Response to the referees

We thank both referees for their reports, which improved the manuscript and, in two cases, the released dataset itself. Below we respond to each comment in turn, first to Referee 1 and then to Referee 2. Section 3 then lists every change made in the manuscript, ordered by section, and states which comment prompted it. Section and table numbers refer to the revised manuscript, and a marked-up version showing all changes is uploaded alongside this response.

Two comments led us to recompute and re-upload parts of the dataset: the replacement of the per-cell standard error of the mean by an inter-track dispersion measure (Referee 1, comment 9) and the addition of `n_tracks` as a provenance variable. The L3 nadir and SWOT collections on Hugging Face have been updated accordingly.

1 Response to Referee 1

We thank the reviewer for a careful report that improved both the manuscript and the released dataset.

General comments

Novelty / "mainly a STAC catalog". The reviewer is right that the component procedures (regridding, metadata standardisation, `int16`+`zlib`) are not inherently novel. We have more clearly reframed the contribution throughout as the harmonised, provenance-preserving integration of SWOT-era L3, L4, reanalysis, and Argo into one sample-invariant, queryable specification, and we draw an explicit line between this internal sample organisation and STAC-style asset discovery. (Abstract, Introduction, framework contrast, STAC discussion, Conclusions.)

Organisation / scattered resolution and depth. We have consolidated temporal cadence and depth/coverage into Table 1 and the appendix variable dictionary (Table A1), with a pointer in Section 2.1, so that both source and stored resolution are given for every product (GLORYS surface subset with 15 m currents; Argo full 0 to 2000 m; L4 wind stored daily from hourly).

Detailed comments

1. Why 2015 to 2025? The sources reach back to 1993 or the early 2000s, so 2015 is not a data-availability limit. The start date follows the salinity record: the multi-observation salinity product we use combines SMAP, SMOS, satellite SST, and in situ salinity through multivariate optimal interpolation, and reaches this configuration in 2015 when SMAP becomes available. From 2015 onward the sensor constellation is constant apart from the L3 SSH missions and SWOT from 2023, which add sampling without removing any modality. Earlier years would append data from a thinner observing system rather than extend this one. Storage supports the choice (581 GB for the extended period alone; reaching back to 1993 would roughly triple it), and the horizon matches recent work such as Martin et al. (2025). The choice is not a hard boundary: all of the processing code is open source, so users with more storage and processing power can append earlier years for back-compatible sources through the public code and TACO folder mode without restructuring the archive. (Changed in the Dataset description; Conclusions.)

2./3. Resolution with the data (Sec. 2.1; L150 to 151). This is covered by the consolidation above; source and stored resolution are now given for all modalities in Tables 1 and A1.

4. Stamps preserved or interpolated? (L158 to 159). We clarified this: L4 and reanalysis products keep their native grids and are only encoded, L3 nadir and SWOT observations are

conservatively binned rather than interpolated (with `obs_mean_lon/lat`, `n_obs`, and `n_tracks` retained), and Argo keeps its native coordinates and timestamps.

5. Products outside the period (L163 to 164). The list describes the DUACS product's mission heritage, not the missions active during 2015 to 2025; we corrected the wording.

6. Full profiles or surface? (L189 to 190). Argo profiles are kept in full over 0 to 2000 m, and GLORYS is included as a surface subset (15 m currents), following Martin et al. (2025). This is now stated in the depth column of the table.

7. "Up to 4" tiles (L221). The reviewer is correct: a query at a shared corner can overlap four tiles, and it is served transparently across all of them. We fixed the wording.

8. Section numbering 2.3.2/2.3.3. We resolved this structurally by converting the per-variable descriptions into run-in labels under the Level 4 and Level 3 data subsections.

9. Eq. 2, SEM at about one observation per cell. The reviewer's concern was well founded. The previous within-track quantity was not meaningful at this sampling, so we replaced it with an inter-track dispersion measure, σ_{track} (the observation-weighted standard deviation of per-track means), reported only for crossover cells where two or more passes overlap, and added `n_tracks` as a provenance variable. It is left undefined (NaN) elsewhere. Where it is defined (2.26 % of cells for L3 SSH and 4.04 % for SWOT), σ_{track} is physically meaningful but modest, with a median of 0.65 to 0.79 cm, about 7 % of the field standard deviation. We recomputed and re-uploaded the L3 nadir and SWOT collections. (Regridding section: the inter-track dispersion equation and surrounding prose; variable dictionary in Table A1.)

