TURBULENCE AND SURFACE WAKES IN A COASTAL AQUACULTURE AREA: INSIGHTS FROM THE SIRENA PROJECT

J.S. Torres[1]*, A. Rodríguez-Santana[1], A.J. González-Ramos[1], A.M. Mancho[2], A. García[1], L. Yubero[2] and A. Marrero-Díaz[1]

[1]ECOAQUA, Universidad de Las Palmas de Gran Canaria, Edificio de Ciencias Básicas, 35017, Las Palmas de Gran Canaria, Spain
[2]CSIC, Instituto de Ciencias Matemáticas, C/ Nicolás Cabrera 13-15, Campus Cantoblanco 28049 Madrid, Spain
Email: jacob.torres@ulpgc.es

Introduction

The SIRENA project aims to develop innovative monitoring tools for aquaculture farms in Gran Canaria, focusing on external and internal environmental processes. This study compares two oceanographic scenarios—one with an island wake and one without—using a multidisciplinary approach combining satellite imagery, high-resolution model data, and in situ turbulence measurements.

Methods

Four seasonal oceanographic campaigns were conducted: March 5, April 7, June 30 and July 28; employing a free-falling turbulence profiler (VMP-250) and a multiparameter EXO2 probe to collect vertical profiles of physical and biogeochemical variables. Surface fields and subsurface profiles were obtained from the Puertos del Estado operational ocean forecast model (≈250 m resolution). Sentinel-2 imagery was used to detect and delineate island wake structures (Figure 1).

Results

Vertical profiles of conservative temperature from in situ measurements and Puertos del Estado model outputs at Station 5 showed discrepancies across seasons (Figure 2). The model generally overestimated temperatures by about 0.3°C in March and surface temperatures in April but underestimated temperatures below 10 meters. On June 30, observed temperatures were up to 2°C higher than the model in upper layers, likely due to inaccuracies in the model's delineation of the island wake boundary. A similar, though less pronounced, pattern was seen on July 28, outside the wake area. Observations showed a warming trend toward summer with stronger stratification later in the year.

Profiles of potential density and turbulent dissipation rate (ε) indicated stronger stratification and higher turbulence below 20 meters on June 30 (with wake) compared to July 28 (without wake). Satellite imagery and model outputs differed notably in the spatial extent of the island wake, affecting sea surface temperature predictions. However, in situ turbulence data were consistent with the wake location observed in Sentinel-2 imagery, supporting the satellite's spatial positioning of the wake.

Conclusions

The Puertos del Estado model shows limitations in accurately resolving island wake features, impacting environmental assessments. These findings underscore the necessity for higher-resolution models, such as the one developed in the SIRENA project, to improve monitoring and management of aquaculture farms in complex coastal environments.

Acknowledgements

This project is developed with the collaboration of the Biodiversity Foundation of the Ministry for Ecological Transition and the Demographic Challenge, through the Pleamar Programme, and is co-financed by the European Union through the European Maritime, Fisheries and Aquaculture Fund.

(Continued on next page)

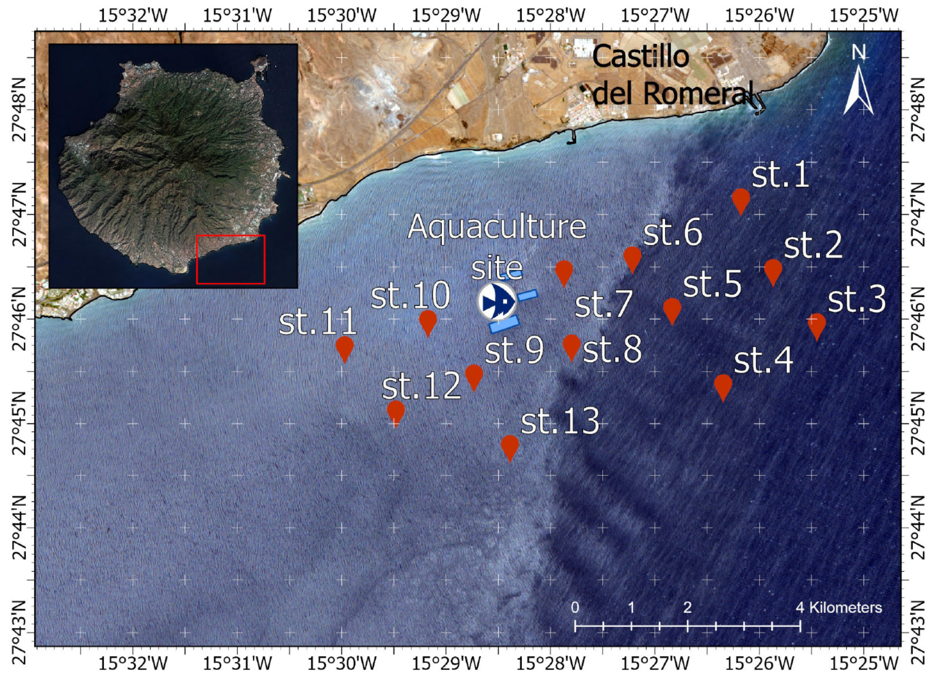


Figure 1. Sentinel-2 image showing surface features associated with the island wake south of Gran Canaria on July 28. Sampling stations and aquaculture sites are marked for reference.

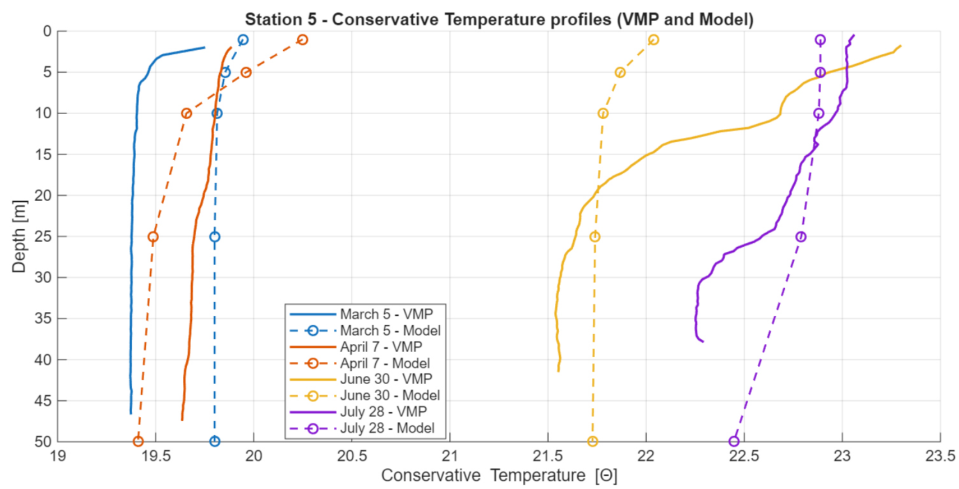


Figure 2. Comparison of in situ conservative temperature profiles (solid lines) and model daily outputs (dashed lines with markers) at Station 5. Colors indicate dates: March 5 (blue), April 7 (orange), June 30 (yellow), and July 28 (purple).

Ocean Modelling 203 (2026) 102782

Contents lists available at ScienceDirect

Ocean Modelling

journal homepage: www.elsevier.com/locate/ocemod

Deep learning weather models for subregional ocean forecasting: A case study on the Canary current upwelling system

Giovanny A. Cuervo-Londoño [a,*], Javier Sánchez [b], Ángel Rodríguez-Santana [a]

[a] Instituto Universitario de Investigación en Acuicultura Sostenible y Ecosistemas Marinos (ECOAQUA), Universidad de Las Palmas de Gran Canaria, Edificio de Ciencias Básicas, Las Palmas de Gran Canaria, 35017, Spain

[b] Instituto Universitario de Cibernética, Empresas y Sociedad (IUCES), Universidad de Las Palmas de Gran Canaria, Edificio de Informática y Matemáticas, Las Palmas de Gran Canaria, 35017, Spain

ARTICLE INFO

Dataset link: https://doi.org/10.48670/moi-00153, https://doi.org/10.17882/46219, https://doi.org/10.48670/moi-00021, https://github.com/gacuervol/regional-graphcast-sst.git

Keywords:
Sea surface temperature forecasting
Graph neural networks
Canary current upwelling system
Data-driven ocean prediction
Operational oceanography

ABSTRACT

Oceanographic forecasting impacts various sectors of society by supporting environmental conservation and economic activities. Based on global circulation models, traditional forecasting methods are computationally expensive and slow, limiting their ability to provide rapid forecasts. Recent advances in deep learning offer faster and more accurate predictions, although these data-driven models entail significant training costs and are often trained with global data from numerical simulations, which may not reflect reality. The emergence of such models presents great potential for improving ocean prediction at a subregional domain. However, their ability to predict fine-scale ocean processes, like mesoscale structures, remains largely unknown. This work aims to adapt a graph neural network initially developed for global weather forecasting to improve subregional ocean prediction, specifically focusing on the Canary Current upwelling system. The model is trained with satellite data and compared to state-of-the-art physical ocean models to assess its performance in capturing ocean dynamics.

Our results demonstrate that GraphCast significantly outperforms conventional baselines, achieving RMSE reductions of up to 76% in 5-day forecasts compared to the GLORYS reanalysis and PSY4V3R1 forecasts, particularly in dynamic coastal zones like Cape Ghir, Cape Bojador, and Blanc (up to 26.5% relative to ConvLSTM). However, we identify that this statistical superiority is partly driven by the model's implicit strategy to minimize the "double penalty" effect — common in chaotic flows — by prioritizing phase accuracy over high-frequency energy preservation. While this approach yields robust error metrics and superior stability, it highlights a critical trade-off between statistical optimization and physical fidelity. These findings validate the disruptive potential of adapting meteorological deep learning models for ocean forecasting but underscore the necessity of physics-informed strategies to bridge the gap between predictive skill and the full dynamic consistency.

1. Introduction

Oceanographic prediction is crucial for understanding climate change and supporting sectors like maritime transport, fisheries, and natural disaster management (Bell et al., 2009). It relies heavily on accurately forecasting mesoscale processes due to their environmental and economic impacts. These processes give rise to distinct structures, influencing mean currents and transporting key ocean properties (Falkowski et al., 1991). Despite its importance, predicting mesoscale processes remains challenging for operational forecasts (Treguier et al., 2017; Mourre et al., 2018), as evidenced by the difficulty in forecasting the Gulf of Mexico Loop Current eddy during the 2010 Deepwater Horizon oil spill (Adcroft et al., 2010; Liu et al., 2013).

Traditional oceanographic prediction techniques rely on numerical models that solve physics-based equations. While early theoretical frameworks utilized quasi-geostrophic (QG) theory, modern operational systems typically discretize the primitive equations, a fundamental approximation of the full Navier–Stokes equations specific to geophysical fluid dynamics. Although these primitive equation models significantly improve upon QG limitations by accounting for variable bathymetric features and surface density gradients (Cushman-Roisin et al., 1990), they still face significant challenges.

* Corresponding author.
E-mail address: giovanny.cuervo101@alu.ulpgc.es (G.A. Cuervo-Londoño).

https://doi.org/10.1016/j.ocemod.2026.102782
Received 27 May 2025; Received in revised form 10 June 2026; Accepted 13 June 2026
Available online 18 June 2026
1463-5003/

In this context, numerical ocean prediction (NOP) models — such as NEMO (Madec et al., 2024), which forms the core of reanalysis systems like GLORYS (Jean-Michel et al., 2021) and operational forecast systems like PSY4V3R1 (Lellouche et al., 2018) — remain the current standard for short-term deterministic forecasting. Nevertheless, they cannot accurately represent reality due to diverse constraints such as incomplete understanding of subgrid-scale parameterizations, poorly known forcing fields, insufficient knowledge of interactions with other Earth system components, and restricted computational resources (Sommer et al., 2018).

Additionally, incomplete observations with spatiotemporal gaps and the limitations of data assimilation schemes — still in continuous improvement — prevent these models from fully capturing the ocean's state. Furthermore, these models do not fully utilize extensive historical data and lack optimization for modern hardware, such as GPUs, which further reduces their efficiency. These limitations highlight the need for novel approaches to improve ocean prediction capabilities.

In recent years, short-term machine learning weather prediction (MLWP) models have emerged in global atmospheric forecasting, surpassing the efficiency and accuracy of traditional numerical systems (Bouallègue et al., 2024). Models such as Pangu Weather (Bi et al., 2023), GraphCast (Lam et al., 2023), Aurora (Bodnar et al., 2024), NeuralGCM (Kochkov et al., 2023), Gencast (Price et al., 2024), or AIFS (Lang et al., 2024) have demonstrated the power of this approach, significantly reducing inference times and computational costs. By leveraging historical data and focusing on spatiotemporal patterns, MLWP models bypass the constraints of incomplete physical understanding and adapt easily to new prediction tasks without architectural modifications (Dueben and Bauer, 2018; Scher, 2018; Bi et al., 2023). While this independence from governing equations limits the diagnosis of physical mechanisms — lacking the dynamic consistency and conservation properties of numerical models — it provides a crucial advantage for operational applications, where rapid and accurate foresight of the ocean state is the primary objective.

Nevertheless, MLWP models heavily depend on data quality and availability, facing challenges with heterogeneous, sparse, spatially discontinuous, or noisy datasets, which can limit performance in certain contexts. Additionally, the physical consistency of these models deteriorates over long timescales due to spectral bias, which can result in numerical instabilities or unrealistic hallucinations (Chattopadhyay et al., 2023).

While MLWP models perform well in atmospheric systems, their application to oceanography presents distinct challenges. Unlike atmospheric data, which is abundant, spatially continuous, and less influenced by external factors, oceanographic data is constrained by atmospheric interference, which affects observational accuracy and consistency, spatial discontinuities, predominantly induced by the presence of continental landmasses, which significantly disrupt data continuity and turn the training of these models complex.

Recent advancements in machine learning ocean prediction (MLOP) have led to the development of innovative models at global and regional scales. Xihe (Wang et al., 2024), for instance, has been introduced for global ocean forecasting, while SeaCast (Holmberg et al., 2024) and OceanNet (Chattopadhyay et al., 2024) focus on regional applications. Similarly, progress in mesoscale ocean forecasting, including studies on eddy shedding predictability in the Gulf of Mexico (Zeng et al., 2015) and data-driven turbulence forecasting using autoregressive techniques (Chattopadhyay et al., 2021), has shown promising results.

The Canary current upwelling system (CCUS), the only eastern boundary upwelling system with islands, presents a unique prediction challenge. In this region, intense mesoscale activity results from the interplay between oceanic features, coastal regimes, and bathymetry, driving strong mesoscale stirring and creating a highly dynamic environment (Sangrà et al., 2009; Arístegui et al., 1994; Barton et al., 1998). These dynamics create a complex environment for sub-regional ocean prediction. State-of-the-art MLOP models have not been applied to medium-term forecasts in this region, highlighting the need for novel approaches.

This work adapts the GraphCast model (Lam et al., 2023) for sub-regional ocean forecasting to evaluate its performance under ocean-specific conditions. Using sea surface temperature (SST) as a case study, we introduce modifications to handle challenges like spatial discontinuities and satellite-based data, including spatially masked loss functions. We assess whether the model can capture mesoscale features such as upwelling, potentially addressing the high computational demands required by traditional ocean models to accurately resolve frontal dynamics and subgrid-scale processes.

In this context, SST is useful for studying the surface expression of mesoscale structures (Hausmann and Czaja, 2012), as eddies, fronts, and filaments create distinct thermal signatures, i.e., warm/cold rotating cores. This study uses the Copernicus Marine Service L4 SST reprocessed product (1982–2020), spanning 39 years, for a subdomain within the IBI region (Iberia, Biscay, Ireland) to train and validate ocean models. The dataset provides daily, gap-free sea surface temperature fields, derived from multiple intercalibrated satellite sources.

