https://doi.org/10.5194/essd-2026-232
Preprint. Discussion started: 10 April 2026

Open Access Earth System Science Data Discussions

OceanTACO: A Multi-Sensor Global Ocean Sea Surface State Dataset

Nils Lehmann[1,2], Cesar Aybar[3], Ando Shah[4], Marcello Passaro[5], Jonathan L. Bamber[6,7], and Xiao Xiang Zhu[1,2]

[1]Data Science in Earth Observation, Technical University of Munich (TUM), 80333 Munich, Germany
[2]Munich Center for Machine Learning (MCML), 80333 Munich, Germany
[3]Image Processing Lab (IPL), University of Valencia, 46980 Valencia, Spain
[4]School of Information, University of California, Berkeley, Berkeley, CA 94720, USA
[5]Deutsches Geodätisches Forschungsinstitut (DGFI-TUM), Technical University of Munich (TUM), 80333 Munich, Germany
[6]School of Geographical Sciences, University of Bristol, Bristol BS8 1SS, UK
[7]Institute for Advanced Study, Technical University of Munich (TUM), 80333 Munich, Germany

Correspondence: Nils Lehmann (n.lehmann@tum.de), Jonathan L. Bamber (J.Bamber@bristol.ac.uk), and Xiao Xiang Zhu (xiaoxiangzhu@tum.de)

Abstract. We present OceanTACO, a harmonised global collection of sea surface state datasets designed to support reproducible Earth system research. The collection integrates satellite altimetry, sea surface temperature, salinity, surface winds, reanalysis fields, and Argo in situ observations within a unified cloud-optimised specification based on Transparent Access to Cloud-optimised datasets (TACO). It includes Level-3 observations, Level-4 gap-filled products, and reanalysis outputs while preserving native spatial and temporal resolution. The core dataset spans 29 March 2023 to 1 August 2025, covering the Surface Water and Ocean Topography (SWOT) mission, with an extended record from 1 January 2015 until 29 March 2023 for non-SWOT sources.

Datasets are harmonised through standardised metadata, spatial referencing, and temporal indexing, enabling consistent spatiotemporal queries across sensors and processing levels. A uniform internal structure reduces product-specific preprocessing and allows the same data-access routines to be applied across regions, sensors, and studies. This supports Earth systems analyses workflows such as validation against in situ observations, comparisons between observation and mapped products, observation system experiments, and multivariate sensor analyses.

Example applications demonstrate cross-product collocation with Argo, analysis of sea surface height variability during extreme events, and relationships between surface variables relevant for data-driven reconstruction. OceanTACO improves accessibility to coordinated multi-source analyses while preserving data provenance and native observation characteristics, and can be extended with new missions without restructuring the dataset. The core and extended dataset are available at https://doi.org/10.57967/hf/8171 (Lehmann and Aybar, 2026a) and https://doi.org/10.57967/hf/8172 (Lehmann and Aybar, 2026b) respectively.

https://doi.org/10.5194/essd-2026-232
Preprint. Discussion started: 10 April 2026

Open Access Earth System Science Data Discussions

1 Introduction

Ocean dynamics play a central role in regulating Earth's climate system, influencing air–sea fluxes, energy transport, and large-scale circulation patterns (Morrow et al., 2019). Mesoscale eddies (50–300 km) and smaller-scale processes contain a large fraction of the ocean's kinetic energy and strongly impact heat, carbon, and nutrient transport (Klein et al., 2019). Quantifying these dynamics requires accurate and temporally consistent observations of sea surface height (SSH) and related surface variables.

Satellite radar altimetry has provided continuous, high accuracy global SSH observations since 1992 (Le Traon et al., 2025). Because altimeters sample the ocean along sparse ground tracks, spatially complete maps are generated using interpolation and data assimilation systems such as DUACS (Taburet et al., 2019). While these products provide essential large-scale coverage, their effective resolution is limited by sampling constraints and smoothing inherent to mapping procedures (Ballarotta et al., 2019). Complementary datasets, including sea surface temperature (SST), sea surface salinity (SSS), surface winds, wide-swath altimetry from SWOT, and in situ Argo profiles, provide essential ancillary information about ocean surface variability but are distributed across independent archives and formats.

Recent advances in data-driven reconstruction and hybrid data assimilation approaches have demonstrated the scientific value of integrating multi-sensor observations. These approaches range from statistical interpolation to machine learning methods (Martin et al., 2023; Le Guillou et al., 2025; Archambault et al., 2023) and require harmonised, well-documented, and reproducible access to heterogeneous datasets. However, assembling such collections remains technically demanding and time-consuming. Researchers must retrieve data from multiple platforms, reconcile spatial grids and temporal coverage, handle heterogeneous metadata, and define consistent preprocessing pipelines across products distributed in independent archives and formats. As a result, multi-sensor studies often rely on substantial custom preprocessing, undocumented collocation decisions, and varying data handling routines (Johnson et al., 2024; Aouni et al., 2025). This methodological variability limits comparability across studies and complicates efforts to reproduce or extend published analyses. Recent assessments across the geosciences emphasise that data availability alone does not ensure reproducible research and that structured data organisation, transparent provenance, and interoperable access mechanisms are equally critical (Algarabel et al., 2023; Coca-Castro et al., 2025).

To address this gap, we introduce OceanTACO, a globally consistent collection of sea surface state datasets spanning satellite observations, reanalysis products, and in situ measurements (Fig. 1). OceanTACO harmonises Argo profiles, L3 observational data, L4 gap-filled products, and reanalysis outputs within a unified specification based on the TACO framework (Aybar et al., 2025). The dataset preserves native observation characteristics while enabling structured querying by region, time, sensor, and variable. By providing unified access to multi-sensor data, OceanTACO supports reproducible research across oceanography, data assimilation, statistical analysis, and emerging machine learning applications. Figure 1 illustrates a temporal snapshot of the various data included in OceanTACO.

Our proposed OceanTACO dataset makes the following contribution:

https://doi.org/10.5194/essd-2026-232
Preprint. Discussion started: 10 April 2026

1. A global, multi-source sea surface state dataset integrating satellite altimetry, wide-swath SWOT observations, sea surface temperature, sea surface salinity, surface winds, reanalysis products, and in situ Argo profiles across a common temporal and spatial framework.

2. Standardised preprocessing and harmonisation procedures, including regridding of L3 altimetry and SWOT observations, consistent regional tiling, and documented compression schemes that preserve numerical fidelity while reducing storage requirements.

3. Programmatic interfaces for reproducible computational analysis, supporting scientific workflows such as statistical analysis, multi-sensor comparison, data assimilation experiments, and machine learning studies.

4. Comprehensive documentation and examples, facilitating transparent reuse and extension of the dataset.

1.1 Existing dataset frameworks

Johnson et al. (2024) introduced the OceanBench framework, which provides standardised preprocessing and evaluation procedures for sea surface height (SSH) interpolation tasks. OceanBench aggregates curated datasets derived from high-resolution NATL60 simulations together with simulated and real nadir altimetry observations, and includes benchmarking pipelines for method comparison. The framework has contributed substantially to improving reproducibility in regional SSH reconstruction studies, particularly in the Gulf Stream region. Extending this effort, Aouni et al. (2025) provide a more comprehensive benchmark for evaluating global data-driven ocean forecasting across sea surface height, temperature, salinity, and currents. The framework introduces curated datasets from GLORYS12, GLO12, and ECMWF Integrated Forecasting System (IFS) into three standardised evaluation tracks to ensure physical consistency and reproducibility in deep learning-based ocean modelling (Aouni et al., 2025).

While these frameworks represent important steps toward standardised analyses-ready data collections and benchmarking, they are primarily designed for predefined challenge configurations or focus on specific geographic regions. In contrast, OceanTACO is conceived as a global, multi-source data collection that harmonises observational, reanalysis, and auxiliary surface variables within a unified and queryable specification. Rather than prescribing fixed evaluation tracks, OceanTACO provides flexible spatiotemporal and sensor-level subsetting capabilities, enabling researchers to configure region-specific studies, cross-mission validation experiments, and forecasting setups within a consistent data framework. To our knowledge, no existing framework harmonises L3 observations including SWOT, L4 gridded products, reanalysis fields, and in situ profiles from diverse ocean missions within a single, uniformly organised and queryable specification for the SWOT era.

1.2 Multi-source coordination

When multi-sensor ocean analyses are combined from independently distributed archives, each dataset introduces its own unique processing requirements. These choices must be resolved at the preprocessing stage and, when undocumented, introduce sources of variation between studies. Because such decisions are rarely reported in full in published methods sections,

https://doi.org/10.5194/essd-2026-232
Preprint. Discussion started: 10 April 2026

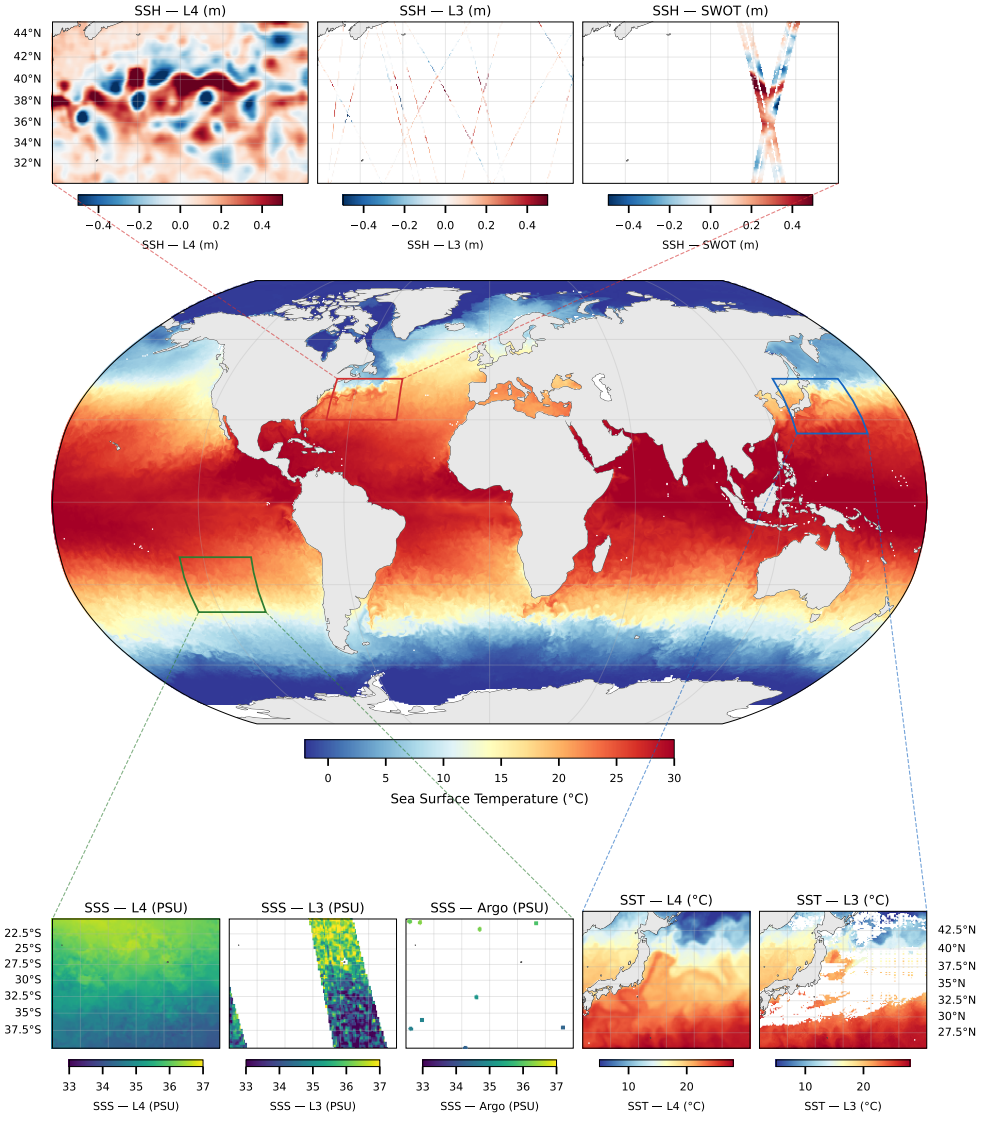

Figure 1. Overview of the OceanTACO dataset illustrated for a single snapshot. *Centre:* Global sea surface temperature from the GLORYS reanalysis (Robinson projection); coloured boxes indicate the three focus regions. *Top:* Gulf Stream region (70–40°W, 30–45°N) showing the gridded L4 DUACS SSH anomaly, L3 conventional along-track altimetry, and L3 SWOT wide-swath observations on a shared colour scale (m), demonstrating the progressive increase in spatial resolution. *Bottom left:* South Pacific region (130–100°W, 40–20°S) comparing L4 and L3 sea surface salinity (10^{-3}) with co-located Argo float profiles. *Bottom right:* Kuroshio Current region (130–160°E, 25–45°N) showing L4 and L3 sea surface temperature (°C). All panels are retrieved via a single spatiotemporal query to the OceanTACO catalogue, illustrating the dataset's ability to collocate heterogeneous observation types within a common regional tiling and storage framework.

https://doi.org/10.5194/essd-2026-232
Preprint. Discussion started: 10 April 2026

nominally comparable analyses may differ in ways that are not apparent from the description alone (Johnson et al., 2024; Aouni et al., 2025). The resulting methodological variability limits the comparability of results across studies and complicates efforts to reproduce or extend existing analyses.

