



Morlenn: a three-hourly North Atlantic dataset of mixed-layer depth and subsurface stratification, merging ARGO, satellite observations, and atmospheric forcing

Lucile Gaultier^{1,★}, Parth Tripathi^{1,★}, Fabrice Collard¹, and Clément de Boyer Montégut²

¹OceanDataLab, Locmaria-Plouzané, France

²Univ Brest, Ifremer, CNRS, IRD, Laboratoire d'Océanographie Physique et Spatiale (LOPS), F-29280 Plouzané, France

★These authors contributed equally to this work.

Correspondence: Lucile Gaultier (lucile.gaultier@oceandatalab.com)

Abstract. Mixed-layer depth (MLD) and the stratification beneath it are key to ocean–atmosphere interactions, influencing how the upper ocean responds to atmospheric forcing. Their variability remains sparsely observed, particularly at subdaily timescales. Here we present Morlenn, a dataset of MLD and maximum buoyancy (Brunt-Väisälä) frequency squared, $N_{\max}^2$ covering the North Atlantic and adjacent equatorial waters from 2011 to 2023, at 0.25° spacing and three-hourly intervals.

5 The dataset is generated from ocean surface satellite observations, atmospheric fluxes and wave forcing history using a neural network trained on Argo individual observations of MLD and stratification. Two versions are provided, one using surface inputs alone and one additionally incorporating neighbouring Argo data. Evaluation combines the broad coverage of Argo with independent glider observations. Scored against Argo stations over 2011–2023, the version using neighbouring data reaches an MLD RMS error of 38 m, against 45 m for ARMOR3D and 56 m for GLORYS on the same stations. It characterises MLD and
 10 $N_{\max}^2$ errors across regions and depth ranges, while spectral and diurnal-cycle analyses assess the variability represented by the product. Validation of subsurface stratification is less extensive. Morlenn provides new estimates of surface ocean mixed-layer depth and stratification for investigating upper-ocean variability, with documented limitations for shallow layers, extreme depths and very low salinity. The dataset is available at <https://doi.org/10.17882/120058> (Gaultier et al., 2026).

1 Introduction

15 The ocean surface mixed layer, the quasi-homogeneous layer in direct contact with the atmosphere, mediates exchanges of heat, momentum and gases between the atmosphere and the ocean interior. Its depth determines the volume of water over which surface inputs are mixed, while the stratification beneath it influences the resistance to further vertical mixing (Sallée et al., 2021). Describing both properties is therefore important for understanding how the upper ocean responds to surface cooling, heating and storms (e.g. Neetu et al., 2012). Two mixed layers of similar depth can respond differently to the same forcing
 20 if the stratification below them differs. Mixed-layer depth (MLD) and the maximum squared buoyancy frequency beneath it, denoted $N_{\max}^2$, provide complementary diagnostics of this vertical structure: the extent of the surface layer and the strength of the underlying stratification (Sérazin et al., 2023).



Observing the evolution of these properties requires measurements across both space and time. Argo profiles provide sparse coverage, good vertical sampling of the ocean interior, but their time separation leaves much of the variability associated with individual storms and the diurnal cycle unresolved. Gliders provide higher frequency sampling along trajectories, over a smaller geographic area. Satellite observations and atmospheric reanalyses offer sub-daily global coverage of the ocean surface and its forcing, but do not directly resolve the vertical structure needed to diagnose MLD and $N_{\max}^2$. The reconstruction problem is consequently to combine the vertical information supplied by profiles with the more extensive surface information, and to establish how mixed-layer variability and stratification can be recovered between in-situ observations.

Existing products address this problem through different combinations of observations and modelling. Profile-based climatologies describe the seasonal distribution of MLD (de Boyer Montégut et al., 2004). ARMOR3D combines satellite and in-situ observations to reconstruct the ocean interior (Guinehut et al., 2012); GLORYS12 provides an ocean reanalysis with data assimilation (Lellouche et al., 2021); and OSnet learns relationships between surface observations and subsurface properties in the Gulf Stream (Pauthenet et al., 2022). These products provide valuable reference estimates, but their sampling, MLD definitions and use of in-situ information differ, and none resolves MLD together with subsurface stratification at subdaily intervals. Here, we focus on providing a joint reconstruction of MLD and subsurface stratification at subdaily intervals over the North Atlantic, together with an assessment of its accuracy and represented variability.

The approach is motivated by the accumulated response of the upper ocean to forcing. Mixed-layer properties depend on the history of wind, waves and air–sea heat exchange, as well as on the ocean conditions in which that forcing acts. An instantaneous surface observation captures only part of this information. A sequence of forcing and surface observations can describe both short events and sustained changes, while contemporary surface temperature, salinity and sea level provide information about the ocean state. Location, bathymetry and season further describe the setting in which mixing and restratification occur. Combining these histories and contextual variables offers a way to estimate subsurface diagnostics without requiring a contemporaneous vertical profile at every prediction location.

We implement this approach in Morlenn (Mixed-layer Ocean Reconstruction from Learning Neural Network), developed within the UpperDyn project (<https://upperdyn.odl.bzh/project>). A neural network is trained on individual Argo-derived MLD and $N_{\max}^2$ estimates paired with collocated input histories and surface conditions. Successive temporal filters represent the dependence on preceding forcing, and a context-dependent weighting combines information from different parts of the history. Separate processing pathways allow MLD and stratification to have different sensitivities to the same inputs. The model estimates each location and time separately and is then applied on a regular grid. Two configurations are provided: a baseline using surface fields and forcing, and an assimilating configuration that additionally uses neighbouring Argo profiles. Together, they allow users to examine the reconstruction supplied by surface information and the contribution of local profile observations.

The resulting dataset covers 2011–2023 at 0.25° spatial spacing and three-hourly intervals over the North Atlantic and adjacent waters. It is intended to support analyses of seasonal, synoptic and subdaily upper-ocean variability using co-located estimates of layer depth and stratification. The output spacing and interval specify the sampling of the dataset; the variability actually represented must be established through validation.



This paper documents the input data, reconstruction procedure, dataset contents and conditions of use. Argo observations support comparisons across regions, seasons and mixed-layer depth ranges, including a temporal validation experiment in which 2023 is excluded from model fitting. Glider observations, which are not used to train Morlenn or supplied to its profile-assimilation branch, provide a complementary assessment along independently sampled trajectories. Comparisons with AR-MOR3D, GLORYS12 and OSnet place the MLD estimates in the context of existing products, accounting for differences in diagnostic definitions and observational dependence. Stratification scores, spectral analyses and comparisons of the diurnal cycle further assess the quality of the second diagnostic and the scales and amplitudes represented by the reconstruction. These evaluations establish the uses supported by the dataset and the limitations that remain for its interpretation.

2 Input data and preliminary study

2.1 Mixed-layer depth and subsurface stratification

In-situ MLD and stratification estimates are derived from ARGO profiles (Argo, 2000) following de Boyer Montégut et al. (2004) (https://mld.ifremer.fr/Data_Description.php). The MLD used here is the shallower of two MLDs based on threshold criteria: one based on a temperature change of 0.2°C , and one based on the density change (equivalent to a cooling of $\Delta T = 0.2^{\circ}\text{C}$ under $\Delta\rho \simeq \left(\frac{\partial\rho}{\partial T}\right)_{S,p} \Delta T$). Such a minimum MLD criterion identifies the depth over which we are sure that both temperature and density (and hence salinity) remain homogeneous to the extent of the threshold values, i.e. 0.2°C in temperature and, from about 0.06 kg m^{-3} (for $T \approx 25^{\circ}\text{C}$) to about 0.01 kg m^{-3} (for $T \approx 0^{\circ}\text{C}$) in density. This criterion deals with areas of barrier layer and vertically compensated layers (de Boyer Montégut et al. (2004)), while being smaller in density in high latitudes cold waters that are less stratified. The reference depth to compute temperature and density variations is set at 10 m.

The second diagnostic, N_{max}^2 , is the largest squared buoyancy frequency below the MLD and above 1000 m depth (and below the MLD if the latter is greater than 800 m), expressed in s^{-2} . It describes the strongest stratification below the MLD and is most of the time right below it (Sérazin et al. (2023)).

In the Argo database version used in this study, profiles were quality controlled, and MLD and N_{max}^2 were computed following de Boyer Montégut et al. (2004) over the period 2010–2024.

2.2 Surface fields and forcing histories

The sources of the surface fields and forcing histories are summarised below. The data are available from the indicated repositories, with the specified versions and access dates.

- Atmospheric and wave fields come from ERA5 Global Reanalysis: ERA5: Fifth generation of ECMWF atmospheric reanalyses of the global climate . Copernicus Climate Change Service Climate Data Store (CDS), accessed on 2025 August 15th (Hersbach et al., 2020, 2023).
- Sea Surface temperature (SST) from ESA CCI (Climate Change Initiative), version 3.0 (Good (2024))



90

- Sea Surface Salinity (SSS) comes from L4 product produced by CNR (Consiglio Nazionale delle Ricerche): "Multi Observation Global Ocean Sea Surface Salinity and Sea Surface Density. E.U. Copernicus Marine Service Information (CMEMS). Marine Data Store (MDS). DOI: <https://doi.org/10.48670/moi-00051> (Accessed on 2025 October 15th)"
- Sea-Level Height and anomalies comes from L4 altimetry derived reconstruction by CLS (Collecte Localisation Satellites): "Global Ocean Gridded L4 Sea Surface Heights And Derived Variables Reprocessed. E.U. Copernicus Marine Service Information (CMEMS). Marine Data Store (MDS). DOI: <https://doi.org/10.48670/moi-00148> (Accessed on 2025 October 15th)"

The input datasets and their sampling are summarised in Table 1.

Table 1. Input datasets and sampling of the fields supplied to the reconstruction.

Input family	Source	Sampling	Variables
Atmospheric forcing	ERA5 from CDS	Hourly, 0.25°	heat fluxes, radiation, 10 m neutral wind
Sea state	ERA5 from CDS	Hourly, 0.25°	mean wave and wind-wave heights and periods
Surface temperature	ESA CCI SST	Daily, 0.25°	sea surface temperature and its gradient
Surface salinity and density	CMEMS L4 (CNR)	Daily, 0.125°	sea surface salinity and its gradient
Sea level	CMEMS L4 altimetry	Daily, 0.125°	absolute dynamic topography, sea-level anomaly and its gradient
Location	Coordinates, bathymetry, calendar		longitude, latitude, day of the year
Bathymetry	GEBCO, version 2025	15 arc-sec (450 m)	elevation

95

The training dataset is computed as follows: At each profile location and time, data are linearly interpolated on the station point in time and space and include 30 days in the past for atmospheric and wave forcing and 10 days for ocean observations. Ten-day ocean histories were considered sufficient to represent recent changes within the surface state. Longer forcing histories are necessary for atmospheric and wave effects to account for the latency of the ocean.

100

The selection of the input variables and the duration of the forcing histories is based on physical understanding of the processes governing mixed-layer evolution, and on an exploratory correlation analysis at Argo profile locations. The following sections summarise this analysis and its results.

2.3 Physical basis and correlations of candidate predictors

105

Predictor selection combines understanding of the processes governing mixed-layer evolution with an exploratory correlation analysis at Argo profile locations. Surface buoyancy exchange, mechanical mixing and upper-ocean preconditioning define the physical input families. Even if the response of the mixed layer to forcings is strongly non-linear, correlations with MLD



and stratification characterise how these associations vary with season, their relative importance and the duration of preceding forcing.

Pearson correlations were calculated on collocated profiles over 2010–2024, using the instantaneous fields and retrospective averages of atmospheric and wave variables. The analysis retains profiles with positive MLD and finite values for the variables included in each calculation. Figures 1 and 2 show the signed instantaneous correlations with MLD and $N_{\max}^2$ by calendar month for the full sample. ERA5 fluxes follow the convention of positive downward (into the ocean) values, so a negative flux correlation with MLD indicates deeper mixed layers under stronger oceanic heat loss. The analysis uses the original variables, with their seasonal and geographic variability retained. The coefficients consequently reflect both local covariation and the distribution of oceanic regimes in the sample. To account for this dependence, correlations are also computed separately in four regions with contrasting regimes: the tropics, the Gulf Stream, the northern seas and the Mediterranean Sea. Correlations are calculated for each calendar month to describe their seasonality.

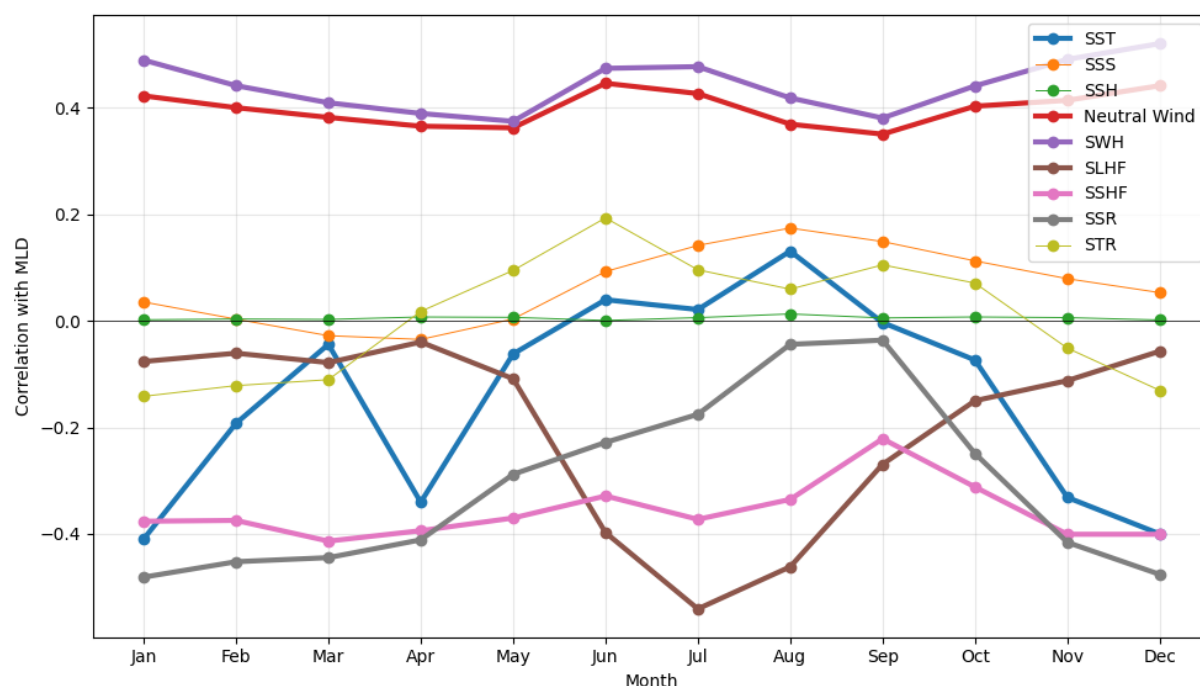


Figure 1. Signed Pearson correlations between collocated instantaneous predictors and MLD, by calendar month, for the full sample. SST, SSS and SSH denote sea surface temperature, salinity and height; Neutral wind is the wind norm at 10 m; SWH is significant wave height; SLHF (Surface Latent Heat Flux), SSHF (Surface Sensible Heat Flux), SSR (Surface Solar Radiation, short-wave) and STR (Surface Thermal Radiation, long-wave).

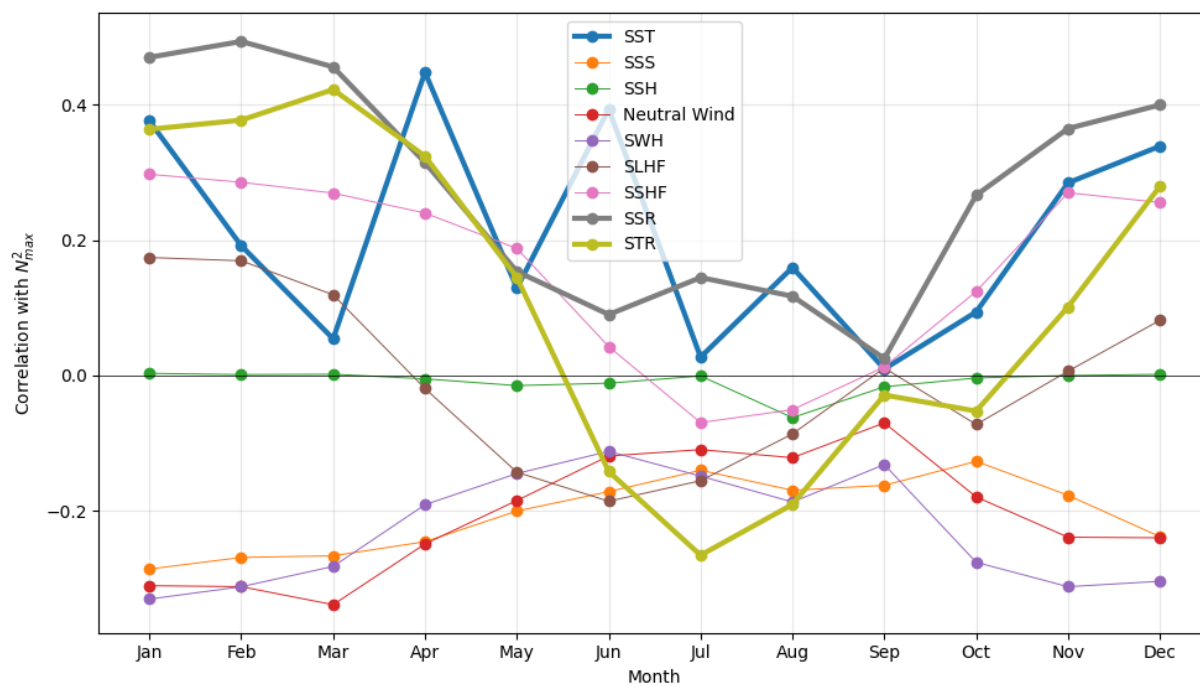


Figure 2. As Fig. 1, for the maximum subsurface squared buoyancy frequency N_{max}^2 .

2.3.1 Buoyancy forcing.

Surface cooling promotes convective mixing, while heating favours near-surface restratification. Sensible and latent heat fluxes (SSHF and SLHF) contribute to this thermal buoyancy forcing. Net short-wave and long-wave radiation (SSR and STR) provide its radiative components. Among these components, latent heat flux has the strongest association with MLD. In the pooled sample, this association is strongest in summer (Fig. 1). Its seasonal cycle differs between regions: it is strongest in boreal winter in the tropics, in late summer in the Gulf Stream, and in late autumn and winter in the northern seas. The pooled summer maximum therefore partly reflects the regional composition of the sample. For stratification, net short-wave radiation has the strongest association, mainly in winter and early spring (Fig. 2). The separate flux components are retained as inputs to describe the radiative cycle.

