To Reviewer 3

First, we appreciate the reviewer's valuable comments. For your comments, we gave our corresponding explanations and responses below:

**1. Validation methodology remains insufficiently justified**

The authors explain that DINEOF-type methods are adaptive reconstruction approaches and therefore do not require conventional training, validation, and testing datasets. While this clarification is helpful, it does not fully address the central concern regarding validation rigor.

The primary issue is not whether machine-learning-style training is used, but whether the reconstruction evaluation framework provides sufficiently robust evidence of generalization and avoids overly optimistic estimates arising from strong spatial and temporal autocorrelation.

The current evaluation still appears to rely primarily on reconstruction performance at observed locations rather than demonstrating reconstruction under realistic missing-data structures. While this may follow common DINEOF practice, adherence to previous practice alone does not necessarily demonstrate methodological robustness.

The manuscript would be significantly strengthened through additional validation experiments using more realistic missing-data scenarios, such as:

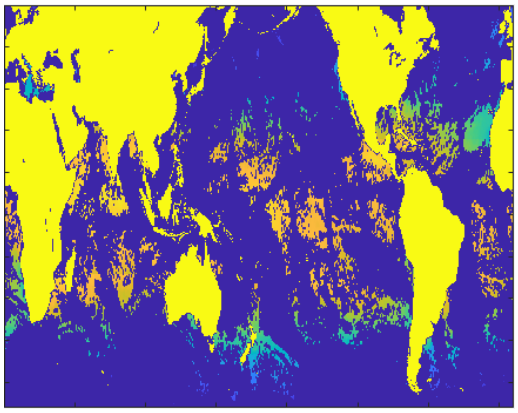
- contiguous spatial masking,
- withheld scenes or temporal holdouts,
- cloud-like masking structures,
- block-based validation approaches.

At minimum, the limitations of the current validation framework should be discussed more explicitly.

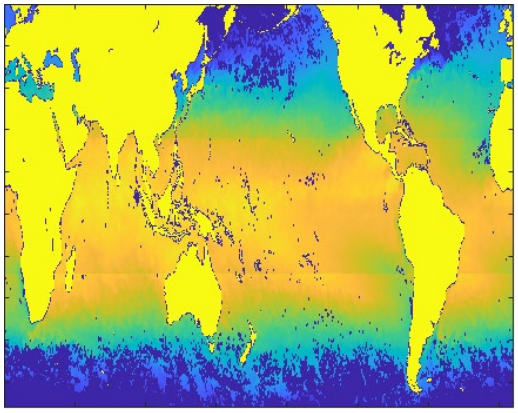
Response: Following the reviewer's suggestion, we generated a more challenging reconstruction dataset by applying the missing-value masks from the daily SST, SCHL, and SSW observations during 1 January–31 March 2022 to the original monthly datasets. Under this setting, the missing-data ratios increased to 89.88% (SST), 94.82% (SCHL), and 52.29% (SSW) in Subregion 1; 79.94%, 96.36%, and 64.31% in Subregion 2; and 89.28%, 95.76%, and 57.49% in Subregion 3, respectively. Figures below provide examples for April 2015, where the original monthly fields (right panels) are compared with the masked fields (left panels). The substantial increase in missing pixels demonstrates the severity of the imposed missing-data conditions and provides a rigorous test for reconstruction performance. In addition, since each subregion was normalized independently, discontinuities may appear along subregion boundaries.

It should be noted that the high proportion of missing data, particularly for SCHL, where the missing rate reaches approximately 95%, poses a substantial challenge for reconstruction. The accuracy is evaluated based on the reconstructed values of daily masked pixels and the corresponding monthly values at the same locations. As expected, this dramatic increase in missing-data ratios results in a pronounced decline in reconstruction performance. Consequently, the

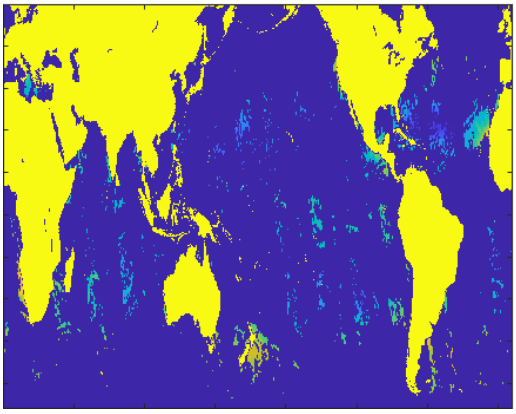
accuracies obtained from the masked datasets are significantly lower than those derived from the original datasets. This limitation affects not only DINEOF-type methods but also data-driven approaches, including machine learning and deep learning. We evaluated the reconstruction accuracy of T-DINEOF and Multi-DINEOF methods, as summarized in Table below. The results show that T-DINEOF consistently outperforms Multi-DINEOF, even under these extreme missing-data conditions, demonstrating the robustness of the T-DINEOF approach in handling highly incomplete datasets.