Format-level standards such as Zarr and NetCDF4/HDF5 eliminate heterogeneity in binary encoding but do not prescribe how variables, coordinates, or metadata are organised within a file or across a collection. Catalogue standards such as STAC address the related but distinct problem of discovery: they specify where data can be found, not how different products are internally organised or how they relate to one another across sensors and processing levels. As a result, multi-source preprocessing workflows remain data product-specific even when all constituent files conform to a common format or are discoverable through a shared catalogue.

The TACO specification addresses this by enforcing a uniform internal organisation at the sample level: every entry in the catalogue replicates an identical subfolder and file-type layout, regardless of which sensors it contains or at which processing level they were acquired (Aybar et al., 2025). Because the internal structure is invariant across products, data access procedures developed for one sample apply without modification to any other sample in the collection.

The naming of the specification reflects two complementary design principles. *Transparent* refers to data access being explicit and auditable: data are referenced through explicit file paths, so every retrieval operation can be traced, replicated, or independently verified. *Cloud-optimised* refers to the storage layout: NetCDF4/HDF5 files with internal chunking support partial, byte-range reads from remote object storage, so that the same access procedure operates identically on locally held data or on data hosted in cloud object stores, without requiring full-file retrieval before analysis can begin.

Together, these properties mean that analysis workflows developed for one region or sensor transfer to other regions and sensors without modification to the access procedure, improving reproducibility or extensions by other research groups.

2 Dataset description

2.1 Data sources

OceanTACO integrates ten data sources spanning satellite altimetry, sea surface temperature, sea surface salinity, surface winds, ocean reanalysis, and in situ Argo profiles (Table 1).

https://doi.org/10.5194/essd-2026-232
Preprint. Discussion started: 10 April 2026

Open Access Earth System Science Data Discussions

Data Source	Level	Resolution (deg/km)†	Key Variables	Availability
GLORYS-12	Reanalysis	$1/12^\circ$ (≈ 9.3 km)	SSH, Temperature, Salinity, U current, V current	Core + Extended
DUACS SSH Product	L4	0.125° (≈ 13.9 km)	SLA, SLA error, U geostrophic anomaly, U geostrophic error, V geostrophic anomaly	Core + Extended
OSTIA SST (Met Office)	L4	0.05° (≈ 5.6 km)	SST, SST analysis error, Sea-ice fraction, Surface mask	Core + Extended
Multi Observation SSS	L4	0.125° (≈ 13.9 km)	SSS, Surface density, SSS error, Surface density error, Sea-ice fraction	Core + Extended
Global Ocean Daily Wind	L4	0.25° (≈ 27.8 km)	U10 wind, V10 wind, U10 wind SD, V10 wind SD, U10 wind min	Core + Extended
Altimetry Along-Track	L3	$\sim 0.06^\circ \times 0.09^\circ$	SLA (filtered), SLA uncertainty, MDT, MDT uncertainty, ADT	Core + Extended
SWOT	L3	$\sim 0.02^\circ \times 0.03^\circ$	SSHA (filtered), SSHA uncertainty, SSHA (unfiltered), SSHA uncertainty (unfiltered), MDT	Core only
L3 SST	L3	0.10° (≈ 11.1 km)	SST time offset, SST, SST (bias corrected), SST uncertainty bias, SST uncertainty SD	Core + Extended
L3 Salinity (SMOS)	L3	$\sim 0.23^\circ \times 0.26^\circ$	SSS, SSS (rain corrected), SSS uncertainty, SSS uncertainty (rain corrected), Swath cross-track distance	Core + Extended
Argo	In situ	–	Temperature, SSS, Pressure, Float cycle number, Profile direction	Core + Extended

Table 1. Overview of the primary data sources used in OceanTACO. The core dataset spans 29 March 2023 to 1 August 2025, while the extended dataset covers 1 January 2015 to 29 March 2023 preceding the SWOT mission. A full description of all variables can be found in Table A1. Key variables here are shown with simplified aliases for readability. †Horizontal kilometer resolutions are approximate values estimated at the equator.

https://doi.org/10.5194/essd-2026-232
Preprint. Discussion started: 10 April 2026

2.1.1 Reanalysis data

The Global Ocean Reanalysis and Simulation (GLORYS12) data product provides a global ocean reanalysis at $1/12°$ horizontal resolution and is available from 1993 to present, covering the full satellite altimetry era (Jean-Michel et al., 2021). GLORYS12 assimilates satellite and in situ observations into the NEMO ocean ice model using a hybrid Kalman filter combined with a three-dimensional variational (3D-Var) bias correction scheme (Jean-Michel et al., 2021).

At its native resolution, GLORYS12 resolves large-scale and mesoscale circulation features but does not fully capture submesoscale variability. Reanalysis products may exhibit regional discrepancies relative to direct satellite observations, particularly in dynamically active mesoscale regions (Martin et al., 2025). Nevertheless, they provide physically consistent, dynamically balanced fields that are valuable for multi-sensor analyses and data-driven studies (Martin et al., 2025).

Although GLORYS12 provides three-dimensional fields with depth, OceanTACO includes a subset of surface variables to maintain consistency with the other sea surface datasets in the collection. Specifically, we include sea surface height, sea surface temperature, sea surface salinity, and surface currents at 15 m depth, following Martin et al. (2025). This selection facilitates joint analyses of surface state variability while preserving the physical coherence of the reanalysis product.

2.1.2 Level 4 data

L4 products provide spatially and temporally complete surface fields derived through data assimilation and optimal interpolation of satellite and in situ observations. In contrast to L3 observations, which retain native sampling characteristics, L4 products represent gap-filled, gridded estimates of ocean surface variables. OceanTACO includes L4 datasets for sea surface height (SSH), sea surface temperature (SST), sea surface salinity (SSS), and surface winds. All L4 products were obtained from the Copernicus Marine Service using the Copernicus Marine Toolbox (Mercator Ocean / Copernicus Marine Service, 2025).

2.1.3 Sea Surface Height (SSH)

We include the DUACS L4 gridded sea level product (Taburet et al., 2019; Copernicus Marine Service, 2025c), provided at $0.125° \times 0.125°$ resolution. DUACS merges multi-mission L3 along-track altimetry observations using optimal interpolation. Contributing missions include Sentinel-3A/B, Sentinel-6A, Jason-3, SARAL/AltiKa, CryoSat-2, OSTM/Jason-2, Jason-1, TOPEX/Poseidon, Envisat, GFO, ERS-1/2, and Haiyang-2A/B. Within OceanTACO, we include the gridded sea level anomaly (SLA) and associated standard variables provided in the product.

2.1.4 Sea Surface Temperature (SST)

Sea surface temperature is dynamically linked to sea surface height in many regions, particularly in western boundary currents such as the Gulf Stream (Martin et al., 2023; Le Guillou et al., 2025). To provide a spatially complete SST field, we include the OSTIA L4 SST product (Met Office, 2025). This dataset provides global daily mean SST on a $0.05° \times 0.05°$ grid. It merges

https://doi.org/10.5194/essd-2026-232
Preprint. Discussion started: 10 April 2026

observations from multiple infrared and microwave satellite sensors (AVHRR, ATSR, SLSTR, AMSR-E, AMSR2; Embury et al. 2024) using the OSTIA optimal interpolation system.

2.1.5 Sea Surface Salinity (SSS)

Sea surface salinity contributes to regional sea level variability through halosteric effects (Llovel and Lee, 2015). We include the Copernicus Marine multi-observation L4 salinity product (Copernicus Marine Service, 2025e), provided at $0.125° \times 0.125°$ resolution. This dataset combines measurements from NASA's Soil Moisture Active Passive (SMAP) mission and ESA's Soil Moisture and Ocean Salinity (SMOS) mission, together with satellite SST and in situ salinity observations. The fields are generated using a multivariate optimal interpolation framework (Droghei et al., 2016; Buongiorno Nardelli et al., 2016; Droghei et al., 2018) to produce gap-free global maps.

2.1.6 Surface Winds

Surface wind stress influences short-term variability in upper ocean dynamics (Li et al., 2022). We incorporate the Global Ocean Hourly Sea Surface Wind L4 product (Copernicus Marine Service, 2025g), provided at $0.25° \times 0.25°$ resolution. The dataset is based on the ECMWF ERA5 reanalysis and is bias-corrected using scatterometer observations from multiple satellite platforms. For consistency with the daily temporal resolution adopted in OceanTACO, hourly wind fields are aggregated to daily means, while also computing daily minimum, maximum, and standard deviation values.

2.1.7 Level 3 data

In contrast to reanalysis and L4 products, L3 datasets retain the native sampling characteristics of satellite and in situ observations without gap filling or large-scale interpolation. As such, they provide observation-based constraints that are essential for evaluating gridded products, assessing sampling effects, and supporting data assimilation studies.

2.1.8 Along-track altimetry

We include multi-mission L3 sea surface height (SSH) along-track altimetry data from the Copernicus Marine Service (Copernicus Marine Service, 2025d, a). These datasets provide geophysical corrections and cross-calibrated measurements at an effective spatial sampling of approximately 7 km along track.

Where available, we prioritise the reprocessed (multi-year) product tailored for data assimilation applications (Copernicus Marine Service, 2025a), which ensures improved cross-mission consistency relative to near-real-time products. For periods not covered by the reprocessed archive, near-real-time products (Copernicus Marine Service, 2025d) are incorporated to maintain temporal continuity. The temporal coverage of each contributing mission across both products is listed in Table B1 in the Appendix.

Mission identifiers are preserved within OceanTACO, enabling filtering by individual satellite platforms. This facilitates cross-mission consistency studies and controlled validation experiments based on sensor-level subsetting.

https://doi.org/10.5194/essd-2026-232
Preprint. Discussion started: 10 April 2026

2.1.9 Wide-swath altimetry (SWOT)

The Surface Water and Ocean Topography (SWOT) mission (Fu et al., 2024) represents a major advancement in satellite altimetry by providing two-dimensional wide-swath sea surface height observations at an approximate resolution of 2 km (Morrow et al., 2019). Using the Ka-band Radar Interferometer (KaRIn), SWOT measures sea surface height across a ~ 120 km swath by estimating phase differences between two spatially separated antennas. This configuration resolves fine-scale ocean variability beyond the reach of conventional nadir altimeters. We include the L3 SWOT Low Rate SSH product distributed by AVISO (AVISO/DUACS, 2024).

2.1.10 L3 sea surface temperature and salinity

To complement the gridded L4 surface products, OceanTACO incorporates observation-based L3 SST and SSS datasets. The L3 SST product (Copernicus Marine Service, 2025f) provides intercalibrated measurements from multiple polar orbiting and geostationary satellites. The L3 SSS dataset (Copernicus Marine Service, 2025b; Boutin et al., 2018) is derived primarily from SMOS observations and includes quality-controlled retrievals prior to spatial interpolation.

Together, these L3 datasets preserve the native sampling structure and measurement characteristics of the underlying observing systems. Including them lets researchers distinguish observation-driven variability from mapping-induced smoothing.

2.1.11 In situ data

To complement satellite-based surface observations, OceanTACO includes in situ measurements from the Argo programme. Argo is a global array of autonomous profiling floats that measure temperature, salinity, and pressure throughout the upper 2000 m of the ocean (Wong et al., 2020). Since its establishment in the early 2000s, the program has collected more than two million high-quality profiles, substantially improving observational coverage of the subsurface ocean, which cannot be directly observed from space (Wong et al., 2020). Argo data are accessed and processed using the `argopy` Python library version 1.4 (Maze and Balem, 2020). We use the research-quality data access mode, which automatically applies quality control filters and excludes profiles that do not meet delayed-mode or real-time quality standards. Within OceanTACO, temperature, salinity, and pressure profiles are retained in their original vertical resolution and linked to the corresponding temporal and spatial indices of the dataset. The inclusion of Argo observations provides an independent in situ reference for evaluating surface products and enables analyses that connect surface variability with subsurface structure.