2.3.2 Mechanical forcing.

Wind and waves contribute information about mechanically driven mixing and its history (Belcher et al., 2012; Li et al., 2017). Significant wave height is the predictor most strongly associated with MLD throughout the year, ahead of neutral wind speed (Fig. 1). Both are associated with weaker stratification, mainly in winter (Fig. 2). The selected family includes neutral wind speed, total significant wave height and mean period, and wind-wave height and period. Wave height and period carry information about the development, persistence and propagation of the wave field. Sea state thus carries information about



preceding and remotely generated forcing alongside the local instantaneous wind. The period and wind-wave descriptors are not shown in Figs. 1–2.

135 2.3.3 Surface state

SST, salinity and surface density constrain the thermohaline surface state. Sea level provides complementary information associated with circulation and vertically integrated density structure. In the pooled sample, SST is the surface variable most strongly associated with MLD, mainly in winter, when colder surface waters coincide with deeper mixed layers (Fig. 1). SST is also associated with stronger stratification, although this relationship varies from month to month (Fig. 2). Salinity and sea level have weaker pooled associations, but both become leading predictors in several regions, salinity in particular for stratification. These results motivate their joint use as indicators of preconditioning. Geography, bathymetry and season identify the environmental setting in which the forcing acts.

Horizontal gradients of SST, salinity and sea-level anomaly extend this physical description to fronts and mesoscale structure. Atmospheric pressure contributes information on the meteorological setting. These descriptors are physically motivated additions to the correlated input families; the exploratory plots quantify marginal associations for a subset of the final predictors. The selected variables are defined in Appendix A.

Regional monthly correlations differ substantially from the pooled ones. For MLD, latent heat flux (-0.39 to -0.72), neutral wind and salinity have the strongest associations in the tropical subset, with maxima in boreal winter. In the Gulf Stream subset, the sea-level and salinity correlations increase from spring to autumn, reaching about 0.5 in November, while the latent heat-flux correlation peaks in September. In the northern-seas subset, SST and sea level dominate in winter (about 0.55 in February) and wind and waves from June to September. Wind and wave correlations in the Mediterranean peak in May and July. For stratification, salinity is the strongest regional predictor: its correlation stays near -0.5 throughout the year in the tropics, reaches -0.6 in late summer in the northern seas and in late winter in the Mediterranean. In the Gulf Stream, salinity, sea level and SST are all negatively correlated with stratification in summer, down to about -0.45 in August. Pooling regimes therefore masks strong regional relationships, notably the haline control of stratification.

2.4 Correlations with preceding forcing over different durations

For each profile, the analysis averages the preceding values of a forcing variable over a chosen time window. It then calculates the Pearson correlation between these averages and the profile diagnostic across the sample. Repeating the calculation with windows from 1 to 24 days shows how the association changes with the duration of preceding forcing. Figure 3 compares winter (December–February) and summer (June–August).

For MLD, neutral wind and wave height have the strongest correlations in both seasons. In winter, their correlations increase with averaging duration and reach a plateau after about ten days. In summer, they peak after a few days and then decline gradually. Among the heat-flux terms, short-wave radiation and sensible heat flux have the strongest correlations in winter, and latent heat flux in summer. These correlations strengthen mainly over the first few days.



165 The stratification curves show a different organisation. Neutral wind and wave height remain associated with weaker stratification, and their correlations strengthen over most of the 24-day range. This increase is steady in summer, whereas in winter it slows after about two weeks. In winter, short-wave radiation has the strongest heat-flux correlation, which saturates after a few days. Heat-flux correlations with stratification are weak in summer. The averaging durations associated with the strongest correlations therefore differ between MLD and stratification and between seasons.

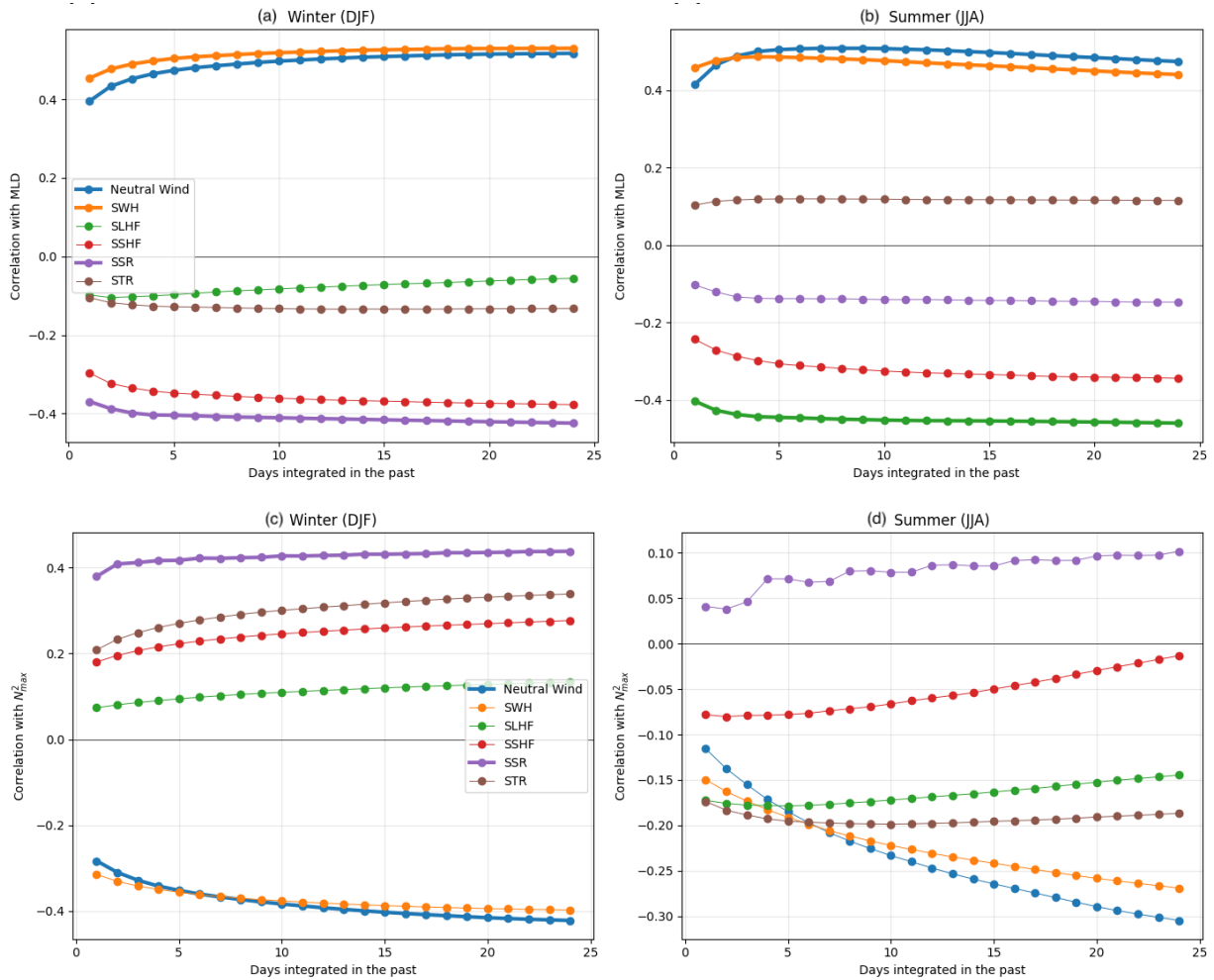


Figure 3. Signed correlations between MLD (a, b) or $N_{\max}^2$ (c, d) and forcing averaged over the preceding 1–24 days, for winter (December–February, a and c) and summer (June–August, b and d). Variable names follow Fig. 1.

170 Several MLD correlation curves flatten after one to two weeks, while the wind and wave correlations with stratification continue to strengthen over the sampled interval. The analysis therefore does not identify a single duration applicable to both diagnostics in all regimes, and supports retaining multiweek forcing histories. The selected 30-day atmospheric and wave histories encompass these daily-to-multiweek associations and the persistence of mixed-layer responses.



These marginal correlations also reflect covariance among predictors, their persistence and regional sampling. They support
 175 the physically motivated selection of input families and history lengths without isolating individual processes.

3 Physical motivation and reconstruction method

3.1 Physical basis and network formulation

MLD is diagnosed from the temperature and density profiles of the water column. For the density threshold, the increase from
 the reference depth is the integral of the vertical density gradient. With depth d positive downwards, a Boussinesq approxima-
 180 tion using potential density gives $\int_{d_{\text{ref}}}^{h_\rho} N^2(d, t) dd \simeq g\Delta\rho/\rho_0$. Here h_ρ is the first depth at which the density increase reaches
 $\Delta\rho$, g is gravitational acceleration and ρ_0 is a constant reference density. The approximation depends on the thermodynamic
 convention (IOC, SCOR and IAPSO, 2010). Morlenn’s MLD also includes the temperature criterion defined in Sect. 2. The
 maximum N_{max}^2 describes the strongest stratification below the layer; it does not specify the full vertical profile.

Surface forcing affects mixed-layer properties through successive processes, including wave development, turbulent mixing
 185 and entrainment. Each process responds to the preceding stages over its own timescales, with a dependence on the geophysical
 setting. A schematic representation of this accumulated response is a cascade of nonlinear integral operators:

$$(h, N_{\text{max}}^2)(t) \simeq \mathcal{G}([\mathcal{A}_{L,\Theta} \circ \dots \circ \mathcal{A}_{1,\Theta}]F](t), \Theta_t), \quad (\mathcal{A}_{k,\Theta}u)(t) = \mathcal{R}_k \left(\int_{t' < t} K_k(t, t'; \Theta_t) u(t') dt' \right). \quad (1)$$

Here F denotes the surface forcing and Θ_t the geophysical setting. At stage k , the kernel K_k weights the preceding history
 of the incoming signal u , and $\mathcal{R}_k$ describes a nonlinear response. The first stage acts on the forcing; each subsequent stage
 190 integrates the response produced by the preceding stage. The estimator $\mathcal{G}$ combines the resulting response with the geophysical
 setting to obtain MLD and stratification.

Morlenn uses a learned, finite-history version of this cascade. Successive temporal filters and nonlinear transformations form
 the operators $\mathcal{A}_{k,\theta,X_t}$. The context X_t combines the geophysical setting with contemporary surface fields and season, provid-
 ing information about the oceanic regime and preconditioning. This context modulates the temporal operators. A temporal
 195 summary P combines their outputs before the final estimator G_θ . The estimate using surface inputs alone is

$$(\widehat{h}, \widehat{N_{\text{max}}^2})(t) = G_\theta(P_{\theta,X_t}[(\mathcal{A}_{L,\theta,X_t} \circ \dots \circ \mathcal{A}_{1,\theta,X_t})F_{t-T(F):t}], X_t). \quad (2)$$

$F_{t-T(F):t}$ denotes the input histories, with duration $T(F)$ depending on the variable: 30 days for atmospheric and wave forcing,
 and 10 days for SST and salinity. X_t contains the contemporary surface fields, position, bathymetry and season. The L filtering
 stages $\mathcal{A}$ extract temporal features; P combines them over time; and G converts the summary and context into MLD and
 200 stratification estimates. The parameters θ are learned from the training profiles. The cascade in Eq. (1) motivates the successive
 temporal layers. Conditioning these layers on X_t allows their response to depend on both the geophysical setting and the
 observed ocean state.

The input variables and their preparation are detailed in Appendix A.



3.2 Pointwise reconstruction

205 Morlenn predicts each location and time separately. It uses the local input history and current context, with neighbouring profiles added in the assimilating configuration. Convolutions operate along time. Coordinates and surface gradients supply spatial information; a horizontal image patch is not required.

Training pairs individual profiles with input fields sampled at their locations and times, without first mapping the targets onto a grid. This preserves the observed profile diagnostics as the learning targets and avoids introducing a separate spatial
 210 interpolation of those targets. Predictions can be made at observation locations, along trajectories or on a chosen grid. Other observational products or numerical models can supply the inputs if their variables, units, histories and scaling are prepared consistently. The GLORYS-input experiment in Appendix E tests this possibility. Skill must be reassessed when the input source or ocean regime changes.

3.3 A compact response model

215 The architecture combines a representation of accumulated forcing with information about the ocean conditions in which it acts (Fig. 4). The blue blocks represent the cascade of temporal responses in Eq. (1); the green block describes the ocean state and geophysical setting; and the orange block combines them to estimate MLD and stratification. The red branch adds neighbouring Argo observations. MLD and stratification have separate temporal branches because their associations with forcing and history length differ (Sect. 2.4). Each branch predicts a central value and a spread, the width of the predicted distribution after target
 220 transformation and scaling. Tables A1 and A2 give the technical settings; Appendix A defines the inputs.

A Accumulated response to forcing. Mixed-layer depth and subsurface stratification depend on how forcing has acted through successive stages of mixing and entrainment. A temporal convolutional network (TCN) (Bai et al., 2018) represents this dependence through a cascade of dilated temporal convolutions and nonlinear transformations. Residual connections (He et al., 2015) preserve information between successive layers. Each layer combines the preceding history of the signals produced by
 225 the previous layer. In discrete time, these weighted sums provide the counterpart of the integral operators in Eq. (1). Increasing the spacing between sampled times through the layers extends their temporal reach, allowing the representation to include brief events and sustained forcing. The layers combine multiple variables, so winds, waves, heat exchange and changes in surface properties can contribute jointly. Individual layers are not assigned to specific physical processes.

P Adaptive weighting of the history. The influence of a past event depends on its intensity, its timing and the ocean conditions.
 230 Context-conditioned additive attention (Bahdanau et al., 2014) makes the temporal summary depend on these conditions: it calculates weights from the filtered history and the context, normalises them across time, and uses them to combine the temporal signals. The weights therefore change between predictions. This allows a recent storm and a prolonged cooling episode to receive different emphasis as the history and ocean state change.

Three attention modules process the temporal signals with different filter widths, providing complementary summaries.
 235 Their filters act on signals already transformed by the TCN, providing summaries at different temporal scales. A second set of learned weights combines the three summaries using the context and the mean, maximum and final value of the temporal

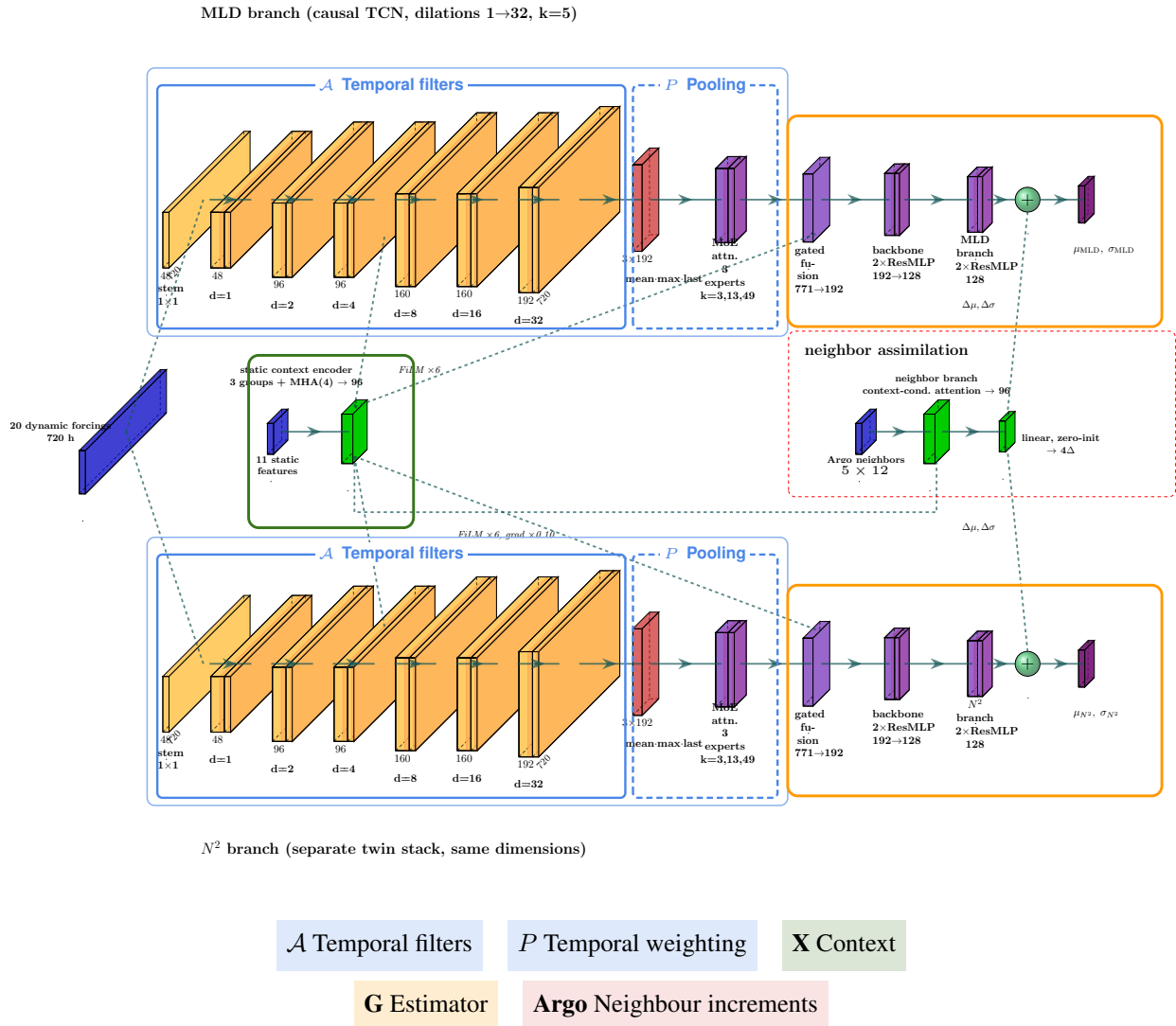


Figure 4. Network architecture. Within each blue memory group, solid sub-blocks identify temporal filtering ($\mathcal{A}$), and dashed sub-blocks identify time averaging with learned weights and other temporal summaries ($\mathcal{P}$). Green denotes the processing of surface state, location and season ($\mathbf{X}$); orange combines this information with the history to produce the estimates ($\mathbf{G}$). The red dashed branch supplies corrections based on neighbouring Argo profiles. Each of the five neighbour slots contains MLD and 11 context variables.



signals. These three statistics also enter the estimator directly. Attention thus supplies an adaptive summary alongside measures of average conditions, peaks and the end of the history. Its weights describe the use of learned signals and do not directly quantify contributions from individual physical processes.