We focus on the complex CCUS region, which features strong mesoscale turbulence, persistent upwelling fronts, and energetic eddies. While data-driven approaches do not explicitly resolve the causal barotropic or baroclinic instabilities driving these processes, we evaluate how well ML-based models capture their resulting spatiotemporal manifestations over different forecast times, assessing their sensitivity to mesoscale features in challenging areas like oceanic islands and coastal capes. In the experimental results, we compare the ocean-adapted GraphCast-based graph neural network (GNN) and ConvLSTM-based model (Shi et al., 2015) against NEMO-based reanalysis and forecast products (Jean-Michel et al., 2021; Lellouche et al., 2018)—specifically GLORYS and PSY4V3R1—in terms of predictive skill, computational efficiency, and suitability for short- to medium-term SST forecasting. We focus on the complex CCUS region, which features strong mesoscale turbulence, persistent upwelling fronts, and energetic eddies. While data-driven approaches do not explicitly resolve the causal barotropic or baroclinic instabilities driving these processes, we evaluate how well ML-based models capture their resulting spatiotemporal manifestations over different forecast times, assessing their sensitivity to mesoscale features in challenging areas like oceanic islands and coastal capes.

Additionally, the study assesses the ability of these models to capture seasonal and interannual variability in coastal upwelling systems, where mesoscale processes often dominate. Model sensitivity to observational errors and initial conditions is also analyzed, highlighting their influence on forecast degradation. Finally, the work explores operational implications and proposes strategies to bridge the gap between data-driven approaches and traditional numerical paradigms, emphasizing integration of physical constraints and improving resolution in coastal environments.

Our model's superior performance in SST forecasting is evidenced by >76% and 48% reductions in RMSE over GLORYS at 5- and 10-day lead times, respectively, and by a computational speed nearly ~100 × faster (20-day forecasts in 2.3 min vs. 7-day forecasts in 4 h). However, deeper analysis reveals that this statistical precision partially stems from a trade-off where the model prioritizes phase accuracy over energetic sharpness, effectively mitigating the "double penalty" effect that hampers numerical systems. Coupled with identified sensitivities to initial condition errors — 8.3% increase in days exceeding instrumental error thresholds (2017–2020) — and mesh-induced artifacts, these findings underscore that achieving operational maturity requires moving toward physics-informed training strategies capable of bridging the gap between statistical minimization and true dynamic fidelity.

Section 2 presents the study area of the CCUS and describes the L4 SST dataset used in our experiments. Section 3 introduces the predictive approaches we evaluate in this work, including the GLORYS12V1 reanalysis, ConvLSTM architecture, and our proposed graph neural network. The experimental results, presented in Section 4, assess the performance of the models, with particular emphasis on regions exhibiting intense mesoscale activity and the principal capes. Section 5 analyzes the strengths and limitations of the methods, addresses systematic errors in coastal versus open-ocean zones, and evaluates model robustness under dynamic oceanic conditions. Finally, Section 6 synthesizes the key contributions of this work, outlines practical implications for operational oceanography, and proposes future research directions.

2. Area of study and dataset

2.1. The Canary current upwelling region

The study focuses on a subdomain within the Canary Current Large Marine Ecosystem, specifically the Moroccan subregion (Arístegui et al., 2009), which extends from 21°S to 33°N between Cape Sim and Cape Blanc; see Fig. 1. This region includes two distinct meridional upwelling zones, as described by Cropper et al. (2014): (i) the 21–26°N zone, characterized by strong, permanent upwelling throughout the year; and (ii) the 26–35°N zone, where upwelling remains permanent but is weaker, intensifying during summer due to the seasonal migration of trade winds. The contrast between these two upwelling zones, both in intensity and seasonal variability, highlights the complexity of the Moroccan subregion and underscores its importance for accurately predicting sea surface temperature dynamics.

The region experiences upwelling-favorable winds year-round, with peak intensity during summer due to the northward migration of the Azores High (Wooster et al., 1976). Additionally, the region exhibits pronounced mesoscale oceanographic variability driven by geographic heterogeneity, including variations in continental shelf width, prominent capes, and perturbations induced by the Canary Islands, which generate filaments and eddies.

The interplay of interconnected physical processes governs the coastal upwelling dynamics off northwest Africa, particularly around Cape Ghir. Prominent capes, such as Ghir, Sim, and Cantin, serve as critical control points where topography and bathymetry modulate atmospheric and oceanic flows. Wind forcing, the primary driver of upwelling, intensifies Ekman transport between Cape Ghir and Cape Sim, injecting positive relative vorticity (Troupin et al., 2012) and enhancing the upwelling of cold, nutrient-rich subsurface waters through shear-driven turbulent mixing (Estrada-Allis et al., 2023). The Cape Ghir Plateau, a submarine projection of the High Atlas orogeny on land, deflects the coastal jet offshore, inducing a potential vorticity imbalance that drives the formation of the characteristic filament (Hagen et al., 1996).

This filament displays a dual structure: a cold, surface-intensified core with temperature minima and chlorophyll maxima, surrounded by a broader domain of less dense water influenced by anticyclonic interthermocline eddies (ITEs) (Sangrà et al., 2015). These recurrent ITEs, located north of the filament, primarily affect the subsurface structure but significantly influence the surface dynamics by compressing the upwelling front and intensifying lateral thermal gradients. Similarly, irregular bathymetry triggers bifurcations in the coastal jet that evolve into cyclonic eddies (Hagen et al., 1996), manifesting in SST imagery as distinct, detached cold-core anomalies. Furthermore, while the filament's subduction highlights its role in deep mass export (Sangrà et al., 2015), this process is identified at the surface by the abrupt termination of the cold water tongue and the associated erosion of its thermal signature.

Other capes, such as Jubi and Bojador, exhibit similar dynamics where the coastal wind angle and bathymetric irregularities locally intensify upwelling. Atmospheric forcing, topographic constraints, and vorticity adjustments collectively sustain the biological productivity and mesoscale variability unique to the CCUS (Pelegrí et al., 2005).

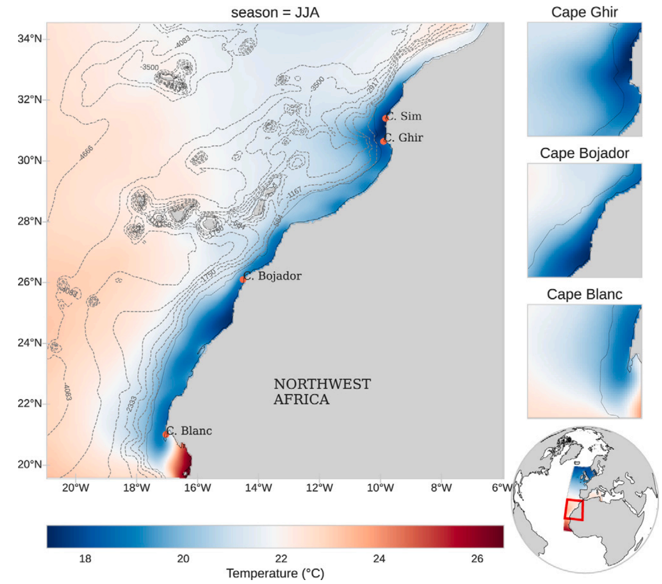


Fig. 1. Summer (JJA) climatology of sea surface temperature (°C) in the Northwest Africa region, highlighting the coastal upwelling system. The main panel displays the spatial distribution of temperature along the coast, with three prominent upwelling centers: Capes Sim and Ghir (~30°N), Cape Bojador (~26°N), and Cape Blanc (~21°N). Dashed gray contours represent isobaths, indicating bathymetric features relevant to upwelling dynamics. Insets on the right provide zoomed views of each cape to better illustrate localized thermal gradients. The bottom-right globe indicates the IBI (Iberian-Biscay-Ireland) domain of the Copernicus dataset, with a red bounding box marking the subdomain corresponding to the study area shown in the main panel.

2.2. High-resolution L4 sea surface temperature reprocessed

SST data is invaluable for validating models as it evaluates air–sea interactions and diagnoses the consequences of vertical mixing while providing insights into the accuracy of model parameterizations and external forcing fields (Mourre et al., 2018). In this study, we use the L4 SST reprocessed product (SST_ATL_SST_L4_REP_OBSERVATIONS_010_026) from the European Union Copernicus Marine Service (CMEMS) for the Atlantic Ocean around Iberia, Biscay, Ireland (IBI), and the northwestern European shelf domain (CMEMS, 2024). It covers nearly 40 years of daily SST data collected by satellites from 1982 to 2020. This high-resolution product, at 0,05 degrees resolution (≈5,55 km), covers the entire IBI domain (≈17 442 538 km^2), ranging from 8,93° to 61,98° latitude and −20,97° to 12,98° longitude. It is provided in NetCDF-4 format and is represented in a standard coordinate system (WGS 84/World Mercator). This resource is produced by Ifremer in France and is updated annually.

The L4 product is built from the L3S product SST_ATL_PHY_L3S_MY_010_038 using the inter-calibration method described in Piollé and Autret (2023). Satellite measurements of the SST come from various sources, such as NASA, NOAA, EUMETSAT OSI-SAF, and ESA. These sources are combined using this inter-calibration to create a unified dataset. However, it is important to note that since the L4 dataset is a blended product derived through spatial interpolation, it inherently smooths high-frequency mesoscale features compared to the original Level-3 (L3) single-sensor observations.

For each day, a correction for SST values is estimated to account for discrepancies between the satellites to ensure a consistent daily dataset. For each satellite, a large-scale bias field is determined by comparing the observations of the satellite with the daily reference field. This bias is then smoothed using a Gaussian filter, which helps reduce noise and other irregularities. The smoothed bias is subtracted from the original SST values, resulting in adjusted temperatures that are more accurate and reliable. This adjustment is essential for correcting systemic errors and improving the overall quality of the data.

Once the SST values are adjusted, the single-sensor composite files, L3C, are combined into a multi-sensor composite file, L3S. This merging process ranks sensors based on their accuracy, determined through comparisons with direct measurements. Each cell in the final grid contains data from the best sensor available, ensuring the highest quality SST values.

Our study area is a subdomain within the IBI region, ranging from 19,55°N to 34,525°N and 20,97°W to 5,975°W, covering an area of approximately 2 462 475 km^2. This region is represented by a grid of 300 × 300 cells. The temporal range of the data used spans from January 1, 1982, to December 31, 2020, corresponding to a total of 14 245 frames (daily images) and a storage size of 10,25 GB. The final L4 product aims to depict a gap-free daily mean sub-skin SST field in Kelvin at a depth of 20 cm. Therefore, there are no considerable differences between the surface and potential temperatures at this depth.

We preprocessed the dataset by handling missing values for two purposes: first, we used them to calculate a static binary land-water mask, which we smoothed using a Gaussian filter; second, we filled the missing values with the average SST value. We used the average because the model requires complete input data without gaps, and the mean provides a neutral value that avoids introducing strong gradients or artificial patterns.

2.3. In-situ validation dataset: EasyCORA

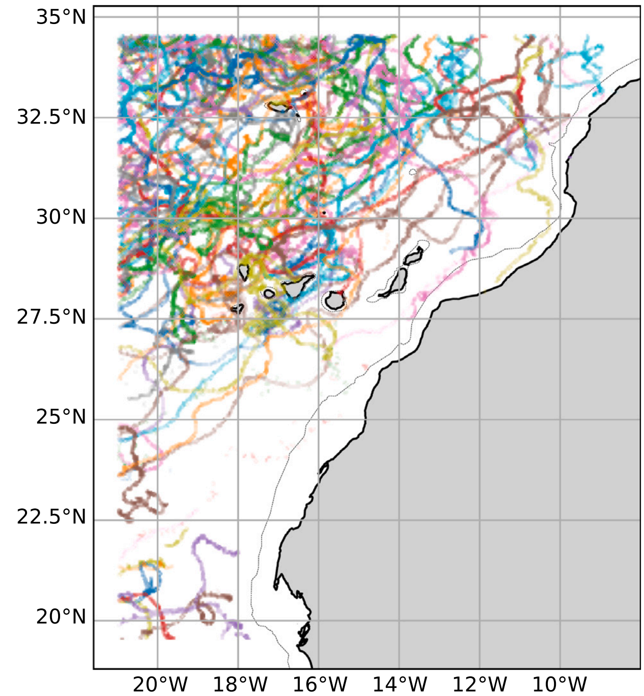
To guarantee a fair and rigorous assessment of the predictive skill across the different forecasting methods, it is essential to validate the model outputs against an independent ground truth, distinct from the satellite-derived L4 product used for training. Relying solely on the L4 dataset for verification could introduce biases, favoring models that replicate the specific interpolation patterns or smoothing inherent to the satellite processing rather than capturing physical reality. Therefore, to complement the satellite-derived observations and establish an unbiased baseline for comparison, in-situ sea surface temperature measurements were obtained from the CMEMS productINSITU_GLO_PHY_TS_DISCRETE_MY_013_001 (CMEMS, 2025).

This study specifically utilizes the EasyCORA dataset, a refined subset of the CORA-GLOBAL reanalysis, optimized for scientific applications through standardized vertical and temporal subsampling. Unlike the satellite data used for model training, the in-situ dataset was specifically retrieved for the evaluation period, spanning from January 2017 to December 2020. The spatial domain of this dataset matches the study area described in Section 2.1, covering the CCUS region.

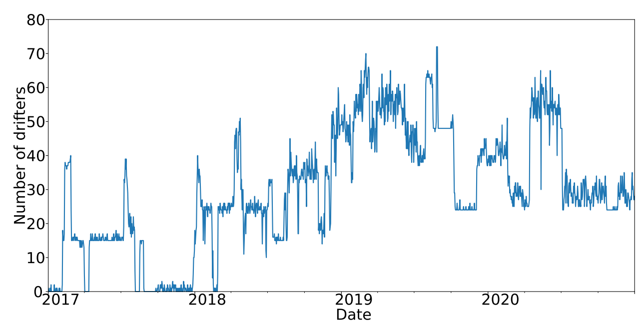
The data selection process targeted the EasyCORA drifter time series files (TS_DR). To mitigate data redundancy and temporal autocorrelation in the validation set, a 1-hour temporal subsampling window was applied, selecting a single representative measurement per window based on a quality weighting algorithm. Given that the primary variable for evaluation is SST, measurements were extracted from the first vertical bin of the EasyCORA grid, corresponding to a 1-meter thickness for the upper layer of the water column, which aligns with the sub-skin temperature representations of the L4 products used in this work.

Rigorous quality control was maintained by filtering observations according to the CMEMS quality flags. Only data points with a Quality Flag of 1 ("good data") or 2 ("probably good data") were retained, while measurements flagged as 3 or 4 were discarded. In accordance with the dataset's technical standards, the adjusted temperature variable (TEMP_ADJUSTED) was prioritized to ensure that the ground truth incorporates the most accurate delayed-mode calibrations provided by the data assembly centers.

It is crucial to acknowledge, however, that the Lagrangian nature of the dataset introduces inherent spatiotemporal inhomogeneities. Spatially, drifters tend to accumulate in physical convergence zones



(a) Position map.



(b) Number of drifters.

Fig. 2. Spatiotemporal distribution of the drifting buoy dataset used for validation (2017–2020). (a) Trajectories of the drifters within the study region, illustrating the coverage of coastal and open ocean areas. (b) Time series of the number of active drifters available per day during the evaluation period. The fluctuation in data availability highlights the irregular temporal sampling inherent to Lagrangian observations.