10. Compression vs Reichelt / ClimateBenchPress. We carried out the suggested benchmark on the five gridded products, using a ClimateBenchPress-style evaluation with SZ3, SPERR, ZFP, EBCC, and bitround variants. Under matched strict bounds these codecs reach higher ratios than the deployed encoding (for example L4 SSH 5.16× against 16.11×, and L4 SST 7.00× against 17.27×), and we report this in a new benchmark table. We have chosen to keep the deployed `int16`+`zlib` scheme, however, for a practical reason: these error-bounded codecs are typically deployed as array-store codecs for Zarr, where the decoder is applied transparently on read, and they are not directly compatible with the TACO framework, which stores self-contained files that any standard reader can open without an additional codec dependency. The deployed scheme keeps the archive simple, transparent, and directly readable, and preserves each field at full numerical resolution rather than degrading it to any single user's error tolerance, so we present the benchmark as a separate, additive result rather than as a replacement.

We also took the opportunity to separate two things that the original text conflated: the lossy-encoding error of the `int16` scheme, and the disk-space savings reported in the storage table. The two measure different quantities. The storage-table savings also include product restructuring rather than encoding alone, for example the consolidation of native hourly L4 wind fields into daily OceanTACO variables and of many individual L3 track files into a single regridded product, which is why their ratios are larger than the encoding step by itself. We revised the table captions and footnotes accordingly. We also corrected the L4 SST benchmark setup, since the source is natively `int16` and error bounds below that quantisation step are not meaningful. (Compression section; appendix Table A5; storage-table caption and footnotes.)

11. Fig. 4, `int16` significant digits. We added this: the caption now states that the absolute error is bounded by the `int16` quantisation step.

12. Scatter / relative error. We added an "RMSE / field std." column to the appendix compression table and a second figure panel mapping the relative error ($|\text{orig} - \text{comp}|/\sigma_{\text{local}}$). As

the reviewer anticipated, this shows the largest relative error in low-variability regions while the absolute error stays at the floor.

13. Other eddy-rich regions (L324 to 325). We added the Southern Ocean and Antarctic Circumpolar Current alongside the Gulf Stream and Kuroshio.

14. PSD example (L332 to 336). The reviewer's comment was correct: the original comparison did not enforce identical sampling, so the long-wavelength offset was a sampling artifact rather than a resolution difference. We replaced it with a matched L4-DUACS-against-SWOT comparison, in which the L4 field is sampled at the exact SWOT positions. The spectra now coincide at large scales (above roughly 100 to 150 km) and diverge only at short wavelengths, which isolates the genuine resolution difference, and constructing this matched comparison is itself straightforward with OceanTACO. To avoid over-interpreting the short-wavelength tail, we use the denoised SWOT L3 field and frame the divergence as a genuine resolution difference only over the roughly 30 to 150 km band, noting in the caption and text that the L3 spectrum flattens at the shortest scales near the effective-resolution limit of the product rather than resolving indefinitely.

15. Table A1, temporal resolution and GLORYS surface. Addressed in the updated table.

16. Table B1 (now Table A4), noise-to-signal and SSS units. We added the noise-to-signal column (see comment 12). On units, we have kept practical salinity but conceded the labelling: the product is practical salinity (PSS-78, dimensionless), inherited from CMEMS, whereas g/kg is absolute salinity (TEOS-10), a different quantity. We relabelled salinity consistently as PSS-78 throughout.

We are grateful for the reviewer's comments and believe they have improved the manuscript.

2 Response to Referee 2

We thank the reviewer for a positive and constructive report. We are glad the reviewer finds the data system useful for many purposes, and we agree that its value lies not in new algorithms but in the harmonised, provenance-preserving organisation of heterogeneous sources. We have taken the opportunity to sharpen the usage examples and to make the dataset's quality-control stance explicit.

General comments

Challenges of combining data from various sources. We fully agree, and this friction is central to the contribution. Harmonising these products required reconciling heterogeneous grids and geometries onto a common reference, aligning differing temporal cadences, and unifying incompatible fill-value, metadata, and licensing conventions across ten sources. These steps are documented in the Introduction and the Processing methods section. OceanTACO removes this friction once, under a single sample-invariant, provenance-preserving structure.

Detailed comments

Noise and spurious data, was a check performed within OceanTACO? Each constituent product is ingested at the published, quality-controlled processing level provided by its authoritative source. We rely on this provider-level quality control and deliberately do not layer additional processing on top of it. Different downstream tasks or use cases impose conflicting requirements. Some analyses call for aggressive outlier rejection, while data assimilation and machine-learning

reconstruction often need the raw observations and their error characteristics preserved, so a single fixed filtering step would suit some applications and silently degrade others. We have added a sentence to Section 2.4 making this stance explicit.