240 **X Geophysical regime and preceding ocean state.** The context combines location, bathymetry, season and contemporary surface-ocean fields. Position identifies regional regimes, and latitude supplies information about the local Coriolis parameter and associated rotational timescales. Bathymetry describes depth constraints, while season locates the prediction within the annual cycle of mixing and restratification. These descriptors allow the response to a given forcing history to vary across oceanic regimes.

245 Contemporary SST, salinity and sea level also provide partial information about ocean preconditioning. Their values can reflect the accumulated effects of forcing and circulation before the explicit input histories begin, including atmospheric forcing older than 30 days. They therefore help distinguish ocean states with similar recent forcing but different earlier evolution. This information is limited by the surface observations and does not fully specify the subsurface state. The term static refers to a single context value per input at each prediction; the surface fields and seasonal descriptors change between predictions.

250 The context encoder processes surface-ocean properties, geography and season as three groups, combining them through multi-head self-attention (Vaswani et al., 2017). The resulting representation influences both the temporal transformations and their final weighting. Feature-wise linear modulation (FiLM) (Perez et al., 2018) applies context-dependent multiplicative factors and additive offsets to the TCN features. The context also enters the attention scores and the weights combining the three summaries. The model can consequently adjust both its representation of the forcing response and the relative importance
 255 of past episodes to the local ocean conditions.

G Joint use of history and context. A gated fusion layer combines the temporal summary and context, learning their relative contributions to the combined representation. Residual fully connected layers then produce the estimates for each diagnostic. Separate output branches allow MLD and stratification to respond differently while sharing information about the geophysical regime. Predicted MLD is limited to the local water depth.

260 3.4 Learning depth and stratification as two targets

MLD and stratification are learned jointly from the reference profiles, with separate error terms to account for their different units and ranges.

MLD training combines errors after a logarithmic transformation with errors measured in meters. The transformation reduces the numerical contrast between shallow and deep layers. Errors in meters retain sensitivity to the magnitude of deep-layer
 265 errors. The Huber loss uses squared errors for small departures and grows linearly for larger departures, limiting the influence of individual extreme errors. Underestimated depths receive extra weight in this term. An additional root-mean-square error (RMSE) term emphasises the deepest reference layers in each training batch. These choices give more weight to the sparsely sampled deep layers. Other terms compare the standard deviation and correlation of predictions and observations within these



samples, encouraging the reconstruction to retain the observed variability and covariation. Stratification has its own error,
standard-deviation and correlation terms. Appendix A gives the coefficients.

A Gaussian likelihood term jointly adjusts the central value and spread for each transformed and scaled target. It penalises both errors that are large relative to the predicted spread and unnecessarily broad distributions, allowing the spread to vary with the inputs. Calibration of the resulting intervals is discussed in Sect. 5.

3.5 Reconstructions with and without neighbouring profiles

Argo Corrections from neighbouring profiles.

Surface fields provide incomplete information about subsurface structure. Neighbouring Argo profiles add direct MLD observations, whose relevance depends on both spatiotemporal proximity and oceanic regime. Across a front, or between an eddy and the surrounding water, nearby profiles can sample different mixed-layer conditions. The model using surface inputs alone (baseline) and the model incorporating these profiles are trained separately.

Up to five candidate profiles are selected by proximity in space and time. Each supplies its measured MLD and the same 11 context descriptors used at the prediction point: surface-ocean properties, position, bathymetry and season. The neighbour branch encodes these measurements and descriptors, then applies additive attention conditioned on the context of the target point (Bahdanau et al., 2014). Normalised scores determine the relative weight of each profile. This weighting allows the model to reduce the influence of a nearby profile whose oceanic conditions differ from those at the target, and to favour observations from a more similar regime.

Coordinates and day of year describe the relative position and seasonal timing of each neighbour. Surface temperature and salinity provide information about contrasts between water masses, while sea level and bathymetry help characterise the circulation and local setting. Together, these descriptors allow attention to account for differences associated with fronts, eddies and other regional structures when weighting neighbouring observations. The weighted representation produces corrections to the predicted central values and spreads. Since the observed diagnostic supplied by this branch is MLD, its influence on stratification is indirect.

Neighbour information is randomly withheld during training so that the model also learns to reconstruct conditions without these observations. Profiles are selected both before and after the prediction time, providing a retrospective reconstruction. Differences between the two configurations reflect both neighbour information and separate training.

3.6 Training, transformation and model selection

MLD is represented through a $\log(1+h)$ transform after limiting training depths to 2000 m. Each input variable and transformed target is scaled by subtracting its median and dividing by its interquartile range, the difference between its 75th and 25th percentiles. Scaling puts variables with different units on comparable numerical ranges; the median and interquartile range reduce the influence of extreme values on these reference scales.



To assess the effect of including 2023 observations in training, both configurations were trained with and without that year. The versions excluding 2023 are identified as `baseline` and `assim`; those trained on all available years are `baseline_prod` and `assim_prod` and are used to generate Morlenn. The 2023 Argo comparison evaluates the sensitivity of the predictions to this retraining. Argo observations also contribute to selecting saved model versions and to the sets of neighbouring profiles. Glider profiles excluded from training and neighbour inputs provide an independent evaluation of the production models.

4 Dataset description

Morlenn is organised into daily NetCDF files with the coverage and variables summarised in Table 2. Quality information includes bathymetry, input-validity and assimilation masks, neighbour diagnostics, the fraction of interpolated inputs. These fields describe where the reconstruction is supported by definite surface inputs, neighbouring observations and temporal interpolation. MLD is capped at local ocean depth. Two additional flags (low salinity and clamped to bathymetry) have been included to filter unreliable $N_{\max}^2$. Data have been masked using AMSR2 L3 sea ice concentration maps (Spreen et al. (2008)), with a threshold of 15% to remove ice-covered regions.

Two datasets are provided, one using surface inputs alone (refer to as **baseline**) and one additionally incorporating neighbouring Argo observations (refer to as **assimilation**). The assimilation configuration is expected to provide more accurate estimates in regions with sufficient Argo coverage, while the baseline configuration provides a more uniform product performance across the domain.

Table 2. Scientific content and nominal sampling of Morlenn.

Property	Description
Period	2011–2023
Domain	North Atlantic, nominal study domain 10°S–80°N, 100°W–44°E
Sampling	0.25° grid; three-hourly outputs
MLD	Mixed temperature–density threshold depth, in metres
Stratification	Maximum squared buoyancy frequency in the reference below-MLD interval, in s^{-2}
Configurations	Surface inputs alone (baseline); surface inputs and neighbouring Argo profiles (assimilation)
File format	NetCDF fields <code>mld</code> , <code>mld_std</code> , <code>N2</code> , <code>N2_std</code> , with time and horizontal coordinates

4.1 Joint depth and stratification fields

Joint MLD and $N_{\max}^2$ fields distinguish the depth of the surface layer from the strength of the stratification beneath it. Layers of similar depth can overlie different stratification maxima, so MLD alone does not describe their resistance to further mixing. The paired fields provide complementary information about this vertical structure.



320 Figure 5 shows winter and summer fields of the baseline configuration. Deep winter mixed layers occur in parts of the northern domain. The stratification panels show stronger maxima in parts of the tropical domain than across much of the northern open ocean, with regional contrasts between the two dates.

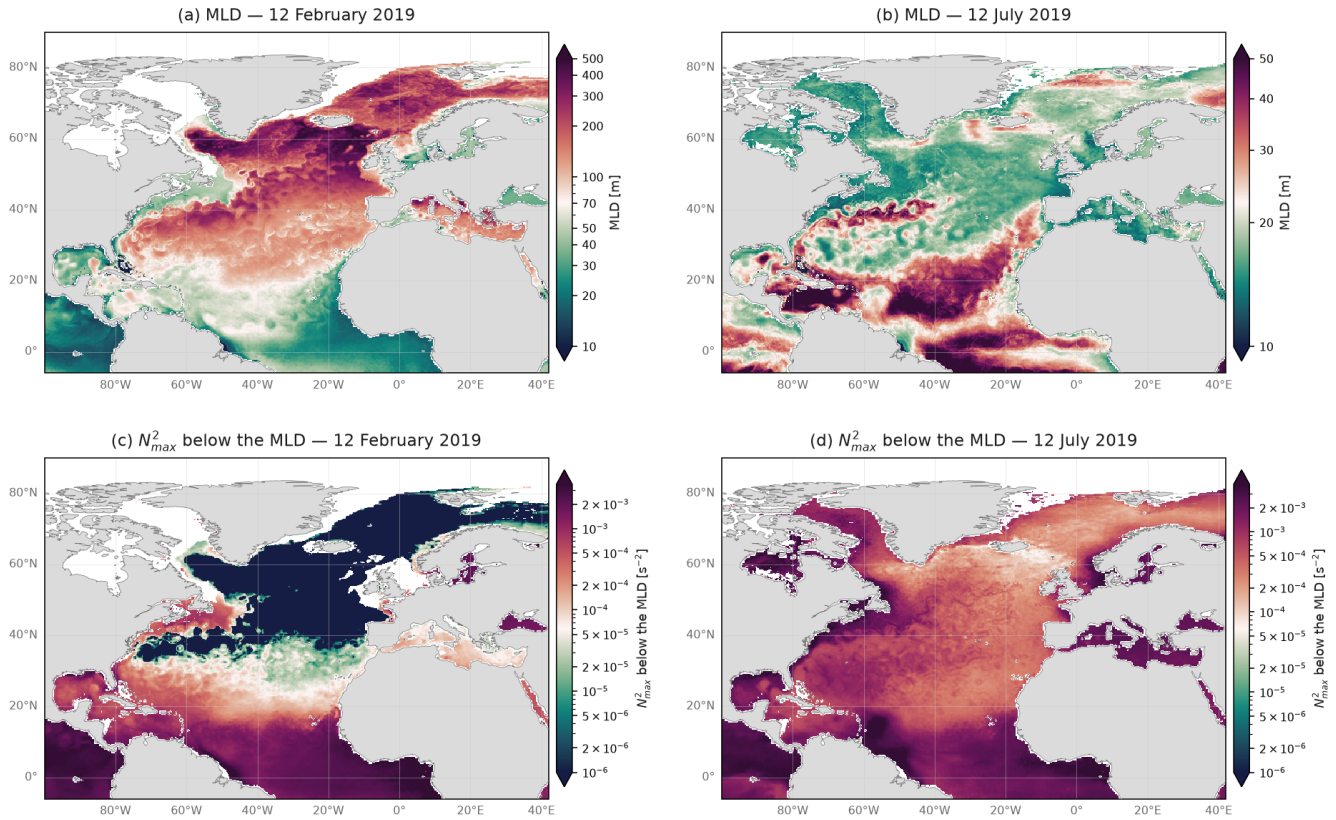


Figure 5. Baseline MLD (a, b) and $N^2_{\max}$ (c, d) on 12 February (a, c) and 12 July 2019 (b, d). MLD is expressed in metres and stratification in s^{-2} . Each panel uses its own colour scale.

Additional seasonal comparisons for assimilation, ARMOR3D, GLORYS and OSnet, together with paired synoptic MLD– $N^2_{\max}$ fields, are shown in Appendix D.

325 5 Technical validation

5.1 Evaluation design and observational independence

Two complementary in-situ observation datasets are used to evaluate Morlenn: Argo floats and autonomous gliders. The Argo comparison uses the broadest geographic coverage and the largest number of stations, while the glider comparison provides an independent test of generalisation and subdaily variability. The evaluation protocol is summarised in Table 3.



330 Argo provides the principal observational basis for evaluating performance across the North Atlantic. The full-period comparison characterises MLD errors across the sampled regions and depth regimes. Gliders provide a complementary independent test of MLD generalisation because their profiles are not used by any product in the benchmark. Their frequent profiles also allow analysis of subdaily variability, but cover fewer regions.

Morlenn is compared with three existing mapped MLD products: ARMOR3D, GLORYS12 (hereafter GLORYS) and, within
 335 its Gulf Stream domain, OSnet. These products use different approaches to reconstruct the upper ocean.

ARMOR3D combines satellite surface fields with in-situ temperature and salinity profiles (Guinehut et al., 2012). Surface temperature, salinity and sea-level anomalies are first projected onto subsurface temperature and salinity through multiple linear regression. The resulting synthetic profiles are then combined with in-situ profiles by optimal interpolation (Greiner et al., 2025). MLD is diagnosed from the reconstructed profiles. It is computed under the same criterium as Morlenn (de Boyer Montégut et al. (2004)).
 340

GLORYS is a global ocean and sea-ice reanalysis based on the NEMO model, driven by atmospheric reanalysis forcing (Lellouche et al., 2021). It assimilates satellite sea-level anomalies, sea surface temperature and sea-ice concentration, together with in-situ temperature and salinity profiles. The assimilation uses a reduced-order Kalman filter, supplemented by a variational correction of large-scale temperature and salinity biases. Its upper-ocean structure therefore reflects both model dynamics and
 345 observational constraints.

OSnet is a regional neural reconstruction of temperature, salinity and MLD in the Gulf Stream (Pauthenet et al., 2022). It is trained on historical in-situ profiles, including Argo. Its inputs include satellite sea surface temperature and sea-level anomalies, mean dynamic topography, bathymetry, position and time. The network jointly predicts temperature and salinity profiles and a mixed-layer mask. This mask is then used to adjust the vertical gradients of the predicted profiles and improve their consistency
 350 with the predicted mixed layer.

All three comparison products use Argo observations, either to constrain the reconstructed fields or to train the reconstruction. Both Morlenn configurations also learn from Argo, and the assimilating configuration additionally uses neighbouring profiles during inference. Argo-based evaluation therefore measures agreement with the shared observing network under different methods of using it.

355 Scoring uses stations valid for all compared products, so differences in geographical sampling do not determine the ranking within a comparison. Each product is evaluated against its corresponding reference criterion. The benchmark currently specifies a mixed temperature–density criterion (de Boyer Montégut et al. (2004)) for Morlenn, a 0.01 kg m^{-3} density criterion for GLORYS, and a 0.03 kg m^{-3} density criterion for OSnet. ARMOR3D’s documented criterion uses the temperature-equivalent density threshold described above. For each product, the reference MLD is then recomputed from the in-situ profiles with the
 360 same criterion, in the tables as in the maps. The criteria are summarised in Table 3.

The common Argo sample contains 374 746 stations for 2011–2023. The glider comparison contains 41 349 daily samples, called glider-day stations, derived from 896 587 profiles and 279 gliders. Morlenn fields are sampled at profile locations and times before daily aggregation. This reduces the influence of differing within-day profiling rates; deployment duration continues to affect the number of daily samples. Figure 6 shows the geographic support of both reference datasets.



Table 3. Sampling, reconstruction methods and MLD criteria of the evaluated product versions.

Product	Grid	Time step	Method and inputs	MLD criterion
Morlenn baseline	1/4°	3 h	Neural reconstruction from surface fields and forcing histories	Temperature–density
Morlenn assimilation	1/4°	3 h	Baseline input families plus neighbouring Argo profiles	Temperature–density
ARMOR3D	1/8°	Daily	Statistical projection of surface fields followed by optimal interpolation with in-situ profiles	Temperature–density
GLORYS	1/12°	Daily	Ocean-model reanalysis with atmospheric forcing and assimilated observations	Density
OSnet	1/4°	Daily	Regional neural reconstruction from surface fields and geographic context, followed by mixed-layer adjustment	Density

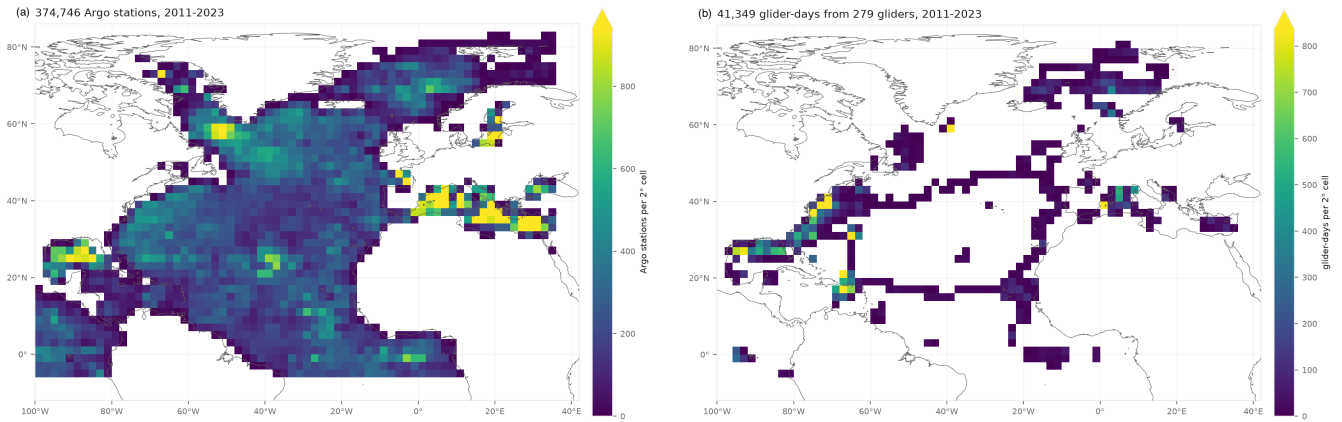


Figure 6. Validation locations: number of Argo stations (a) and glider-day stations (b) per 2° cell in the multi-year comparison, 2011–2023. The two platforms sample different geographic distributions.

365 5.2 Metrics

Errors are calculated as the reconstructed value minus the observed value, using the samples valid for all compared products. For N samples with reconstructed values m_i and observed values o_i , the error is $e_i = m_i - o_i$, and overbars denote means over the samples. Bias is the mean error; positive bias indicates overestimation. Mean absolute error (MAE) averages the absolute errors. Root-mean-square error (RMSE) gives greater weight to large errors. Centred RMSE (cRMSE) removes the mean bias before calculating RMSE:

$$\text{bias} = \frac{1}{N} \sum_{i=1}^N e_i, \quad \text{MAE} = \frac{1}{N} \sum_{i=1}^N |e_i|, \quad \text{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^N e_i^2}, \quad \text{cRMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^N (e_i - \text{bias})^2}. \quad (3)$$

It follows that $\text{RMSE}^2 = \text{bias}^2 + \text{cRMSE}^2$.