2.2 Temporal and spatial coverage

OceanTACO spans the common temporal intersection of all included data sources. The core dataset spans 29 March 2023 to 1 August 2025, covering the currently available SWOT period, including the calibration phase. Spatially, OceanTACO provides global coverage between $90°$S and $90°$N in a WGS 84 (EPSG:4326) projection, with variable-specific native spatial resolutions ranging from approximately 2 km (SWOT) to $0.25°$ (surface winds). While the WGS 84 projection is not area-preserving, and will lead to significant distortions towards the poles, it is the native projection of both GLORYS and L4 data and was therefore

https://doi.org/10.5194/essd-2026-232
Preprint. Discussion started: 10 April 2026

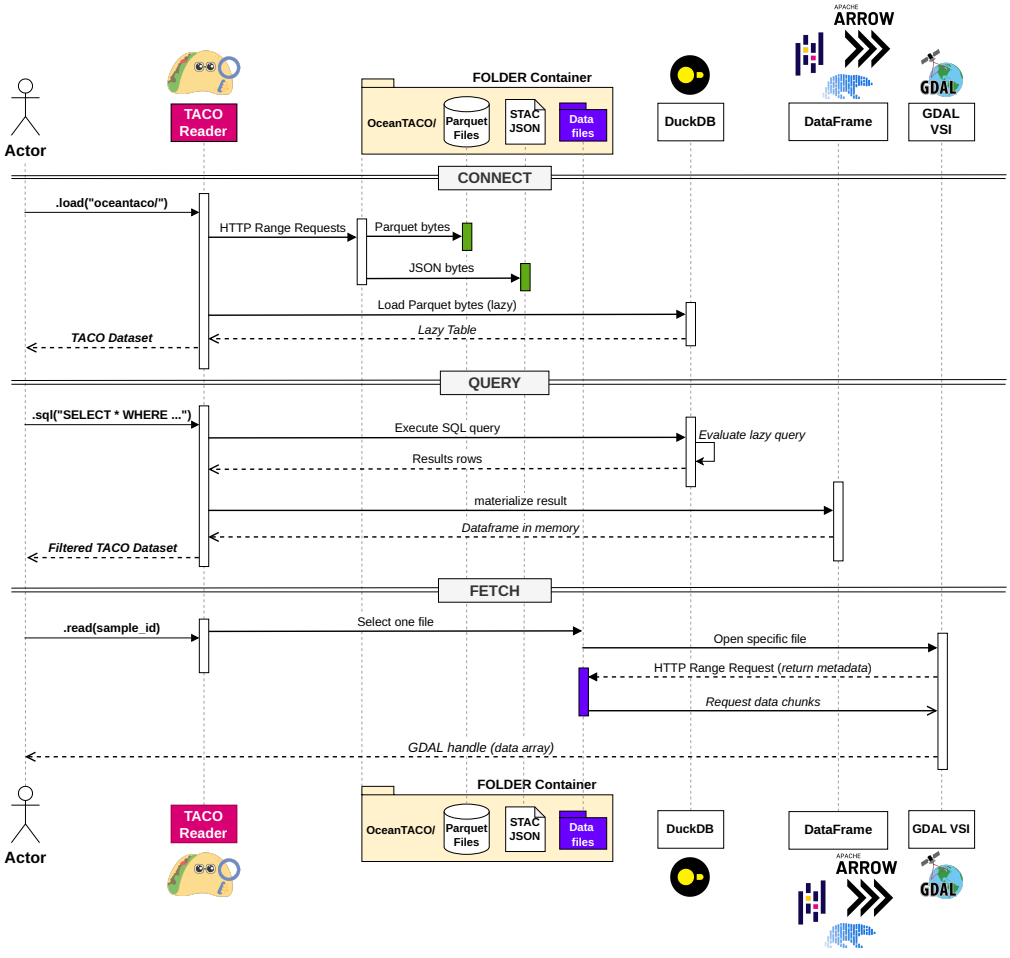

Figure 2. TACO architecture showing the three-phase data access workflow for the OceanTACO folder container. The CONNECT phase retrieves the Parquet metadata files and STAC-compliant JSON collection descriptor from the dataset root, then lazily loads the Parquet tables into DuckDB for evaluation. The QUERY phase filters these tables by geographic extent, observation period, or data source via SQL, materializing results as in-memory DataFrames. The FETCH phase retrieves individual data files on demand through GDAL VSI, which issues HTTP range requests to read only the requested portions of remote files without downloading the complete dataset.

https://doi.org/10.5194/essd-2026-232
Preprint. Discussion started: 10 April 2026

used as the basis for regridding of altimetry track and SWOT data. Additionally, we provide an extended dataset version spanning 1 January 2015 to 29 March 2023 that precedes the SWOT period but allows longer time-series analysis of the remaining data sources; it includes all modalities except SWOT, which was not operational during this period. The extended and core datasets share the boundary date of 29 March 2023, allowing seamless concatenation.

The complete core dataset comprises 856 daily temporal indices, subdivided into eight regional tiles per day. The extended dataset comprises 3009 daily temporal indices. The total storage volume is approximately 324 GB in compressed form, accounting for all stored variables across modalities; Table A1 reports the storage for the primary variable of each modality as a representative indicator of compression performance. OceanTACO is publicly accessible through *Hugging Face* for direct cloud access.

2.3 Processing methods

Because OceanTACO integrates heterogeneous data products with differing spatial resolutions, sampling geometries, and file formats, we implemented a standardised processing workflow to harmonise the collection while preserving native observation characteristics.

The processing pipeline consists of three primary steps:

1. **Regional tiling:** Global daily data files are partitioned into eight equally sized geographic regions in their native WGS 84 projection to facilitate efficient storage and selective retrieval as depicted in Fig. 3.

2. **Observation binning:** L3 along-track altimetry and wide-swath SWOT observations are projected onto regular grids at their native spatial resolutions using binning procedures.

3. **Numerical compression:** Floating-point variables are encoded using scaled `int16` representations combined with lossless zlib compression to reduce storage volume while preserving information.

2.3.1 Regional processing

We process the global data into eight separate regions per day, as shown in Fig. 3. Through the streaming and query possibilities, users can download or interact with a specific region if requested. This reduces memory requirements for regional studies, and also improves data access speeds for smaller patches, since only a regional file has to be opened to extract a geospatial patch.

2.3.2 Regridding

2.3.3 Along-track altimetry L3

L3 along-track sea surface height (SSH) observations from multiple satellite missions are mapped onto a regular 7 km grid using a conservative binning approach. Grid cell boundaries are placed at the midpoints between adjacent cell centres, and each observation is assigned to the cell whose boundaries contain its geographic coordinates.

https://doi.org/10.5194/essd-2026-232
Preprint. Discussion started: 10 April 2026

Open Access Earth System Science Data Discussions

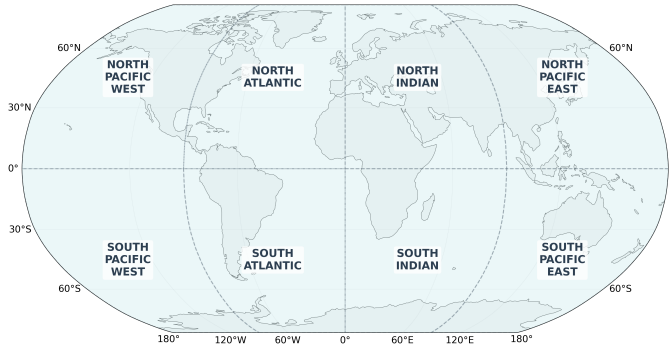

Figure 3. The eight spatial regions provided in OceanTACO for region-specific data access. Each region is stored as an independent subset of all data sources, enabling efficient spatially targeted queries without loading the global dataset.

For each contributing satellite track k, the arithmetic mean $\bar{z}_k$ and within-track sum of squared deviations $\mathrm{SS}_k = \sum_{i=1}^{n_k}(z_i - \bar{z}_k)^2$ are computed per cell, where z_i denotes the SLA observation (in metres) from satellite i within track k, and n_k is the number of observations in that cell from track k. Across all tracks that contribute to a cell, the observation-weighted mean is

$$\bar{z} = \frac{\sum_k n_k \bar{z}_k}{\sum_k n_k}, \tag{1}$$

and the pooled standard error of the mean (SEM), quantifying within-track measurement dispersion, is

$$\mathrm{SEM} = \frac{\sqrt{\sum_k \mathrm{SS}_k}}{\sqrt{n\,(n-K)}}, \tag{2}$$

where $n = \sum_k n_k$ is the total number of observations and K is the number of contributing tracks. The SEM is defined only for cells containing at least two observations in total and reflects within-track variability. The stored SEM therefore represents a lower bound on total per-cell uncertainty; between-track variance, which reflects temporal and spatial sampling offsets between missions, is not included. Cells containing no observations are masked as missing. No spatial smoothing, interpolation, or gap filling is applied beyond this cell-based aggregation.

To further preserve sampling information and data provenance, auxiliary metadata layers are retained for each grid cell: (i) the total number of contributing observations n, (ii) a primary track identifier, (iii) an overlap mask indicating multi-track intersections, and (iv) the observation-mean geographic position within each cell. This structure enables per-mission separation and assessment of sampling density.

2.3.4 Wide-swath altimetry (SWOT) L3

The L3 SWOT product is regridded onto a 2 km regional grid that maintains the native resolution of the KaRIn instrument, using the same conservative binning procedure described above for along-track altimetry. No spatial smoothing is applied; cells without observations remain masked. The swath geometry is projected with identical cell-boundary conventions, and the observation-weighted mean, pooled standard error of the mean, and auxiliary metadata layers (primary pass identifiers,

https://doi.org/10.5194/essd-2026-232
Preprint. Discussion started: 10 April 2026

overlap indicators, observation counts, and mean observation positions) are retained to support analyses of crossover regions and temporal sampling offsets. Redistribution of native SWOT L3 files in their original form is restricted under the AVISO+ data licence (Issue 19, February 2024); however, that licence explicitly permits the creation and redistribution of Derivative Works for any purpose. Confirmed by the AVISO team, OceanTACO's regridding constitutes such a Derivative Work: the irreversible coordinate transformation and spatial binning remove the original swath/pixel structure, so the native AVISO+ files cannot be reconstructed from this dataset. Full licensing details, including the specific AVISO+ clauses relied upon, are provided in the dataset card accompanying the Hugging Face repository.

2.3.5 Compression

OceanTACO applies scaled int16 encoding combined with zlib compression to reduce storage requirements while preserving numerical fidelity. To quantify the impact of this transformation, we evaluate reconstruction errors relative to the original float32 source fields prior to encoding. For each variable and data source, we decode the stored representation and compute the absolute difference with respect to the original floating point data. We report the 99th-percentile error (P99), root-mean-square error (RMSE), and mean bias over the full spatial domain and representative time period, which can be found in Table B1 in the Appendix. Fig. 4 demonstrates that the encoding introduces negligible reconstruction error relative to the original `float32` representation. The overall compression factor results from two components: (i) the deterministic packing ratio from `float32` to `int16`, and (ii) the zlib deflation ratio, which depends on spatial sparsity and field smoothness. Fig. C1 demonstrates that an example compression analysis for a fully gridded L4 product.

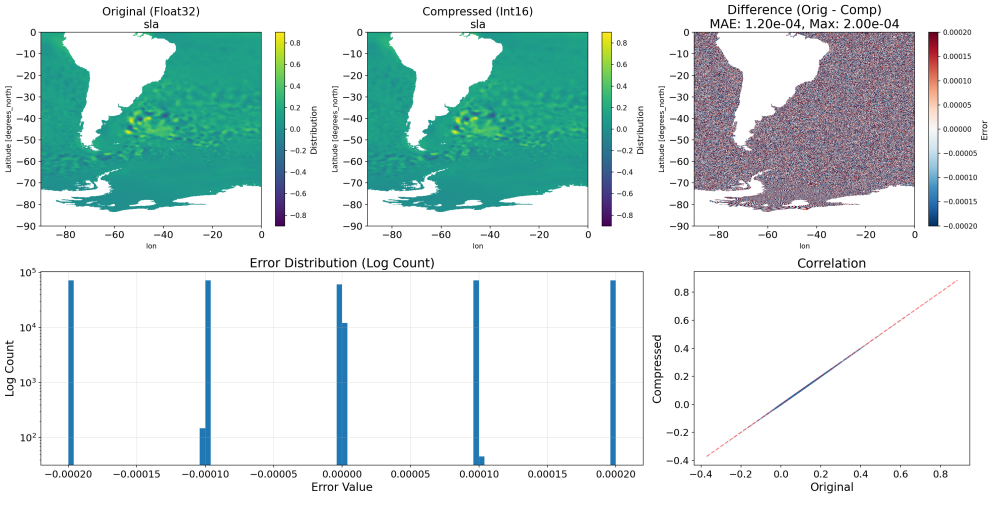

Figure 4. Encoding quality of the `int16+zlib` compression scheme applied to the L4 SSH sea level anomaly (SLA) over the South Atlantic on 2 April 2023, representative of the compression quality across all gridded products. *Top row, left to right:* original `float32` SLA field (m), `int16`-decoded field (m), and their absolute difference (m). *Bottom row:* error distribution across all valid ocean pixels (log-scaled counts; left) and scatter plot of compressed versus original values with the identity line (right).

https://doi.org/10.5194/essd-2026-232
Preprint. Discussion started: 10 April 2026

2.4 TACO dataset access and interaction

Building reproducible workflows from independently archived ocean datasets demands substantial custom preprocessing. The SpatioTemporal Asset Catalog (STAC) standard is widely used for geospatial dataset discovery and cataloguing; however, it addresses only data discovery rather than data organisation. As a result, each dataset requires custom loading code, forcing practitioners to inspect each dataset prior to developing data access routines. This is especially challenging when combining heterogeneous data types, including satellite altimetry, sea surface temperature, salinity, winds, and in situ profiles, each of which originates from distinct archives and formats.

OceanTACO implements the TACO (Transparent Access to Cloud-optimised) specification (Aybar et al., 2025) to resolve this. TACO is fully compliant with STAC, and any TACO dataset can be mapped to a valid STAC catalog without information loss. Additionally, TACO directly follows the FAIR principles proposed by Wilkinson et al. (2016). Beyond cataloging, TACO enforces consistent internal organisation across all samples within a dataset. For example, if one sample contains folders `GLORYS/`, `SWOT/`, and `ARGO/` with specific file types, every other sample replicates this exact structure. TACO references data through GDAL Virtual File System (VSI) paths rather than HTTP URLs. VSI provides a unified file access abstraction, so the same reading code operates transparently on local filesystems, cloud object stores such as S3 or GCS, and compressed archives, with only the path prefix varying between backends. Consequently, a data loading pipeline developed for a local copy of OceanTACO functions without modification when the dataset is hosted remotely, and partial reads of large files are managed natively by the GDAL driver, eliminating the need for full downloads. TACO supports two storage modes: ZIP mode, which packages the dataset as a single archive for static distribution, and folder mode, which stores data as a directory tree on disk or cloud object storage. OceanTACO utilises folder mode, which allows the dataset to be extended with new satellite passes, reprocessed products, or additional variables by simply writing files into the directory tree and appending rows to the metadata catalog, without modifying existing components.