Usage examples in 3.3 and 3.5 are superficial. We agree the examples could be more concrete, and we have expanded both sections with specific pointers and a fuller account of the relevant methods (detailed below). We note one boundary: the ESSD data-description manuscript type scopes out "extensive analysis and interpretation of data," which "should seek publication in an appropriate regular journal." This applies to running a complete observation-system experiment or a full model-development study, not to describing how the dataset supports them or to situating them in the literature, both of which we have expanded here.

Impact of very large regions on subregional phenomena (3.3). This is an important point, and we realise the text was unclear: the eight "regions" are purely a storage-and-retrieval efficiency mechanism (regional tiling for selective I/O), not analysis regions. Any arbitrarily sized bounding box, or any user-defined subregion, can be retrieved, served transparently across up to four tiles where a box straddles tile boundaries (Section 2.4.1), and each product is preserved at its native resolution, down to about 2 km for SWOT. Subregional and fine-scale phenomena therefore remain fully accessible. The Hurricane Milton case study (Section 3.4) already demonstrates a localised subregional analysis at full resolution. We have added a clarifying sentence to Section 3.3 to state this explicitly.

Observation system experiments (3.3). A genuine OSE requires running a data-assimilation system with observing systems selectively withheld, which is a research study in its own right and falls under the "extensive analysis" that ESSD places out of scope for this manuscript type. What the dataset provides is the machinery that makes such experiments configurable, and we have made this concrete in Section 3.3: OceanTACO preserves per-mission identifiers and sensor-level indexing; the per-mission temporal coverage that determines which sensors are available for a given experiment is tabulated in the Appendix (Table A2); and the matched SWOT-versus-mapped spectral comparison in Section 3.2 already illustrates the kind of resolved-scale, mission-impact diagnostic these queries support.

Machine learning methods and specifics (3.5). We have expanded Section 3.5 to describe, at a high level, what each cited method does rather than referring to them collectively: supervised networks that fuse multiple satellite variables into gridded SSH, end-to-end learned variational assimilation schemes for nadir and wide-swath altimetry, dynamically constrained joint reconstruction of SSH and temperature from multi-sensor observations, generative assimilation of multi-modal satellite observations, and unsupervised spatio-temporal interpolation of altimetry maps. The sea surface temperature and salinity work is predominantly supervised super-resolution and gap filling. Their common requirement is consistently collocated multi-variable, multi-sensor inputs, which is exactly the harmonisation OceanTACO supplies, and the accompanying documentation shows how to build ML-ready PyTorch datasets from the same spatiotemporal queries.

We thank the reviewer again for the helpful and encouraging assessment.

3 List of changes made in the manuscript

The list follows the order of the manuscript. Referee comments are abbreviated as R1-n and R2, and all changes are marked in the uploaded track-changes version.

Abstract and Introduction

- Abstract: the sentence on the uniform internal structure now states that the contribution is a sample-invariant, provenance-preserving organisation rather than a new regridding or compression method. (R1 general)
- Introduction: the contribution list was rewritten from four items to three, dropping the claim about documentation and examples and restating the remaining items around the harmonised collection, the sample-invariant specification, and the reproducible access pattern. (R1 general)
- Framework contrast (Section 2 preamble): the comparison against OceanBench and related benchmarks now separates internal sample organisation from asset discovery. (R1 general)

Section 2.1, Data sources

- Added a sentence pointing to the consolidated resolution, cadence, coverage, and vertical extent in Table 1 and Table A1. (R1 general, R1-2, R1-3)
- Table 1 gained a *Temporal* column and a vertical-extent entry in the resolution column: GLORYS as a surface subset with currents at 15 m, Argo over the full 0 to 2000 m, and L4 wind stored daily from a native hourly product. (R1-2, R1-3, R1-6, R1-15)
- The four L4 and four L3 product descriptions were converted from `\subsubsection` headings to run-in bold labels, which removes the numbering inconsistency. (R1-8)
- DUACS: the mission list is now described as the product's mission heritage rather than as the missions active during the OceanTACO period. (R1-5)

Section 2.2, Temporal and spatial coverage

- Rewritten to present the release as two contiguous segments and to give the reason for the 2015 start: the multi-observation salinity product reaches its SMAP, SMOS, satellite SST, and in situ configuration in 2015, and the sensor constellation is constant from that year onward. The paragraph also states that earlier years can be appended through the public pipeline. (R1-1)
- Corrected the temporal index counts (857 core, 3010 extended) and split the storage figure into 311 GB for the core period and 581 GB for the extended period, replacing the single 324 GB figure. (arithmetic correction)