Pearson correlation measures how closely the reconstructed and observed variations follow one another. Relative RMSE is RMSE divided by the reference mean, expressed as a percentage. Nash–Sutcliffe efficiency (EV in the tables) is 100 times one minus the ratio of summed squared errors to the summed squared departures of observations from their mean:

$$r = \frac{\sum_i (m_i - \bar{m})(o_i - \bar{o})}{\sqrt{\sum_i (m_i - \bar{m})^2 \sum_i (o_i - \bar{o})^2}}, \quad \text{RMSE\%} = 100 \frac{\text{RMSE}}{\bar{o}}, \quad \text{EV} = 100 \left(1 - \frac{\sum_i e_i^2}{\sum_i (o_i - \bar{o})^2} \right). \quad (4)$$

A perfect reconstruction scores 100%; predicting the reference mean scores zero. Negative values indicate larger squared errors than that constant prediction. Product-specific MLD definitions can change the reference mean and variance. Correlation describes covariation and can remain high when amplitudes are inaccurate. The depth-class table reports median error; the aggregate tables report mean bias.

5.3 Evaluation of maximum sub-mixed-layer stratification N_{max}^2

Stratification is evaluated against its Argo-derived reference diagnostic (Table 4). As no other mapped product provides a comparable stratification diagnostic, the evaluation is limited to Morlenn. No specific tuning of hyperparameters was applied to stratification, as it was not initially the main variable of interest. The evaluation therefore provides a preliminary assessment of the model’s ability to predict this variable from surface inputs and forcing history.

Table 4. Preliminary aggregate Argo-based N_{max}^2 scores. Bias, RMSE, cRMSE and MAE are in s^{-2} . Stations with surface salinity below 25 are excluded as unreliable.

Product	bias	RMSE	RMSE%	cRMSE	MAE	r	EV%
Morlenn assim	-4.30×10^{-5}	8.84×10^{-4}	87	8.83×10^{-4}	4.33×10^{-4}	0.729	52
Morlenn baseline	-4.97×10^{-5}	8.81×10^{-4}	87	8.80×10^{-4}	4.25×10^{-4}	0.730	53

Both configurations have similar scores, with RMSE near $8.8 \times 10^{-4} \text{ s}^{-2}$, correlation of 0.73 and efficiency of 52–53%. The bias is small compared with RMSE, and cRMSE is almost equal to RMSE: errors are dominated by scatter rather than by a systematic offset. MAE is about half the RMSE, indicating a substantial contribution from large departures. Relative RMSE reaches 87% of the reference mean, much higher than for MLD, so stratification is less well constrained by the surface inputs than the layer depth.

The spatial distribution of the relative RMSE is shown in Fig. 7. Relative errors are lowest in the tropical band, in the equatorial Atlantic and the eastern tropical Pacific, where a strong and persistent thermocline maintains high stratification below the mixed layer. They are highest in the subpolar gyre, the Labrador and Irminger seas and the Nordic seas, where stratification is weak and eroded by deep winter mixing, so that a given absolute error represents a larger fraction of the mean. Relative errors are also high near the Amazon and Orinoco plumes and along the northwest African upwelling, where stratification depends on processes weakly constrained by the surface inputs. The subtropical gyre and the Mediterranean Sea have intermediate values. The two configurations show the same pattern; assimilation slightly reduces relative errors in the subpolar gyre and the Nordic seas.

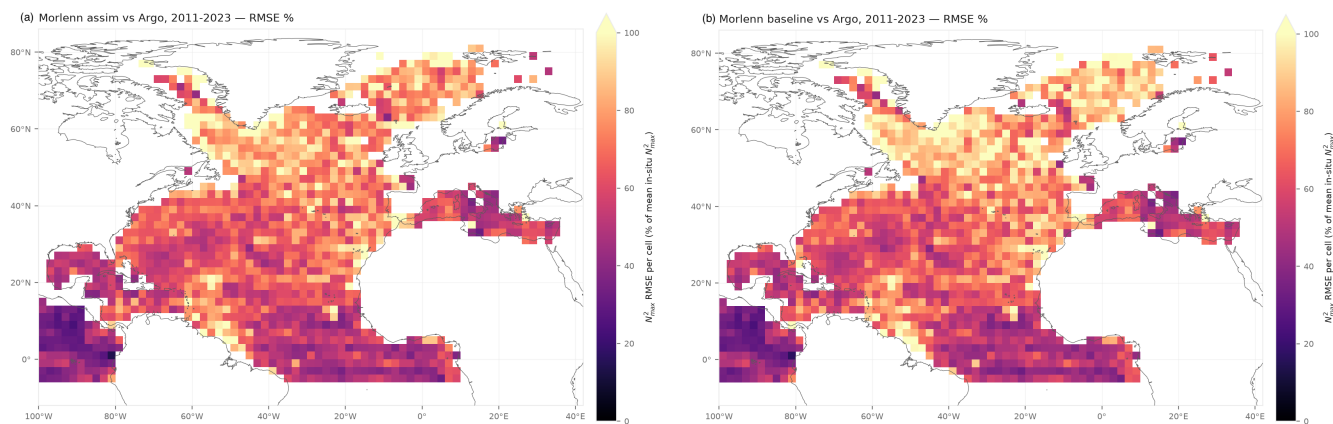


Figure 7. Relative RMSE of $N_{\max}^2$ against Argo per 2° cell, 2011–2023, expressed as a percentage of the mean in-situ $N_{\max}^2$ in the cell, for the assimilating (a) and baseline (b) configurations. The colour scale spans 0–100%, with larger values assigned the over-range colour. Only cells with at least 30 stations are shown.

5.4 MLD evaluation against ARGO over 2011–2023

400 The production models are trained over the whole evaluation period. However, models trained without 2023 score almost the same on that year as models trained with it (Appendix B). The Argo validation therefore uses the full 2011–2023 period, which provides the most extensive observational coverage of the domain.

On the 374 746 common stations, the assimilating configuration has the lowest RMSE (38.08 m), the highest correlation (0.871) and the highest efficiency (73%) (Table 5). Baseline follows with RMSE of 43.41 m and correlation of 0.824, close
 405 to ARMOR3D (44.96 m and 0.817). GLORYS has the largest errors, with RMSE of 55.81 m and efficiency of 41%. Both Morlenn configurations overestimate MLD on average, with bias of +5.95 and +9.68 m, whereas ARMOR3D and GLORYS underestimate it. ARMOR3D has the smallest bias magnitude and the lowest MAE, 14.15 m versus 17.13 m for assimilation. It is therefore more accurate for typical layers, whereas Morlenn has fewer large departures. The ranking depends on the weighting of large errors relative to typical absolute error. The good ARMOR3D scores against Argo are expected: its fields are
 410 built by optimal interpolation of in-situ profiles, in which Argo has a large weight, so the reconstruction is strongly constrained at and near the evaluation stations. The comparison against independent gliders, where ARMOR3D errors are much larger (Sect. 5.5), shows that this advantage does not extend away from the assimilated profiles.

Geographic aggregation shows where the basin-wide statistics are representative. The full-period Argo inventory contains 379 658 stations; the common valid four-product sample used for aggregate scores contains 374 746 stations, distributed un-
 415 evenly across the domain (Fig. 6a). Broad open-ocean coverage contrasts with locally concentrated sampling in marginal seas, the Gulf of Mexico and parts of the subpolar North Atlantic. The overall scores therefore give more weight to densely sampled locations.



Table 5. Gridded MLD products scored against ARGO stations, 2011–2023. All rows are scored on the same 374 746 stations. Each product uses its reference MLD criterion. Dimensional scores are in metres. Metrics are defined in section 5.2.

Product	bias	RMSE	RMSE%	cRMSE	MAE	r	EV%
Morlenn assim	+5.95	38.08	69	37.61	17.13	0.871	73
Morlenn baseline	+9.68	43.41	79	42.32	19.89	0.824	65
ARMOR3D	-3.77	44.96	82	44.81	14.15	0.817	62
GLORYS*	-8.06	55.81	102	55.23	21.37	0.753	41

Figure 8 shows where these differences arise, as the difference in RMSE and correlation between assimilation and each comparison product in 2° cells. The differences with ARMOR3D are patchy, with neighbouring cells often changing sign, particularly in the subpolar gyre, the Nordic seas and the Mediterranean Sea. This patchiness is consistent with the dependence of ARMOR3D on in-situ profiles: it is often closer to Argo in densely sampled regions such as the Mediterranean Sea and the Gulf Stream, and differences vary irregularly where sampling is sparse, as in the Nordic seas (Fig. 6a). The densely sampled Labrador Sea is an exception, where Morlenn remains closer to Argo, consistent with the underestimation of deep convective layers by ARMOR3D (Table 7). Against ARMOR3D, Morlenn has lower errors and higher correlations in the subpolar North Atlantic, the Labrador and Irminger seas and parts of the Nordic seas, where deep winter mixing dominates MLD variability. ARMOR3D has lower errors in the Gulf Stream and its extension and, to a lesser extent, over the subtropical gyre. It also has higher correlations over the subtropical gyre, the Caribbean and the equatorial Atlantic. Against GLORYS, Morlenn has lower errors and higher correlations over almost the whole domain, with the largest error reductions in the subpolar gyre, the Nordic seas and the western equatorial Atlantic; only a few cells, mainly along the Gulf Stream, favour GLORYS. As in the tables, each product is scored against a reference MLD computed with its own criterion.

5.5 MLD Evaluation against independent glider profiles

Glider profiles are not used by any product in the benchmark, neither in Morlenn’s training labels and neighbour inputs nor in the reconstruction of the comparison products, so they provide an independent evaluation of all products and direct evidence of generalisation beyond the Argo observations used in training and assimilation (Table 6). They do not, however, cover the whole domain (Fig. 6). Most glider-days are concentrated along the North American shelf, in the Gulf of Mexico, the Caribbean, the western Mediterranean Sea and the Nordic seas, while the open-ocean interior is sampled only by a few transects. The glider scores are therefore representative of these regions rather than of the whole basin. Baseline and assimilation have RMSE of 39.43 and 39.84 m, compared with 52.19 m for ARMOR3D and 51.83 m for GLORYS under the criterion-specific protocol. Morlenn’s RMSE is approximately 23–24% lower, with efficiencies of 61–62% compared with 33–34%. This ranking differs from the Argo comparison, where ARMOR3D had the lowest MAE (Sect. 5.4) and the smallest errors for shallow layers (Sect. 5.6). Away from the assimilated in-situ profiles, ARMOR3D relies on its statistical projection of surface fields and its skill drops: its RMSE increases from 44.96 m against Argo to 52.19 m against gliders, and its efficiency from 62 to 33%.

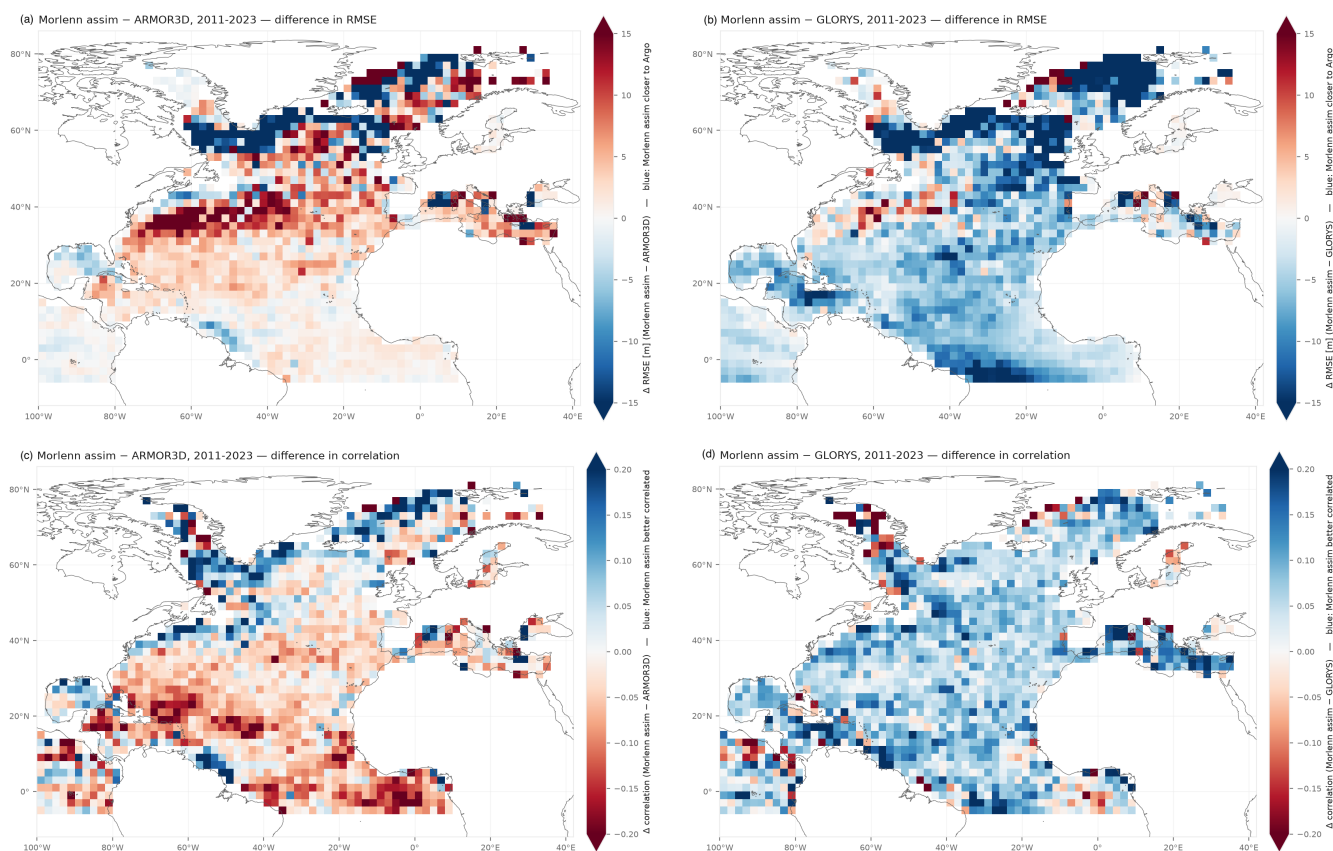


Figure 8. Differences in MLD scores against Argo between assimilating Morlenn and ARMOR3D (a, c) or GLORYS (b, d), 2011–2023, in 2° cells containing at least 30 stations common to all four products. (a, b) Difference in RMSE (m). (c, d) Difference in correlation. In all panels, blue indicates that Morlenn is closer to Argo or better correlated. Each product is scored against a reference MLD computed with its own criterion (Table 3).

Morlenn, which learns the relationship between surface forcing and MLD rather than interpolating profiles, has similar RMSE on both datasets and a much smaller loss of efficiency, from 73 to 61% for assimilation.

445 Baseline has slightly lower RMSE and higher correlation than assimilation on gliders. Assimilation reduces mean bias from +6.46 to +2.84 m and MAE from 17.15 to 15.70 m, while centred RMSE increases from 38.90 to 39.74 m. Assimilation therefore reduces the mean offset and average absolute error, while slightly increasing the scatter of errors. Baseline and ARMOR3D have similar MAE, 17.15 and 17.14 m, respectively.

Figure 9 shows the differences between assimilation and ARMOR3D against gliders in 2° cells. Morlenn has lower errors and
 450 much higher correlations along the North American shelf and in the Gulf of Mexico, where most glider-days are concentrated and Argo floats, which drift in deep water, provide few profiles to constrain ARMOR3D. ARMOR3D is closer to the gliders



Table 6. Gridded MLD products scored against the glider-day stations, 2011–2023. All rows are scored on the same 41 349 stations. Each product uses its reference MLD criterion. Dimensional scores are in metres. Metrics are defined in section 5.2.

Product	bias	RMSE	RMSE%	cRMSE	MAE	r	EV%
Morlenn assim	+2.84	39.84	90	39.74	15.70	0.793	61
Morlenn baseline	+6.46	39.43	89	38.90	17.15	0.797	62
ARMOR3D	-7.70	52.19	118	51.62	17.14	0.680	33
GLORYS*	-6.92	51.83	117	51.37	17.85	0.686	34

* GLORYS: σ_θ MLD criterion.

in the Caribbean, off West Africa and in parts of the Nordic seas and the eastern Mediterranean Sea. As against Argo, the differences are patchy in the Nordic seas and the Mediterranean Sea.

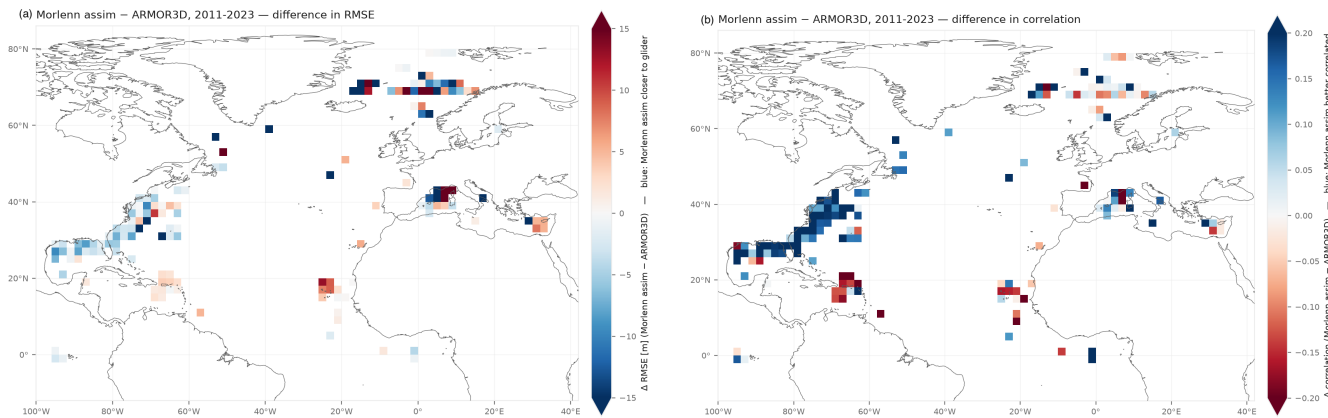


Figure 9. Differences in MLD scores against gliders between assimilating Morlenn and ARMOR3D, 2011–2023, in 2° cells: RMSE in metres (a) and correlation (b). Blue indicates that Morlenn is closer to the gliders or better correlated.

Within the OSnet domain, the benchmark includes 50 597 Argo stations over 2011–2019 and 2 433 glider-day stations from 2012 and 2015–2019 (Appendix C). On gliders, Morlenn assimilation has the lowest RMSE, 34.36 m, while OSnet has the smallest bias magnitude which is easier to achieve in a regional learning. Refer to appendix C for the full validation.