Access is structured in three phases (Fig. 2). The CONNECT phase loads Parquet metadata into DuckDB for lazy evaluation and presents the catalog as relational tables. The QUERY phase filters these tables by geographic extent, observation period, or data source, and materializes results as DataFrames. The FETCH phase retrieves individual files on demand through GDAL VSI handlers, which read from local paths or cloud object storage. Both the core (https://doi.org/10.57967/hf/8171 (Lehmann and Aybar, 2026a)) and extended (https://doi.org/10.57967/hf/8172 (Lehmann and Aybar, 2026b)) dataset are available on Hugging Face.

3 Multi-source earth system use cases with OceanTACO

OceanTACO is designed to support reproducible analysis workflows that combine heterogeneous ocean observations, gridded products, and in situ measurements within a consistent spatiotemporal framework. The following subsections illustrate several representative workflow categories enabled by the dataset. The examples are organised into four broad categories that frequently arise in Earth system research: (i) validation using independent observations, (ii) comparisons across data processing levels, (iii) observation system experiments evaluating the contribution of individual sensors, and (iv) data-driven reconstruction work-

https://doi.org/10.5194/essd-2026-232
Preprint. Discussion started: 10 April 2026

flows that integrate multiple surface variables. In addition, we include a short case study demonstrating how OceanTACO can facilitate rapid analysis of extreme ocean events. These examples illustrate reproducible usage patterns rather than provide a comprehensive scientific evaluation of the underlying datasets. Example workflows and tutorials can be accessed and recreated in Jupyter notebooks (Perez and Granger, 2015) on the dataset documentation page.

3.1 Data validation workflows

Systematic validation of gridded products against independent observations is essential for quantifying structural biases, effective resolution, and regional uncertainty characteristics. Such analyses require reproducible collocation procedures, consistent temporal matching, and transparent spatial aggregation strategies.

Because all OceanTACO components share a common temporal index and regional tiling scheme, data retrieval operations are reproducible across studies. L3 observations, L4 products, reanalysis fields, and Argo profiles can be queried under identical spatial and temporal constraints, reduces ambiguity in matching procedures and reducing methodological variability between studies.

3.1.1 Example workflow: in situ validation of gridded SST products

As a representative application, L4 SST products and reanalysis fields are collocated with Argo profiles over a one-year period. For each Argo profile, we select the shallowest valid measurement within the upper 5 m to approximate near-surface conditions. Satellite and gridded fields are retrieved using same-day temporal matching and nearest-neighbor spatial selection at their native grid resolution. For each matchup, the bias is defined as

$$b_i = M_i - O_i,$$

where M_i and O_i denote the gridded product and Argo value, respectively. Spatially aggregated bias fields are computed on a $2^\circ \times 2^\circ$ grid, masking cells with fewer than three observations. Seasonal and latitudinal variability is examined using monthly latitude-band averages:

$$\bar{b}(\phi, t) = \frac{1}{n_{\phi,t}} \sum_{i \in (\phi,t)} b_i,$$

where ϕ denotes latitude, t time, and $n_{\phi,t}$ is the number of matches in latitude band ϕ at time t. Latitude-dependent mean bias and temporal variability are calculated as

$$\mu_b(\phi) = \frac{1}{T} \sum_{t=1}^{T} \bar{b}(\phi, t), \quad \sigma(\phi) = \sqrt{\frac{1}{T-1} \sum_{t=1}^{T} \left(\bar{b}(\phi, t) - \mu_b(\phi)\right)^2},$$

where T is the total number of monthly time steps.

Fig. 5 shows the spatial bias of L4 SST fields against matched Argo observations, with latitude-band mean bias and temporal variability. The general trend shows a consistent negative bias of the gridded L4 product and also a large variance in the Gulf Stream region.

https://doi.org/10.5194/essd-2026-232
Preprint. Discussion started: 10 April 2026

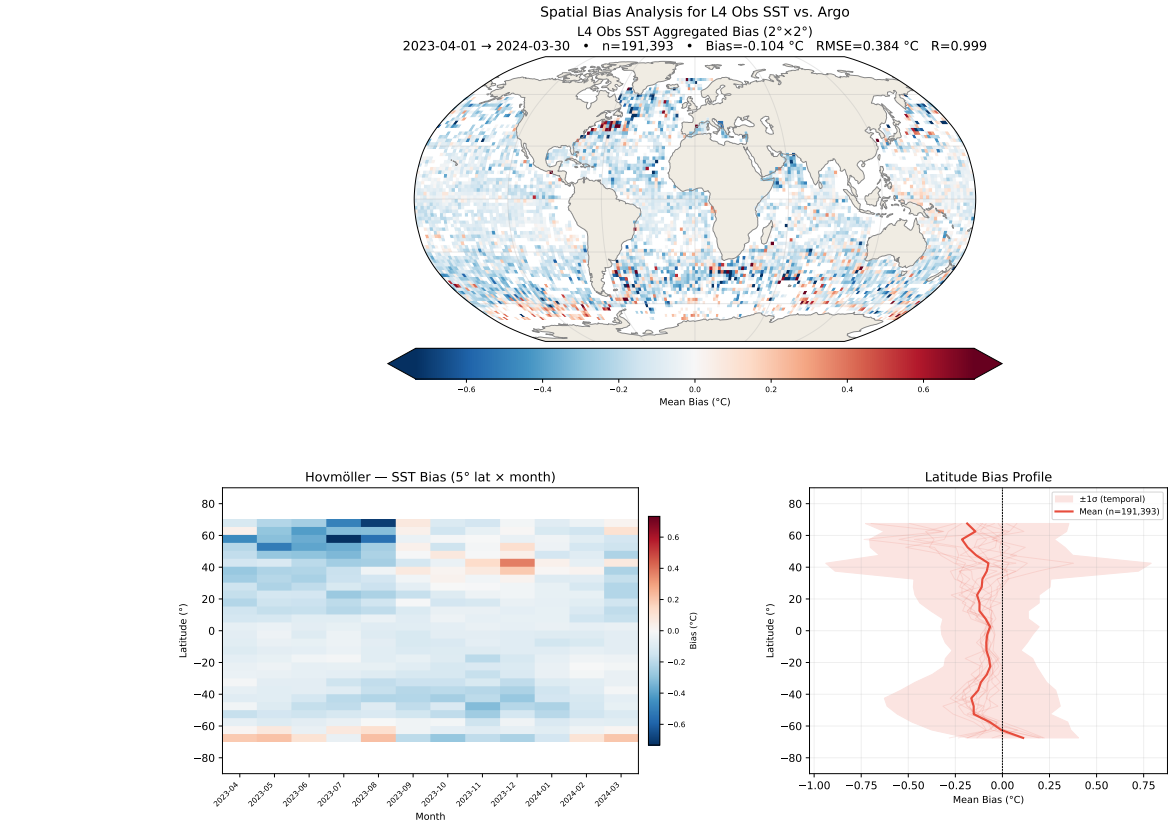

Figure 5. Spatial bias of the Met Office OSTIA Level-4 SST product against co-located Argo in situ observations, computed as the mean difference (OSTIA minus Argo) on a $2^\circ \times 2^\circ$ grid over the period 1 April 2023 to 30 March 2024. Units are $^\circ$C; the colour scale uses a diverging palette normalised to the 95th percentile of absolute bias to highlight spatial structure while limiting the influence of outliers.

3.2 Analyses across processing levels

OceanTACO preserves the native sampling geometry and spatial resolution of each product, enabling controlled comparisons across processing levels without introducing additional harmonisation artifacts. Such comparisons are particularly useful when evaluating differences between L3 observations, which retain native measurement characteristics, and L4 mapped or assimilated products, which introduce spatial smoothing and model-based constraints.

As a representative workflow, OceanTACO enables computation of wavenumber power spectral density (PSD) diagnostics across different observation types. For example, users can retrieve L3 nadir along-track SSH and L3 SWOT wide-swath SSH for identical spatial and temporal domains and compute along-track spectra in dynamically active regions such as the Gulf Stream (80°–40° W, 25°–45° N) or the Kuroshio Current (130°–160° E, 25°–45° N). Fig. 6 illustrates an example comparison produced using OceanTACO from 1 March 2025 to 29 May 2025. The comparison is intended as an illustrative diagnostic rather than a comprehensive spectral analysis.

https://doi.org/10.5194/essd-2026-232
Preprint. Discussion started: 10 April 2026

Open Access Earth System Science Data Discussions

Previous studies have shown that conventional nadir altimetry spectra often exhibit artificially shallow slopes at short wavelengths because the instrument noise floor contaminates the mesoscale band unless explicit noise correction is applied (Vergara et al., 2019). In contrast, SWOT wide-swath observations resolve substantially smaller spatial scales and provide improved signal-to-noise characteristics for mesoscale and submesoscale variability. Independent analyses indicate that SWOT observations can resolve ocean variability at spatial scales on the order of tens of kilometers, representing a substantial improvement over conventional nadir altimetry (Krishna and Sreejith, 2025; Wang et al., 2025).

This example illustrates how OceanTACO enables reproducible cross-product comparisons under identical spatiotemporal constraints. Such workflows are increasingly relevant in the SWOT era for evaluating scale-dependent variance and interpreting differences between observation-level measurements and mapped products.

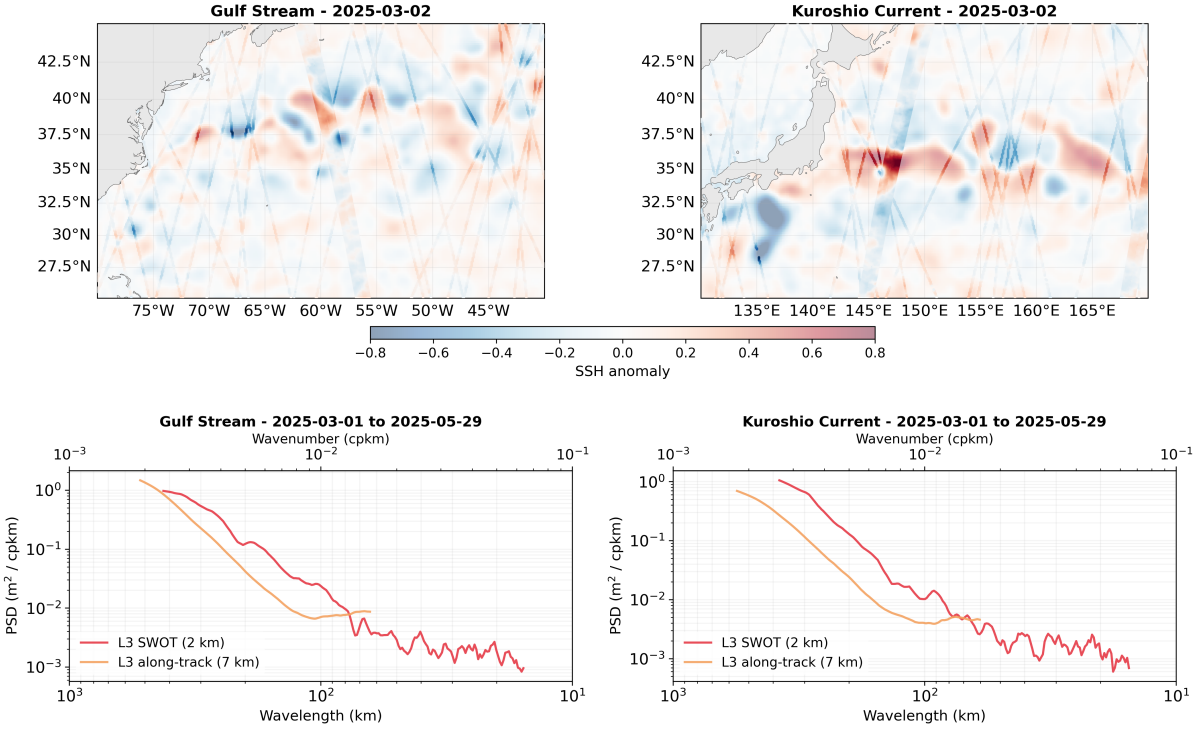

Figure 6. Spatial snapshots and wavenumber power spectral density (PSD) of sea surface height products for the Gulf Stream (left column) and Kuroshio Current (right column) regions. *Top row:* L4 DUACS SSH anomaly field overlaid with co-located L3 conventional along-track and L3 SWOT wide-swath observations for a representative snapshot. *Bottom row:* Along-track PSD curves for L3 conventional altimetry and L3 SWOT, computed from daily passes accumulated over the period 1 March 2025 to 29 May 2025, illustrating the difference in effective spatial resolution between the two products.

Beyond comparisons across processing levels, many studies seek to quantify the contribution of individual observing systems to reconstructed ocean fields. OceanTACO facilitates these controlled configurations through explicit sensor-level indexing and mission identifiers.

https://doi.org/10.5194/essd-2026-232
Preprint. Discussion started: 10 April 2026

3.3 Observation system experiments and mission impact studies

The introduction of new observing systems, such as the Surface Water and Ocean Topography (SWOT) mission, enables new assessments of their incremental contribution relative to established nadir altimetry and assimilated products. Early SWOT analyses emphasise its capability to resolve fine-scale ocean variability and improve characterization of mesoscale and submesoscale dynamics relative to nadir altimetry, while also highlighting the need for careful cross-comparison with existing mapping systems and climate data records (Ballarotta et al., 2025; Fouchet et al., 2025; Archer et al., 2025).