— such as fronts and mesoscale eddies — while dispersing in divergence regions, leading to an oversampling of specific dynamic features and an undersampling of others (see Fig. 2(a)). Temporally, the observational density exhibits a non-uniform trend, characterized by a progressive increase in the number of active buoys over the evaluation period, as illustrated in Fig. 2(b). This shifting sampling regime implies that interannual performance comparisons must be interpreted with caution.

Finally, the validation takes into account the inherent instrumental uncertainty of the drifter network. Drifting buoys typically exhibit temperature measurement accuracy ranging from 0.002°C to 0.01°C. These high-precision in-situ observations provide a robust and reliable baseline for determining the predictive accuracy of the deep learning model in comparison to established Mercator Ocean products like PSY4V3R1.

3. Forecast methods

This section describes the forecasting approaches used to predict SST in the CCUS region. To ensure a robust assessment, all models are validated against both the satellite-derived L4 SST data (serving as the gridded ground truth) and the independent in-situ drifter measurements described in Section 2.3. We compare our graph neural network against three baselines: two numerical ocean models and one machine learning approach. While bespoke regional numerical configurations tuned for the CCUS could theoretically offer specific improvements, we selected the global systems PSY4V3R1 and GLORYS12V1 as our primary physical benchmarks. These models represent the widely established, state-of-the-art operational standards provided by the CMEMS, thereby offering the most robust and relevant baseline for evaluating the potential of ML models to enhance currently accessible public forecasting services. First, PSY4V3R1, the operational numerical forecast system that generates daily 10-day predictions using the NEMO platform, represents the state-of-the-art deterministic ocean forecasting. Second, GLORYS, a high-resolution global ocean reanalysis based on the same NEMO framework but enhanced through reanalyzed atmospheric forcing data. Regarding machine learning baselines, we include a ConvLSTM neural network, which combines convolutional layers with long short-term memory (LSTM) (Hochreiter and Schmidhuber, 1997) units to capture spatiotemporal patterns in SST evolution.

Then, we explain our graph neural network in detail, a model that leverages a multiscale graph representation to improve predictions while reducing computational cost. We discuss the modifications to adapt the model for regional oceanography, including adjustments to the spatial structure and loss function optimization. The specific configurations and hyperparameters for the graph neural network and ConvLSTM models are detailed in Appendix A.

3.1. PSY4V3R1 forecast

Since October 2006, the Mercator Ocean PSY4V3R1 system has provided high-resolution global ocean monitoring and forecasting under CMEMS. With a $1/12°$ ($\sim$9 km) horizontal resolution and 50 vertical levels — offering fine-scale detail in the upper ocean — it captures essential oceanic processes for operational use. Built on NEMO v3.1 (Madec et al., 2008), it integrates high-frequency atmospheric forcing from the European Centre for Medium-Range Weather Forecasts (ECMWF) (Lellouche et al., 2018). Its data assimilation scheme, which combines a reduced-order Kalman filter with 3D-VAR bias correction (Brasseur and Verron, 2006), ingests satellite altimetry, SST, sea ice concentration, and in situ TS profiles (Lellouche et al., 2013).

PSY4V3R1 introduces several key upgrades over its predecessor, PSY4V2. It corrects atmospheric forcing with satellite data, incorporates freshwater runoff from ice sheet melt, and applies a time-varying steric effect to improve sea level representation. The system refines mean dynamic topography with GOCE geoid data (Rio et al., 2011), enhances coastal accuracy with adaptive observational errors — which modulate the observation error variance to mitigate representativeness issues in shallow waters — and reduces deep ocean drifts by integrating WOA13v2 climatology.

It also strengthens data reliability through improved TS profile quality control, optimized SSH increments, and assimilation of CMEMS OSI SAF sea ice concentrations (Lellouche et al., 2013). By aligning a 2.2 mm/yr global mass trend with contemporary sea-level rise estimates and deriving background error covariance from bias-corrected simulations, the system improves forecasting stability (Chambers et al., 2017).

The system generates 10-day ocean forecasts, including the runtime day itself, meaning it provides projections for nine days into the future (Galloudec et al., 2024). Updated daily at 00 UTC, it uses the PSY4V3R1 model to predict 3D ocean variables (e.g., temperature, salinity, currents) and 2D variables (e.g., sea level, ice thickness, mixed layer). The atmosphere–ocean coupled NEMO model requires at least 5.5 h to complete a 6-day forecast when utilizing 864 processors (Thompson et al., 2021). However, in practice, the data assimilation process used in PSY4V3R1 introduces additional computational overhead, so the total runtime for a full cycle may extend to several more hours, even when using a few hundred cores.

3.2. GLORYS12V1 reanalysis

Jean-Michel et al. (2021) describes GLORYS12 as a high-resolution global ocean reanalysis system based on the ocean and sea ice NEMO models (Madec et al., 2024), starting its simulation in 1991. The system operates NEMO on a quasi-isotropic grid with a $1/12°$ horizontal resolution and 50 vertical levels. The ocean model is coupled with and forced by the ERA-Interim (Dee et al., 2011) atmospheric reanalysis for surface conditions. It also benefits from reanalyzed atmospheric forcing rather than analyses and forecasts, incorporates higher-quality reprocessed observations, and includes refined data assimilation procedures.

GLORYS12 applies the singular evolutive extended Kalman (SEEK) (Brasseur and Verron, 2006) filter method for data assimilation, integrating various sources of information (i.e., satellite sea level anomalies (SLA) (Pujol et al., 2016), satellite SST (Ezraty et al., 2007), and in situ temperature and salinity (T/S) vertical profiles (Cabanes et al., 2013; Szekely et al., 2019)). A 3D-VAR bias correction scheme also estimates large-scale temperature and salinity biases, improving subsurface ocean variability representation.

GLORYS12 utilizes 1296 processors and completes a 7-day simulation in approximately four hours of computational time. However, the total runtime extends to 14 days because the model employs the incremental analysis update (IAU) method (Lellouche et al., 2013) to assimilate corrections.

3.3. ConvLSTM

The Convolutional LSTM (ConvLSTM) cell with peephole connections, introduced by Shi et al. (2015), is a specialized recurrent neural network (RNN) that combines the capabilities of CNNs to extract spatial correlations with the gating mechanisms of peephole LSTMs to capture temporal dependencies. ConvLSTM is widely used in spatiotemporal prediction tasks, including ENSO forecasting (Mu et al., 2019, 2021), nearshore water level prediction (Yang et al., 2024), and tropical cyclone precipitation nowcasting (Yang et al., 2022).

Researchers initially used RNNs for time-series problems but faced challenges with vanishing and exploding gradients in long sequences (Bengio et al., 1994). LSTM networks introduced gating mechanisms to address these issues, enhancing the handling of extended sequential tasks (Hochreiter and Schmidhuber, 1997; Gers et al., 1999; Gers and Schmidhuber, 2000). ConvLSTMs further advanced this by incorporating convolutions into their gates, enabling the simultaneous capture of temporal and spatial features.

However, modern LSTM architectures often omit peephole connections, reducing the parameter number without significantly impacting performance (Greff et al., 2017). This adjustment aligns contemporary ConvLSTM implementations more closely with the standard LSTM cells. ConvLSTM replaces traditional matrix multiplications with convolutional operations in input-to-state and state-to-state transitions. This design enables the model to effectively capture spatial features by analyzing local neighboring inputs and states.

The input and forget gates regulate data flow by controlling how new data integrates with previous states to generate an updated cell state, while the output gate determines the current output. Let $\mathbf{x}^t$ be the SST map for a given time instant t, the ConvLSTM equations are given by:

$$\mathbf{i}^t = \sigma(\mathbf{W}_{ii} * \mathbf{x}^t + \mathbf{W}_{hi} * \mathbf{h}^{t-1} + \mathbf{b}_i),$$
$$\mathbf{f}^t = \sigma(\mathbf{W}_{if} * \mathbf{x}^t + \mathbf{W}_{hf} * \mathbf{h}^{t-1} + \mathbf{b}_f),$$

$$\mathbf{g}^t = \tanh(\mathbf{W}_{ig} * \mathbf{x}^t + \mathbf{W}_{hg} * \mathbf{h}^{t-1} + \mathbf{b}_g),$$
$$\mathbf{o}^t = \sigma(\mathbf{W}_{io} * \mathbf{x}^t + \mathbf{W}_{ho} * \mathbf{h}^{t-1} + \mathbf{b}_o),$$
$$\mathbf{c}^t = \mathbf{f}^t \odot \mathbf{c}^{t-1} + \mathbf{i}^t \odot \mathbf{g}^t,$$
$$\mathbf{h}^t = \mathbf{o}^t \odot \tanh(\mathbf{c}^t),$$

where $\mathbf{i}^t$ represents the input gate, which determines how much new information from the input sequence $\mathbf{x}^t$ and the previous hidden state $\mathbf{h}^{t-1}$ is allowed into the memory cell. The cell state $\mathbf{c}^t$ acts as a memory that retains relevant information over time and is updated based on the input gate and the forget gate $\mathbf{f}^t$, which controls the amount of information from the previous cell state $\mathbf{c}^{t-1}$ to retain or discard. The output gate, $\mathbf{o}^t$, determines how much information from the updated cell state $\mathbf{c}^t$ is passed to the hidden state $\mathbf{h}^t$, which serves as the final output of the cell at the current time step. The activation functions are the sigmoid, $\sigma(\cdot)$, and the hyperbolic tangent, $\tanh(\cdot)$.

Parameters $\mathbf{W}_{ii}$, $\mathbf{W}_{hi}$, $\mathbf{W}_{if}$, $\mathbf{W}_{hf}$, $\mathbf{W}_{ig}$, $\mathbf{W}_{hg}$, $\mathbf{W}_{io}$ and $\mathbf{W}_{ho}$ represent the weight matrices associated with the input, $\mathbf{x}^t$, and the hidden state, $\mathbf{h}^{t-1}$, for each gate, while $\mathbf{b}_i$, $\mathbf{b}_f$, $\mathbf{b}_g$ and $\mathbf{b}_o$ are the biases for the respective gates. The Hadamard product ($\odot$) performs element-wise multiplication, enabling selective gating at each step. While classical LSTMs rely on matrix multiplications, ConvLSTMs replace these operations with convolutions ($*$) within each gate.

3.4. Graph neural network

Our method is based on a GNN model for global medium-range weather forecasting, originally trained on weather reanalysis data, predicting various weather variables globally at a high resolution in under a minute. It is an adaptation of the GraphCast (Lam et al., 2023) model for oceanographic forecasting. This autoregressive model predicts a new state based on two previous time steps. The processing occurs in an underlying multiscale mesh refined from an icosahedron in multiple resolutions. The neural network structure comprises an *encoder*, which embeds the input-grid variables into the mesh nodes, a *processor* that propagates messages through the multiscale mesh, and a *decoder* that maps the forecasting back onto the grid.

The model was originally designed for global atmospheric forecasting and employs two variable types: input and input/target. The model can predict eleven variables: five surface and six atmospheric variables, using data from the ERA5 dataset. These predictions rely on two static input variables, geopotential at the surface and land–sea mask, and five input forcing terms, including solar radiation at the top of the atmosphere and four time-related features.

We adapted this model to predict SST in a local region. This adaptation involved simplifying the model, reducing both the input and input/target variables, and relying on the four time-based features and a single static variable, the land–sea mask. Additionally, to optimize the model for regional use, we replaced the icosahedral mesh with a square curvilinear mesh that represents a curved section of a spherical surface. These modifications highlight the model's versatility and adaptability for different scales and tasks.

The model uses graphs to simulate the relationship between SST values, represented as discrete cells in a grid. It relies on a bipartite graph $\mathcal{G} = \{\mathcal{V}^{\mathrm{g}}, \mathcal{V}^{\mathrm{m}}, \varepsilon^{\mathrm{g}}, \varepsilon^{\mathrm{g2m}}, \varepsilon^{\mathrm{m2g}}\}$, made up of two sets of nodes or subgraphs: $\mathcal{V}^{\mathrm{g}}$, arranged in a grid pattern, and $\mathcal{V}^{\mathrm{m}}$, structured in a planar and regular mesh. Only the nodes in $v_i^m \in \mathcal{V}^{\mathrm{m}}$ have bidirectional connections via edges $SS^m_{s,r} \in \varepsilon^{\mathrm{m}}$. Additionally, these two sets of nodes are connected by edges, $SS^{g2m}_{s,r} \in \varepsilon^{\mathrm{g2m}}$ and $SS^{m2g}_{s,r} \in \varepsilon^{\mathrm{m2g}}$, modeling a directional relationship between the grid and the mesh nodes, and vice versa.

This bipartite graph defines local relationships ($\varepsilon^{\mathrm{g2m}}, \varepsilon^{\mathrm{m2g}}$) between grid node groups $v_i^g \in \mathcal{V}^{\mathrm{g}}$, linked through mesh nodes v_i^m. Additionally, distant multi-scale relationships between these neighborhoods are captured by ε^{m} connections. The number of scales, r, is modeled using a multi-mesh configuration $M^r : \mathcal{V}^{\mathrm{g}}$, defined by embedded mesh refinements $\{M^0, M^1, M^2, \ldots, M^r\}$. Fig. 3 shows an M^2 mesh with red nodes at the coarsest scale with long-range connections, blue nodes at the second scale, and green nodes at the finest scale with short-distant edges.

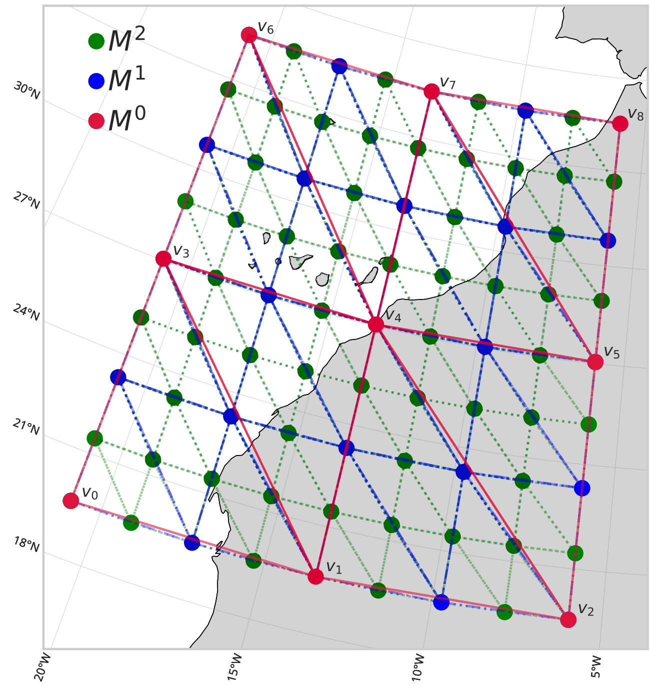


Fig. 3. Representation of a multi-mesh with a refinement factor of $r = 2$, consisting of a total of 81 nodes. The nodes are grouped into three resolution levels: 9 nodes belong to M^0 (the coarsest scale, in red), 16 nodes to M^1 (the intermediate level, in blue), and 56 nodes to M^2 (the finest level, in green). We replaced the traditional icosahedral mesh with a curvilinear mesh based on latitude and longitude coordinates. In this approach, each node's position is explicitly described by its latitudinal and longitudinal values, ensuring a more accurate representation of the Earth's spherical geometry. New nodes are generated by refining the angular midpoint between existing nodes in spherical coordinates, avoiding distortions from planar approximations and improving spatial accuracy for regional-scale modeling.