5.6 Depth dependence and extreme mixed layers

Depth-stratified scores over 2011–2023 distinguish shallow layers from the more sparsely sampled deep layers, against Argo (Table 7 and Fig. 10a, b) and against gliders (Fig. 10c, d). This separates errors relevant to the thin surface layers from those affecting winter mixing and deep convection. Against Argo, ARMOR3D has the smallest RMSE in the two classes shallower than 50 m: 12.5 versus 18.4 m for assimilation in the 0–25 m class, and 19.8 versus 26.9 m in the 25–50 m class. At 50–100 m,



assimilation and ARMOR3D have similar RMSE, 41.6 and 39.1 m. Below 100 m, Morlenn has the smallest errors, and the advantage of assimilation grows with depth.

Table 7. Median bias and RMSE in metres by disjoint in-situ MLD class, ARGO stations, 2011–2023. N is the station count of the class, shared by all products.

Depth class	N	Morlenn assim		Morlenn baseline		ARMOR3D		GLORYS*	
		bias	RMSE	bias	RMSE	bias	RMSE	bias	RMSE
0–25 m	142 671	+2.5	18.4	+7.0	20.3	+0.3	12.5	-2.4	19.9
25–50 m	112 927	-1.1	26.9	+4.6	28.7	-3.3	19.8	-11.5	30.1
50–100 m	72 647	-2.4	41.6	+3.1	43.6	-7.5	39.1	-17.5	54.1
100–300 m	40 738	+6.9	61.4	+3.2	64.0	-13.8	78.5	-29.8	103.3
> 300 m	6 304	-37.3	146.4	-72.9	199.4	-38.9	228.9	-62.5	239.5

For reference MLD exceeding 300 m, assimilation RMSE is 146.4 m, compared with 199.4 m for baseline, 228.9 m for ARMOR3D and 239.5 m for GLORYS. This class contains 6 304 stations. Assimilation has a median error of -37.3 m, indicating persistent underestimation of deep mixed layers, which all products share.

Within-class correlations are lower than the aggregate ones, because conditioning on narrow depth intervals restricts the variance of the reference MLD. They increase with depth for all products. In the classes shallower than 50 m, correlations are weak; Morlenn and ARMOR3D are similar, and GLORYS is lowest. Below 50 m, assimilation has the highest correlation, and its advantage is largest for layers deeper than 300 m. In this class, the baseline correlation falls below those of ARMOR3D and GLORYS, so the improvement of assimilation is consistent with the information provided by neighbouring profiles during deep convection.

Against the independent gliders, the ranking differs (Fig. 10c, d). Morlenn has the smallest RMSE in the classes shallower than 50 m, where ARMOR3D was the most accurate against Argo, and in the classes deeper than 100 m. At 50–100 m, all products have similar errors. Both Morlenn configurations have the highest correlation in every class, whereas ARMOR3D has much weaker correlations for shallow layers than against Argo. Baseline is slightly better correlated than assimilation for layers shallower than 100 m, whereas assimilation is better for deeper layers, as against Argo. For all products, errors in the deepest class are larger than against Argo. The advantage of ARMOR3D for shallow layers is therefore specific to the Argo comparison, consistent with its constraint by the assimilated profiles (Sect. 5.4).

5.7 Mixed-layer response during strong-wind events

The September 2023 diagnostics show reconstructed MLD anomalies during strong-wind events (Fig. 11). On 3 September at 03:00 UTC, the 15 and 20 m s^{-1} wind contours enclose a pronounced positive MLD anomaly south of Greenland, near 60° N . This spatial correspondence is consistent with wind-driven mixing and mixed-layer deepening. Areas of strong wind also overlap positive MLD anomalies in the western subtropical Atlantic.

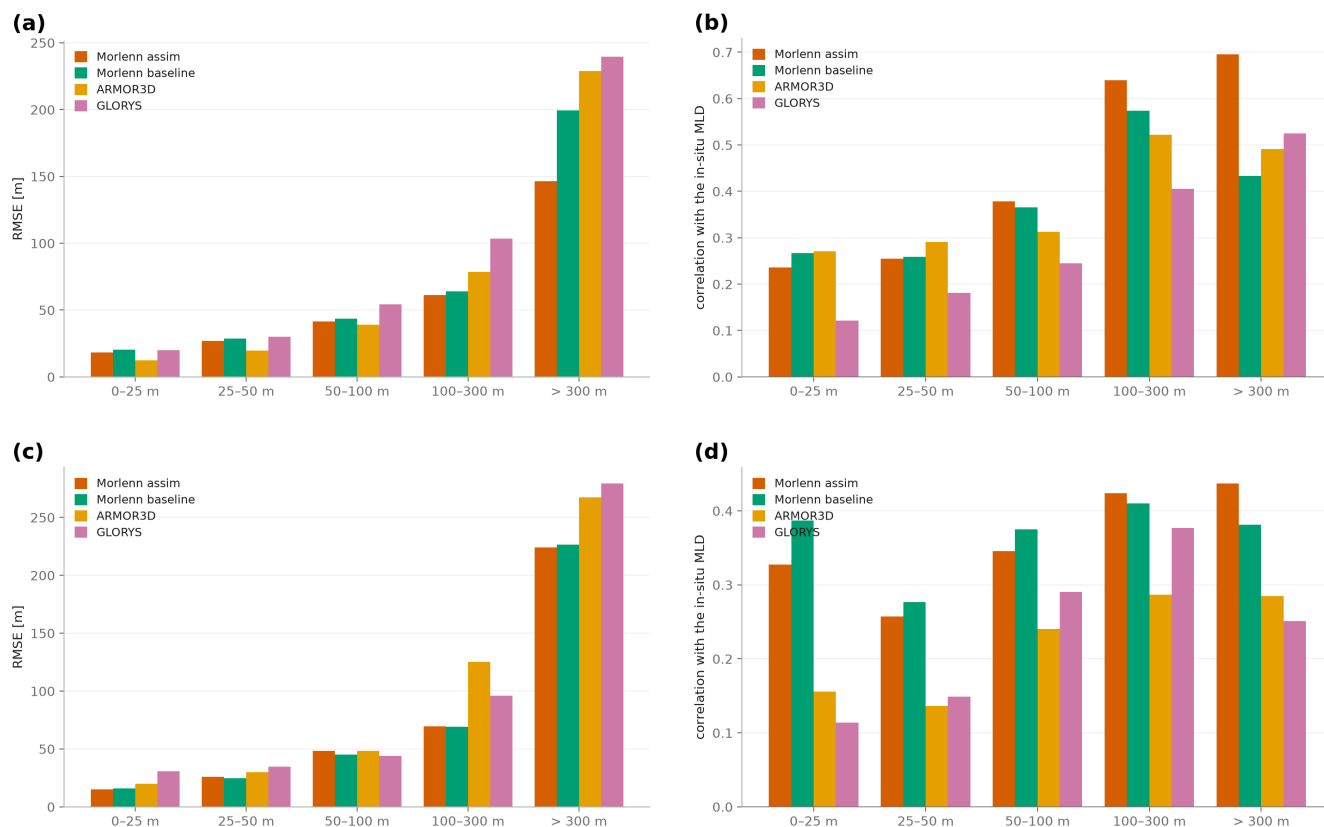


Figure 10. RMSE (a, c) and correlation (b, d) by disjoint in-situ MLD class, 2011–2023, against Argos (a, b) and gliders (c, d), on the stations common to all four products. Argos station counts per class are given in Table 7. GLORYS MLD is diagnosed with a σ_θ criterion.

485 The temporal sequence (Fig. 12) shows how these anomalies develop and persist. On 1 September, strong winds blow over the Labrador Sea, where depth anomalies are still weak or negative, and off the North American coast, where they are positive (Fig. 12a). By 4 September, the strongest positive anomalies, exceeding 20 m, lie beneath the $15\text{--}20\text{ m s}^{-1}$ contours south of Iceland, and positive anomalies extend along the Gulf Stream and into the central North Atlantic near 45° N (Fig. 12b). On 7 September, strong winds have almost disappeared from the domain, but positive anomalies remain south of Iceland, in the central North Atlantic and along the Gulf Stream (Fig. 12c). This persistence after the wind maximum has moved or weakened is physically consistent with an ocean response that depends on preceding forcing and the initial stratification, as represented by the model's forcing-history and context pathways. The response nevertheless varies geographically: elevated wind speed can coincide with weak or negative depth anomalies at a given instant. Anomalies measure departures from the seasonal reference, and their sign depends on the pre-existing mixed layer as well as the recent forcing.

495 Wind is a model input, so this analysis assesses the physical coherence of the MLD response. Event-specific profile collocations would be needed to quantify errors in the magnitude and timing of deepening.

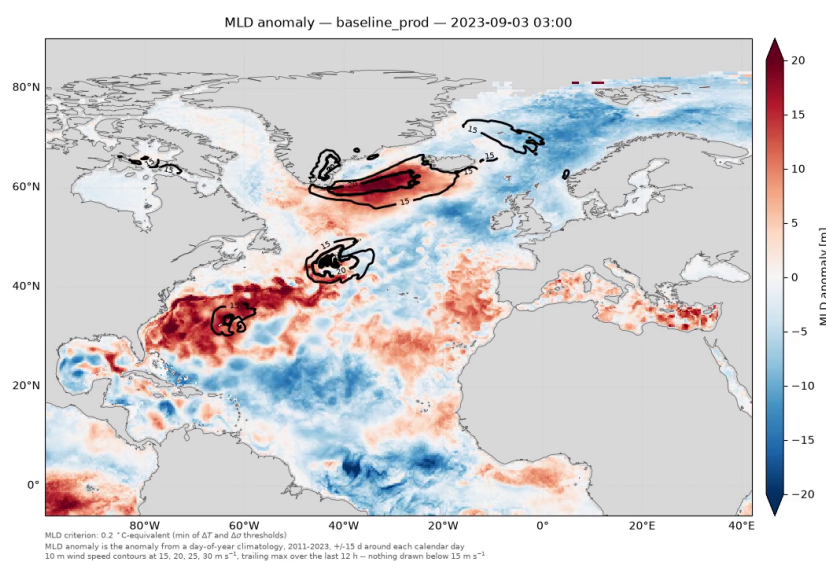


Figure 11. Baseline MLD anomaly on 3 September 2023 at 03:00 UTC, with contours of 10-m wind speed at 15, 20, 25 and 30 m s^{-1} . Wind contours use the maximum over the preceding 12 hours. The depth anomaly is relative to a 2011–2023 day-of-year climatology with a ± 15 -day calendar window. Positive anomalies denote deeper mixed layers. The MLD diagnostic uses the minimum of the temperature and temperature-equivalent density threshold depths.

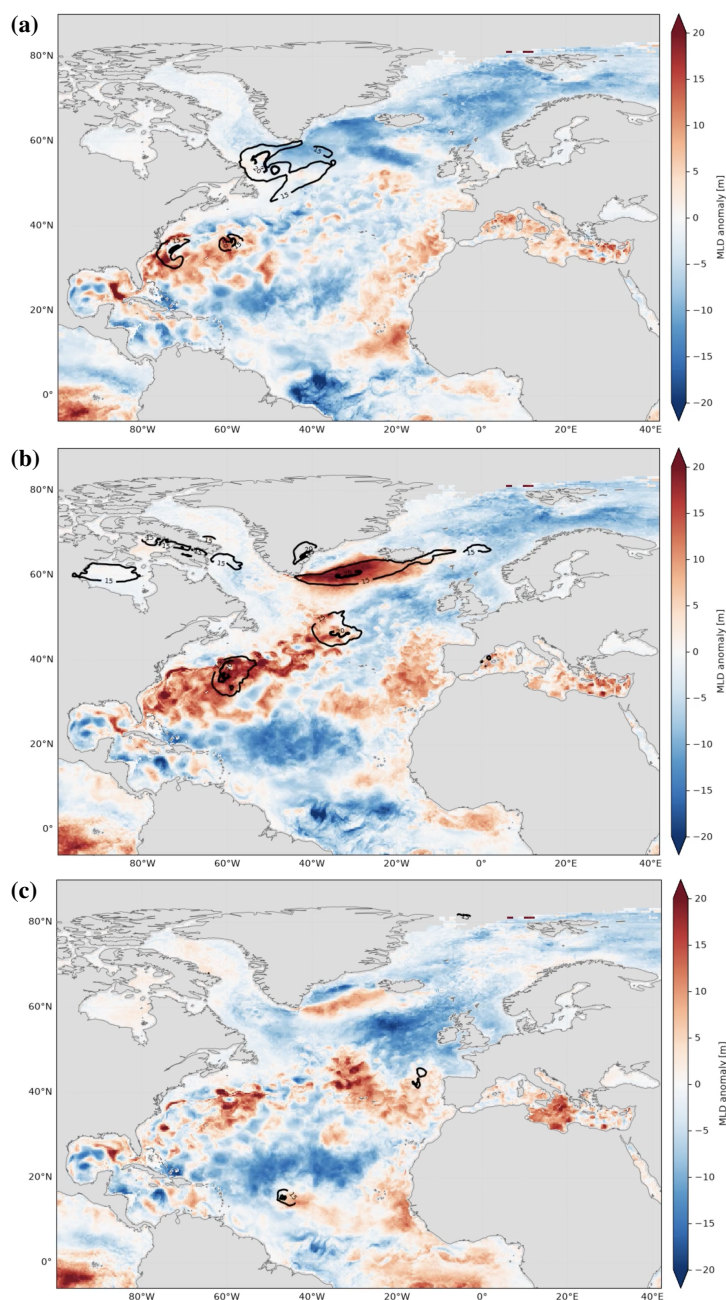


Figure 12. Baseline MLD anomaly (colours) with contours of ERA5 10-m wind speed at 15 and 20 m s^{-1} (black) on (a) 1, (b) 4 and (c) 7 September 2023 at 00:00 UTC. Wind contours use the maximum over the preceding 12 hours. The depth anomaly is relative to the 2011–2023 day-of-year climatology described in Fig. 11; positive anomalies denote deeper mixed layers. An animation showing the passage of the wind events and the MLD response, including the intervening frames, is available at <https://youtu.be/eHZ6OEQwNtw>.



5.8 Spectral content of the reconstructions

Spectra assess how much variability the reconstructions contain at different periods and wavelengths. They are computed for 2016, and each product is analysed on its own grid and at its own output interval. The temporal Nyquist period is therefore
 500 six hours for Morlenn and two days for the daily products. Spectral differences alone do not distinguish attenuated signal from spurious variance.

Over the whole region, Morlenn has more temporal variance than ARMOR3D and GLORYS at periods longer than about a month, including the seasonal cycle (Fig. 13). Between about 3 days and a month, Morlenn has slightly more variance than GLORYS, and the two spectra converge at shorter periods. ARMOR3D has more variance than both at periods shorter than
 505 about 10 days, with a flat spectrum down to its Nyquist period. At subdaily periods, which only Morlenn resolves, the spectra contain sharp peaks at 24, 12 and 8 hours, harmonics of the diurnal cycle. Between these peaks, assimilation has more variance than baseline. The amplitude and phase of the diurnal cycle are assessed against observations in Sect. 5.9.

MLD spectra — Full domain, 2016 (−6.0–90.0N, −100.0–42.0E, 10,595 of 17,794 rectangle cells, 366 days)

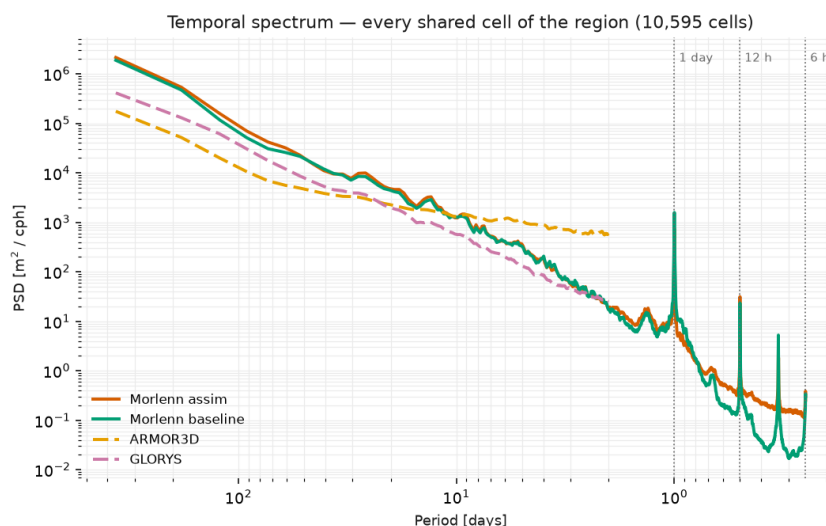


Figure 13. Temporal MLD spectra over the whole region in 2016. Each product is analysed up to its own Nyquist period; vertical guides mark 1 day, 12 hours and 6 hours.

The spatial spectra show a more regional behaviour (Fig. 14). Over the whole region, Morlenn has less variance than ARMOR3D and GLORYS at all wavelengths shorter than about 500 km, by one to two orders of magnitude at the mesoscale: the reconstruction is smoother than the comparison fields. In the Gulf Stream, Morlenn has more variance than the other products at wavelengths longer than about 700 km and similar variance between 300 and 500 km. At shorter wavelengths, it has less variance than ARMOR3D and GLORYS, but more than OSnet, whose spectrum drops steeply below about 150 km. The two
 510 neural reconstructions are therefore smoother at the mesoscale than the products that interpolate or assimilate in-situ profiles.



In the Mediterranean Sea, the spectra are closer: Morlenn has variance similar to the other products down to about 100 km, and more than ARMOR3D at shorter wavelengths, where the ARMOR3D spectrum falls off. The degree of spatial smoothing of Morlenn therefore varies between regions. The spectral energy of Morlenn remains stable over 2011–2023 in all regions, whereas the energy of ARMOR3D increases over the period, following the growth in the number of Argo profiles assimilated in its reconstruction (not shown). This dependence on the density of the observing network is consistent with the patchiness of ARMOR3D errors (Sect. 5.4), and indicates that the spectral content of Morlenn does not depend on the Argo sampling.