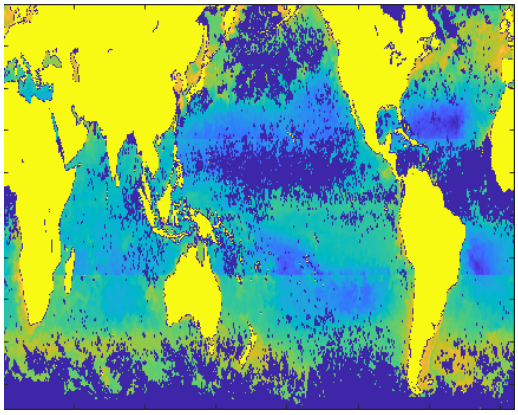
(a) SST with mask



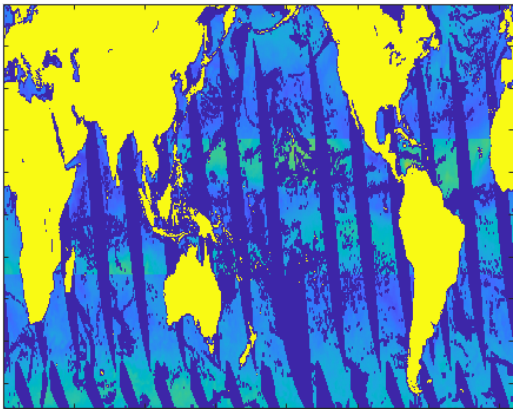
(b) original monthly SST at Apr, 2015



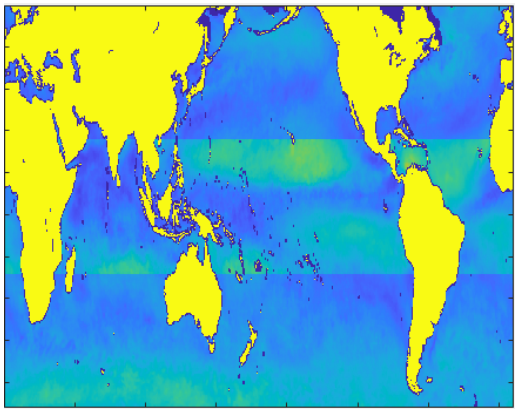
(c) SCHL with mask



(d) original monthly log(SCHL) at Apr, 2015



(e) SSW with mask



(f) original monthly SSW at Apr, 2015

| | T-DINEOF | | Multi-DINEOF | |
|---|---|---|---|---|
| | RMSE | Correlation coefficient | RMSE | Correlation coefficient |
| Subregion1 | 7.9185 | 0.8126 | 7.9624 | 0.8023 |
| Subregion2 | 8.0515 | 0.9738 | 8.0682 | 0.9704 |
| Subregion3 | 9.9562 | 0.6119 | 9.9845 | 0.6072 |

**2. Claims regarding low-correlation performance remain stronger than the evidence currently supports**

The authors clarify that correlations were calculated across the entire input tensor and report relatively weak correlations involving SSW. However, the conclusion that T-DINEOF performs better in low-correlation situations still appears stronger than the presented evidence justifies.

The current evidence is primarily based on a single dataset and a single low-correlation variable rather than multiple independent cases or systematically varying correlation regimes.

Therefore, either:

- stronger supporting analyses should be provided, or
- the conclusions regarding low-correlation performance should be moderated.

Currently, the results support improved performance for the presented dataset but do not yet convincingly demonstrate broader behavior under low-correlation conditions.

Response: As this study does not focus on analyzing the correlations among SST, SCHL, and SSW, we have limited the discussion regarding low-correlation scenarios in the conclusions. This ensures that the conclusions accurately reflect the scope of our work while avoiding potential misunderstandings for the reader.

In the revised manuscript, the original content in the Conclusion section has been modified as follows: “Even though the input SSW exhibits low correlation with SST and CHL, the T-DINEOF method is still able to achieve high-accuracy reconstruction of the SSW field.”

**3. Physical realism is improved but still only partially demonstrated**

The addition of gradient maps represents a meaningful improvement and partially addresses previous concerns regarding physical interpretation.

However, gradient magnitude alone does not fully establish that reconstructed fields preserve oceanographically meaningful structures.

Additional quantitative evidence would strengthen this section, for example:

- variance preservation analyses,

- anomaly structure comparisons,
- feature-preservation metrics,
- spatial spectral analyses,
- evaluation of mesoscale structure retention.

The manuscript has improved substantially in this area, but stronger quantitative evidence would increase confidence that improvements extend beyond pixel-level statistics.

Response: We selected three quantitative metrics to evaluate feature preservation: variance preservation (VP), Structural Similarity Index (SSIM), and anomaly structure (AS). VP was calculated as the ratio of the total variance of the reconstructed field to that of the original field at existing pixels, quantifying the fraction of total variance retained. SSIM was used to assess the structural similarity between reconstructed and original fields, taking into account luminance, contrast, and spatial structure. AS was computed as the correlation between the reconstructed and original anomaly fields, highlighting the ability to preserve fine-scale variations and key spatial features.