The model uses a learnable algorithm called Interaction Network (IN) (Battaglia et al., 2016; Watters et al., 2017) to define how nodes in the graph interact with others. This IN is designed to understand relationships in complex systems (Battaglia et al., 2018; Pfaff et al., 2020; Keisler, 2022). At its core, the IN relies on multilayer perceptrons (MLPs), which in the standard GraphCast implementation typically employ a latent size of 512. However, we use smaller latent sizes since our model forecasts a single oceanographic variable. This reduction in latent dimensionality offers a significant advantage, resulting in a more parsimonious model with fewer parameters. Consequently, the model becomes computationally less demanding, leading to accelerated training and inference.

The IN mechanism facilitates sending messages from one sender node s to another receiver node r, allowing information to flow and update the features of both the nodes and the edges across $\mathcal{G}$. The *encoder*, *processor*, and *decoder* work through specific message-passing steps within different parts of $\mathcal{G}$. In the *encoder*, nodes in $\mathcal{V}^{\mathrm{g}}$ send messages to nodes in $\mathcal{V}^{\mathrm{m}}$, $s^g \rightarrow r^m$. The *decoder* then reverses this process, with nodes in $\mathcal{V}^{\mathrm{m}}$ sending messages back to nodes in $\mathcal{V}^{\mathrm{g}}$, $r^g \leftarrow s^m$. The *processor* handles messages between nodes in $\mathcal{V}^{\mathrm{m}}$, $s^m \rightarrow r^m$, performing this step one or more times. Each additional message-passing step increases the number of model parameters, as each step requires an independent IN.

More details about the architecture of the model are given in Appendix B, with the structure of the bipartite graph and the equations of the *encoder*, *processor*, and *decoder*.

3.5. In-situ validation strategy

To evaluate the forecast skill against the Lagrangian in-situ observations described in Section 2.3, a rigorous match-up procedure was implemented to address the dimensional discrepancies between the gridded model outputs and the point-based drifter measurements. Since the model predictions represent discrete grid cells at a daily resolution, while drifting buoys provide continuous point measurements at hourly intervals, a direct comparison requires precise colocation.

To spatially collocate these datasets, two primary methodologies exist. One approach is to integrate the high-frequency Lagrangian observations into the Eulerian model grid by spatial averaging, effectively filtering out sub-grid scale variations prior to comparison. However, this study employs the inverse approach: interpolating the model outputs to the exact spatiotemporal coordinates of each drifter trajectory before calculating the individual biases. We selected this method because it preserves the fine spatial gradients captured by the drifter prior to any spatial aggregation. Furthermore, by relying on the linearity of the expected value, averaging these point-wise differences a posteriori yields an equivalent mean bias to averaging the observations first.

First, we performed a spatiotemporal interpolation to map the gridded model field to the continuous trajectory of the drifters. For a given drifter measurement $x_{obs,k}$ located at coordinates (ϕ_k, λ_k) and time t_k, the corresponding model prediction $\hat{x}_k$ was derived via trilinear interpolation of the surrounding grid nodes v_i^g. This ensures that the comparison accounts for the specific spatiotemporal path of the buoy. Consequently, the point-wise bias was calculated as $\delta_k = \hat{x}_k - x_{obs,k}$.

Second, to prevent statistical distortions caused by the non-uniform spatial distribution of the drifter network — where physical convergence zones may lead to oversampling in specific areas — a spatial binning strategy was applied. Calculating global metrics directly from raw point data would introduce a sampling density bias, disproportionately weighting regions with more instruments. To mitigate this, we discretized the point-wise biases δ_k back into the model's native grid resolution ($0.05° \times 0.05°$). All individual observations falling within the spatial boundaries of a grid cell v_i^g were aggregated, and their biases were averaged to produce a single representative "super-observation" bias, $\bar{\delta}_i$. This approach ensures that every grid cell contributes equally to the final error metrics, independent of the number of drifters present within it.

4. Results

4.1. Metric-based evaluation and quantitative analysis of forecast skill

The performance of the MLOP and NOP models was verified against the SST L4 satellite data and assessed through a comprehensive forecast verification framework, detailed in Appendix C, based on four metrics: Root Mean Square Error (RMSE), Anomaly Correlation Coefficient (ACC), Bias, and Relative Activity (RA). These metrics were calculated following the methodology described in Bouallègue et al. (2024). the ML and numerical models Each score was computed independently for each lead time, up to 20 days, using daily temperature fields. The evaluation was performed on a gridded domain of 300 × 300 points in latitude and longitude, covering the Morocco subregion. For each grid point, the scores were averaged over 365 × 4 × 20 individual forecast realizations, corresponding to daily forecasts generated from 2017 to 2020, the time range of the test set, as shown in Fig. 4. This methodology yields robust statistical estimates of model performance for each lead time.

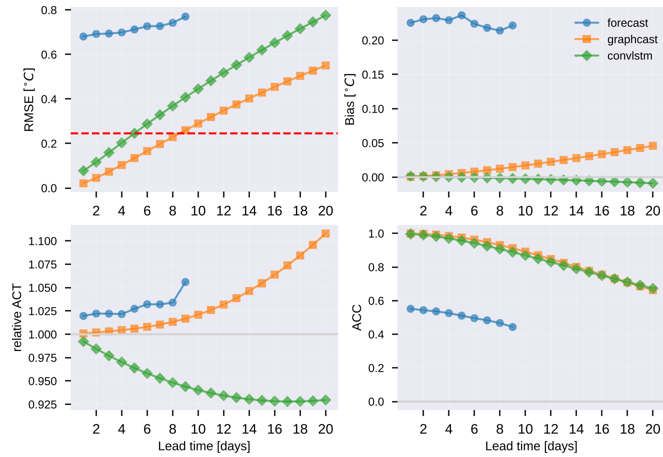


Fig. 4. Performance evaluation of the models for SST, verified against satellite L4 observations. The plots display four key metrics over a 20-day forecast period: RMSE, Bias, Relative ACT, and ACC. The models compared are GraphCast (orange), ConvLSTM (green), and PSY4V3R1 (blue), while the red dashed line in the RMSE panel represents the instrumental error of the satellite data (±0,25°C).

The SST satellite product used for verification contains an instrumental error for each grid cell within the domain and for each day of the test dataset. The average of this instrumental error is ±0,25°C across the entire dataset. This value serves as a reference threshold in the RMSE evaluation, establishing a lower bound for the expected accuracy of the predictive models. In this context, our model surpasses the instrumental error threshold on the eighth day and ConvLSTM on the fifth day, indicating differences in their ability to match the observational accuracy of the satellite product.

A quantitative comparison of the forecast Skill Score (S_{score}) between the two deep learning models was performed using the normalized difference of their scores. This methodology follows the approach described by Geer (2016). The relative RMSE (Eq. (C.2)) and Activity (Eq. (C.3)) percentage errors were calculated as:

$$S_{rmse} = \frac{\left(\overline{\mathrm{RMSE}_B} - \overline{\mathrm{RMSE}_A}\right)}{\overline{\mathrm{RMSE}_B}} \times 100 \tag{1}$$

where A represents the GraphCast model, and B may represents the reference GLORYS or ConvLSTM models. For the ACC (Eq. (C.1)) and bias (Eq. (C.5)) metrics, which can take negative values, the normalized difference was calculated as:

$$S_{acc} = \frac{\left(\overline{\mathrm{ACC}_A} - \overline{\mathrm{ACC}_B}\right)}{\left(\overline{1 - \mathrm{ACC}_B}\right)} \times 100 \tag{2}$$

The spatial averages were computed on the 300 × 300 grid cells in every forecast, and the temporal averages were subsequently calculated for each day. This resulted in 20 values representing the normalized difference for each lead time. Each value represents the average of $300 \times 300 \times 365 \times 4$ data points.

The S_{rmse} and S_{acc} quantify the relative performance of GraphCast compared to ConvLSTM. Those percentages indicate that our model achieves a higher or lower score than ConvLSTM at a given lead time.

Looking at Fig. 4, we observe that, for the initial five lead times, GraphCast demonstrates a 81,7% to 38,4% higher S_{acc} relative to ConvLSTM, while for the subsequent ten lead times (6 to 10), the S_{acc} is 30,9% to 7,8% higher. In terms of S_{rmse}, GraphCast exhibits an error reduction between 74,6% and 30,0% in the first ten lead times, and an improvement of 28,4% to 21,7% in the last ten lead times. These results highlight the enhanced predictive skill of the graph model across short and extended forecast horizons.

Table 1
Percentage of forecasts exceeding the instrumental error threshold for GraphCast and ConvLSTM (2017–2020). The comparison is based on 28 719 RMSE values, above the instrumental error as the baseline.

Year	N	GraphCast %	ConvLSTM %
2017	7070	51,0	72,0
2018	7300	52,0	71,0
2019	7300	57,0	73,1
2020	7049	59,3	75,2

In terms of RA, GraphCast shows overactivity with a monotonically increasing level relative to observations, while ConvLSTM shows underactivity, with a consistently lower and decreasing level across lead times. Despite these differences, both models maintain a low mean bias across all forecast lead times. ConvLSTM's bias remains virtually zero (< 0009 °C), whereas GraphCast's bias increase gradually up to 0,05°C at the last lead time. In both cases, the bias stays well within the instrumental uncertainty. Note that these biases are averages over the spatial domain and all initialization dates yielding a naively results.

4.2. Interannual model performance

We can analyze the precision of the methods using barrier plots, where the y-axis represents the lead time, the x-axis represents the forecast day, and the color scale indicates the RMSE value. These plots depict 365 × 4 × 20 predictions, where each point represents the latitude-weighted spatial average of the 300 × 300 grid cells for each day. Fig. 5 compares the barrier plots of the models from 2017 to 2020.

We can use these barrier plots to assess the percentage of forecast days in which model errors exceeded the mean satellite instrumental error across different lead times and forecast realizations from January 3, 2017, to December 26, 2020. We quantified the proportion of forecast lead time days where RMSE values surpassed the instrumental error threshold (cf. Table 1). Higher percentages indicate periods of reduced model skill, and lower values indicate improved performance.

A notable anomaly occurred on January 1, 2020, during which both models demonstrated unusually low skills. This anomaly coincided with a significant shift in satellite instrumental error, suggesting a potential link between the two events, as illustrated in Fig. 5. The magnitude of this instrumental error shift was quantified to range between 3,2σ to 5,2σ and was compared against the time series of instrumental error variability. This comparison supports the hypothesis that observational uncertainties contributed to the degraded model performance on that day.

Both models recorded the highest percentage in 2020, with 59,3% of RMSE values above the threshold for GraphCast and 75,2% for ConvLSTM. This achieved its lowest value in 2018, at 71%, while GraphCast recorded its lowest in 2017 at 51%. This highlights a period of relatively better performance for GraphCast, particularly in 2017, where it was significantly better than ConvLSTM. The results underscored GraphCast's improved consistency and accuracy over ConvLSTM across the evaluated years.

GraphCast produced a four-year RMSE average of 0,42°C with a standard deviation of 0,13°C. The model showed a slight reduction in 2018 (0,40°C) and the highest increment in 2020 (0,43°C). In contrast, ConvLSTM yielded higher RMSE values, with a four-year average of 0,53°C and a standard deviation of 0,2°C. The model exhibited its lowest RMSE in 2019 (0,52°C) and the highest value in 2018 (0,54°C).

Additionally, we calculated the skill RMSE of GraphCast compared to ConvLSTM, following the same methodology described in the previous section. The S_{rmse} shows that GraphCast consistently outperformed ConvLSTM in the test period, with improvements ranging from approximately 19,4% to 26,5%, with the largest improvement in 2018. GraphCast obtained lower RMSE and more values under the instrumental error threshold, consistently indicating superior predictive performance compared to ConvLSTM in all years.

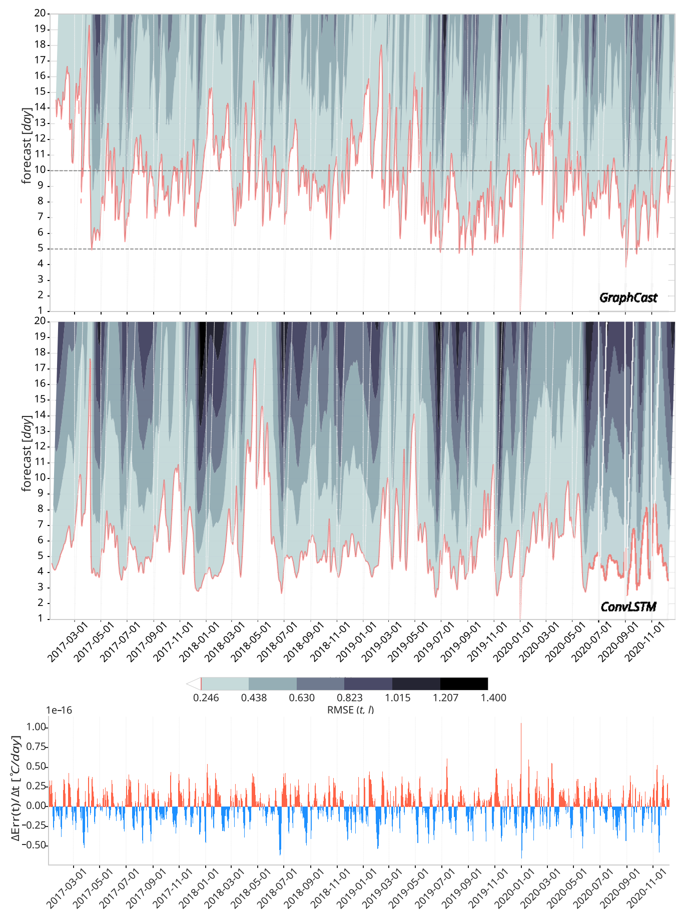


Fig. 5. Predictability barrier plots of daily RMSE for SST forecasts at lead times of 1–20 days from 2017 to 2020, comparing GraphCast (top panel) and ConvLSTM (middle panel). Color shading—from pale (low RMSE) to dark (high RMSE)—represents the evolution of forecast error as a function of lead time and forecast initialization date. The solid red line indicates the mean instrumental-error threshold, while dashed gray lines mark 5- and 10-day lead-time references. All RMSE values are validated against L4 satellite SST data. The bottom panel displays time series of daily instrumental-error anomalies ($\Delta Err/\Delta t$, °C day^{-1}), with red (blue) bars indicating increasing (decreasing) errors. Notably, on 1 January 2020, both models exhibit a simultaneous RMSE spike exceeding the mean instrumental-error threshold, coinciding with a sharp positive anomaly in instrumental error—indicating a transient but substantial loss of predictability.

We compare the RMSE time series of GraphCast for 5-day and 10-day forecast lead times with the RMSE of the GLORYS reanalysis data in the test period, as illustrated in Fig. 6. The results demonstrate significant improvements in forecast skill: for the 5-day lead time, GraphCast achieves a 75,5% reduction in RMSE compared to the reanalysis; similarly, for the 10-day lead time, GraphCast shows a 47% lower RMSE than the reanalysis. These findings highlight GraphCast's superior performance in reducing prediction errors across the test period. Additionally, GraphCast generates a 20-day SST forecast in approximately 140 s on a Quadro RTX 4000 GPU with 8 GB RAM, showcasing its remarkable computational efficiency. In contrast, GLORYS requires approximately 4 h to complete a 7-day simulation using 1296 processors (Jean-Michel et al., 2021). However, this comparison must be normalized by the task complexity: GLORYS solves the full primitive equations for a comprehensive suite of variables (e.g., salinity, currents) across 50 vertical levels, whereas our model targets a

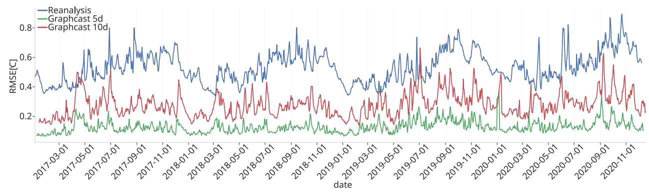


Fig. 6. Time series of RMSE for SST predictions over the test period. RMSE values from GraphCast forecasts at 5-day (green) and 10-day (red) lead times are shown alongside those from the GLORYS reanalysis (blue). GraphCast consistently exhibits lower RMSE throughout the period, indicating reduced prediction error relative to the reanalysis.