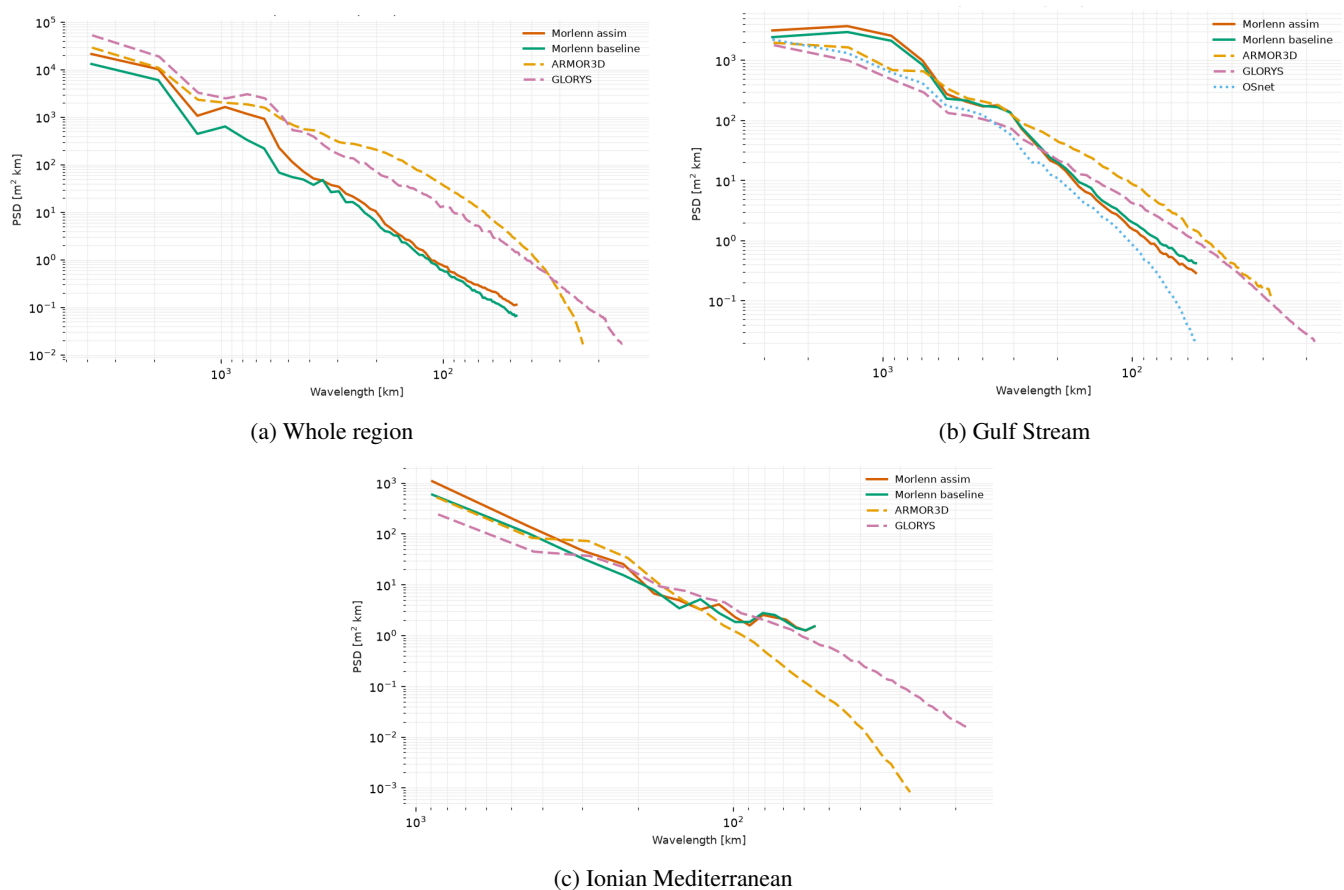


Figure 14. Isotropic spatial MLD spectra in 2016 over (a) the whole region, (b) the Gulf Stream and (c) the Ionian Mediterranean. OSnet is included in the Gulf Stream only. Each product is analysed on its own grid.



5.9 Diurnal cycle in high-frequency profile subsets

The diurnal cycle is assessed with platforms that profile several times a day in the Tropics: 25 833 glider profiles from 1 815 glider-days and 31 platforms over 2019–2020, and 2 721 Argo profiles from 332 float-days and 15 floats over 2011–2022. Only ARGO that sample profiles at least every 6 hours are included. MLD anomalies are calculated relative to each platform-day mean and averaged in hourly bins of local solar time, aligning the daily heating cycle across longitudes. Reconstructed fields are sampled at the profile times and processed identically. The amplitude of the 24-hour harmonic fitted to each composite
 525 are sampled at the profile times and processed identically. The amplitude of the 24-hour harmonic fitted to each composite summarises the strength of the diurnal cycle. The same process is applied on Morlenn baseline and assimilation products on Glider or ARGO profiles location.

In the glider subset, the observed mixed layer is deepest in the early morning, around 06:00–08:00, and shallowest in the late afternoon, around 17:00, with a 24-hour harmonic amplitude of 2.06 m (Fig. 15). Both reconstructions reproduce this phase:
 530 the layer shoals from mid-morning and reaches its minimum depth at the same time as observed. The baseline amplitude, 2.10 m, matches the observations, whereas assimilation reduces it to 1.53 m, with a weaker night-time deepening. The main discrepancy occurs in the evening: the observed layer remains shallow until about 22:00, whereas both reconstructions deepen from about 19:00.

The Argo composite is noisier, because it contains fewer profiles, particularly at night. Its 24-hour harmonic amplitude is
 535 1.57 m, and the observed layer is also shallowest around 17:00. Both reconstructions reproduce the afternoon shoaling. Baseline slightly overestimates the amplitude, at 1.87 m, and assimilation slightly underestimates it, at 1.33 m.

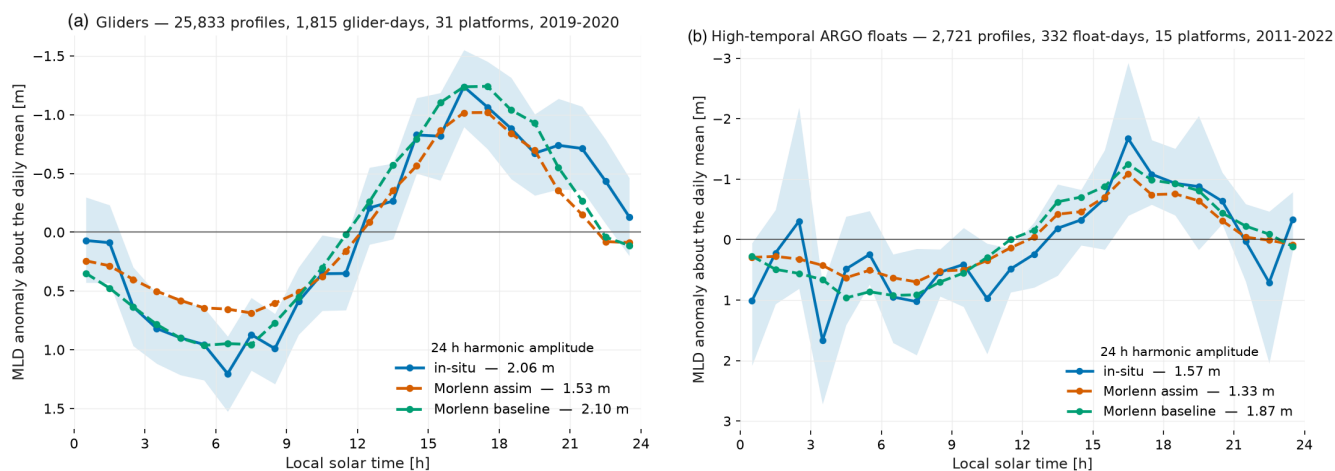


Figure 15. Local-solar-time MLD anomaly composites relative to the daily mean for gliders (a) and high-frequency Argo floats (b). Negative anomalies, plotted upwards, indicate shallower layers. Legends give the 24-hour harmonic amplitude.

In both subsets, the reconstructions capture the phase of the diurnal cycle and its amplitude to within about 25%. Assimilation consistently reduces the amplitude, possibly because neighbouring profiles, collected at other times of day, damp the daily signal. The 10-m reference depth limits sensitivity to shallower near-surface warm layers.



540 6 Usage and limitations

6.1 Intended uses

Morlenn provides mixed-layer depth and the strongest stratification beneath it on a common three-hourly grid. The fields allow users to examine how the two diagnostics change together, from the diurnal cycle to seasonal and regional variations, and how they respond to atmospheric events such as storms.

545 6.2 Choice of configuration

The two configurations serve different purposes. The assimilating configuration uses neighbouring Argo profiles and gives the most accurate MLD against Argo; on gliders, it also reduces MLD bias and MAE. The baseline uses surface inputs and forcing histories only. It reproduces the amplitude of the diurnal cycle more closely, and its estimates do not depend on the local availability of Argo profiles. The baseline is therefore recommended for studies of the diurnal cycle and for trend analyses, which also require an assessment of the temporal stability of the upstream surface products. The two configurations should not be combined within a time series, as switching between them can introduce discontinuities.

6.3 Known limitations of the MLD

Morlenn overestimates MLD on average, by 6 to 10 m against Argo. For layers shallower than 50 m, errors remain large relative to layer depth, and deep convective layers are underestimated. The diurnal cycle is reproduced with the observed timing and amplitudes within about 25%, but assimilation damps it, and the evening deepening starts earlier than observed. The 10-m reference depth of the training diagnostic also limits sensitivity to shallower near-surface warm layers. The fields are smoother than the comparison products at the mesoscale over most of the domain, including the Gulf Stream, which can affect horizontal gradients near fronts. Applications focused on near-surface warm layers, extreme depths, sharp fronts or diurnal amplitudes therefore require a specific evaluation in the region of interest.

560 6.4 Known limitations of the stratification

Stratification was not specifically tuned, and its evaluation is preliminary. It has a moderate correlation with the Argo reference but large amplitude errors. Relative errors are largest in the weakly stratified subpolar and Nordic seas and near river plumes. Stations with surface salinity below 25 are excluded from the evaluation, so stratification in these low-salinity waters is not validated. Analyses of strong maxima or of resistance to mixing depend on the depth interval in which $N_{\max}^2$ is searched and on the vertical sampling of the reference profiles. Both configurations have similar stratification skill. Additional parameters such as the N^2 right below the MLD or at a given depth below MLD could also give us a better insight of the stratification process.



6.5 Structural limitations

The model estimates each grid point independently from its inputs and does not enforce heat or salt budgets. Map gradients inherit the interpolation of the inputs and the statistical smoothing of the reconstruction, which can affect transport estimates and budget residuals. MLD and $N_{\max}^2$ summarise layer depth and peak stratification without reconstructing a complete density profile; pycnocline thickness, integrated stratification and entrainment require additional vertical information. The model also predicts a spread alongside each central estimate, but its calibration has not yet been established, so the corresponding probability intervals have no verified observational coverage.

6.6 Coverage of the validation

Argo provides the validation across the whole domain. The independent glider evaluation covers mainly the North American shelf, the Gulf of Mexico, the Caribbean, the Mediterranean Sea and the Nordic seas; in the open-ocean interior, performance rests on the Argo comparison alone. Because the model estimates each point from its inputs, it can be applied to other regions or with finer surface inputs, provided these inputs are prepared consistently and performance is evaluated for the new regime.

7 Conclusions

Morlenn is, to our knowledge, the first observation-based dataset to provide mixed-layer depth at subdaily resolution, and the first gridded dataset of maximum subsurface stratification. It covers the North Atlantic and adjacent equatorial waters for 2011–2023, at 0.25° spacing and three-hourly intervals, and delivers both diagnostics jointly. The three-hourly sampling resolves the diurnal cycle of the mixed layer, which daily products cannot represent. The reconstruction combines surface-ocean fields with atmospheric and wave forcing histories, selected from the physics of mixed-layer evolution and an exploratory correlation analysis. A temporal cascade modulated by ocean conditions represents the accumulated response of the mixed layer to forcing. Two configurations are released: a baseline using surface inputs alone, and an assimilating version that also uses neighbouring Argo profiles.

Morlenn outperforms the existing MLD products evaluated here. Against Argo over 2011–2023, the assimilating configuration has the lowest RMSE and the highest correlation and efficiency. Withholding 2023 from training leaves the scores for that year almost unchanged, indicating no substantial overfitting. ARMOR3D has smaller typical errors for shallow layers against Argo, which it assimilates, but this advantage does not extend to independent observations. Against gliders, which none of the products use, Morlenn has RMSE about 25% lower than ARMOR3D and GLORYS, with the smallest errors in both shallow and deep classes. Its skill is similar against Argo and gliders, and its spectral content does not depend on the density of the observing network, whereas ARMOR3D follows the growth of Argo sampling. In the Gulf Stream, Morlenn also better reproduces the variability of glider MLD than the regional OSnet reconstruction, which has a smaller bias.

The diurnal cycle is a distinctive feature of the dataset. In high-frequency glider and Argo subsets, Morlenn reproduces the timing of the afternoon shoaling and the night-time deepening, with amplitudes within about 25% of the observations; the



baseline matches the observed glider amplitude. Stratification, although not specifically tuned, has a preliminary skill with a correlation of 0.73 against Argo; its relative errors are lowest in the tropics and highest in the weakly stratified subpolar and Nordic seas.

Several limitations remain. Deep convective layers are underestimated, MLD is overestimated on average, and the fields are smoother than the comparison products at the mesoscale. Assimilation damps the diurnal cycle, and the predicted spreads are not yet calibrated. The reference depth of the MLD criterion also sets how much of the diurnal cycle the dataset can carry. It is taken at 10 m, following de Boyer Montégut et al. (2004), which filters part of the near-surface diurnal variability and leaves the part that reaches below that depth, a regime documented from satellite and in-situ observations (Gille, 2012; Gasparin et al., 2025). This residual diurnal signal is what Sect. 5.9 measures. A shallower reference, at 5 m, would give access to more of the diurnal amplitude and to finer temporal structure, at the cost of the Argo profiles whose sampling starts below it, so the gain in resolved amplitude has to be weighed against the loss of training data. Regional versions trained on limited domains, specific training for stratification and calibration of the predictive uncertainty are the main planned developments.

8 Data availability

Morlenn is publicly available on SEANOE at <https://doi.org/10.17882/120058> (Gaultier et al., 2026), under the Creative Commons Attribution 4.0 International (CC BY 4.0) licence.

9 Code availability

The software for model training and dataset generation is available on GitHub: <https://github.com/oceandatalab/morlenn.git>

Video supplement. An animation of baseline MLD anomalies and ERA5 10-m wind speed during September 2023 is available at <https://youtu.be/eHZ6OEQwNtw>.

Appendix A: Model configuration

Tables A1 and A2 specify the layers corresponding to Fig. 4. The MLD and stratification pathways have separate parameters for the TCN, temporal attention, fusion, residual blocks and output heads. They share the context encoder. The assimilating configuration adds the neighbour pathway; the two configurations are trained independently.

Dimensions are given per sample, with temporal arrays written as time steps $\times$ channels. FC denotes a fully connected linear layer, and Conv1D a one-dimensional convolution. GeLU is the Gaussian error linear unit (Hendrycks and Gimpel, 2016); GN and LN denote group normalisation (Wu and He, 2018) and layer normalisation (Ba et al., 2016). Dropout probability is 0.10 wherever dropout is specified. Softmax converts scores into non-negative weights summing to one; sigmoid gates have values between zero and one.



Table A1. Layers and dimensions of the production network. The temporal and estimation rows apply separately to MLD and $N_{\max}^2$. Block letters match the architecture figure.

Layer or block	Input $\rightarrow$ output	Operations
A: temporal convolutional network		
Input projection	$720 \times 20 \rightarrow 720 \times 48$	Conv1D, kernel 1; GeLU.
TCN block 1	$720 \times 48 \rightarrow 720 \times 48$	Two causal convolutions; kernel 5; dilation 1.
TCN block 2	$720 \times 48 \rightarrow 720 \times 96$	Two causal convolutions; kernel 5; dilation 2.
TCN block 3	$720 \times 96 \rightarrow 720 \times 96$	Two causal convolutions; kernel 5; dilation 4.
TCN block 4	$720 \times 96 \rightarrow 720 \times 160$	Two causal convolutions; kernel 5; dilation 8.
TCN block 5	$720 \times 160 \rightarrow 720 \times 160$	Two causal convolutions; kernel 5; dilation 16.
TCN block 6	$720 \times 160 \rightarrow 720 \times 192$	Two causal convolutions; kernel 5; dilation 32.
TCN block operations	$720 \times C_{\text{in}} \rightarrow 720 \times C_{\text{out}}$	GN (one group), GeLU, causal Conv1D, dropout; GN (one group), GeLU, causal Conv1D, dropout; residual addition. The skip uses a kernel-1 projection when channel counts differ.
FiLM after each block	Context $96 \rightarrow 2C_{\text{out}}$	FC $96 \rightarrow 96$, GeLU, FC $96 \rightarrow 2C_{\text{out}}$. Tanh bounds multiplicative gains to 1 ± 0.1 and offsets to ± 0.1 . Factors are shared over time within each prediction.
X: shared context encoder		
Three group encoders	$(5, 4, 2) \rightarrow 3 \times 96$	Surface fields (5), position and bathymetry (4), season (2). Each group: FC to 96, GeLU, LN, dropout, FC $96 \rightarrow 96$; add a learned group embedding.
Self-attention	$3 \times 96 \rightarrow 3 \times 96$	Four heads, 24 dimensions per head; attention dropout; residual addition and LN.
Feed-forward layer	$3 \times 96 \rightarrow 3 \times 96$	FC $96 \rightarrow 192$, GeLU, dropout, FC $192 \rightarrow 96$; residual addition and LN.
Group weighting	$288 \rightarrow 3$	Concatenate the three tokens; FC $288 \rightarrow 96$, GeLU, dropout, FC $96 \rightarrow 3$, softmax across groups.
Context output	$192 \rightarrow 96$	Concatenate the weighted group sum and componentwise group maximum; FC $192 \rightarrow 96$, GeLU, dropout, FC $96 \rightarrow 96$.

The stratification gradient entering the shared context encoder is multiplied by 0.10. Training uses 32 samples per parameter update, an initial learning rate of 10^{-4} , weight decay of 10^{-4} and a gradient-norm clipping threshold of 1.0.

A1 Physical input variables

630 Table A3 defines the physical variables supplied as hourly histories. Sea-surface temperature, salinity (sos), absolute dynamic topography and sea-level anomaly also enter the contemporary context. Salinity is duplicated in the production input arrays, giving 20 history channels and five surface-context entries. Each input array contains 720 hourly entries. SST and salinity supply 10 days of history, padded by repeating an endpoint value; atmospheric and wave inputs supply 30 days. Daily SST and salinity are interpolated over the most recent 240 hourly entries; the remaining entries repeat an endpoint value. The duration



Table A2. Continuation of Table A1: temporal attention, estimation and neighbour pathway.