For the three study regions, T-DINEOF yielded the following results:

| Subregion | VP | SSIM | AS |
|---|---|---|---|
| 1 | 0.9987 | 0.5626 | 0.9992 |
| 2 | 0.9996 | 0.6011 | 0.9998 |
| 3 | 0.9981 | 0.5677 | 0.9991 |

These quantitative results, together with visual inspection of gradient maps, indicate that T-DINEOF can preserve the dominant spatial structures and major fine-scale features despite some smoothing of local details. The moderate SSIM values likely reflect the intrinsic smoothing behavior of DINEOF-type reconstructions during iterative low-rank approximation.

In the revised manuscript, we additionally included the following analyses and discussions in the revised manuscript:

“In addition to gradient-map comparisons, we introduced three quantitative metrics to evaluate feature preservation: variance preservation (VP), Structural Similarity Index (SSIM), and anomaly structure (AS). VP was calculated as the ratio of the total variance of the reconstructed field to that of the original field at existing pixels, quantifying the fraction of total variance retained. SSIM was used to assess the structural similarity between reconstructed and original fields by considering luminance, contrast, and spatial structure. AS was computed as the correlation between the reconstructed and original anomaly fields, highlighting the ability to preserve fine-scale variations and key spatial features. For the three study subregions, T-DINEOF yielded VP values of 0.9987, 0.9996, and 0.9981, SSIM values of 0.5626, 0.6011, and 0.5677, and AS values of 0.9992, 0.9998, and 0.9991, respectively.

Although the SSIM values (0.56–0.60) indicate that some local-scale differences remain between reconstructed and original fields, the consistently high VP and AS values demonstrate that the dominant variance and anomaly structures are effectively preserved. Combined with the gradient-map comparisons, these results suggest that T-DINEOF retains the major spatial features and fine-scale structures reasonably well, although some degree of smoothing is still present, which is consistent with the intrinsic characteristics of DINEOF-type low-rank reconstructions.”

**4. Computational characterization remains limited**

The authors acknowledge that T-DINEOF is computationally more expensive and discuss hardware specifications and relative convergence behavior.

However, the manuscript still lacks quantitative characterization of computational performance.

For a methodological contribution introducing a more computationally demanding tensor framework, readers would benefit from:

- approximate runtime comparisons,
- memory usage,
- scaling behavior,
- practical computational limitations.

Even approximate benchmarks would substantially improve the practical relevance of the study.

Response: We appreciate the reviewer's suggestion regarding computational characterization. In the revised manuscript, we recalculated and reported the approximate computational time of the proposed T-DINEOF and Multi-DINEOF methods based on data generated from the daily mask. For subregions 1–3, the running times of T-DINEOF were 29.27h, 35.83h, and 40.68h, respectively, whereas those of Multi-DINEOF were 17.44h, 16.01h, and 13.41h, respectively. Since the total amount of data processed is the same, with the only difference being its organization as a tensor in T-DINEOF and as a matrix in Multi-DINEOF, both methods required similar memory, approximately 6GB per subregion. However, we would like to clarify several important points regarding the interpretation of these benchmarks.

First, the primary objective of this study is methodological development rather than software engineering or computational optimization. The current implementation should therefore be considered a research-oriented prototype intended to demonstrate the feasibility and reconstruction capability of the proposed tensor-based extension. Considerable room for computational optimization likely remains, including memory management, parallelization strategies, and code-level acceleration.

Second, during the reconstruction experiments, multiple regions and/or different algorithms (e.g., T-DINEOF and Multi-DINEOF) were often executed simultaneously to reduce the overall experimental time. Furthermore, the experiments were performed on a single workstation that was simultaneously used for routine research and office work. Consequently, the reported computational costs cannot be interpreted as strict isolated benchmarks for a single region or a single algorithm, and the measured runtime and memory usage should therefore be regarded as approximate rather than rigorous performance metrics.

Third, similar to other DINEOF-type methods, the computational burden increases progressively as the number of retained modes increases because more information participates in the iterative reconstruction process. In the present study, both T-DINEOF and Multi-DINEOF were configured to compute up to 100 modes before selecting the optimal solution, despite the optimal mode number typically occurring around 30–40 modes. This choice was made to prioritize reconstruction capability evaluation and methodological consistency rather than computational efficiency. In practice, strategies such as early-stop criteria could substantially reduce runtime and memory consumption and represent an important direction for future optimization.

Overall, while the computational benchmarks reported in this study provide a preliminary indication of practical cost, they should not be interpreted as fully optimized performance estimates. Our current focus is primarily on extending the DINEOF-type framework and providing new methodological perspectives for multivariate ocean reconstruction. Further engineering-level optimization and software-oriented implementation could be pursued in future work, potentially in collaboration with computational specialists. Therefore, we did not include the exact computational times in the revised manuscript, in order to avoid potentially misleading the readers.