Table 2
Seasonal RMSE for GraphCast and ConvLSTM models. The table presents the mean (μ) and standard deviation (σ) of RMSE values for each season: December–January–February (DJF), March–April–May (MAM), June–July–August (JJA), and September–October–November (SON). GraphCast and ConvLSTM performance is compared across seasons, highlighting variations in prediction accuracy and consistency.

Season	GraphCast		ConvLSTM	
	μ (°C)	σ (°C)	μ (°C)	σ (°C)
DJF	0,35	0,07	0,58	0,20
MAM	0,38	0,08	0,40	0,10
JJA	0,43	0,11	0,60	0,20
SON	0,44	0,10	0,50	0,15

single 2D surface variable. Thus, while the numerical model bears a significantly higher computational load per grid point, the data-driven approach demonstrates remarkable efficiency for applications where rapid surface forecasting is the priority.

4.3. Seasonal analysis

To further investigate the seasonal patterns in the model, we used RMSE barrier plots based on a three-month moving window average centered on each day of the year. We apply a 90-day smoothing window to the daily spatially latitude-weighted RMSE values from the barrier plots of each model, spanning the four years of the test set. The window included 45 days before and after the target day, accumulating 90 days, repeated across the four years. Each value represents an average of 300×300×90×4 RMSE values, resulting in 365 × 20 data points per day of the year and per lead time. Fig. 7 shows the seasonal barrier plots.

The seasonal performance of the models was evaluated by calculating the percentage of days within each season where the RMSE exceeded the mean instrumental error. ConvLSTM recorded the highest rate in DJF and JJA, with 19,1% and 20,2% of RMSE values above the threshold, respectively. GraphCast, on the other hand, showed its peak percentage in JJA and SON, with 15,4% and 15,6% of values above the threshold, respectively. GraphCast demonstrated its lowest rate in DJF, at 12%, while ConvLSTM achieved its best performance in MAM, with 17% of values above the threshold. These results highlight seasonal variations in model performance, with GraphCast showing relatively better consistency across seasons than ConvLSTM.

Additionally, the mean RMSE for each season was computed to assess the seasonal performance of the models (cf. Table 2). GraphCast exhibited the highest RMSE of 0,44°C with a standard deviation of 0,1°C during SON, while its lowest RMSE of 0,35°C with a standard deviation of 0,07°C occurred in DJF. In contrast, ConvLSTM produced its highest RMSE of 0,6°C with a standard deviation of 0,2°C in JJA, and its lowest RMSE of 0,4°C with a standard deviation of 0,1°C in MAM. These findings align with the previous analysis, reinforcing that GraphCast consistently outperforms ConvLSTM across seasons, particularly in winter (DJF), where it achieves the lowest rate.

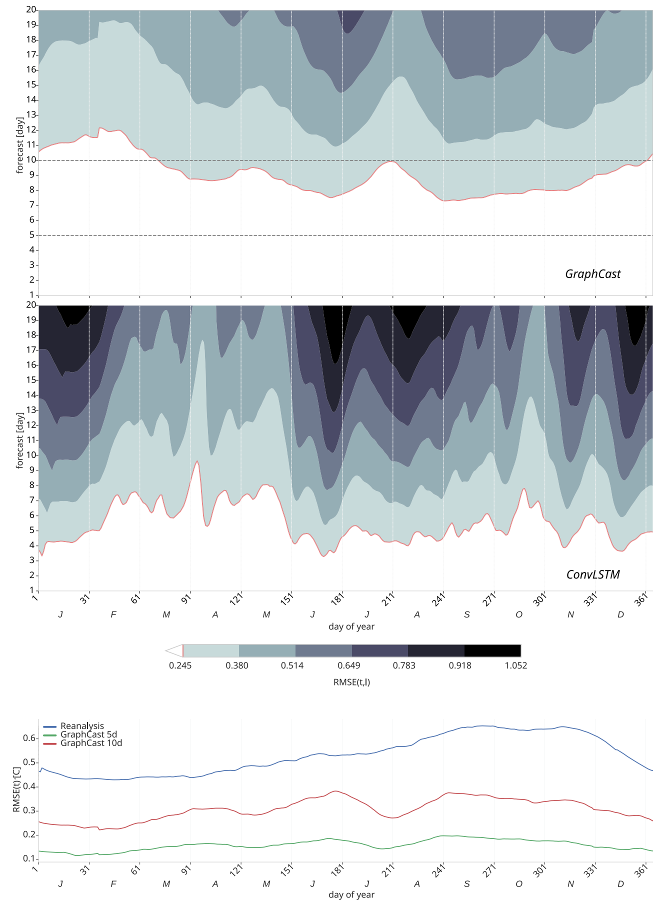


Fig. 7. Seasonal predictability barrier plots showing daily RMSE for SST forecasts at lead times of 1–20 days, averaged using a 90-day moving window (±45 days) to highlight seasonal patterns in forecast skill. The top panel corresponds to GraphCast and the middle panel to ConvLSTM. Color shading — from pale (low RMSE) to dark (high RMSE) — indicates forecast accuracy as a function of lead time and day of year. The solid red line marks the mean instrumental-error threshold, and dashed gray lines indicate 5- and 10-day lead-time references. All RMSE values are validated against Level-4 satellite SST data. The x-axis spans the full calendar year (days 1–365), with each point representing the center of a 90-day moving window, averaged across four years (2017–2020) to capture the seasonal cycle of forecast skill. The bottom panel presents the seasonal cycle of RMSE at selected lead times (5 and 10 days) for GraphCast, along with the corresponding RMSE from the reanalysis (blue), providing a reference for error magnitude across the year. These time series correspond to horizontal slices through the top panel and emphasize periods of relatively higher or lower model skill.

Next, we calculate the relative improvement of GraphCast to ConvLSTM by estimating the S_{rmse} for each season and lead time. The results demonstrate that GraphCast consistently outperformed ConvLSTM across all seasons. The most significant improvement was observed in DJF, with a 38,5% reduction in RMSE, while the smallest improvement occurred in MAM, with a rate of 7%. GraphCast also achieved lower RMSE values in the remaining seasons, with 28,4% in JJA and 10,6% in SON compared to ConvLSTM. These findings align with seasonal performance trends identified earlier, further emphasizing GraphCast's superior accuracy and consistency across all seasons, particularly in winter and summer.

Additionally, the same moving window averaging procedure was applied to the RMSE of the GLORYS reanalysis dataset, verified against satellite SST, to establish a reference climatology for the reanalysis. The seasonal RMSE time series for GraphCast at 5-day and 10-day lead times was then compared to the GLORYS reference, providing a

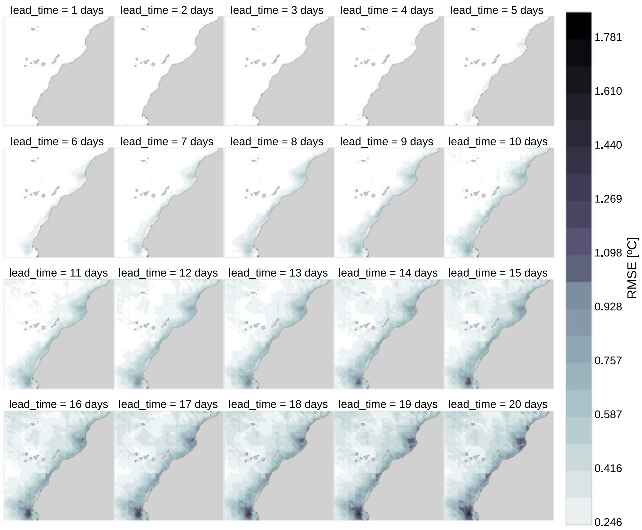


Fig. 8. Point-wise RMSE average for 20 lead times forecasts of the GraphCast model. Each map represents the average RMSE at each point of our area of study for 365 days and 4 years of the test set. The images show the results of the 1 to 20 lead times from left to right and top to bottom.

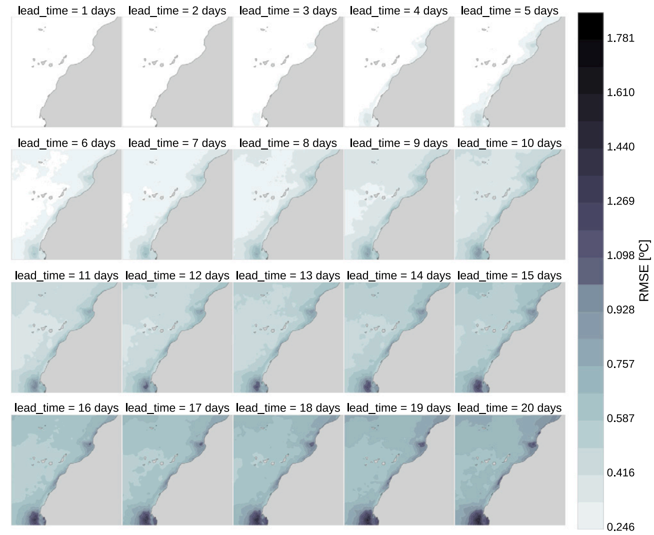


Fig. 9. Point-wise RMSE average for 20 lead times forecasts of the ConvLSTM model. Each map represents the average RMSE at each point of our area of study for 365 days and 4 years of the test set. The images show the results of the 1 to 20 lead times from left to right and top to bottom.

comparative view of how well these forecast lead times reproduced the seasonal cycle, as shown in Fig. 7c. The normalized difference between GraphCast and GLORYS was calculated to quantify the relative skill of GraphCast in capturing the seasonal cycle compared to the reanalysis data.

We computed the S_{rmse} of GraphCast relative to the GLORYS reanalysis for each forecast lead time to evaluate model performance further. The results demonstrate significant improvements in forecast accuracy: for the 5-day lead time, GraphCast achieved the highest reduction in RMSE during DJF and SON, with a rate of 77,4% for both seasons, while a rate of 73,0% and 75,3% were observed for MAM and JJA, respectively. Similarly, for the 10-day lead time, GraphCast showed the highest skill in DJF and SON, at 51,2% and 50,7%, respectively, with 42,4% and 46,0% for MAM and JJA. This comparison highlights the seasonal variability in model performance and emphasizes the extent to which GraphCast outperforms GLORYS across different seasons and forecast horizons.

4.4. Analysis of the spatial accuracy of models forecasts

In this section, we analyze the accuracy of models in the area of study. We generated RMSE maps for each lead time, where each grid cell represents the temporal average of 365 × 4 RMSE values, corresponding to the four years of the test set. This averaging procedure results in 20 RMSE maps. Fig. 8 shows the corresponding error maps for GraphCast and Fig. 9 the maps for the ConvLSTM model.

The first step of the analysis quantifies the percentage of ocean grid cells, out of a total of 49 061, where the RMSE exceeded the mean instrumental error for each lead time and model. This metric provides insight into the spatial extent of regions with significant prediction errors. From the first lead time, ConvLSTM showed 0,9% of cells above the instrumental error threshold, while GraphCast had no cells exceeding this error. By the fourth lead time, GraphCast surpassed the instrumental error in 1,05% of cells, whereas ConvLSTM had a rate of 12,54% of cells above the threshold. Notably, ConvLSTM reached 100% of ocean cells exceeding the instrumental error by the eighth lead time, while GraphCast remained significantly lower at 22,96% for the same lead time. By the twentieth lead time, GraphCast produced 99,01% of ocean cells exceeding the instrumental error. Therefore, GraphCast maintains lower prediction errors across a larger spatial extent and longer lead times than ConvLSTM.

The spatial RMSE average reveals that GraphCast consistently exhibits lower values than ConvLSTM across all lead times. At the first lead time, GraphCast achieves μ = 0,02°C and σ = 0,008°C, while ConvLSTM shows μ = 0,05°C and σ = 0,054°C. However, the standard deviation of GraphCast grows more rapidly than that of ConvLSTM, reaching μ = 0,21°C and σ = 0,094°C by the eighth lead time, compared to ConvLSTM's μ = 0,36°C and σ = 0,088°C. Beyond the eighth lead time, GraphCast's standard deviation remains higher than ConvLSTM, while its mean is lower. At the last lead time, GraphCast achieves a smaller mean (μ = 0,5°C) than ConvLSTM (μ = 0,76°C), but its standard deviation (σ = 0,218°C) is larger (σ = 0,13°C). Therefore, GraphCast maintains lower prediction errors despite its higher variability in later lead times.

We also computed the spatial average S_{rmse} relative to ConvLSTM. The mean reduction in RMSE across all lead times is 43,86%, with the highest reduction of 62,48% at the first lead time and the lowest reduction of 35,73% at the last lead time. This indicates that GraphCast provides consistently small errors over ConvLSTM across all forecast horizons.

The spatial analysis also allows us to identify regions with high RMSE values. Visual inspection of the maps revealed that the areas with the highest values for both models were concentrated near prominent capes, which collectively accounted for more than 18,6% of the total RMSE across all lead times relative to the entire domain. Specifically, Cape Ghir contributed with a rate of 6,8%, Cape Bojador with 4,0%, and Cape Blanco with 7,8% to the overall RMSE. Despite these challenges, GraphCast demonstrated superior performance in these regions compared to ConvLSTM. On average, GraphCast achieved significant reductions in RMSE relative to ConvLSTM, with improvements of 22,2% at Cape Ghir, 24,9% at Cape Bojador, and 28,5% at Cape Blanco. Therefore, GraphCast can handle complex regional dynamics better than ConVLSTM, particularly in areas with high prediction errors.

Finally, we computed the spatial S_{rmse} relative to the time-average GLORYS reanalysis, with an full domain average RMSE of 0,48°C. This metric quantifies the domain-averaged skill in RMSE between GraphCast and the reanalysis. For the 5-day lead time, the GraphCast average RMSE is smaller by a rate of 74,2% relative to the reanalysis across

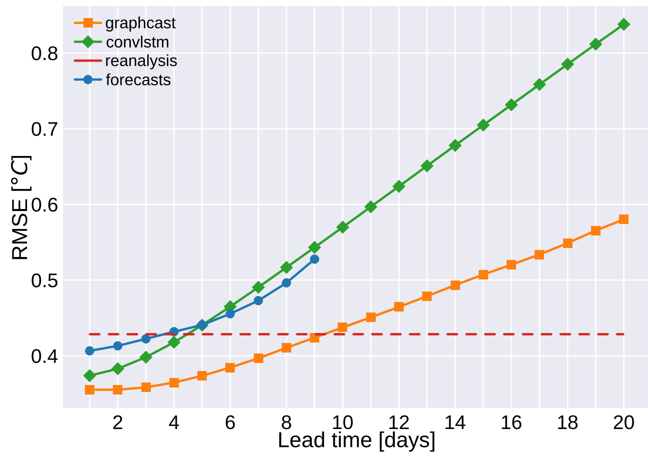


Fig. 10. Evolution of the RMSE of SST forecasts validated against independent in-situ drifter observations. The plot compares the performance of GraphCast (orange squares), ConvLSTM (green diamonds), and the operational forecast PSY4V3R1 (blue circles) over a 20-day lead time. The GLORYS12V1 reanalysis (red dashed line) is included as a baseline reference for error magnitude. GraphCast consistently exhibits the lowest RMSE, outperforming both the deep learning baseline and the numerical operational system.

the entire domain. At the 10-day lead time, the error is reduced by a 44,1%, reflecting diminished but still significant improvements. These results demonstrate GraphCast's ability to outperform the reanalysis benchmark, with higher performance at shorter lead times. The same analysis was applied to each coastal cape and, for the 5-day lead time, GraphCast improved the average RMSE by a 69,7% at Cape Ghir, 78,6% at Cape Bojador, and 78,5% at Cape Blanc relative to the reanalysis. At the 10-day lead time, the reductions were 36,8%, 53,5%, and 51,1%, respectively. This shows that GraphCast provides consistently better estimates than the reanalysis data.