Layer or block	Input → output	Operations
P: temporal attention and summary		
Three temporal filters	$720 \times 192 \rightarrow 3 \times 720 \times 192$	Depthwise Conv1D with kernels 3, 13 and 49, centred padding, no bias; weights initialised to $1/k$. Each filtered sequence is averaged with its unfiltered input.
Attention scores	Sequence 192, context $96 \rightarrow 1$ per time step	Separate FC projections $192 \rightarrow 96$ and $96 \rightarrow 96$; add, apply GeLU, then FC $96 \rightarrow 1$. Softmax over the 720 steps; weighted sum gives 192 values per module.
Module weighting	$672 \rightarrow 3$	Concatenate context and temporal mean, maximum and final value; FC $672 \rightarrow 192$, GeLU, dropout, FC $192 \rightarrow 3$, softmax. Weighted combination of the three module outputs.
Temporal summary	$720 \times 192 \rightarrow 771$	Concatenate mean (192), maximum (192), final value (192), combined attention output (192), and module weights (3).
G: fusion and target-specific estimation		
History and context projections	$771 \rightarrow 192$; $96 \rightarrow 192$	Separate FC layers, each followed by GeLU.
Fusion gate	$867 \rightarrow 192$	Concatenate history summary and context; FC $867 \rightarrow 192$, GeLU, dropout, FC $192 \rightarrow 192$, sigmoid. Componentwise mixture of the projected history and context.
Fusion output	$384 \rightarrow 192$	Concatenate the mixture and the difference between the projected history and context; FC $384 \rightarrow 192$, GeLU, dropout.
Residual backbone	$192 \rightarrow 128$	Two residual fully connected blocks at width 192, followed by FC $192 \rightarrow 128$ and GeLU.
Target branch	$128 \rightarrow 128$	FC $128 \rightarrow 128$, GeLU, dropout; two residual fully connected blocks at width 128.
Fully connected residual block	$D \rightarrow D$	FC $D \rightarrow D$, GeLU, dropout, FC $D \rightarrow D$; residual addition followed by GeLU.
Central-value head	$128 \rightarrow 1$	FC $128 \rightarrow 128$, GeLU, FC $128 \rightarrow 1$.
Spread head	$128 \rightarrow 1$	Separate FC $128 \rightarrow 128$, GeLU, FC $128 \rightarrow 1$; softplus plus 10^{-3} after any neighbour correction. Values are in transformed, normalised target coordinates.
Neighbour pathway: assimilating configuration only		
Neighbour encoder	$5 \times 12 \rightarrow 5 \times 96$	Each profile supplies MLD and 11 context entries. Shared FC $12 \rightarrow 96$, GeLU, LN.
Neighbour attention	5×96 , target context $96 \rightarrow 96$	Project target context with FC $96 \rightarrow 96$; add to each neighbour embedding, tanh, FC $96 \rightarrow 1$. Mask missing neighbours; softmax and weighted sum. A learned 96-vector handles the absence of valid neighbours.
Neighbour correction	$96 \rightarrow 4$	FC $96 \rightarrow 96$, GeLU, dropout, FC $96 \rightarrow 96$; final FC $96 \rightarrow 4$, initialised to zero. Add corrections to the two central values and the two spread logits before softplus.
Neighbour dropout	Whole neighbour set	During training, all neighbours of a sample are withheld with probability 0.5.



635 $T(F)$ in Eq. (2) refers to the available history before padding. The context supplies one value per descriptor at the target location and time. Each history variable uses the same scaling at every time step.

Table A3. Definitions of the dynamic inputs. An asterisk marks variables also supplied as contemporary context. Identifiers provide correspondence with the input files.

Input identifier	Physical quantity
<code>swh, mwp</code>	Significant height (m) and mean period (s) of combined wind waves and swell.
<code>shww, mpww</code>	Significant height (m) and mean period (s) of wind waves.
<code>sshf, slhf</code>	Time-integrated surface sensible and latent heat fluxes, in J m^{-2} .
<code>wind_speed_neutral</code>	Neutral-equivalent 10-m wind speed, in ms^{-1} , $\sqrt{u_{10n}^2 + v_{10n}^2}$, where u_{10n} and v_{10n} are its zonal and meridional components.
<code>msl</code>	Mean sea-level atmospheric pressure, in Pa.
<code>ssr, str</code>	Accumulated surface net short-wave and long-wave radiation, in J m^{-2} .
<code>tsr</code>	Accumulated top-of-atmosphere net short-wave radiation, in J m^{-2} .
<code>analysed_sst*</code>	ESA CCI analysed sea-surface temperature, in kelvins.
<code>grad_analysed_sst</code>	Magnitude of the horizontal SST gradient, in kelvins per metre.
<code>sos*</code>	CNR sea-surface salinity used in production.
<code>dos</code>	CNR sea-surface density, in kg m^{-3} .
<code>grad_sss</code>	Magnitude of the horizontal gradient of <code>sos</code> , per metre.
<code>adt*</code>	Absolute dynamic topography, in metres: sea-surface height relative to the geoid.
<code>sla*</code>	Sea-level anomaly, in metres, relative to the 1993–2012 mean.
<code>grad_sla</code>	Horizontal gradient descriptor of sea-level anomaly, supplied by the input archive.

A2 Geographic, bathymetric and seasonal context

The remaining six context variables describe location, seafloor elevation and annual phase (Table A4). Together with the five surface-context entries described above, they form the 11-dimensional context input. These variables allow the response to the
 640 same forcing history to depend on location and season.

The three geographic components are dimensionless and satisfy $\text{geo_x}^2 + \text{geo_y}^2 + \text{geo_z}^2 = 1$. They embed location on the unit sphere, maintaining continuity across the longitude seam. The seasonal pair represents annual phase on the unit circle. The implementation uses a fixed 365-day denominator, including in leap years. Both encodings are calculated before the feature-wise scaling described in Sect. 3.

645 A3 Loss coefficients and output transformation

The loss configuration assigns MLD coefficients of 0.6 to log-space error, 0.25 to physical-space Huber error, 1.5 to RMSE, 2.0 to tail RMSE, 80 to the dispersion penalty and 10 to the correlation penalty. The Huber scale is 25 m; underpredictions have a



weight multiplier of 1.5 in that term. The tail threshold is the batch 95th percentile. The N^2 RMSE, dispersion and correlation coefficients are 10, 50 and 5; its base loss has weight 2.0. The N^2 terms start at 20% of their configured weight and reach full weight over five passes through the training data. The Gaussian likelihood coefficient is 0.02. The dispersion penalties compare the standard deviations of predictions and reference values within each batch. These terms operate in different units and scalings, so coefficient magnitudes alone do not express their relative contributions.

Exponentiating the predicted mean in log coordinates gives the median depth after reversing the log transform. The mean depth also depends on the predicted variance. This distinction matters when interpreting the central value and spread in metres.

Appendix B: Overfitting test: withholding 2023

A model that overfits its training data performs markedly better on the profiles it has learned than on new ones. To test this, the baseline and assimilating models were first trained with all 2023 Argo profiles withheld (`assim` and `baseline`), then evaluated on the 31 889 Argo stations of that year. Their scores are compared with those of the production models trained on all years, 2023 included (`assim_prod` and `baseline_prod`), on the same stations (Table B1). If the models memorised their training profiles, including 2023 in training would substantially reduce the errors on these stations.

The scores are nearly identical. For the assimilating model, RMSE is 32.40 m without 2023 and 32.97 m with it, and correlation remains 0.882. Bias decreases from 7.44 to 5.13 m and MAE from 16.54 to 16.07 m. For the baseline model, RMSE is 38.01 m without 2023 and 37.73 m with it. Performance on a year absent from training therefore matches performance on a year included in training, which indicates no substantial overfitting of MLD. The baseline test is the stricter one, because the withheld assimilating model still receives neighbouring 2023 Argo profiles as inputs during inference.

Glider profiles excluded from training and neighbour inputs provide a complementary independent test of the production models. On 41 349 glider-day stations, the assimilating model trained on all years achieves RMSE of 39.84 m, correlation of 0.793 and efficiency of 61% (Sect. 5.5). Its bias and MAE are lower than those of the model using surface inputs alone.

Table A4. Definitions of the geographic, bathymetric and seasonal context variables. Latitude ϕ and longitude λ are in radians; d is the integer day of year, with 1 January numbered 1.

Identifier	Definition	Interpretation
<code>geo_x</code>	$\cos \phi \cos \lambda$	Equatorial component towards 0° longitude.
<code>geo_y</code>	$\cos \phi \sin \lambda$	Equatorial component towards 90° E.
<code>geo_z</code>	$\sin \phi$	Component along the north polar axis.
<code>elevation</code>	Seafloor elevation, in metres	Bathymetric setting of the local column.
<code>doy_sin</code>	$\sin(2\pi d/365)$	Sine component of the annual phase.
<code>doy_cos</code>	$\cos(2\pi d/365)$	Cosine component of the annual phase.



Table B1. Gridded MLD products scored against ARGO stations from 2023. `Morlenn assim` and `Morlenn baseline` are the runs from which 2023 was withheld; `Morlenn assim` and `Morlenn baseline` are separately refitted production versions. All rows are scored on the same 31 889 stations. Each product uses its reference MLD criterion; 2023 is excluded from target fitting of the held-out runs only; the held-out assimilating run still uses neighbouring 2023 profiles as inputs. Dimensional scores are in metres. Metrics are defined in section 5.2.

Product	bias	RMSE	RMSE%	cRMSE	MAE	r	EV%
Morlenn assim (2023 held out)	+7.44	32.40	64	31.53	16.54	0.882	76
Morlenn baseline (2023 held out)	+9.74	38.01	75	36.74	18.81	0.844	66
Morlenn assim (2023 in training)	+5.13	32.97	65	32.57	16.07	0.882	75
Morlenn baseline (2023 in training)	+11.28	37.73	75	36.00	19.42	0.849	67
ARMOR3D	-4.54	40.19	79	39.93	13.41	0.819	62
GLORYS*	-7.11	39.41	78	38.77	18.78	0.827	64

Together, the stable 2023 scores and the glider skill provide evidence against substantial MLD overfitting in the evaluated regimes.

Appendix C: Regional comparison with OSnet in the Gulf Stream

OSnet is a regional product, so its comparison is restricted to its Gulf Stream domain and to the period it covers. The benchmark includes 50 597 Argo stations over 2011–2019 (Table C1) and 2 433 glider-day stations from 2012 and 2015–2019 (Table C2). As in the basin-wide comparisons, each product is scored against a reference MLD computed with its own criterion, on the stations common to all five products. Argo profiles are used by all products, for training (`Morlenn` and `OSnet`) or assimilation (`ARMOR3D`, `GLORYS` and `Morlenn assimilation`), whereas the gliders are independent of all of them.

Against Argo, `ARMOR3D` has the best scores in this domain, with RMSE of 29.79 m, MAE of 13.87 m and efficiency of 72%. The Gulf Stream is densely sampled by Argo, so this result is consistent with the strong constraint of `ARMOR3D` by the assimilated profiles (Sect. 5.4). The two `Morlenn` configurations have similar RMSE, near 35.3 m, and assimilation has the same correlation as `ARMOR3D`, 0.852. Both overestimate MLD by 10–12 m. `GLORYS` has the smallest bias, and `OSnet` has the largest RMSE, 37.73 m, and a lower correlation of 0.776.

Figure C1 shows how these differences are distributed within the domain against Argo. `Morlenn` has lower relative errors over the subtropical gyre south of about 34°N and in the northeastern part of the domain, north of 40°N and east of 50°W. `OSnet` has lower relative errors along the path of the Gulf Stream and its extension, between about 35 and 42°N and west of 50°W, where MLD gradients are strongest. This band is where a regionally trained product benefits most from its adjustment to the local MLD distribution. `Morlenn` is better correlated with Argo than `OSnet` almost everywhere in the domain, including along the Gulf Stream, with only a few isolated cells favouring `OSnet`.



Table C1. Gridded MLD products scored against ARGO stations in the Gulf Stream within the OSnet domain, 2011–2019, each against its own MLD criterion. All rows are scored on the same 50 597 stations. Metrics are defined in section 5.2.

Product	bias	RMSE	RMSE%	cRMSE	MAE	r	EV%
Morlenn assim	+10.14	35.47	60	33.99	19.30	0.852	60
Morlenn baseline	+11.89	35.20	59	33.13	20.40	0.829	60
ARMOR3D	-4.60	29.79	50	29.43	13.87	0.853	72
GLORYS [vs dr001]	+1.41	33.28	75	33.25	18.02	0.752	53
OSnet [vs dr003]	+5.39	37.73	67	37.34	21.77	0.776	57

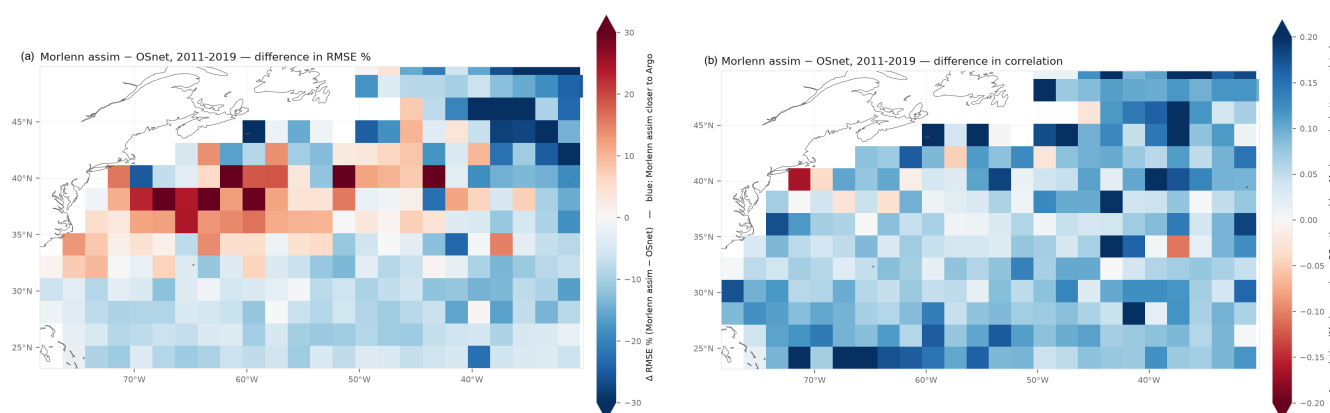


Figure C1. Differences in MLD scores against Argo between assimilating Morlenn and OSnet within the OSnet domain, 2011–2019, in 2° cells: relative RMSE (a, in % of the mean in-situ MLD) and correlation (b). Blue indicates that Morlenn is closer to Argo or better correlated. Each product is scored against a reference MLD computed with its own criterion.

The glider comparison changes the ranking. Assimilation has the lowest RMSE, 34.36 m, the highest correlation, 0.710, and the highest efficiency, 40%, followed by baseline. OSnet has the smallest bias and MAE, 0.34 and 17.49 m, but a lower correlation, 0.570, and an efficiency of 31%. ARMOR3D and GLORYS have the lowest skill, with efficiencies of 3 and 5%. As over the whole basin (Sect. 5.5), the advantage of ARMOR3D against Argo does not extend to the independent gliders. Efficiencies are lower than basin-wide for all products, reflecting the strong mesoscale variability of the Gulf Stream and the small glider sample, concentrated on the shelf and slope (Fig. 6b).

Among the two neural reconstructions, OSnet is less biased, whereas Morlenn better reproduces the variability of the glider MLD. The smaller bias of OSnet is expected from a product trained on a single region, which can adjust to the local MLD distribution, whereas Morlenn is trained over the whole basin. Regional versions of Morlenn trained on a limited domain are therefore expected to reduce its bias substantially. The additional forcing histories used by Morlenn, which OSnet does not use, are consistent with this better representation of MLD variations.



Table C2. Gridded MLD products scored against the glider-day stations inside the OSnet box, 2012 and 2015–2019 each against its own MLD criterion. All rows are scored on the same 2 433 stations. Metrics are defined in section 5.2.