Observation system experiments (OSEs) and mission impact studies are therefore essential for studying how new sensors improve resolved spatial scales, dynamical consistency and downstream reconstruction performance. These studies typically require selective inclusion or exclusion of specific sensors, consistent regional subsetting, and reproducible definition of validation datasets across multiple processing levels.

OceanTACO preserves mission identifiers and full product provenance within a unified indexing framework. Researchers can configure controlled experiments that isolate nadir-only configurations, incorporate wide-swath SWOT observations, or evaluate L4 fields with and without specific observational inputs. Because these configurations are defined through explicit spatiotemporal queries rather than bespoke preprocessing pipelines, mission impact assessments can be reproduced consistently across regions and time periods.

3.4 Extreme events: case study of Hurricane Milton

In addition to structured workflow categories such as validation or observation system experiments, OceanTACO can facilitate rapid exploratory analyses of individual oceanographic events. The common temporal index means multi-sensor data of any event can be retrieved with a single spatiotemporal query. As an illustrative example, we examine sea surface height variability during Hurricane Milton, a major hurricane that occurred in the Gulf of Mexico during October 2024.

Satellite altimetry has been widely used to investigate extreme sea-level events, including storm surges, by measuring sea surface height (SSH) anomalies along satellite ground tracks. Previous studies have used along-track altimetry to detect and analyse storm surges across large ocean regions (Ji et al., 2019; Li et al.). However, the narrow sampling geometry of traditional nadir altimeters limits their ability to resolve the full spatial structure of these events (Abdalla et al., 2021). The wide-swath measurements from the SWOT mission provide two-dimensional SSH fields at high spatial resolution, enabling more detailed observation of the spatial variability associated with extreme sea-level signals (Srinivasan and Tsontos, 2023).

Vega-Gimenez et al. (2025) used SWOT observations to examine SSH anomalies during Hurricane Milton, one of the strongest and most destructive hurricanes recorded in the Gulf of Mexico, which subsequently tracked across central Florida into the Atlantic Ocean (Smith, 2025). Fig. 7 shows snapshots of L3 Altimetry and SWOT data, as well as L4 wind data along the hurricane track. On 9 October 2024, SWOT directly overpassed the hurricane eye, allowing Vega-Gimenez et al. (2025) to analyse the spatial structure of the observed anomaly. Fig. 8 presents a cross-product comparison and correlations between SWOT and conventional altimetry with the L4 DUACS product. For each L3 observation within the domain, the collocated

https://doi.org/10.5194/essd-2026-232
Preprint. Discussion started: 10 April 2026

L4 DUACS value was obtained by bilinear interpolation of the gridded field onto the along-track measurement position, and Pearson correlation coefficients and root-mean-square errors were computed over all valid collocation pairs.

While the previous examples focus on observational analysis and product intercomparison, an increasing number of Earth system studies rely on data-driven models that learn relationships between multiple surface variables. These approaches require consistent collocation of heterogeneous datasets across space and time, which can be technically challenging when data originate from independent archives. OceanTACO directly supports such workflows by providing aligned multi-variable samples across observations, gridded products, and reanalysis fields.

3.5 Data-driven and machine learning reconstruction workflows

Reconstruction of spatially complete ocean surface fields from sparse and irregular observations is a core challenge in Earth system science. For sea surface height (SSH), operational optimal interpolation approaches provide global coverage but may not capture mesoscale variability and reduces effective resolution below that of the observational inputs (Taburet et al., 2019; Ballarotta et al., 2019). In response, numerous works have explored variational, hybrid, and machine learning methods to improve spatiotemporal reconstruction skill (Beauchamp et al., 2023; Martin et al., 2023; Le Guillou et al., 2025; Archambault et al., 2024).

Similar developments have emerged for other ocean surface variables. Deep learning approaches have demonstrated improved gap filling and super-resolution for sea surface temperature (Fanelli et al., 2024; Zou et al., 2023; Zhao et al., 2025), while analogous strategies have been proposed for sea surface salinity (Liang et al., 2025). These approaches increasingly rely on multi-variable learning, where relationships between ocean surface variables provide additional constraints for reconstruction and forecasting. However, building such datasets requires coordinated access to heterogeneous observations and gridded products that must be consistently collocated in space and time.

OceanTACO directly supports such analyses by enabling consistent retrieval of L3 observations (e.g., nadir altimetry, SWOT, SST, and SSS), L4 gridded products, reanalysis fields, and in situ measurements under identical spatial and temporal constraints. As an illustrative example, Fig. 9 shows the spatial correlation between L4 SSH and both SST and SSS in two dynamically distinct regions: the Gulf Stream and the North Indian Ocean. The analysis, performed for September 2023, highlights strong spatial variability and region-dependent relationships between surface variables. Such nonlinear and spatially heterogeneous coupling suggests that data-driven approaches are well suited to exploiting correlated information in reconstruction tasks.

Beyond exploratory analysis, OceanTACO also supports the generation of machine learning–ready training and evaluation datasets. Based on the same spatiotemporal queries used in the correlation analysis, the accompanying documentation demonstrates how OceanTACO can be used to retrieve aligned multi-variable samples and construct PyTorch (Paszke et al., 2019) dataloaders for model training. Because input-target configurations are defined through the spatiotemporal queries used throughout the collection, training and evaluation schemes of data-driven reconstruction experiments become more reproducible.

https://doi.org/10.5194/essd-2026-232
Preprint. Discussion started: 10 April 2026

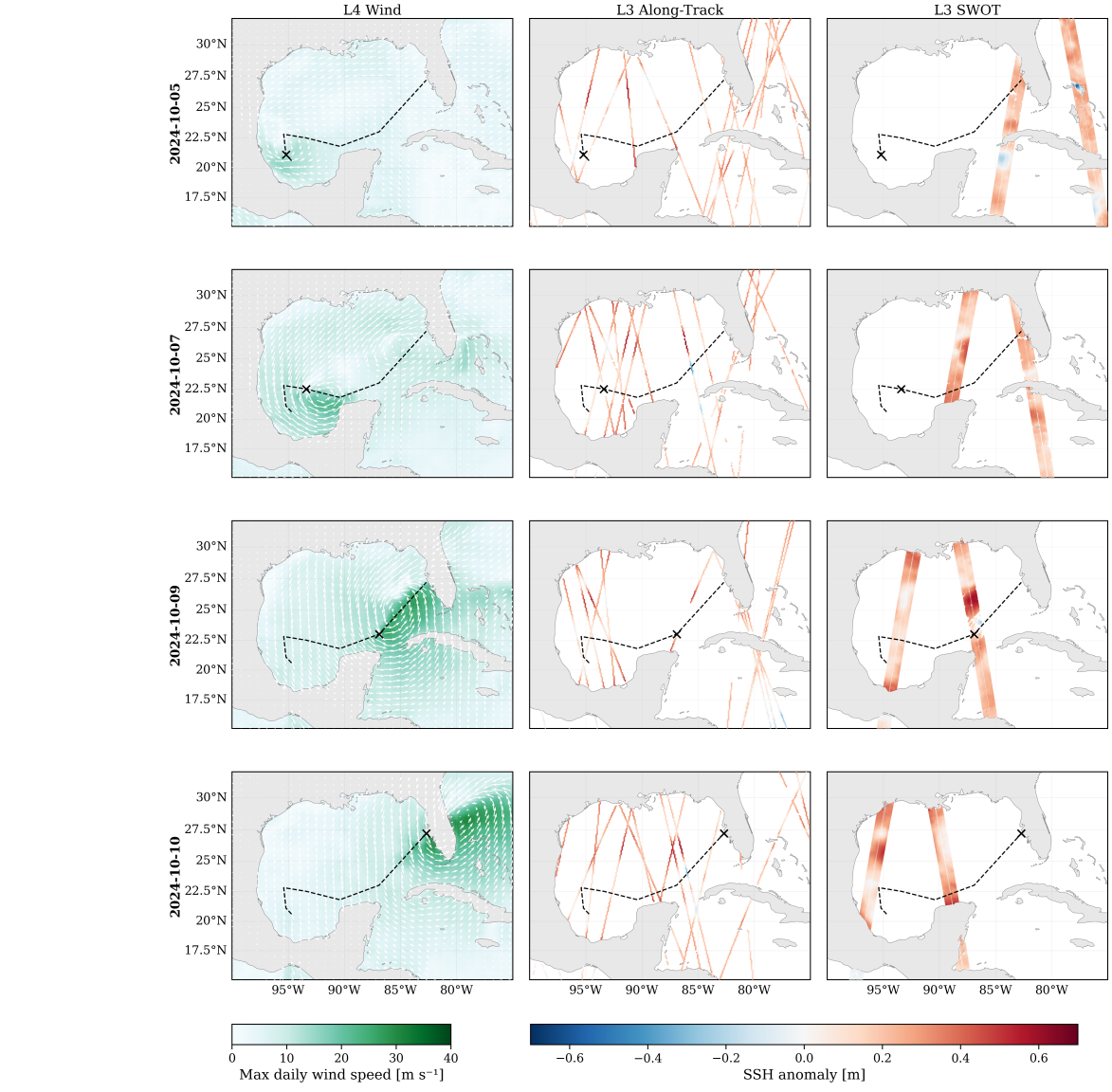

Figure 7. SSH anomaly snapshots during Hurricane Milton (Gulf of Mexico, 100–75°W, 15–32°N) for four dates between 5–10 October 2024. Columns show the gridded L4 Wind product with wind direction components and maximum daily wind speed, L3 conventional along-track altimetry (7 km), and L3 SWOT wide-swath observations (2 km). The dashed line and cross marker indicate the IBTrACS (Knapp et al., 2010) best-track hurricane path and the daily eye position, respectively. Example code that produces this figure is available in this notebook.

https://doi.org/10.5194/essd-2026-232
Preprint. Discussion started: 10 April 2026

Open Access Earth System Science Data Discussions

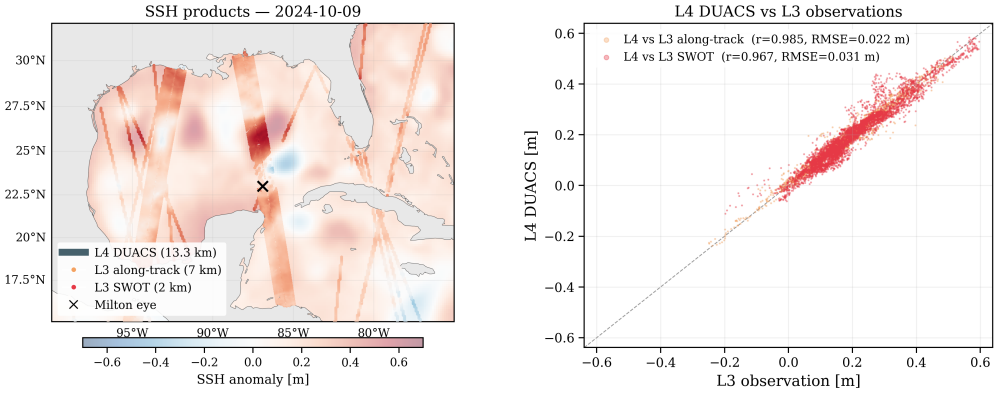

Figure 8. Cross-product SSH comparison over the Gulf of Mexico on 9 October 2024, when SWOT passed directly over the Hurricane Milton eye. *Left:* L4 DUACS field (semi-transparent background) with L3 along-track and L3 SWOT observations overlaid; the cross marks the hurricane eye position. *Right:* Collocated scatter of L3 observations (m) against the bilinearly interpolated L4 DUACS value (m) at each valid measurement location; Pearson R and RMSE are computed over all collocation pairs.

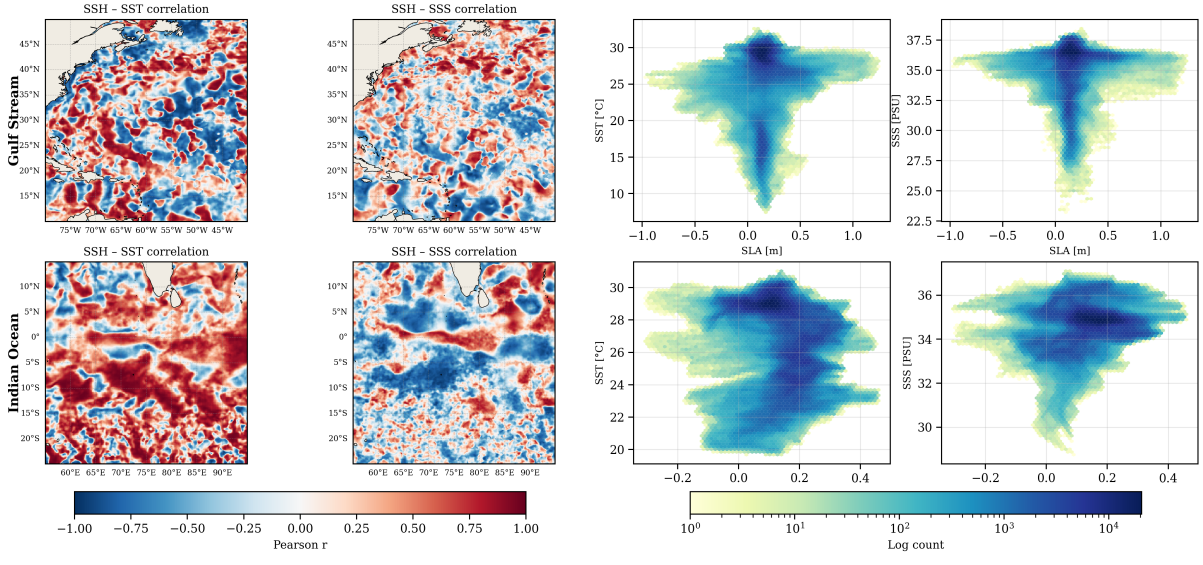

Figure 9. Pixel-wise Pearson correlation between Level 4 SSH anomaly (SLA; m) and sea surface temperature (SST, °C; columns 1 and 3) and sea surface salinity (SSS, PSU; columns 2 and 4) over the Gulf Stream (top row) and North Indian Ocean (bottom row), computed from daily L4 DUACS, OSTIA, and CMEMS SSS fields over all 30 days of September 2023. Correlation maps use a diverging colour scale centred at zero. SST and SSS grids were interpolated to the SSH grid prior to computing per-pixel correlations. Scatter panels show the full distribution of daily SLA–SST and SLA–SSS pairs at all valid pixels, displayed as log-scaled density plots.

https://doi.org/10.5194/essd-2026-232
Preprint. Discussion started: 10 April 2026

4 Conclusions

We have presented OceanTACO, a harmonised global collection of sea surface state datasets structured under the TACO specification. The dataset integrates L3 observations, L4 gridded products, reanalysis fields, and in situ Argo profiles within a unified spatiotemporal indexing and storage framework while preserving native sampling characteristics and full data provenance.