4.5. Evaluation against In-Situ observations

The performance of the models was further validated against independent in-situ measurements from drifting buoys, as detailed in the methodology. Fig. 10 presents the evolution of the Root Mean Square Error (RMSE) as a function of forecast lead time.

Consistent with the validation against satellite data, GraphCast (blue line) exhibits the lowest RMSE across all forecast horizons. Specifically, at the first lead time, GraphCast achieves an RMSE of approximately 0.36°C, maintaining a systematic advantage over the operational physical forecast PSY4V3R1 (0.41°C) and even outperforming the point-wise intrinsic error of the GLORYS reanalysis benchmark (0.43°C). The error growth for GraphCast is monotonic, reaching roughly 0.58°C by day 20. This contrasts sharply with the ConvLSTM model (orange line), which, despite starting with a comparable initial error ($\approx$0.37°C), degrades rapidly to exceed 0.84°C at the end of the forecast period, indicating a limited capacity to retain predictive skill over medium-range horizons.

To disentangle the components of these errors, Fig. 11 displays a Taylor Diagram summarizing the statistical relationship between the model predictions and the drifter observations (Taylor, 2001). The radial distance represents the normalized standard deviation (σ_n), the angular position corresponds to the Pearson correlation coefficient (ρ), and the distance from the reference point (red circle at $\sigma_n = 1, \arccos\rho = 0$) represents the centered RMS difference (RMSD). These metrics are geometrically linked by the Law of Cosines relation $\text{RMSD}^2 = \sigma_m^2 + \sigma_o^2 - 2\sigma_m\sigma_o\rho$, where σ_m and σ_o denote the model and observation standard deviations, respectively. Intuitively, this diagram synthesizes model performance into a geometric proximity: a prediction is considered

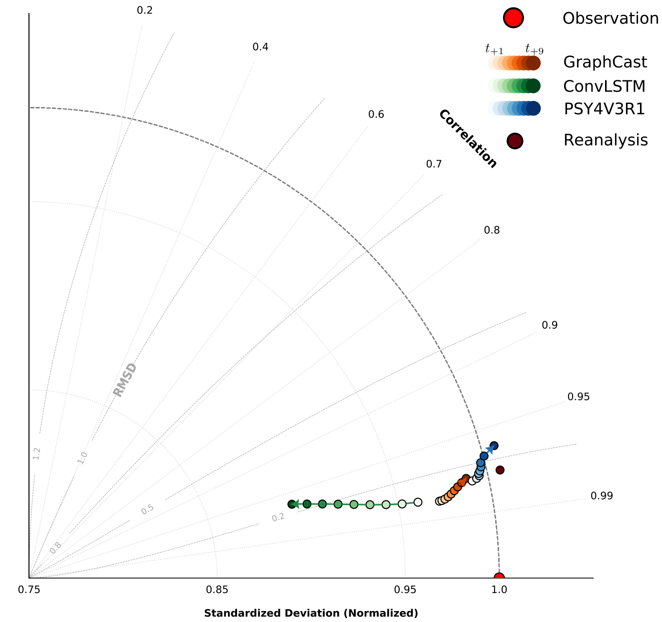


Fig. 11. Taylor Diagram summarizing the statistical performance of the forecast models against in-situ drifter observations. The angular position corresponds to the Pearson correlation coefficient (ρ), the radial distance represents the normalized standard deviation (σ_n), and the distance from the reference point (Observation) indicates the centered Root Mean Square Difference (RMSD). GraphCast (orange) shows the highest correlation, indicating superior phase accuracy, while the numerical models (GLORYS and PSY4V3R1) align closer to the unit circle, indicating a better preservation of physical variance despite larger phase errors.

more accurate the closer it lies to the reference point, which signifies perfect correlation and identical variability amplitude to the observations. Conversely, points situated further away indicate larger errors, stemming either from phase mismatches (lower angular correlation) or from an inability to reproduce the intensity of the observed fluctuations (radial deviations from the unit arc).

The diagram reveals that GraphCast, GLORYS, and PSY4V3R1 cluster relatively close to the reference observation, whereas ConvLSTM disperses significantly as lead time increases. A closer inspection highlights a key distinction: GraphCast consistently achieves the highest correlation with the observations, with coefficients remaining above $\rho > 0.96$ for the initial lead times, which drives its superior RMSE performance. However, the physical models (GLORYS and PSY4V3R1) are positioned closer to the unit circle (dashed arc at $\sigma_n = 1.0$), implying that their variance is nearly identical to that of the observations ($\sigma_n \approx 1$). In contrast, the GraphCast points fall inside the unit circle with normalized standard deviations fluctuating between 0.97 and 1.03, suggesting a slight underestimation of the true variability of the system.

4.6. Visual assessment of forecast quality

To complement the metric-based evaluation, we performed a visual inspection of the spatial SST structures predicted by the models. Fig. 12 displays a qualitative comparison for a specific forecast initialized on November 16, 2020. This date was strategically selected to capture a period of high variability characterized by the southward migration of the trade winds during late autumn. This atmospheric forcing facilitates intense upwelling events along the southern African coast, particularly near Cape Blanc (Mauritania), making it an ideal case study to observe complex thermal dynamics.

The comparison includes the L4 satellite observations (ground truth), GraphCast, ConvLSTM, and the operational numerical forecast (PSY4V3R1). Note that while our deep learning models forecast up to 20 days (t_{+1} to t_{+20}), the comparison with the numerical system is limited to its operational horizon of 9 days (t_{+1} to t_{+9}).

The satellite observations (top row) reveal the characteristic thermal front of the coastal upwelling. Comparing the deep learning models, GraphCast and ConvLSTM exhibit similar large-scale patterns, which is expected as both were trained on the same dataset. However, a closer inspection reveals that ConvLSTM (third row) produces a slightly more smoothed field, losing detail in the fine-scale thermal structures compared to GraphCast (second row). This smoothing leads to a broader, less defined representation of the upwelling front, contributing to the higher bias values observed in the difference maps (right column). Regarding the triangular artifacts discussed in Section 5, it is worth noting that they are not yet apparent at lead time 9 in this case study, as these distortions typically emerge at longer forecast horizons (e.g., t_{+15}).

The numerical forecast (PSY4V3R1, fourth row) presents a distinct behavior. It generates highly defined mesoscale structures that are visually sharper than those in the satellite product or the deep learning forecasts. However, these structures often lack spatial coherence with the observations; for instance, the precise positioning of eddies and filaments does not align with the satellite ground truth.

This discrepancy does not necessarily imply a failure in the model's physics; rather, it highlights the challenges of deterministic prediction in a turbulent flow regime. While the numerical model correctly resolves the hydrodynamical instabilities governing the inverse energy cascade and mesoscale turbulence, the chaotic intrinsic variability of the ocean imposes strict limits on the predictability of the phase (location) of these features. Consequently, a physically realistic but spatially displaced eddy incurs a "double penalty" in standard error metrics: the model is penalized once for missing the observed feature and again for predicting a feature where none exists in the observations (Crocker et al., 2020; Smith and Fortin, 2022).

In contrast, the deep learning approaches, by learning from the smoothed L4 data, tend to filter out these unpredictable high-frequency dynamics. This results in a texture that, while less physically detailed in terms of sharp gradients, minimizes phase errors and effectively avoids the double penalty effect, leading to lower RMSE and bias scores despite the loss of fine-scale structure.

5. Discussion

The results presented in the previous section highlight key differences in performance between the two MLOP models — GraphCast and ConvLSTM — and the GLORYS reanalysis, along with the instrumental error baseline. GraphCast shows consistently lower RMSE and higher ACC values in all lead times, indicating superior predictive capability in the medium-range forecasts.

Between 2017 and 2020, both deep learning models showed an increasing number of days when RMSE surpassed the instrumental error threshold, though with different patterns. GraphCast outperformed ConvLSTM overall but had a sharper rise in exceedances (from 51.0% to 59.3%; $+\Delta 8.3\%$) compared to ConvLSTM's slower increase (from 72.0% to 75.2%; $+\Delta 3.2\%$). This suggests a notable limitation of GraphCast, especially in 2019 ($+\Delta 5.0\%$ interannual change): although it captures fine-scale features well, it is more sensitive to initial errors. Small inaccuracies, such as those from upwelling zones, can grow unpredictably over time, especially under strong mesoscale activity. This pattern corresponds with observed upwelling trends in the California Current Upwelling System (CCUS) (Cropper et al., 2014), where summer upwelling has intensified in the permanent (21–26°N) and weak permanent (26–35°N) zones.

In terms of forecast quality, ConvLSTM reduces sensitivity to upwelling-driven variability by generating smoother outputs through anomaly averaging. This behavior is evident in the RA score: ConvLSTM tends to be underactive, while GraphCast produces noisier, overactive forecasts. However, this smoothing limits ConvLSTM's ability to resolve fine-scale dynamics, making it less suitable for high-resolution applications. By contrast, the GNN captures spatial dependencies more effectively, achieving lower RMSE and higher ACC across most lead times. Yet, GraphCast's higher standard deviation at later lead-times reveals sensitivity to local ocean variability, while ConvLSTM's convolutional structure inherently suppresses small-scale fluctuations. This trade-off prioritizes stability over accuracy—ConvLSTM's visually coherent forecasts lack the granularity to capture localized dynamics, whereas GraphCast's graph-based connectivity preserves fine-scale patterns.

Integrating GraphCast fine-scale representation with ConvLSTM stabilizing smoothing could mitigate error propagation while retaining critical dynamic structures. Their divergent performances stem from fundamental architectural differences: GraphCast graph-based approach resolves ocean patterns more precisely, while ConvLSTM convolutions blur spatially nuanced features.

This trade-off between smoothness and spatial fidelity is further clarified by the combined insight from the Taylor Diagram analysis and the visual assessment. While GraphCast excels in phase accuracy — evidenced by the highest correlation with drifter anomalies — it exhibits a reduced standard deviation compared to the in-situ measurements, positioning it inside the unit circle. This indicates that the model underestimates the high-frequency variability and total energy of the system.

In contrast, the numerical models (GLORYS and PSY4V3R1), despite larger phase errors, align more closely with the observed normalized standard deviation. This suggests they better preserve the physical energy of the ocean, generating sharp, energetic mesoscale structures. However, as shown in the visual assessment (Section 4.6), these structures are often spatially displaced due to the chaotic nature of the flow, leading to a low correlation and a high RMSE driven by the "double penalty" effect (Ebert, 2008). Essentially, the numerical model provides a physically correct image of the ocean dynamics, but one that does not correspond to the actual state of the system at the observed coordinates.

Our deep learning models, conversely, implicitly learn to mitigate this penalty by outputting a smoothed field — driven by the spectral characteristics of the L4 training data — that effectively filters out sub-mesoscale components. By prioritizing phase consistency over structural sharpness, GraphCast "hedges" against the chaotic divergence that penalizes the more physically explicit numerical forecasts.

Therefore, the current limitation in physical consistency is not intrinsic to the deep learning architecture but rather a consequence of the training strategy. To overcome this disadvantage and achieve predictions that are both spatially and physically coherent, future work must move beyond simple point-wise evaluations. Standard metrics like RMSE and MAE remain strictly necessary to ensure comparability against global baselines and standard evaluation frameworks such as WeatherBench (Rasp et al., 2020) or OceanBench (El Aouni et al., 2025) benchmarks. Nevertheless, relying solely on them inherently encourages spatial smoothing to minimize double penalty errors. Future studies should complement these standard metrics with more advanced spatial verification techniques, such as the Earth Mover's Distance (Wasserstein distance) (Vissio et al., 2020; Hyun et al., 2022) or fuzzy verification frameworks (Ebert, 2008). Integrating these spatially-aware metrics into the loss functions of deep learning models will be crucial to ensure that data-driven models learn to replicate not just the location of ocean features, but also their structural realism and sharp spatial gradients.

GraphCast demonstrates the transformative potential of MLOP, achieving 75,5% and 47% RMSE reductions compared to GLORYS at 5-day and 10-day lead times while generating 20-day SST forecasts in just 2,3 minutes. In contrast, GLORYS provides a comprehensive multi-variable ocean state representation, including salinity and velocity components, but requires 4 h for a 7-day simulation. Despite its broad scope, GLORYS incorporates uncertainties stemming from data assimilation methods, parameterizations, and model biases. GraphCast's

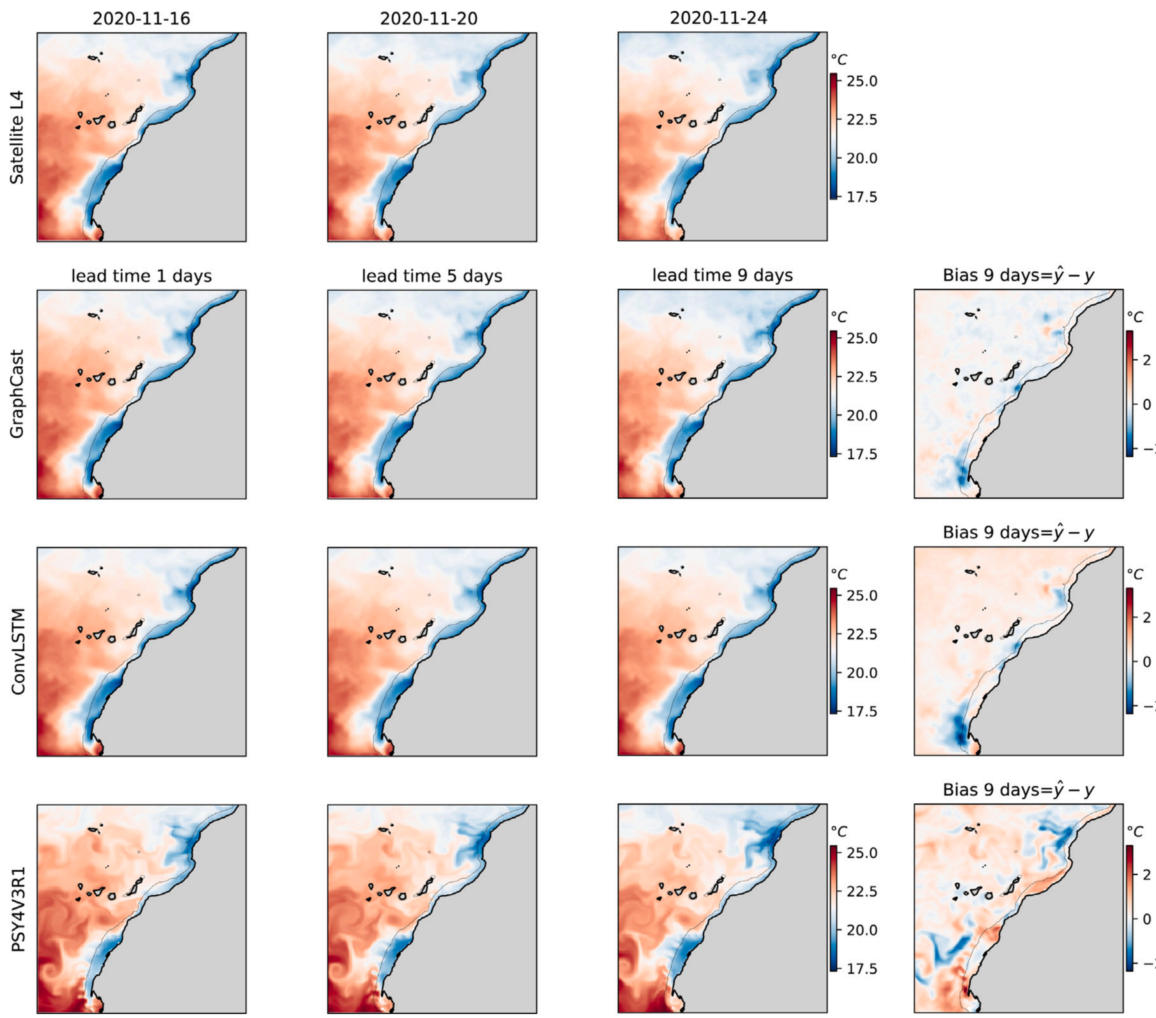


Fig. 12. Visual comparison of SST forecasts at lead times of 1, 5, and 9 days, initialized on November 15, 2020. The rows display: Satellite L4 observations (top), GraphCast, ConvLSTM, and the operational numerical forecast PSY4V3R1 (bottom). The selected period highlights active upwelling near Cape Blanc driven by the seasonal migration of trade winds. The rightmost column specifically displays the spatial bias ($\hat{y} - y$) relative to the Satellite L4 ground truth for the 9-day forecast (valid for November 24, 2020). While GraphCast and ConvLSTM show similar patterns, ConvLSTM produces a notably smoother field. In contrast, the numerical model displays sharp but spatially displaced mesoscale structures. These phase errors result in large localized bias dipoles (blue/red features) on this specific date, illustrating the trade-off between physical sharpness and spatial accuracy.

exceptional speed makes it a powerful tool for real-time forecasting. Expanding its capabilities to include variables like salinity and currents is crucial for broader oceanographic applications.

