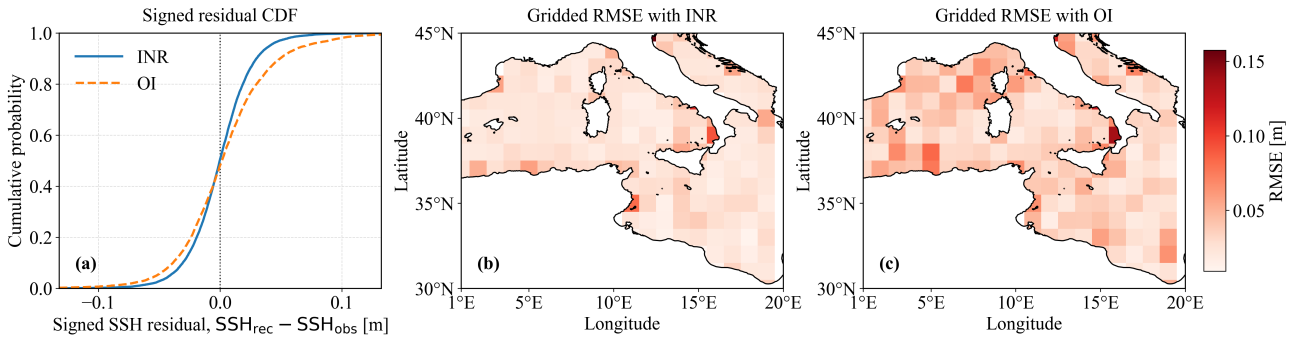

Figure 9: Combined error diagnostics for the INR-based reconstruction and OI. (a) Cumulative distribution function (CDF) of signed SSH residuals, defined as $\mathrm{SSH}_{\mathrm{rec}} - \mathrm{SSH}_{\mathrm{obs}}$, at withheld along-track observation locations. (b, c) Gridded RMSE maps for INR and OI, respectively, computed from withheld residuals within each spatial grid cell. Coastlines are shown in panels (b) and (c).

Comment 17: "*The caption for Figure 5 seems confused. The first panel is not the along-track SSH time series, as far as I can tell. It seems to be a count of daily observations in that time range. In Figure 5, the performance of the INR is always above the OI, but this figure prompts the desire for error bars on the differences. How variable are the performances at different times (for the same training and testing rubric) and/or with different training and testing ranges?*"

Response: We thank the reviewer for this comment. We agree that the caption was misleading: the first panel was not an along-track SSH time series, but the daily number of available along-track observations used in the evaluation. We corrected both the caption and the main text.

To address the variability of the INR–OI difference, we reorganized the figure and placed the revised version directly below. The figure now contains three panels: (a) the daily number of available along-track observations, (b) the daily RMSE skill scores for INR and OI, and (c) the paired daily difference in RMSE skill score, defined as $\Delta = \mathrm{RMSE}_{\mathrm{SS}}^{\mathrm{INR}} - \mathrm{RMSE}_{\mathrm{SS}}^{\mathrm{OI}}$. This layout makes it easier to see whether the INR improvement persists across the evaluation period or is driven by a few isolated dates. The corresponding revised figure and caption in Section 5.4.3 (Detailed Comparison with Optimal Interpolation) are as follows.