OceanTACO addresses a persistent challenge in Earth system research: the technical fragmentation of multi-sensor ocean datasets across independent archives, formats, and preprocessing conventions. Consistent spatial tiling, temporal alignment, and organisation make multi-product analyses deterministic without custom preprocessing. This supports validation studies, scale-dependent diagnostics, observation system experiments, and multivariate surface state investigations across dynamically diverse ocean regions.

The core dataset spans the SWOT calibration and science phase from 29 March 2023 until 1 August 2025 and can easily be extended when future data become available. For further experiments, we also provide an extended dataset version beginning on January 1 2015 until 29 March 2023, preceding the SWOT era. Additional time periods, reprocessed products, and future satellite missions can be incorporated without restructuring existing components. By combining harmonised data organisation with cloud-native access mechanisms, OceanTACO provides a structured analytical foundation for reproducible ocean surface research within the broader Earth system context.

5 Code and data availability

The complete code to generate the dataset and all included figures can be found at https://github.com/nilsleh/oceanTACO. Additionally, we provided a documentation page at https://oceantaco.readthedocs.io/en/latest/index.html, with installation instructions, dataset descriptions and tutorial notebooks for mentioned workflows. The core dataset is hosted on Hugging Face with DOI https://doi.org/10.57967/hf/8171 (Lehmann and Aybar, 2026a) and released under CC-BY-4.0 License with complete license information in the dataset card. The extended dataset version is also available on Hugging Face under the same license with DOI https://doi.org/10.57967/hf/8172 (Lehmann and Aybar, 2026b).

Appendix A

A1 Data variables

https://doi.org/10.5194/essd-2026-232
Preprint. Discussion started: 10 April 2026

Open Access Earth System Science Data Discussions

Table A1: OceanTACO full variable dictionary across Core and Extended releases. Time coverage is 1 January 2015 to 1 August 2025 (Extended: 1 January 2015 to 29 March 2023; Core: 29 March 2023 to 1 August 2025). Only the Core dataset includes L3-SWOT data. All datasets are time-stamped. Time is stored as coordinate metadata and is not listed as a data variable. Units follow NetCDF/CF metadata as stored in the files.

Data Source	Level	Resolution	Variables (with definition)	Reference
GLORYS-12	Reanalysis	$1/12^\circ$ (≈ 9.3 km)	**zos** – Sea surface height. (m) **thetao** – Temperature. ($^\circ$C) **so** – Salinity. (10^{-3}) **uo** – Eastward velocity. ($\mathrm{m\,s^{-1}}$) **vo** – Northward velocity. ($\mathrm{m\,s^{-1}}$)	(Jean-Michel et al., 2021)
DUACS SSH Product	L4	0.125° (≈ 13.9 km)	**sla** – Sea level anomaly. (m) **err_sla** – Formal mapping error. (m) **ugosa** – Geostrophic velocity anomalies: zonal component. ($\mathrm{m\,s^{-1}}$) **err_ugosa** – Formal mapping error on zonal geostrophic velocity anomalies. ($\mathrm{m\,s^{-1}}$) **vgosa** – Geostrophic velocity anomalies: meridional component. ($\mathrm{m\,s^{-1}}$) **err_vgosa** – Formal mapping error on meridional geostrophic velocity anomalies. ($\mathrm{m\,s^{-1}}$) **adt** – Absolute Dynamic Topography. (m) **ugos** – Absolute geostrophic velocity: zonal component. ($\mathrm{m\,s^{-1}}$) **vgos** – Absolute geostrophic velocity: meridional component. ($\mathrm{m\,s^{-1}}$) **flag_ice** – Ice Flag for a 15**tpa_correction** – TOPEX-A instrumental drift correction derived from altimetry and tide gauges global comparisons (WCRP Sea Level Budget Group, 2018). (m)	(Taburet et al., 2019)
OSTIA SST (Met Office)	L4	0.05° (≈ 5.6 km)	**analysed_sst** – Daily analysed sea surface temperature. ($^\circ$C) **analysis_error** – Estimated standard deviation of analysed sea surface temperature error. (K) **sea_ice_fraction** – Sea-ice area fraction. (%) **mask** – Land-sea-ice-lake classification mask.	(Met Office, 2025)
Multi Observation SSS	L4	0.125° (≈ 13.9 km)	**sos** – Sea surface salinity. (10^{-3}) **dos** – Sea surface density. ($\mathrm{kg\,m^{-3}}$) **sos_error** – Sea surface salinity error. (10^{-3}) **dos_error** – Sea surface density error. ($\mathrm{kg\,m^{-3}}$) **sea_ice_fraction** – Sea-ice area fraction. (%)	(Copernicus Marine Service, 2025e)
Global Ocean Daily Wind	L4	0.25° (≈ 27.8 km)	**eastward_wind** – Daily mean of the stress-equivalent eastward wind component at 10 m. ($\mathrm{m\,s^{-1}}$) **northward_wind** – Daily mean of the stress-equivalent northward wind component at 10 m. ($\mathrm{m\,s^{-1}}$) **eastward_wind_std** – Daily standard deviation of the stress-equivalent eastward wind component at 10 m. ($\mathrm{m\,s^{-1}}$) **northward_wind_std** – Daily standard deviation of the stress-equivalent northward wind component at 10 m. ($\mathrm{m\,s^{-1}}$) **eastward_wind_min** – Daily minimum of the stress-equivalent eastward wind component at 10 m. ($\mathrm{m\,s^{-1}}$) **northward_wind_min** – Daily minimum of the stress-equivalent northward wind component at 10 m. ($\mathrm{m\,s^{-1}}$) **eastward_wind_max** – Daily maximum of the stress-equivalent eastward wind component at 10 m. ($\mathrm{m\,s^{-1}}$) **northward_wind_max** – Daily maximum of the stress-equivalent northward wind component at 10 m. ($\mathrm{m\,s^{-1}}$)	(Copernicus Marine Service, 2025g)

https://doi.org/10.5194/essd-2026-232
Preprint. Discussion started: 10 April 2026

Data Source	Level	Resolution	Variables (with definition)	Reference
Altimetry Along-Track	L3	$\sim 0.06^\circ \times 0.09^\circ$	**sla_filtered** – Sea Level Anomaly (filtered). (m) **sla_filtered_sem** – Standard Error of Mean of SLA (filtered). (m) **mdt** – Mean Dynamic Topography. (m) **mdt_sem** – Standard Error of Mean of MDT. (m) **adt** – Absolute Dynamic Topography. (m) **adt_sem** – Standard Error of Mean of ADT. (m) **obs_mean_lon** – Mean longitude of observations per grid cell. ($^\circ$E) **obs_mean_lat** – Mean latitude of observations per grid cell. ($^\circ$N) **n_obs** – Number of observations per grid cell. **primary_track** – Index of the first contributing track per grid cell. **is_overlap** – Flag indicating that multiple tracks contributed to the grid cell. **track_ids** – Per-track source file identifier. **track_times** – Per-track representative acquisition timestamp. **track_platforms** – Per-track satellite platform name.	(Copernicus Marine Service, 2025a)
SWOT	L3	$\sim 0.02^\circ \times 0.03^\circ$	**ssha_filtered** – Sea Surface Height Anomaly (filtered). (m) **ssha_filtered_sem** – Standard Error of Mean of SSHA (filtered). (m) **ssha_unfiltered** – Sea Surface Height Anomaly (unfiltered). (m) **ssha_unfiltered_sem** – Standard Error of Mean of SSHA (unfiltered). (m) **mdt** – Mean Dynamic Topography. (m) **mdt_sem** – Standard Error of Mean of MDT. (m) **adt_filtered** – Absolute Dynamic Topography (filtered). (m) **adt_filtered_sem** – Standard Error of Mean of ADT (filtered). (m) **adt_unfiltered** – Absolute Dynamic Topography (unfiltered). (m) **adt_unfiltered_sem** – Standard Error of Mean of ADT (unfiltered). (m) **obs_mean_lon** – Mean longitude of observations per grid cell. ($^\circ$E) **obs_mean_lat** – Mean latitude of observations per grid cell. ($^\circ$N) **n_obs** – Number of observations per grid cell. **primary_track** – Index of the first contributing track per grid cell. **is_overlap** – Flag indicating that multiple tracks contributed to the grid cell. **track_ids** – Per-track source file identifier. **track_times** – Per-track representative acquisition timestamp.	(AVISO/DUACS, 2024)
L3 SST	L3	0.10° (≈ 11.1 km)	**sst_dtime** – Time difference from reference time. (seconds) **sea_surface_temperature** – Sea surface temperature. ($^\circ$C) **adjusted_sea_surface_temperature** – Bias-adjusted sea surface temperature. ($^\circ$C) **sses_bias** – Sensor-specific error statistic bias estimate. (K) **sses_standard_deviation** – Sensor-specific error statistic standard deviation. (K) **quality_level** – Per-pixel quality level. **sources_of_sst** – Source sensor identifier for sea surface temperature. **bias_to_reference_sst** – Bias to the reference sea surface temperature used for cross-calibration. (K)	(Copernicus Marine Service, 2025f)
L3 Salinity (SMOS)	L3	$\sim 0.23^\circ \times 0.26^\circ$	**Sea_Surface_Salinity** – Practical sea surface salinity. (10^{-3}) **Sea_Surface_Salinity_Rain_Corrected** – Practical sea surface salinity corrected from rain instantaneous freshening effect (bulk salinity in rainy conditions). (10^{-3}) **Sea_Surface_Salinity_Error** – Random uncertainty on sea surface salinity. (10^{-3}) **Sea_Surface_Salinity_Rain_Corrected_Error** – Random uncertainty on sea surface salinity corrected from rain effect. (10^{-3}) **X_Swath** – Distance of the grid point to the center of the swath. (km) **Mean_Acq_Time** – Dwell line mean acquisition time. (nanoseconds) **Sea_Surface_Salinity_QC** – Quality flag for sea surface salinity.	(Copernicus Marine Service, 2025b)

https://doi.org/10.5194/essd-2026-232
Preprint. Discussion started: 10 April 2026

Data Source	Level	Resolution	Variables (with definition)	Reference
Argo	In situ	–	**CYCLE_NUMBER** – Float cycle number. **DIRECTION** – Direction of the station profiles. **PLATFORM_NUMBER** – Float unique identifier. **PRES** – Sea Pressure. (dbar) **PRES_ERROR** – ERROR IN Sea Pressure. (dbar) **PSAL** – PRACTICAL SALINITY. (psu) **PSAL_ERROR** – ERROR IN PRACTICAL SALINITY. (psu) **TEMP** – SEA TEMPERATURE IN SITU ITS-90 SCALE. (°C) **TEMP_ERROR** – ERROR IN SEA TEMPERATURE IN SITU ITS-90 SCALE. (°C)	(Wong et al., 2020)

https://doi.org/10.5194/essd-2026-232
Preprint. Discussion started: 10 April 2026

Open Access Earth System Science Data Discussions

A1 L3 SSH mission temporal coverage

Table B1. Comparison of mission temporal coverage between reprocessed (MY_008_062) and near-real-time (NRT_008_044) sea level products