On January 1, 2020, both MLOP models experienced a sharp drop in forecast skill, coinciding with a significant increase in satellite instrumental error. This highlights the strong dependence of data-driven ocean forecasts on the quality of initial conditions. As noted by Bouallègue et al. (2024), initializing with higher-quality data, such as operational IFS analyses instead of ERA5, can significantly improve forecast accuracy. The substantial observational errors introduced at initialization likely degraded forecast performance, depending on each model's sensitivity to error growth, where the ocean's chaotic nature amplifies small discrepancies. This event underscores the vulnerability of machine learning models to sporadic observational anomalies and reinforces the need for robust quality-control procedures, improved error modeling, and advanced data assimilation strategies.

The seasonal analysis using a three-month moving window average shows that GraphCast consistently outperforms both ConvLSTM and the GLORYS reanalysis, although performance varies by season. The deep learning models exhibit their weakest performance during summer (JJA), when upwelling intensifies due to the seasonal shift of the Azores High and the Inter-Tropical Convergence Zone (ITCZ) (Wooster et al., 1976; Cropper et al., 2014). Despite these challenging conditions, GraphCast reduces RMSE by 28.4% compared to ConvLSTM and by 75.3% relative to GLORYS, suggesting that its graph-based architecture more effectively captures oceanographic processes linked to intensified upwelling.

In contrast, ConvLSTM performs notably worse in JJA due to its convolutional architecture, which smooths the forecasts and filters out high-frequency variability. While this reduces noise, it also suppresses the sharp gradients and nonlinear interactions characteristic of active upwelling regimes, limiting its ability to resolve fine-scale summer dynamics. In winter (DJF), when upwelling is weaker and ocean conditions are more stable, this smoothing becomes less detrimental, and GraphCast's advantage becomes even more pronounced, excelling at capturing steady-state patterns. During spring (MAM) and autumn (SON), both models face similar challenges, and their performance converges.

The spatial distribution of errors reveals high RMSE values near Cape Ghir, Bojador, and Blanco—areas characterized by intense upwelling, complex bathymetry, and coastal current interactions. These conditions pose significant challenges for both numerical and deep learning models, which still struggle to fully resolve filament generation processes driven by wind forcing, coastline irregularities, and mesoscale instabilities. The absence of explicit atmospheric forcings (e.g., wind stress) and bathymetric inputs in the deep learning framework likely limits its ability to disentangle these mechanisms. Additionally, the satellite-derived L4 data used for training inherently smooths mesoscale features due to interpolation and lacks the resolution to capture submesoscale dynamics (<10 km), potentially introducing systematic biases. This aligns with findings from recent studies showing that global climate models underestimate upwelling intensity due to coarse spatial resolution (Bindoff et al., 2019; Docquier et al., 2019), highlighting the need for regional downscaling and high-resolution ocean–atmosphere coupled models (Roberts et al., 2018).

A drawback of GraphCast is the emergence of triangular artifacts at longer lead times, visible as mesh-like patterns superimposed on SST predictions. Visual analysis links these artifacts to the decoder's mesh structure, where each triangular element updates SST values within its influence area. Since neighboring triangles share only one edge and two nodes, inconsistencies arise — especially in dynamic regions like coastal upwelling zones — where adjacent elements may experience different physical processes. This highlights the difficulty of ensuring coherence across mesh elements under heterogeneous ocean conditions.

Mitigating these artifacts presents both computational and architectural challenges. Refining the mesh can improve spatial consistency by reducing each triangle's influence area, as seen in higher-resolution graph-based models (Lam et al., 2023; Oskarsson et al., 2023; Holmberg et al., 2024), but at significant computational cost. Alternatively, increasing node connectivity in the decoder, drawing information from more than three nodes using attention mechanisms, could smooth transitions across triangles mitigating artifacts by distributing updates across a broader mesh neighborhood. A third strategy involves integrating convolutional layers into the decoder to blend localized element outputs, leveraging their proven spatial smoothing effect. This approach, supported by ConvLSTM's artifact-free outputs and lower variability, has been effective in reducing grid artifacts in transformer-based models (Couairon et al., 2024). Future work should explore hybrid architectures that combine adaptive mesh connectivity with convolutional smoothing, preserving GraphCast's ability to resolve large-scale dynamics while enhancing robustness to localized errors.

GraphCast outperforms ConvLSTM and GLORYS in coastal upwelling zones, reducing errors by up to 28.5% in upwelling regions, highlighting its superior ability to capture atmosphere–ocean interactions and fine-scale spatial gradients. However, persistent errors in these areas show that neither numerical nor deep learning models achieve optimal performance without targeted improvements. Despite these challenges, GraphCast's strong SST forecasting skill demonstrates the potential of machine learning for ocean prediction. Advancing toward operational viability will require reducing sensitivity to initial condition errors — especially under strong mesoscale activity — through improved data assimilation or hybrid modeling strategies. Architectural enhancements such as convolutional layers or attention mechanisms may help mitigate artifacts efficiently, while scaling to multivariate forecasts will demand balancing complexity and speed, possibly via latent representations and physics-informed regularization. Addressing regional and seasonal variability will also require adaptive graph structures and real-time forcing assimilation, with focused validation in key regions like upwelling systems to ensure robustness and reliability.

6. Conclusion

This study presented the adaptation and evaluation of two deep learning architectures, GraphCast and ConvLSTM, for forecasting SST in the complex Canary Current Upwelling System. Our results demonstrate that data-driven models, particularly GraphCast, can significantly outperform traditional numerical baselines (GLORYS and PSY4V3R1) in standard error metrics, achieving RMSE reductions of up to 75.5% at a 5-day lead time while reducing computational cost by several orders of magnitude.

However, a critical analysis reveals that this superior performance is driven by a fundamental trade-off between statistical optimization and physical fidelity. Visual and spectral assessments indicate that GraphCast excels at phase accuracy — correctly positioning thermal features — by learning a smoothed representation of the ocean state. This strategy effectively "hedges" against the intrinsic chaotic variability of the flow, minimizing the double penalty effect that disproportionately punishes the sharper, more energetic structures generated by NOP models. Consequently, while operational numerical systems suffer from high RMSE due to spatial misalignments of mesoscale eddies, they arguably preserve a more realistic energy budget than current deep learning models, which tend to filter out sub-mesoscale dynamics to maximize metric performance.

Beyond the spectral analysis, our findings highlighted distinct seasonal and regional performance patterns. While GraphCast excelled under winter steady-state conditions, it exhibited increased sensitivity to initial condition errors during high-variability summer periods. This limitation, coupled with pronounced error concentrations near coastal capes (Ghir, Bojador, Blanco), exposes the vulnerability of data-driven models when emergent dynamics — driven by complex bathymetry and coastal forcing — move beyond the statistical scope of the interpolated L4 training data. Furthermore, while GraphCast's multi-scale connectivity captured spatial dependencies better than ConvLSTM, it developed characteristic triangular artifacts at longer lead times, highlighting the need for architectural refinements in the mesh-grid interaction.

Looking forward, the path toward operational maturity lies in a paradigm shift regarding training strategies and model scope. To bridge the gap between the high phase correlation of MLOP models and the physical realism of numerical systems, future research must move beyond simple RMSE minimization. We propose the integration of physics-informed loss functions and the curation of hybrid training datasets to ensure dynamic consistency. Crucially, such advancements must not only aim for statistical accuracy but also adhere to fundamental physical principles — such as geostrophy, conservation of momentum, and energy — while expanding the prediction scope beyond SST to include currents and salinity.

In summary, this work positions GraphCast as a transformative step forward, offering rapid, high-quality forecasts valuable for real-time decision-making. However, for deep learning models to evolve from statistical approximations into true successors of numerical engines, they must bridge the gap between statistical accuracy and physical fidelity, while scaling their architecture to handle the full systemic complexity of multivariate ocean dynamics. By solving these dual challenges, data-driven approaches have the potential to define the next generation of operational oceanography.

CRediT authorship contribution statement

Giovanny A. Cuervo-Londoño: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Javier Sánchez:** Writing – review & editing, Writing – original draft, Supervision, Resources, Project administration, Methodology, Formal analysis, Conceptualization. **Ángel Rodríguez-Santana:** Writing – review & editing, Supervision, Conceptualization.

Computational library

The methodology is implemented in Python using the JAX library for efficient multi-GPU deep learning training. Atmospheric and oceanographic datasets are processed using Xarray, enabling scalable handling of multi-dimensional arrays, while evaluation metrics are computed with NumPy. The deep learning models are constructed using Google DeepMind Flax and Haiku frameworks, which are optimized for JAX, enabling high-performance model development.

Declaration of Generative AI and AI-assisted technologies in the writing process

The authors used ChatGPT, Gemini, and Deepseek to improve the readability and language of the manuscript during the preparation of this work. After using these tools, the authors reviewed and edited the content as needed and took full responsibility for the content of the published article.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors would like to express their gratitude to Javier J. Sánchez-Medina from Centro de Innovación para la Sociedad de la Información (CICEI) at the Universidad de Las Palmas de Gran Canaria for providing access to their computing servers, which were essential for training the models. We also extend our sincere thanks to Mercator Ocean International (MOi) for sharing the PSY4V3R1 model forecasting data, which was provided upon request. Their support and collaboration were invaluable to the success of this research. Furthermore, this work was supported by the projects SIRENA and SIRENA 2, funded by the collaboration of the Biodiversity Foundation of the Ministry for the Ecological Transition and the Demographic Challenge, through the Pleamar Program, and are co-financed by the European Union through the EMFAF (European Maritime, Fisheries and Aquaculture Fund).

Appendix A. Model configuration, training protocols and experimental setup

A.1. Training details

The dataset was split into training, validation, and test sets using an 80/10/10 ratio, based on years rather than individual days. This temporal division preserves the representativeness of seasonal cycles in each subset. The training set covers 1982 to 2012 (31 years), the validation set spans from 2013 to 2016 (4 years), and the test set includes 2017 to 2020 (4 years). This scheme prioritizes training with historical data while testing is conducted with the most recent information.

The samples in the training set are organized into windows of three consecutive dates, with each window overlapping the next one by a single day. This approach maintains continuity in prediction trajectories while maximizing the use of available data.

We used each set to generate time forcings, providing the model with temporal context for the current day and year. As in Lam et al. (2023), we calculated these forcings as seconds starting at the UNIX epoch from January 1, 1970, and encoded them using sine and cosine transformations to capture cyclical patterns.

For training, we organized the dataset into batches of eight samples. Each batch consists of three components: (i) the input data, which includes the initial conditions for SST over two days; (ii) the target data, representing the SST values for the lead times (days to predict); and (iii) the forcing data, which includes the time-generated forcings for the lead times and the static land–sea mask. To ensure a stable optimization process, we applied shuffling exclusively to the training set, while the validation and test sets remained unshuffled to maintain their temporal structure.

We computed normalization factors, such as the mean, standard deviation, and standard deviation of temporal changes, using the training set. These factors were then applied to normalize the dataset, ensuring consistent scaling across all samples. This preprocessing step improved model stability and enhanced training performance by standardizing the input distribution.

We adopted the training methodologies from Lam et al. (2023), excluding the initial warm-up phase because our experiments did not indicate a significant impact. Our training strategy comprised a 150-epoch training phase, completed in approximately 64 h. We adjusted the learning rate using a half-cosine decay function, gradually reducing it from 10^{-3} to zero and updating it after each iteration. The training phase was oriented to forecasting one lead time with samples of three-time instants, $(\mathbf{x}^{t-1}, \mathbf{x}^{t}) \rightarrow \mathbf{x}^{t+1}$.

We used gradient descent to optimize the loss function. For adaptive moment estimation and weight decay management, we utilized the AdamW optimizer with parameters $\beta_1 = 0{,}9$, $\beta_2 = 0{,}95$, $\epsilon = 10^{-8}$, and $\lambda = 0{,}1$. Furthermore, we implemented gradient clipping with a maximum norm value of 32 to ensure stable training dynamics.

A.2. Model configuration

The ConvLSTM model employed in this study was specifically designed for spatiotemporal ocean prediction. To ensure a fair comparison with GraphCast, the model was configured and trained under similar configurations. It comprised two stacked ConvLSTM layers: the first layer contained 8 feature channels, and the second layer reduced this to 1 feature channel. Both layers utilized a 3×3 kernel with a stride of 1 and padding of 1. The input sequences had a shape corresponding to a batch size of 8, two time steps, and spatial dimensions of 256×256 with a single-channel input. The model was optimized using the AdamW optimizer, with a learning rate of 10^{-2} and a weight decay of 0,1. Training was conducted over 150 epochs, with updates for single-time-step predictions.

The architecture of the graph-based approach employed three levels of multi-mesh refinement, M^3, implemented by recursively subdividing triangular mesh elements into smaller components. The grid structure operated at a resolution of $0{,}05°$ derived from the satellite L4 product resolution in the training data. Spatial connectivity was regulated by a parameter that defined the neighborhood radius for each mesh node, set at 0.6 times the edge length. Within this range, grid nodes were connected to the corresponding mesh node in the encoder, ensuring localized information aggregation. The processor module performed 6 message-passing steps within an 8-dimensional MLP latent space, with all MLPs containing a single hidden layer. In the decoder, edge normalization for mesh-to-grid transitions was performed using the maximum edge length. This configuration effectively balanced model complexity and computational efficiency.