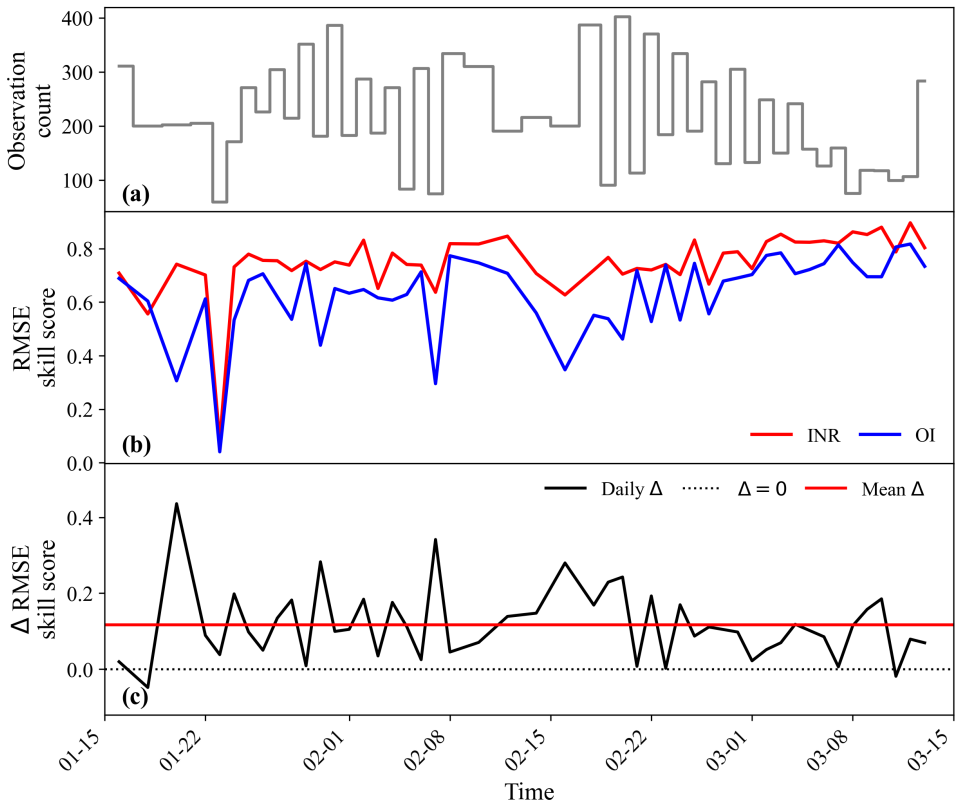

Figure 10: Temporal variability of the reconstruction skill. (a) Daily number of available along-track observations used for the evaluation. (b) Daily RMSE skill scores for the INR-based reconstruction and OI. (c) Daily paired difference in RMSE skill score, $\Delta = \mathrm{RMSE}_{\mathrm{SS}}^{\mathrm{INR}} - \mathrm{RMSE}_{\mathrm{SS}}^{\mathrm{OI}}$; the dotted line denotes $\Delta = 0$, and the red line denotes the mean difference. Only valid paired evaluation days are shown.

The corresponding revised text in Section 5.4.3 (Detailed Comparison with Optimal Interpolation) is as follows.

- **(2) Temporal performance: RMSE skill score.** The temporal behaviour of the reconstruction is assessed using the RMSE skill score, denoted as $\mathrm{RMSE}_{\mathrm{SS}}$. Figure 10 summarizes the temporal variability of the evaluation data and the corresponding reconstruction skill. Figure 10a shows the daily number of available along-track observations used for the evaluation. The observation count varies noticeably over time, reflecting the non-uniform temporal sampling of the along-track data.

  Figure 10b shows the daily RMSE skill scores for the INR-based reconstruction and OI. The INR method yields higher $\mathrm{RMSE}_{\mathrm{SS}}$ values than OI over most valid paired evaluation days. Across 49 paired evaluation days, the mean daily $\mathrm{RMSE}_{\mathrm{SS}}$ is 0.748 for INR and 0.631 for OI, with standard deviations of 0.120 and 0.151, respectively. This indicates that INR improves the average daily validation skill and also shows a slightly smaller day-to-day spread than OI.

Figure 10c further reports the paired daily difference between INR and OI. We define this difference as $\Delta = \mathrm{RMSE}_{\mathrm{SS}}^{\mathrm{INR}} - \mathrm{RMSE}_{\mathrm{SS}}^{\mathrm{OI}}$. Positive values indicate days when INR outperforms OI. The paired daily difference is 0.117 ± 0.095, with an approximate 95 % confidence interval of $[0.090, 0.143]$, and INR outperforms OI on 95.9 % of the evaluated days. The positive confidence interval and the large fraction of positive daily differences indicate that the INR improvement is not driven by a few isolated dates, but is maintained over most of the OSE evaluation period. The remaining day-to-day variability indicates that the improvement is not identical on every date. Some dates are more difficult because the withheld tracks provide fewer samples or sample more challenging parts of the domain.

Comment 18: "*For Table 1 and its discussion, they use the notation* $u(RMSE_S)$ *without comment, but the original discussion in section 5.1 just uses* $RMSE_S$*. I don't see the point of the extra* $u()$*.*"

Response: We thank the reviewer for pointing out this inconsistency. We removed the unexplained $u(\cdot)$ notation. In the revised text, $\mathrm{RMSE}_{\mathrm{SS}}(t_k)$ denotes the RMSE skill score on an individual valid evaluation day, while the quantity reported in the multi-method comparison table is its temporal mean, $\overline{\mathrm{RMSE}_{\mathrm{SS}}}$. This mean value is now defined explicitly before the table. The corresponding revisions in Section 5.4.4 (Comprehensive comparison with multiple methods) are as follows.