Mission	MY_008_062 (Reprocessed)	NRT_008_044 (Near-Real-Time)
TOPEX/Poseidon	13 Oct 1992 – 24 Apr 2002	
TOPEX/Poseidon (new orbit)	10 Sep 2002 – 03 Oct 2005	
ERS-1	23 Oct 1992 – 15 May 1995	
ERS-1 (geodetic phase)	10 Apr 1994 – 21 Mar 1995	
ERS-2	15 May 1995 – 14 May 2002	
Envisat	17 May 2002 – 18 Oct 2010	
Envisat (new orbit)	26 Oct 2010 – 08 Apr 2012	
GFO	07 Jan 2000 – 07 Sep 2008	
Jason-1	24 Apr 2002 – 19 Oct 2008	
Jason-1 (new orbit)	10 Feb 2009 – 03 Mar 2012	
Jason-1 (geodetic orbit)	07 May 2012 – 01 Jun 2013	
Jason-2	19 Oct 2008 – 26 May 2016	
Jason-2 (interleaved orbit)	17 Oct 2016 – 17 May 2017	
Jason-2 (long-repeat orbit)	11 Jul 2017 – 14 Sep 2017	
Jason-3	26 May 2016 – 29 Dec 2021	
Jason-3 (interleaved)	25 Apr 2022 – 19 Nov 2024	03 May 2022 – 08 Jan 2025
Jason-3 (long-repeat orbit)		19 Jun 2025 – 24 Oct 2025
Saral/AltiKa	14 Mar 2013 – 31 Mar 2015	25 Oct 2023 – 24 Oct 2025
Saral/AltiKa (geodetic orbit)	31 Mar 2015 – 01 May 2025	
CryoSat-2	16 Jul 2010 – 31 Jul 2020	25 Oct 2023 – 24 Oct 2025
CryoSat-2 (new orbit)	01 Aug 2020 – 01 May 2025	
Sentinel-3A	28 Jun 2016 – 01 May 2025	25 Oct 2023 – 24 Oct 2025
Sentinel-3B	27 Nov 2018 – 01 May 2025	25 Oct 2023 – 24 Oct 2025
Sentinel-6A (SAR)	29 Dec 2021 – 01 May 2025	25 Oct 2023 – 24 Oct 2025
HaiYang-2A	12 Apr 2014 – 15 Mar 2016	
HaiYang-2A (geodetic orbit)	31 Mar 2016 – 09 Jun 2020	
HaiYang-2B	20 Dec 2019 – 01 May 2025	22 Oct 2023 – 21 Oct 2025
HaiYang-2B (5 Hz)		07 Jul 2024 – 21 Oct 2025
SWOT nadir CalVal	16 Jan 2023 – 09 Jul 2023	
SWOT nadir	21 Jul 2023 – 01 May 2025	09 Nov 2023 – 24 Oct 2025

https://doi.org/10.5194/essd-2026-232
Preprint. Discussion started: 10 April 2026

Open Access Earth System Science Data Discussions

B1 Disk space compression

For all processed region files we choose one of the primary variables stored as `int16` with `zlib` compression (level 4) applied via the NetCDF4/HDF5 layer. To quantify the resulting storage savings we compared, for each data source, (i) the uncompressed baseline size against (ii) the actual on-disk compressed size, measured directly from the HDF5 chunk storage via `h5py`'s `get_storage_size()`. For gridded L4 products and GLORYS, the uncompressed baseline is the size of the raw source files read at their native floating-point precision (`dtype.itemsize` × number of elements); for the along-track swath products L3 SWOT and L3 SSH, whose raw files are individual track files, we choose a conservative `float32` estimate (4 bytes per element). Sizes were accumulated over all eight ocean regions and the core dataset period, then averaged per day. Sizes are reported for the primary variable of each modality; modalities with multiple stored variables (e.g., GLORYS: SSH, SST, SSS, uo, vo; L4 Wind: eastward and northward components plus standard deviations) have correspondingly larger per-day footprints in the full dataset.

Across all ten data sources the OceanTACO format achieves an overall compression ratio of $20.5\times$. Individual ratios range from $2.2\times$ for the sparse L3 SSS SMOS ascending/descending passes to $60.3\times$ for L4 Wind, which contains large homogeneous low-wind regions that compress extremely well. Gridded SSH, SST, and GLORYS records yield ratios of 8.6–$11.6\times$, while the along-track SWOT and L3 SSH products achieve $26.9\times$ and $33.1\times$, respectively, owing to the prevalence of fill values outside the narrow swath. The full per-source breakdown is given in Table A1.

https://doi.org/10.5194/essd-2026-232
Preprint. Discussion started: 10 April 2026

Open Access Earth System Science Data Discussions

Modality	Primary variable	Uncompressed (MB day^{-1})	Compressed (MB day^{-1})	Ratio
GLORYS	`zos`	70.5	6.9	10.3×
L4 SSH	`sla`	35.8	3.1	11.6×
L4 SST	`analysed_sst`	60.4	7.0	8.6×
L4 SSS	`sos`	33.0	7.8	4.2×
L4 Wind	`eastward_wind`	306.9	5.1	60.3×
L3 SST	`adjusted_sea_surface_temperature`	23.0	1.9	12.3×
L3 SWOT[a]	`ssha_filtered`	546.7	20.3	26.9×
L3 SSH[a]	`sla_filtered`	46.4	1.4	33.1×
L3 SSS SMOS Asc	`Sea_Surface_Salinity`	1.6	0.7	2.2×
L3 SSS SMOS Desc	`Sea_Surface_Salinity`	1.6	0.7	2.2×
Argo	`TEMP`	0.0	0.0	1.0×
TOTAL (primary vars.)		**1125.9**	**55.0**	**20.5**×

Table A1. Per-modality storage footprint of the OceanTACO dataset. For each data source the table lists the primary variable stored, the average daily uncompressed size, the average daily on-disk size after `int16` quantisation and `zlib` compression, and the resulting compression ratio. Sizes are averaged over the full dataset period across all eight ocean regions. The TOTAL row sums compressed sizes for the listed primary variables only; the full dataset footprint (all variables per modality) is approximately 324 GB for the combined core and 556GB for the extended periods.

[a] Uncompressed size estimated as `float32` (4 bytes per element) because raw L3 SWOT and L3 SSH source files contain overlapping swath segments that are not directly comparable to the gridded region files. All other modalities use the actual raw-file size at native floating-point precision (`dtype.itemsize` × number of elements).

https://doi.org/10.5194/essd-2026-232
Preprint. Discussion started: 10 April 2026

Open Access Earth System Science Data Discussions

Table B1. Compression-loss statistics for each data source after `int16+zlib` encoding. For gridded products, errors are computed against a pre-encoding regional reference produced by the same formatting pipeline; for L3 track products, source swaths are conservatively binned to the target grid (no smoothing) and compared on overlapping valid cells only. L3 SST natively comes as int16 data already, and is therefore not included in the table.

Modality	Variable	Unit	RMSE (mean±std)	Bias (mean±std)	P99 Error (mean±std)
GLORYS	SSH	m	$1.44 \times 10^{-4} \pm 7.6 \times 10^{-8}$	$8.97 \times 10^{-9} \pm 2.3 \times 10^{-7}$	$2.48 \times 10^{-4} \pm 8.6 \times 10^{-8}$
	SST	°C	$2.89 \times 10^{-4} \pm 2.2 \times 10^{-7}$	$2.62 \times 10^{-7} \pm 7.8 \times 10^{-7}$	$4.95 \times 10^{-4} \pm 5.3 \times 10^{-7}$
	SSS	10^{-3}	$2.89 \times 10^{-4} \pm 1.9 \times 10^{-7}$	$-2.79 \times 10^{-6} \pm 5.9 \times 10^{-7}$	$4.94 \times 10^{-4} \pm 2.6 \times 10^{-7}$
	uo	$\mathrm{m\,s^{-1}}$	$2.89 \times 10^{-4} \pm 1.7 \times 10^{-7}$	$-7.17 \times 10^{-9} \pm 5.9 \times 10^{-7}$	$4.94 \times 10^{-4} \pm 7.4 \times 10^{-7}$
	vo	$\mathrm{m\,s^{-1}}$	$2.89 \times 10^{-4} \pm 1.7 \times 10^{-7}$	$-2.53 \times 10^{-8} \pm 4.5 \times 10^{-7}$	$4.94 \times 10^{-4} \pm 7.4 \times 10^{-7}$
L4 SSH	SLA	m	$1.41 \times 10^{-4} \pm 1.7 \times 10^{-7}$	$3.36 \times 10^{-8} \pm 3.3 \times 10^{-7}$	$2.00 \times 10^{-4} \pm 1.1 \times 10^{-17}$
L4 SST	SST	°C	$9.46 \times 10^{-6} \pm 4.9 \times 10^{-7}$	$1.22 \times 10^{-6} \pm 1.7 \times 10^{-6}$	$1.59 \times 10^{-5} \pm 0$
L4 SSS	SSS	10^{-3}	$5.77 \times 10^{-4} \pm 4.5 \times 10^{-7}$	$6.94 \times 10^{-8} \pm 1.0 \times 10^{-6}$	$9.90 \times 10^{-4} \pm 1.9 \times 10^{-7}$
L4 Wind	Wind	$\mathrm{m\,s^{-1}}$	0.0029 ± 0.0000	$8.15 \times 10^{-8} \pm 3.8 \times 10^{-6}$	0.0050 ± 0.0000
L3 SSH	SLA	m	$2.05 \times 10^{-4} \pm 4.1 \times 10^{-5}$	$1.04 \times 10^{-7} \pm 1.0 \times 10^{-6}$	0.0110 ± 0.0029
L3 SWOT	SSHA	m	$2.89 \times 10^{-4} \pm 1.2 \times 10^{-4}$	$-2.01 \times 10^{-7} \pm 6.4 \times 10^{-7}$	0.0120 ± 0.0047
L3 SSS SMOS	SSS	10^{-3}	$7.07 \times 10^{-4} \pm 2.8 \times 10^{-6}$	$1.22 \times 10^{-6} \pm 5.6 \times 10^{-6}$	0.0010 ± 0.0000

https://doi.org/10.5194/essd-2026-232
Preprint. Discussion started: 10 April 2026

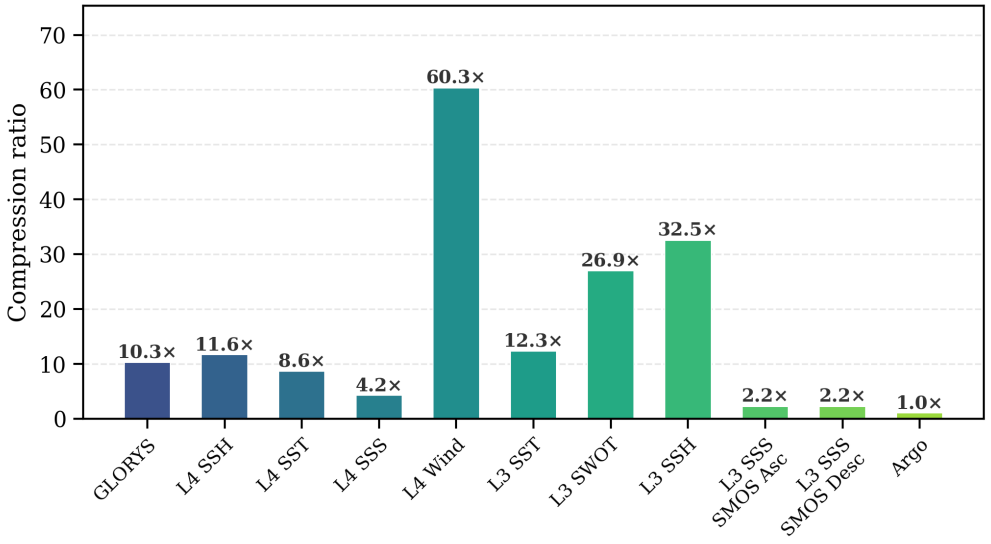

Figure C1. Compression ratios achieved by the OceanTACO processing pipeline for each of the ten data sources. Each bar shows the ratio of the uncompressed source-file size to the on-disk size of the corresponding processed region files (`int16+zlib`). The ratio for L3 SWOT and L3 SSH is computed relative to a `float32` element-count estimate rather than the raw swath files (see Table A1).

https://doi.org/10.5194/essd-2026-232
Preprint. Discussion started: 10 April 2026

Author contributions. The dataset was conceptualized by NL and CA. The curation, construction, and corresponding figures were also carried out by NL and CA. The initial manuscript was drafted by NL with inputs and subsequent revisions by all authors.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. The OceanTACO dataset is generated using E.U. Copernicus Marine Service Information and AVISO+ Products. Argo data were collected and made freely available by the International Argo Program and the national programs that contribute to it. The Argo Program is part of the Global Ocean Observing System. NL and XXZ were supported by German Federal Ministry for Economic Affairs and Climate Action in the framework of the "national center of excellence ML4Earth" (grant number: 50EE2201C) and by Munich Center for Machine Learning. JLB was supported by the Technical University of Munich – Institute for Advanced Study, Germany. We also acknowledge Claude Code, the AI tool developed by Anthropic, as a coding assistant.

https://doi.org/10.5194/essd-2026-232
Preprint. Discussion started: 10 April 2026

References

Abdalla, S., Kolahchi, A. A., Ablain, M., Adusumilli, S., Bhowmick, S. A., Alou-Font, E., Amarouche, L., Andersen, O. B., Antich, H., Aouf, L., et al.: Altimetry for the future: Building on 25 years of progress, Advances in Space Research, 68, 319–363, 2021.

Algarabel, G., Steventon, M., Munafò, M., and Ireland, M.: How reproducible and reliable is geophysical research? A review of the availability and accessibility of data and software for research published in journals, Seismica, 2, https://doi.org/10.26443/seismica.v2i1.278, 2023.

Aouni, A. E., Gaudel, Q., Johnson, J. E., Charly, R., Sommer, J. L., van Gennip, Fablet, R., Drevillon, M., DRILLET, Y., and Traon, P. Y. L.: OceanBench: A Benchmark for Data-Driven Global Ocean Forecasting systems, in: The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track, https://openreview.net/forum?id=wZGe1Kqs8G, 2025.

Archambault, T., Filoche, A., Charantonis, A. A., and Béréziat, D.: Multimodal Unsupervised Spatio-Temporal Interpolation of satellite ocean altimetry maps, in: VISAPP, 2023.