Section 2.3, Processing methods

- The numerical-compression step now distinguishes the dense gridded fields, which are encoded as scaled `int16`, from the sparse L3 along-track and SWOT products, which are retained at full `float32`. (R1-10)
- Added a paragraph stating that L4 and reanalysis products keep their native grids, that L3 nadir and SWOT observations are conservatively binned rather than interpolated, and that Argo retains its original coordinates and timestamps. (R1-4)
- The same paragraph states that each product is ingested at its published, quality-controlled processing level and that OceanTACO performs no additional filtering or outlier removal. (R2)
- Section 2.3.1: added that a bounding-box query can overlap up to four tiles and is served transparently across all of them. (R1-7, R2)

Section 2.3.2, Regridding

- Equation 2 was replaced. The pooled standard error of the mean was removed and replaced by the inter-track dispersion σ_{track}, the observation-weighted standard deviation of the per-track means, defined only where two or more tracks overlap a cell. The surrounding prose explains why the previous quantity was not meaningful at roughly one observation per cell per pass. (R1-9)
- The auxiliary metadata list gained the number of contributing tracks (`n_tracks`) as a fifth provenance layer. (R1-9)
- The SWOT regridding paragraph was updated to refer to σ_{track} rather than the standard error of the mean. (R1-9)
- The L3 nadir and SWOT collections were recomputed and re-uploaded to Hugging Face to reflect the new variable. (R1-9)

Section 2.3.3, Compression

- Added a paragraph separating the deployed archive representation from the controlled codec comparison, and reporting the ClimateBenchPress-style benchmark against SZ3, SPERR, ZFP, EBCC, and bitround variants. (R1-10)
- Figure 4 gained a second panel mapping the relative error $|\text{orig} - \text{comp}|/\sigma_{\text{local}}$, and its caption now states that the absolute error is bounded by the `int16` quantisation step. (R1-11, R1-12)

Section 2.4, TACO dataset access

- The STAC paragraph was reworded to state that STAC addresses data discovery rather than the internal organisation of samples. (R1 general)

Section 3, Use cases

- Section 3.1: the Argo validation result is now quantified in prose, giving the number of near-surface matchups, the mean bias, and the RMSE, in place of the previous qualitative statement about a consistent negative bias. (ESSD requirement for quantitative validation)
- Section 3.2: the PSD example was replaced by a matched comparison in which the L4 DUACS field is sampled at the exact SWOT positions, so that both spectra are evaluated over identical sampling. The text and caption now frame the divergence as a resolution difference only over the roughly 30 to 150 km band and note that the L3 spectrum flattens near the effective-resolution limit of the product. The figure was regenerated. (R1-14)
- Section 3.2: the Southern Ocean and Antarctic Circumpolar Current were added to the eddy-rich regions named alongside the Gulf Stream and Kuroshio. (R1-13)
- Section 3.3: expanded to state that the eight regions are a storage-and-retrieval mechanism rather than analysis regions, that arbitrary subregions are retrievable at native resolution, and that per-mission identifiers and the Appendix coverage table are what make OSE configurations expressible. (R2)
- Section 3.5: expanded to describe individually what each cited reconstruction method does, rather than referring to them collectively. (R2)

Conclusions

- Restated the contribution as a harmonised, sample-invariant organisation rather than a new regridding or compression algorithm. (R1 general)
- Gave the reason for the 2015 start date, consistent with Section 2.2. (R1-1)

Appendix

- Table A1 (variable dictionary) gained *Temporal* and *Depth* columns for every source, and its introductory text explains both. (R1-2, R1-3, R1-6, R1-15)
- Table A1: the `*_sem` variables were renamed to `*_intertrack_std` and their definitions updated to the inter-track dispersion, crossover cells only. (R1-9)
- Table A3 (storage footprint): the caption and footnotes now state that the reported ratios include product restructuring, hourly-to-daily consolidation for L4 wind and track-file consolidation for L3 SWOT and L3 SSH, in addition to numerical encoding. Affected rows carry a footnote marker. The overall ratio was corrected to 20.4× and the L3 SSS rows to 2.0×. (R1-10)
- Table A4 (compression loss): gained an "RMSE / field std." column, the L3 along-track and SWOT rows were removed because those products are stored at full `float32`, and the caption was rewritten accordingly. (R1-10, R1-12)
- New Table A5 (codec benchmark): compares the deployed encoding against the best passing codec under matched strict bounds, for the five gridded products. (R1-10)
- Salinity units were relabelled from 10^{-3} to PSS-78 throughout the tables. (R1-16)
- Corrected the appendix float numbering. Repeated `\appendixfigures` and `\appendixtables` markers had advanced the appendix letter at each call, which produced two tables numbered A1 and two numbered B1. The appendix tables now run A1 to A5 and the appendix figure is A1. (numbering correction)