- This section provides a broader evaluation of the proposed approach by comparing it against several representative reconstruction methods. The assessment relies on the two quantitative metrics introduced in Section 5.2, namely the temporal mean of the RMSE skill score, denoted as $\overline{\mathrm{RMSE}_{\mathrm{SS}}}$, and the PSD-score-based effective spatial scale λ_x. Here, $\overline{\mathrm{RMSE}_{\mathrm{SS}}}$ is defined as

$$\overline{\mathrm{RMSE}_{\mathrm{SS}}} = \frac{1}{N_{\mathrm{day}}} \sum_{k=1}^{N_{\mathrm{day}}} \mathrm{RMSE}_{\mathrm{SS}}(t_k), \tag{12}$$

  where N_{day} denotes the number of valid evaluation days, and $\mathrm{RMSE}_{\mathrm{SS}}(t_k)$ denotes the RMSE skill score computed from the RMSE skill-score definition on the kth valid evaluation day. The results are summarized in Table 2.

  As shown in Table 2, INR achieves a slightly lower $\overline{\mathrm{RMSE}_{\mathrm{SS}}}$ than WaveVar, but it gives the smallest λ_x among all compared methods. This indicates that INR does not provide the best temporal-mean global error score in this comparison, but shows a more favorable PSD-score-based spatial effective-scale diagnostic under the adopted PSD-skill criterion. This behavior is consistent with the characteristics of the proposed framework. The SIREN-based representation provides a continuous coordinate-to-SSH mapping for describing spatially and temporally continuous SSH variability, while the gradient-based regularization constrains unstable or noisy variations in the learned field. Together, these two components help maintain competitive overall reconstruction accuracy while improving the spectral consistency and spatial coherence of the reconstructed SSH field.

Table 2: Results of different methods using satellite altimetry observations in the Mediterranean region. $\overline{\mathrm{RMSE}_{\mathrm{SS}}}$ denotes the temporal mean of the RMSE skill score over the valid evaluation days. Bold values denote best performance.

Method	$\overline{\mathrm{RMSE}_{\mathrm{SS}}}$	λ_x (km)
OI	0.674136	148
$\mathrm{BFN}_{\mathrm{QG}}$	0.720857	125
$\mathrm{BFNQG}_{\mathrm{coast}}$	0.724554	125
WaveVar	**0.759956**	118
INR	0.747545	**106**

Comment 19: "*For Figure 7, they should tell the dates they chose to plot, so the reader can have some idea of the evolution, and say why they chose these times. I can understand choosing arbitrary times, but random is a very specific term and could mean irregularly spaced, which triggers the question, why not pick four regularly spaced dates during the period?*"

Response: We thank the reviewer for this comment. The word "random" was indeed inappropriate, because it could imply a stochastic or irregular choice of dates. We removed this wording and made the snapshot selection reproducible: the evaluation period is divided into four approximately equal sub-periods, and the middle valid snapshot from each sub-period is shown. These snapshots are used only for qualitative illustration; the quantitative scores are still computed over all valid evaluation days.

We also added a note that the full temporal evolution is provided in Supplementary Video S1. The revised figure and caption are shown above with the response to Comment 12, where the physical SSH-gradient diagnostic is discussed. The related main-text description was revised as follows.

The corresponding revisions in Section 5.5.3 (Evaluation of reconstructed sea surface height) are as follows.

- Figure 8 shows qualitative examples at four selected times covering the evaluation period. Specifically, we divide the evaluation period into four approximately equal sub-periods and use the middle valid snapshot from each sub-period. The four columns compare the along-track observations, the reconstructed SSH fields, and the corresponding physical SSH-gradient magnitudes.

  The gradient diagnostic provides a complementary view of the reconstructed SSH field by showing whether rapid spatial transitions, such as fronts and eddy boundaries, are represented as coherent structures.

  While Figure 8 shows four selected snapshots, the full temporal evolution of the along-track observations, reconstructed SSH fields, and corresponding physical SSH-gradient magnitudes is provided in Supplementary Video S1. Together with the snapshot examples, the video provides a more complete visual assessment of the reconstruction behavior over the evaluation period.

Comment 20: "*The method performs well in their two examples, but the lack of UQ and the lack of transparency, e.g., the meaning of the weights learned in the neural net, are barriers to learning about the dynamics, such as the time scales used in the map.*"

Response: We thank the reviewer for this comment. We agree that the lack of uncertainty quantification and the limited interpretability of neural-network weights are important limitations, especially if the goal is to infer dynamical time scales rather than only reconstruct SSH.

We revised the Discussion to make this boundary explicit. The INR is presented as a reconstruction and diagnostic framework: it represents the observed spatio-temporal SSH variability within the target period, but the learned weights are not interpreted as physical parameters or dynamical time scales. The corresponding revisions in Section 6.1 (Methodological characteristics and relation to existing approaches) and Section 6.2 (Applicability and limitations) are as follows.

- 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.

- 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.

Thanks to the great comments/suggestions of the referee, the changes implemented in the revised manuscript have greatly enhanced its readability. As our best effort, we have also carefully worked through the entire manuscript to ensure it is error-free and to improve the readability.