Archambault, T., Filoche, A., Charantonis, A., Béréziat, D., and Thiria, S.: Learning sea surface height interpolation from multi-variate simulated satellite observations, Journal of Advances in Modeling Earth Systems, 16, e2023MS004 047, 2024.

Archer, M., Wang, J., Klein, P., Dibarboure, G., and Fu, L.-L.: Wide-swath satellite altimetry unveils global submesoscale ocean dynamics, Nature, 640, 691–696, 2025.

AVISO/DUACS: The SWOT L3 LR SSH product, derived from the L2 SWOT KaRIn low rate ocean data products (NASA/JPL and CNES), is produced and made freely available by AVISO and DUACS teams as part of the DESMOS Science Team project, https://doi.org/10.24400/527896/A01-2023.017, 2024.

Aybar, C., Contreras, J., Ma, C., Pellicer-Valero, O. J., Mateo-García, G., Gómez-Chova, L., Camps-Valls, G., Lehmann, N., Czerkawski, M., Montero, D., et al.: The Missing Piece: Standardising for AI-ready Earth Observation Datasets, in: TerraBytes-ICML 2025 workshop, 2025.

Ballarotta, M., Ubelmann, C., Pujol, M.-I., Taburet, G., Fournier, F., Legeais, J.-F., Faugère, Y., Delepoulle, A., Chelton, D., Dibarboure, G., et al.: On the resolutions of ocean altimetry maps, Ocean science, 15, 1091–1109, 2019.

Ballarotta, M., Ubelmann, C., Bellemin-Laponnaz, V., Le Guillou, F., Meda, G., Anadon, C., Laloue, A., Delepoulle, A., Faugère, Y., Pujol, M.-I., et al.: Integrating wide-swath altimetry data into Level-4 multi-mission maps, Ocean Science, 21, 63–80, 2025.

Beauchamp, M., Febvre, Q., Georgenthum, H., and Fablet, R.: 4DVarNet-SSH: End-to-end learning of variational interpolation schemes for nadir and wide-swath satellite altimetry, Geoscientific Model Development, 16, 2119–2147, 2023.

Boutin, J., Vergely, J.-L., Marchand, S., d'Amico, F., Hasson, A., Kolodziejczyk, N., Reul, N., Reverdin, G., and Vialard, J.: New SMOS Sea Surface Salinity with reduced systematic errors and improved variability, Remote Sensing of Environment, 214, 115–134, 2018.

Buongiorno Nardelli, B., Droghei, R., and Santoleri, R.: Multi-dimensional interpolation of SMOS sea surface salinity with surface temperature and in situ salinity data, Remote Sensing of Environment, 180, 392–402, https://doi.org/10.1016/j.rse.2015.12.052, 2016.

Coca-Castro, A., Fouilloux, A., Barros Lourenço, R., McDonald, A., Rao, Y., and Hosking, J. S.: Improving the reproducibility in geoscientific papers: lessons learned from a Hackathon in climate science, Environmental Data Science, 4, e6, https://doi.org/10.1017/eds.2024.35, 2025.

Copernicus Marine Service: Global Ocean Along Track L3 Sea Surface Heights Reprocessed 1993 Ongoing Tailored For Data Assimilation, https://doi.org/10.48670/moi-00146, accessed: 2025-10-28, 2025a.

https://doi.org/10.5194/essd-2026-232
Preprint. Discussion started: 10 April 2026

Copernicus Marine Service: SMOS CATDS Qualified (L2Q) Sea Surface Salinity, https://doi.org/10.1016/j.rse.2016.02.061, accessed: 2025-10-28, 2025b.

Copernicus Marine Service: Global Ocean Gridded L4 Sea Surface Heights and Derived Variables Reprocessed 1993 Ongoing, Copernicus Marine Service, https://doi.org/10.48670/moi-00148, accessed: 2025-10-28, 2025c.

Copernicus Marine Service: Global Ocean Along Track L3 Sea Surface Heights NRT, https://doi.org/10.48670/moi-00153, accessed: 2025-10-28, 2025d.

Copernicus Marine Service: Multi Observation Global Ocean Sea Surface Salinity and Sea Surface Density, https://doi.org/10.48670/moi-00150, accessed: 2025-10-28, 2025e.

Copernicus Marine Service: Sea Surface Temperature Multi-sensor L3 Observations, https://doi.org/10.48670/moi-00151, accessed: 2025-10-28, 2025f.

Copernicus Marine Service: Global Ocean Hourly Reprocessed Sea Surface Wind and Stress, https://doi.org/10.48670/moi-00152, accessed: 2025-10-28, 2025g.

Droghei, R., Buongiorno Nardelli, B., and Santoleri, R.: Combining in-situ and satellite observations to retrieve salinity and density at the ocean surface, Journal of Atmospheric and Oceanic Technology, 33, 1211–1223, https://doi.org/10.1175/JTECH-D-15-0194.1, 2016.

Droghei, R., Buongiorno Nardelli, B., and Santoleri, R.: A New Global Sea Surface Salinity and Density Dataset From Multivariate Observations (1993-2016), Frontiers in Marine Science, 5, 84, https://doi.org/10.3389/fmars.2018.00084, 2018.

Embury, O., Merchant, C. J., Good, S. A., Rayner, N. A., Høyer, J. L., Atkinson, C., Block, T., Alerskans, E., Pearson, K. J., Worsfold, M., McCarroll, N., and Donlon, C.: Satellite-based time-series of sea-surface temperature since 1980 for climate applications, Scientific Data, 11, 326, https://doi.org/10.1038/s41597-024-03147-w, 2024.

Fanelli, C., Ciani, D., Pisano, A., and Buongiorno Nardelli, B.: Deep learning for the super resolution of Mediterranean sea surface temperature fields, Ocean Science, 20, 1035–1050, 2024.

Fouchet, E., Benkiran, M., Le Traon, P.-Y., and Remy, E.: Comparison of a global high-resolution ocean data assimilation system with SWOT observations, Frontiers in Marine Science, 12, 1563 934, 2025.

Fu, L.-L., Pavelsky, T., Cretaux, J.-F., Morrow, R., Farrar, J. T., Vaze, P., Sengenes, P., Vinogradova-Shiffer, N., Sylvestre-Baron, A., Picot, N., et al.: The surface water and ocean topography mission: A breakthrough in radar remote sensing of the ocean and land surface water, Geophysical Research Letters, 51, e2023GL107 652, 2024.

Jean-Michel, L., Eric, G., Romain, B.-B., Gilles, G., Angélique, M., Marie, D., Clément, B., Mathieu, H., Olivier, L. G., Charly, R., et al.: The Copernicus global 1/12 oceanic and sea ice GLORYS12 reanalysis, Frontiers in Earth Science, 9, 698 876, 2021.

Ji, T., Li, G., and Zhang, Y.: Observing storm surges in China's coastal areas by integrating multi-source satellite altimeters, Estuarine, Coastal and Shelf Science, 225, 106 224, 2019.

Johnson, J. E., Febvre, Q., Gorbunova, A., Metref, S., Ballarotta, M., Le Sommer, J., et al.: OceanBench: the sea surface height edition, vol. 36, 2024.

Klein, P., Lapeyre, G., Siegelman, L., Qiu, B., Fu, L.-L., Torres, H., Su, Z., Menemenlis, D., and Le Gentil, S.: Ocean-scale interactions from space, Earth and Space Science, 6, 795–817, 2019.

Knapp, K. R., Kruk, M. C., Levinson, D. H., Diamond, H. J., and Neumann, C. J.: The international best track archive for climate stewardship (IBTrACS) unifying tropical cyclone data, Bulletin of the American Meteorological Society, 91, 363–376, 2010.

Krishna, D. and Sreejith, K.: Resolution of SWOT altimetry: Improvements along continental margins, Earth and Space Science, 12, e2025EA004 312, 2025.

https://doi.org/10.5194/essd-2026-232
Preprint. Discussion started: 10 April 2026

Le Guillou, F., Chapron, B., and Rio, M.-H.: VarDyn: Dynamical joint-reconstructions of sea surface height and temperature from multi-sensor satellite observations, Journal of Advances in Modeling Earth Systems, 17, e2024MS004 689, 2025.

Le Traon, P.-Y., Dibarboure, G., Lellouche, J.-M., Pujol, M.-I., Benkiran, M., Drevillon, M., Drillet, Y., Faugère, Y., and Remy, E.: Satellite altimetry and operational oceanography: from Jason-1 to SWOT, Ocean Science, 21, 1329–1347, 2025.

Lehmann, N. and Aybar, C.: OceanTACO, https://doi.org/10.57967/hf/8171, 2026a.

Lehmann, N. and Aybar, C.: OceanTACO Extended, https://doi.org/10.57967/hf/8172, 2026b.

Li, J.-L. F., Tsai, Y.-C., Xu, K.-M., Lee, W.-L., Jiang, J. H., Yu, J.-Y., Fetzer, E. J., and Stephens, G.: Inferring the linkage of sea surface height anomalies, surface wind stress and sea surface temperature with the falling ice radiative effects using satellite data and global climate models, Environmental Research Communications, 4, 125 004, 2022.

Li, X., Han, G., Yang, J., Chen, D., Zheng, G., and Chen, N.: Using satellite altimetry to calibrate the simulation of typhoon Seth storm surge off Southeast China, Remote Sensing, 10.

Liang, Z., Bao, S., Zhang, W., Yan, H., Duan, B., and Wang, H.: Super-resolution reconstruction of SMOS sea surface salinity from multi-variate satellite observations based on deep learning, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2025.

Llovel, W. and Lee, T.: Importance and origin of halosteric contribution to sea level change in the southeast Indian Ocean during 2005–2013, Geophysical Research Letters, 42, 1148–1157, 2015.

Martin, S. A., Manucharyan, G. E., and Klein, P.: Synthesizing sea surface temperature and satellite altimetry observations using deep learning improves the accuracy and resolution of gridded sea surface height anomalies, Journal of Advances in Modeling Earth Systems, 15, e2022MS003 589, 2023.

Martin, S. A., Manucharyan, G. E., and Klein, P.: Generative data assimilation for surface ocean state estimation from multi-modal satellite observations, Journal of Advances in Modeling Earth Systems, 17, e2025MS005 063, 2025.

Maze, G. and Balem, K.: argopy: A Python library for Argo ocean data analysis, Journal of Open Source Software, 5, https://doi.org/10.21105/joss.02425, 2020.

Mercator Ocean / Copernicus Marine Service: Copernicus Marine Toolbox (CLI & Python) [software], https://github.com/mercator-ocean/copernicus-marine-toolbox, version 2.2.3; licensed under EUPL-1.2., 2025.

Met Office: Global Ocean OSTIA Sea Surface Temperature and Sea Ice Reprocessed, https://doi.org/10.48670/moi-00168, accessed: 2025-10-28, 2025.

Morrow, R., Fu, L.-L., Ardhuin, F., Benkiran, M., Chapron, B., Cosme, E., d'Ovidio, F., Farrar, J. T., Gille, S. T., Lapeyre, G., et al.: Global observations of fine-scale ocean surface topography with the surface water and ocean topography (SWOT) mission, Frontiers in Marine Science, 6, 232, 2019.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library, Advances in neural information processing systems, 32, 2019.

Perez, F. and Granger, B. E.: Project Jupyter: Computational narratives as the engine of collaborative data science, Retrieved September, 11, 108, 2015.

Smith, A. B.: US Billion-dollar weather and climate disasters, 1980-present, https://www.ncei.noaa.gov/access/billions/, 2025.

Srinivasan, M. and Tsontos, V.: Satellite altimetry for ocean and coastal applications: A review, Remote Sensing, 15, 3939, 2023.

Taburet, G., Sanchez-Roman, A., Ballarotta, M., Pujol, M.-I., Legeais, J.-F., Fournier, F., Faugere, Y., and Dibarboure, G.: DUACS DT2018: 25 years of reprocessed sea level altimetry products, Ocean Science, 15, 1207–1224, 2019.

https://doi.org/10.5194/essd-2026-232
Preprint. Discussion started: 10 April 2026

Vega-Gimenez, D., Amores, A., Paris, A., and Pascual, A.: Expanding the coastal observation frontier: SWOT reveals the spatial footprint of storm surges, Geophysical Research Letters, 52, e2025GL117 299, 2025.

Vergara, O., Morrow, R., Pujol, I., Dibarboure, G., and Ubelmann, C.: Revised global wave number spectra from recent altimeter observations, Journal of Geophysical Research: Oceans, 124, 3523–3537, 2019.

Wang, Y., Zhang, S., and Jia, Y.: Enhanced resolution capability of SWOT sea surface height measurements and their application in monitoring ocean dynamics variability, Ocean Science, 21, 931–944, 2025.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., et al.: The FAIR Guiding Principles for scientific data management and stewardship, Scientific data, 3, 1–9, 2016.

Wong, A. P. S. et al.: Argo Data 1999–2019: Two Million Temperature-Salinity Profiles and Subsurface Velocity Observations From a Global Array of Profiling Floats, Frontiers in Marine Science, 7, 700, https://doi.org/10.3389/fmars.2020.00700, 2020.

Zhao, E., Goh, E., Yepremyan, A., Wang, J., and Wilson, B.: Multi-satellite U-Net for high-resolution sea surface temperature reconstruction, EGUsphere, 2025, 1–25, 2025.

Zou, R., Wei, L., and Guan, L.: Super resolution of satellite-derived sea surface temperature using a transformer-based model, Remote Sensing, 15, 5376, 2023.