Product	bias	RMSE	RMSE%	cRMSE	MAE	r	EV%
Morlenn assim	+6.77	34.36	71	33.69	18.15	0.710	40
Morlenn baseline	+8.17	36.22	74	35.28	19.11	0.655	33
ARMOR3D	-15.70	43.56	89	40.63	21.93	0.413	3
GLORYS [vs dr001]	+4.77	39.28	107	38.99	19.73	0.482	5
OSnet [vs dr003]	+0.34	38.29	83	38.29	17.49	0.570	31

Appendix D: Seasonal and synoptic field comparisons

700 The assimilating reconstruction supplies MLD and $N^2_{\max}$ fields for the same winter and summer dates as the baseline maps (Fig. D1). ARMOR3D, GLORYS and OSnet provide complementary MLD comparisons (Fig. D2). The products represent

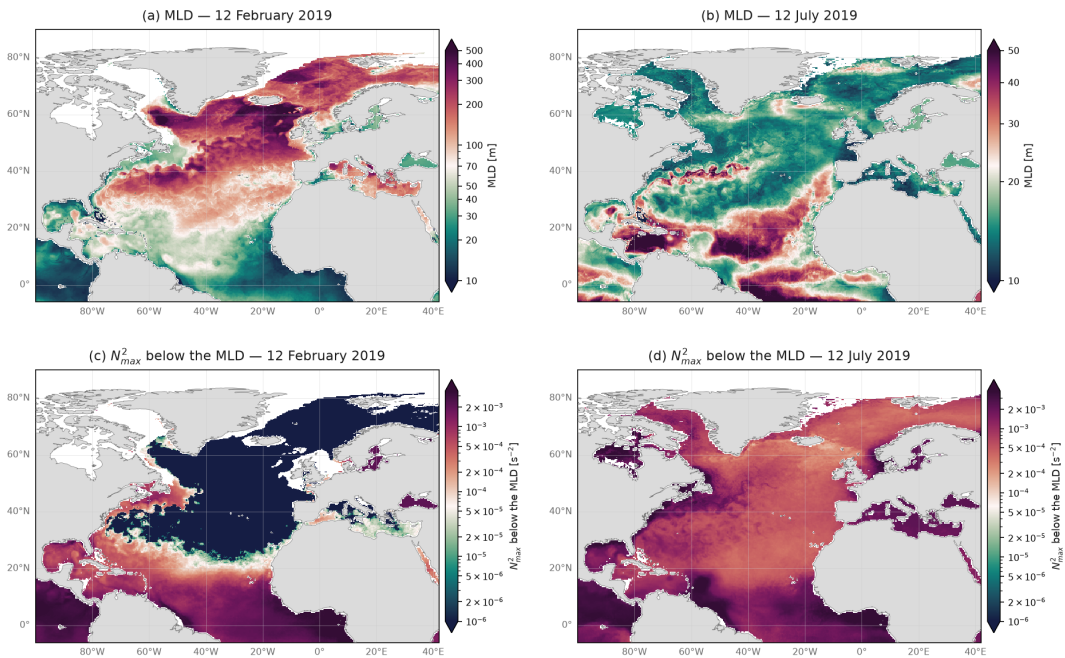


Figure D1. Assimilating MLD and maximum subsurface stratification on 12 February and 12 July 2019. The arrangement follows the baseline comparison in Fig. 5.

the broad winter deepening of the northern ocean and summer shoaling, with differences in frontal detail, local magnitude

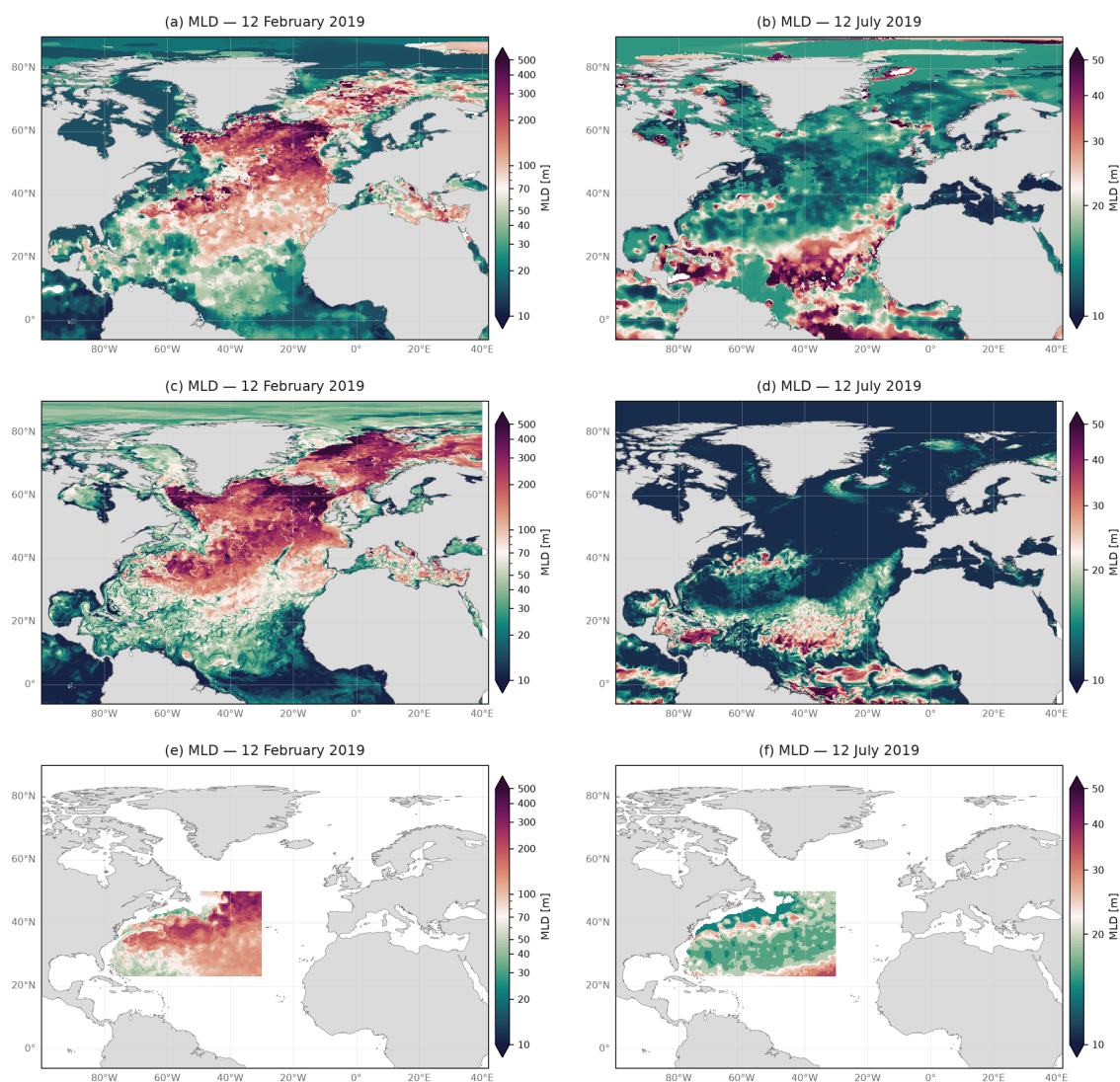


Figure D2. Winter and summer MLD fields for ARMOR3D (a, b), GLORYS (c, d) and OSnet (e, f), on 12 February and 12 July 2019. OSnet covers a regional Gulf Stream domain. Each date retains its displayed scale.



and coverage. Product-specific MLD definitions and colour scales enter the interpretation of these contrasts. The three-hourly snapshots in Fig. D3 illustrate the paired diagnostics in both Morlenn configurations.

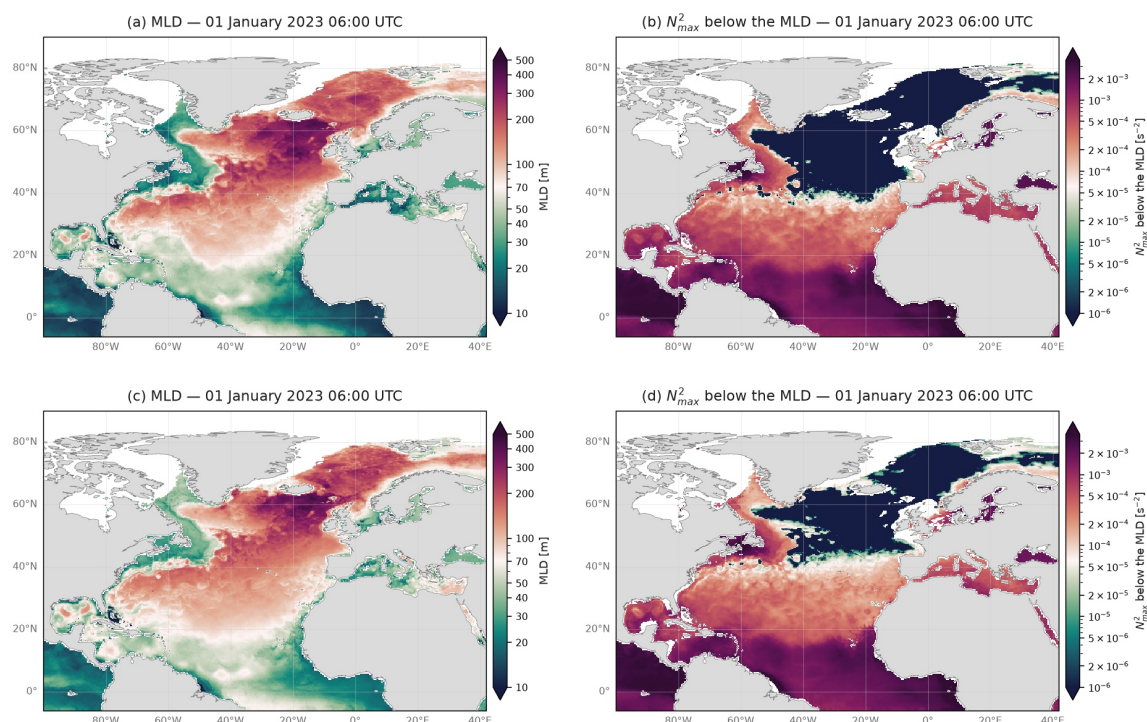


Figure D3. MLD and maximum subsurface stratification at 06:00 UTC on 1 January 2023 for assimilation (a, b) and baseline (c, d). These paired fields provide a synoptic view of the two predicted diagnostics.

705 Appendix E: Preliminary reconstruction from model-derived ocean inputs

An exploratory transfer experiment replaces the observed ocean surface inputs with GLORYS fields, retaining atmospheric and sea-state forcing similar to that used with GLORYS. Reconstruction is performed at 0.25° and six-hour intervals, followed by daily averaging for comparison with GLORYS MLD. This experiment examines whether a network trained with observation-based surface fields retains agreement with the model reference when supplied with model-derived ocean inputs. Dependencies on the reference model and differences between the GLORYS and Morlenn MLD criteria contribute to the resulting scores.

The assimilation configuration has correlation 0.9445, bias +12.04 m, RMSE 28.68 m, centred RMSE 26.03 m, MAE 14.67 m and efficiency 85.7%. The baseline configuration has correlation 0.9102, bias +18.04 m, MAE 24.04 m and efficiency 77.2%. These results indicate substantial covariability with the model reference and a positive depth offset.



715 *Author contributions.* Lucile Gaultier and Parth Tripathi contributed equally to this work and share first authorship. Lucile Gaultier defined and selected the input variables, participated in the design of the reconstruction strategy, designed the validation methodology and implemented the validation framework. Parth Tripathi developed the reconstruction methodology, implemented the algorithm and participated in the validation. Fabrice Collard contributed to scientific discussions on the MLD reconstruction strategy and on the definition and selection of the input variables. Clément de Boyer Montégut provided and prepared the initial Argo-based MLD dataset, including its extension to the most recent years. Lucile Gaultier and Parth Tripathi wrote the manuscript, and Fabrice Collard and Clément de Boyer Montégut reviewed it.

Competing interests. The authors declare that they have no competing interests.

725 *Acknowledgements.* We thank Marie-Hélène Rio for the supervision of the project as ESA Project Officer for UpperDyn. The Argo data were collected and made freely available by the International Argo Program and the national programs that contribute to it (<https://argo.ucsd.edu>, <https://www.ocean-ops.org>). The Argo Program is part of the Global Ocean Observing System. This study contains modified Copernicus Climate Change Service information (2025) and uses E.U. Copernicus Marine Service Information. The authors used LLM tools to assist with the drafting and editing the manuscript text. The authors also used LLM tools to help write the code necessary to implement the Morlenn algorithm. The authors reviewed all content and take full responsibility for it.

Financial support. This research has been supported by the European Space Agency through the UpperDyn project ESA Contract No. 4000146102/24/I-EB .



730 References

- Argo: Argo float data and metadata from Global Data Assembly Centre (Argo GDAC), SEANOE [data set], <https://doi.org/10.17882/42182>, 2000.
- Ba, J. L., Kiros, J. R., and Hinton, G. E.: Layer Normalization, arXiv:1607.06450, <https://doi.org/10.48550/arXiv.1607.06450>, 2016.
- Bahdanau, D., Cho, K., and Bengio, Y.: Neural Machine Translation by Jointly Learning to Align and Translate, arXiv:1409.0473, <https://doi.org/10.48550/arXiv.1409.0473>, 2014.
- Bai, S., Kolter, J. Z., and Koltun, V.: An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling, arXiv:1803.01271, <https://doi.org/10.48550/arXiv.1803.01271>, 2018.
- Belcher, S., Grant, A., Hanley, K., Fox-Kemper, B., Van Roekel, L., Sullivan, P., Large, W., Brown, A., Hines, K., Calvert, D., and Rutgers-
 740 son, A.: A global perspective on Langmuir turbulence in the ocean surface boundary layer, *Geophysical Research Letters*, 39, L18 605, <https://doi.org/10.1029/2012GL052932>, 2012.
- de Boyer Montégut, C., Madec, G., Fischer, A. S., Lazar, A., and Iudicone, D.: Mixed layer depth over the global ocean: An examination of profile data and a profile-based climatology, *Journal of Geophysical Research: Oceans*, 109, C12003, <https://doi.org/10.1029/2004JC002378>, 2004.
- Gasparin, F., Cravatte, S., Jouanno, J., Hans, A. C., Kestenare, E., Gévaudan, M., Llido, J., and Brandt, P.: Diurnal oceanic thermal variability
 745 in the Atlantic cold tongue: dependence on wind speed and background conditions, *Journal of Geophysical Research: Oceans*, 130, e2025JC022 488, <https://doi.org/10.1029/2025JC022488>, 2025.
- Gaultier, L., Tripathi, P., Collard, F., and de Boyer Montégut, C.: Morlenn: three-hourly North Atlantic mixed-layer depth and maximum sub-mixed-layer stratification, 2011–2023, SEANOE [data set], <https://doi.org/10.17882/120058>, 2026.
- Gille, S. T.: Diurnal variability of upper ocean temperatures from microwave satellite measurements and Argo profiles, *Journal of Geophysical
 750 Research: Oceans*, 117, C11 027, <https://doi.org/10.1029/2012JC007883>, 2012.
- Good, S.A.; Embury, O.: ESA Sea Surface Temperature Climate Change Initiative (SST_cci): Level 4 Analysis product, version 3.0, NERC EDS Centre for Environmental Data Analysis, 09 April 2024., <https://doi.org/10.5285/4a9654136a7148e39b7feb56f8bb02d2>, 2024.
- Greiner, E., Verbrugge, N., and Zunino, P.: Quality Information Document: Ocean 3D Temperature, Salinity, Geopotential Heights, Geostrophic Currents and 2D Mixed Layer Depth, Tech. Rep. CMEMS-MOB-QUID-015-012, Copernicus Marine Service,
 755 <https://doi.org/10.48670/moi-00052>, issue 2.0, product MULTIOBS_GLO_PHY_TSUV_3D_MYNRT_015_012, 2025.
- Guinehut, S., Dhomp, A.-L., Larnicol, G., and Le Traon, P.-Y.: High resolution 3-D temperature and salinity fields derived from in situ and satellite observations, *Ocean Science*, 8, 845–857, <https://doi.org/10.5194/os-8-845-2012>, 2012.
- He, K., Zhang, X., Ren, S., and Sun, J.: Deep Residual Learning for Image Recognition, arXiv:1512.03385, <https://doi.org/10.48550/arXiv.1512.03385>, 2015.
- Hendrycks, D. and Gimpel, K.: Gaussian Error Linear Units (GELUs), arXiv:1606.08415, <https://doi.org/10.48550/arXiv.1606.08415>, 2016.
- Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., De Chiara, G., Dahlgren, P., Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hogan, R. J., Hólm, E., Janisková, M., Keeley, S., Laloyaux, P., Lopez, P., Lupu, C., Radnoti, G., de Rosnay, P., Rozum, I., Vamborg, F., Villaume, S., and Thépaut,
 765 J.-N.: The ERA5 global reanalysis, *Q. J. Roy. Meteor. Soc.*, 146, 1999–2049, <https://doi.org/10.1002/qj.3803>, 2020.



- Hersbach, H., Bell, B., Berrisford, P., Biavati, G., Horányi, A., Muñoz Sabater, J., Nicolas, J., Peubey, C., Radu, R., Rozum, I., Schepers, D., Simmons, A., Soci, C., Dee, D., and Thépaut, J.-N.: ERA5 hourly data on single levels from 1940 to present, Copernicus Climate Change Service (C3S) Climate Data Store (CDS) [data set], <https://doi.org/10.24381/cds.adbb2d47>, 2023.
- 770 IOC, SCOR and IAPSO: The international thermodynamic equation of seawater—2010: Calculation and use of thermodynamic properties, Intergovernmental Oceanographic Commission, Manuals and Guides No. 56, https://www.teos-10.org/pubs/gsw/html/gsw_Nsquared.html, 2010.
- Lellouche, J.-M., Greiner, E., Bourdallé-Badie, R., Garric, G., Melet, A., Drévillon, M., Bricaud, C., Hamon, M., Le Galloudec, O., Regnier, C., Candela, T., Testut, C.-E., Gasparin, F., Ruggiero, G., Benkiran, M., Drillet, Y., and Le Traon, P.-Y.: The Copernicus Global 1/12° Oceanic and Sea Ice GLORYS12 Reanalysis, *Frontiers in Earth Science*, 9, 698 876, <https://doi.org/10.3389/feart.2021.698876>, 2021.
- 775 Li, Q., Fox-Kemper, B., Breivik, O., and Webb, A.: Statistical models of global Langmuir mixing, *Ocean Modelling*, 113, 95–114, <https://doi.org/10.1016/j.ocemod.2017.03.016>, 2017.
- Neetu, S., Lengaigne, M., Vincent, E. M., Vialard, J., Madec, G., Samson, G., Ramesh Kumar, M. R., and Durand, F.: Influence of upper-ocean stratification on tropical cyclone-induced surface cooling in the Bay of Bengal, *Journal of Geophysical Research: Oceans*, 117, C12 020, <https://doi.org/10.1029/2012JC008433>, 2012.
- 780 Pauthenet, E., Bachelot, L., Balem, K., Maze, G., Tréguier, A.-M., Roquet, F., Fablet, R., and Tandeo, P.: Four-dimensional temperature, salinity and mixed-layer depth in the Gulf Stream, reconstructed from remote-sensing and in situ observations with neural networks, *Ocean Science*, 18, 1221–1244, <https://doi.org/10.5194/os-18-1221-2022>, 2022.
- Perez, E., Strub, F., de Vries, H., Dumoulin, V., and Courville, A.: FiLM: Visual Reasoning with a General Conditioning Layer, *Proceedings of the AAAI Conference on Artificial Intelligence*, 32, <https://doi.org/10.1609/aaai.v32i1.11671>, 2018.
- 785 Sallée, J.-B., Pellichero, V., Akhondas, C., Pauthenet, E., Vignes, L., Schmidtke, S., Naveira Garabato, A., Sutherland, P., and Kuusela, M.: Summertime increases in upper-ocean stratification and mixed-layer depth, *Nature*, 591, 592–598, <https://doi.org/10.1038/s41586-021-03303-x>, 2021.
- Sérazin, G., Tréguier, A., and de Boyer Montégut, C.: A seasonal climatology of the upper ocean pycnocline, *Frontiers in Marine Science*, 10, 1085 130, <https://doi.org/10.3389/fmars.2023.1085130>, 2023.
- 790 Spreen, G., Kaleschke, L., and Heygster, G.: Sea ice remote sensing using AMSR-E 89 GHz channels, *J. Geophys. Res.*, 113, <https://doi.org/10.1029/2005JC003384>, 2008.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I.: Attention Is All You Need, *arXiv:1706.03762*, <https://doi.org/10.48550/arXiv.1706.03762>, 2017.
- Wu, Y. and He, K.: Group Normalization, *arXiv:1803.08494*, <https://doi.org/10.48550/arXiv.1803.08494>, 2018.