A.3. Spatially-weighted loss function

Most weather prediction studies use the Mean Squared Error (MSE) to optimize the neural network parameters during training. Since the area of a spherical grid cell decreases toward the geographic poles, MSE is often adjusted by latitude to account for this distortion. Common approaches include weighting by the cosine of the latitude (Rasp et al., 2020; Lam et al., 2023) or by using the difference between the sines of the cell's edges (Rasp et al., 2024), both of which serve as relative indicators of grid cell area. Without such adjustments, large errors in small high-latitude cells are treated the same as small errors in much larger equatorial cells, leading to spatial imbalance in the loss. Therefore, many global prediction studies use latitude-weighted MSE.

However, in this study, we adopt a different strategy that focuses the learning process exclusively on the ocean by applying a spatial mask derived from the structure of the data. This spatial mask assigns a weight of zero to grid cells over land and a weight of one to cells over the ocean. These binary weights are used to exclude land areas from contributing to the loss computation, effectively masking out irrelevant regions and ensuring that the loss is computed only over ocean areas. This approach is particularly well suited to our setup, which targets a geographically limited domain. In such cases, the variation in latitude across the domain is relatively small, and the differences in grid cell area are minimal. Therefore, the benefit of applying latitude-based weighting is negligible, and the masking strategy offers a more direct way to focus the training process on the regions of interest.

The loss function is thus defined as:

$$\mathcal{L}_{\text{MSE}} = \frac{1}{N_{\text{Rollout}}} \sum_{t=t_0+1}^{t_0+N_{\text{Rollout}}} \frac{1}{|\mathbb{G}_{\text{ocean}}|} \sum_{v_i \in \mathbb{G}_{\text{ocean}}} \left(\hat{x}_i^t - x_i^t\right)^2,$$

where

- t_0 is the initial forecast time,
- N_{Rollout} is the number of rollout steps used during training,
- v_i is a grid point defined by its latitude (ϕ) and longitude (λ),
- $\mathbb{G}$ is the full spatial grid, and $|\mathbb{G}| = |\phi| \cdot |\lambda|$,
- $\mathbb{G}_{\text{ocean}} \subset \mathbb{G}$ is the subset of grid points located over the ocean, determined using a binary land–ocean mask $m_i \in \{0, 1\}$, where $m_i = 1$ indicates ocean and $m_i = 0$ indicates land.

A.4. Experimental setup

We used two computers to conduct our experiments. The first machine, a local workstation equipped with a single Quadro RTX 4000 GPU with 8 GB of memory, was used for hyperparameter tuning. The second machine, a remote server with 8 Quadro RTX 4000 GPUs and 8 GB of memory, was utilized for training the final model with the full dataset. The training phase on the remote server took approximately 64 h. Our software stack included JAX, Haiku, Jraph, Optax, Jaxline, and xarray (Hoyer and Hamman, 2017) for model customization and training.

During the development phase on the local machine, we conducted a hyperparameter search to explore various configurations. Specifically, we investigated the number of message-passing steps $\{2, 6, 7\}$, latent size $\{2, 4, 8, 16\}$, and the mesh refinement level $\{2, 4, 6\}$ denoted as M^r. We tested batch sizes of 8 and 16, as preliminary results indicated that batch size had a negligible impact on the model's performance. To efficiently cover a wide range of configurations, we performed a grid search over these three hyperparameters, limiting the number of combinations while ensuring a comprehensive exploration.

Once the hyperparameters were determined, we proceeded to train the final model on the remote server using the full dataset. To optimize the training process, we parallelized it using the Distributed Data-Parallel (DDP) strategy (Li et al., 2020). In this setup, each of the 8 GPUs held its copy of the model and processed a separate batch of data. After processing, we used the all-reduce operation to combine the gradients from all GPUs into a unified gradient, which was then used to update the model weights.

To accommodate the parallel processing, we added a new dimension to the data called *device*, with a size equal to 8 GPUs. This modified the data dimensions to 8 devices, 8 batches, 1 to n lead-time predictions, and 300×300 latitude–longitude size. Each GPU was assigned identical initialized weights, and during each training step, a forward pass was performed to calculate the loss and gradients for its assigned batch. At the end of the step, the gradients were summed on the CPU, and the resulting unified gradient was used to update the model weights across all GPUs.

This process was repeated iteratively until only residual batches remained. Since the number of residual batches was smaller than 8 (the number of GPUs), they could not be parallelized and were instead processed sequentially on the CPU.

Appendix B. Graph neural network model

Let $\mathbf{x} : \mathbb{R}^3 \to \mathbb{R}$ be a spatiotemporal variable containing the SST volume for our period of study. Each sequence value is represented as x_i^t, with i standing for node in a grid $v_i \in \mathcal{V}^g$, and t the time instant. For simplicity, let $\mathbf{x}^t$ be the SST bidimensional array for time t and $\mathbf{x}_i$ the temporal evolution of the SST in node i. Each node v_i has a latitude–longitude position given by (ϕ_i, λ_i).

Our method is based on the GraphCast model (Lam et al., 2023), which is defined as an autoregressive model:

$$\hat{\mathbf{x}}^{t+1} = f(\mathbf{x}^t, \mathbf{x}^{t-1}), \tag{B.1}$$

where the $\hat{\mathbf{x}}^{t+1}$ map is a forecast obtained from two previous time instants. This output can be iteratively fed into the model to predict consecutive forecastings in a roll-out manner.

The graph of the model, $\mathcal{G}(\mathcal{V}^g, \mathcal{V}^m, \mathcal{E}^m, \mathcal{E}^{g2m}, \mathcal{E}^{m2g})$, is composed of grid nodes, $\mathcal{V}^g$, mesh nodes, $\mathcal{V}^m$, bidirectional edges connecting mesh nodes, $\mathcal{E}^m$, and directed edges from grid to mesh nodes, $\mathcal{E}^{g2m}$, and vice-versa, $\mathcal{E}^{m2g}$.

Each node in the grid, $\mathbf{v}_i^g \in \mathcal{V}^g$, is built from the SST values in times t and $t-1$, several external forcings in times $t-1$, t and $t+1$, and constant values, i.e., $\mathbf{v}_i^g = [x_i^t, x_i^{t-1}, \mathbf{f}_i^{t-1}, \mathbf{f}_i^t, \mathbf{f}_i^{t+1}, \mathbf{c}_i]$, in a given latitude–longitude position. The forcings are defined by $\mathbf{f} = [\sin(h), \cos(h), \sin(y), \cos(y)]$, with h the local time of day and y the year progress, normalized to $[0, 1)$. The constants are defined as $\mathbf{c}_i = [m_i^{0/1}, \cos(\phi_i), \sin(\lambda_i), \cos(\lambda_i)]$, with $\mathbf{m}^{0/1}$ a binary land–sea mask and no physical forcings.

Each node in the mesh, $\mathbf{v}_i^m \in \mathcal{V}^m$, is defined as $\mathbf{v}_i^m = [\cos(\phi_i), \sin(\lambda_i), \cos(\lambda_i)]$ and the mesh edges, $\mathbf{e}_{s,r}^m \in \mathcal{E}^m$, from the sender node s to the receiver node r, contains the following features: $\mathbf{e}_{s,r}^m = [\text{distance}(s, r), \mathbf{p}_s - \mathbf{p}_r]$, i.e. the edge length and the difference between the 3D spatial locations of nodes s and r. Fig. 3 shows a representation of the multi-mesh with three levels of resolution.

Unidirectional edges from the grid to the mesh are similarly defined by $\mathbf{e}_{s,r}^{g2m} = [\text{distance}(s, r), \mathbf{p}_s - \mathbf{p}_r]$. In this case, the sender node is on the grid and the receiver is on the mesh. An edge is added between two nodes if distance(s, r) is smaller than or equal to 0,6 times the length of the edges at the finest scale.

Similarly, unidirectional edges from the mesh to the grid are defined by $\mathbf{e}_{s,r}^{m2g} = [\text{distance}(s, r), \mathbf{p}_s - \mathbf{p}_r]$. An edge is created with the three mesh nodes of the triangular face that contains the grid node.

B.1. Encoder

The *encoder* maps the input data from the latitude–longitude grid into the multi-scale mesh. It relies on multi-layer perceptrons (MLP) to embed the graph variables, $\mathbf{v}_i^g$, $\mathbf{v}_i^m$, $\mathbf{e}_{s,r}^m$, $\mathbf{e}_{s,r}^{g2m}$ and $\mathbf{e}_{s,r}^{m2g}$, into a latent space:

$$\begin{aligned}
\tilde{\mathbf{v}}_i^g &= \text{MLP}_{\mathcal{V}^g}(\mathbf{v}_i^g), \\
\tilde{\mathbf{v}}_i^m &= \text{MLP}_{\mathcal{V}^m}(\mathbf{v}_i^m), \\
\tilde{\mathbf{e}}_{s,r}^m &= \text{MLP}_{\mathcal{E}^m}(\mathbf{e}_{s,r}^m), \\
\tilde{\mathbf{e}}_{s,r}^{g2m} &= \text{MLP}_{\mathcal{E}^{g2m}}(\mathbf{e}_{s,r}^{g2m}), \\
\tilde{\mathbf{e}}_{s,r}^{m2g} &= \text{MLP}_{\mathcal{E}^{m2g}}(\mathbf{e}_{s,r}^{m2g}).
\end{aligned} \tag{B.2}$$

Then, the information from the grid is transferred to the mesh using interaction networks (IN) (Battaglia et al., 2016). First, the edges are updated with information from the sending and receiving nodes,

$$\mathbf{d}\tilde{\mathbf{e}}_{s,r}^{g2m} = \text{MLP}_{\mathcal{E}^{g2m}}([\tilde{\mathbf{e}}_{s,r}^{g2m}, \mathbf{s}^g, \mathbf{r}^m]), \tag{B.3}$$

and the mesh nodes are updated by aggregating the edges as

$$\mathbf{d}\tilde{\mathbf{v}}_i^m = \text{MLP}_{\mathcal{V}^m}\left([\tilde{\mathbf{v}}_i^m, \sum_{s \in \mathcal{V}^g; r = \tilde{\mathbf{v}}_i^m} \tilde{\mathbf{e}}_{s,r}^{g2m}]\right), \tag{B.4}$$

with s the set of nodes in the grid that have an edge towards $\mathbf{v}_i^m$. The grid nodes are also updated as

$$\mathbf{d}\tilde{\mathbf{v}}_i^g = \mathrm{MLP}_{\mathcal{V}^g}(\tilde{\mathbf{v}}_i^g), \tag{B.5}$$

All MLPs in these equations are independent and do not share their parameters. Finally, the latent variables are updated with residual connections as

$$\begin{aligned}\tilde{\mathbf{v}}_i^g &\leftarrow \tilde{\mathbf{v}}_i^g + \mathbf{d}\tilde{\mathbf{v}}_i^g,\\ \tilde{\mathbf{v}}_i^m &\leftarrow \tilde{\mathbf{v}}_i^m + \mathbf{d}\tilde{\mathbf{v}}_i^m,\\ \tilde{\mathbf{e}}_{s,r}^{g2m} &\leftarrow \tilde{\mathbf{e}}_{s,r}^{g2m} + \mathbf{d}\tilde{\mathbf{e}}_{s,r}^{g2m}.\end{aligned} \tag{B.6}$$

B.2. Processor

The *processor* performs learned message-passing through several layers on the multi-mesh. First, the edges are updated through an MLP, concatenating the mesh edge with the receiver and sender nodes' latent variables as

$$\mathbf{d}\tilde{\mathbf{e}}_{s,r}^m = \mathrm{MLP}_{\mathcal{E}^m}([\tilde{\mathbf{e}}_{s,r}^m, \mathbf{s}^m, \mathbf{r}^m]). \tag{B.7}$$

Mesh nodes are updated by aggregating the edges as

$$\mathbf{d}\tilde{\mathbf{v}}_i^m = \mathrm{MLP}_{\mathcal{V}^m}\left(([\tilde{\mathbf{v}}_i^m, \sum_{s\in\mathcal{V}^m; r=\tilde{\mathbf{v}}_i^m} \tilde{\mathbf{e}}_{s,r}^m])\right), \tag{B.8}$$

and the variables are updated as a residual connection as

$$\begin{aligned}\tilde{\mathbf{v}}_i^m &\leftarrow \tilde{\mathbf{v}}_i^m + \mathbf{d}\tilde{\mathbf{v}}_i^m,\\ \tilde{\mathbf{e}}_{s,r}^m &\leftarrow \tilde{\mathbf{e}}_{s,r}^m + \mathbf{d}\tilde{\mathbf{e}}_{s,r}^m.\end{aligned} \tag{B.9}$$

This process represents one layer of the processor. It contains multiple iterative layers with independent MLP parameters that perform several message-passing operations.

B.3. Decoder

The *decoder* performs the inverse mapping from the mesh to the latitude–longitude grid. First, the edges from the mesh to the grid are updated as

$$\mathbf{d}\tilde{\mathbf{e}}_{s,r}^{m2g} = \mathrm{MLP}_{\mathcal{E}^{m2g}}([\tilde{\mathbf{e}}_{s,r}^{m2g}, \mathbf{s}^m, \mathbf{r}^g]). \tag{B.10}$$

The grid nodes are then updated, aggregating the information of the three edges that arrive at the grid node:

$$\mathbf{d}\tilde{\mathbf{v}}_i^g = \mathrm{MLP}_{\mathcal{V}^g}\left([\tilde{\mathbf{v}}_i^g, \sum_{s\in\mathcal{V}^g; r=\tilde{\mathbf{v}}_i^g} \mathbf{d}\tilde{\mathbf{e}}_{s,r}^{m2g}]\right). \tag{B.11}$$

A residual connection is used to update the information of the grid nodes coming from the embedding in the encoder:

$$\tilde{\mathbf{v}}_i^g \leftarrow \tilde{\mathbf{v}}_i^g + \mathbf{d}\tilde{\mathbf{v}}_i^g, \tag{B.12}$$

and the output prediction is obtained with another MLP as

$$\hat{\mathbf{y}}^t = \mathrm{MLP}_{\mathcal{V}^g}(\tilde{\mathbf{v}}_i^g). \tag{B.13}$$

Finally, the forecasting is obtained as a residual connection with the input data as

$$\hat{\mathbf{x}}^{t+1} = \mathbf{x}^t + \hat{\mathbf{y}}^t$$

All input variables are normalized to zero mean and unit variance. The output variable $\hat{\mathbf{y}}^t$ is multiplied by the average standard deviation of the temporal change $\mathbf{d}\hat{\mathbf{x}} = \hat{\mathbf{x}}^{t+1} - \hat{\mathbf{x}}^t$ computed from the train set.

All MLPs have one hidden layer with a latent dimension of 8, the same as the output layer. The size of the output layer in the last MLP, corresponding to the *decoder*, is one for the prediction of the value of the SST at each node.

Appendix C. Description of the evaluation metrics

In this study, we used four metrics to assess the performance of the ocean prediction models: the Anomaly Correlation Coefficient (ACC), the Root Mean Square Error (RMSE), the Activity (ACT), and the Bias. While there are several ways to define those metrics, we use the approach outlined by Bouallègue et al. (2024) as it aligns with the scores set by the World Meteorological Organization (WMO) and the ECMWF. Below is a detailed description of each metric, including its mathematical formulation and interpretation.

The RMSE quantifies the average magnitude of the prediction error, providing a measure of the overall accuracy of the model. The ACC evaluates the linear association between the predicted and observed anomalies, serving as an indicator of the model's skill in capturing the spatiotemporal variability of the ocean field. The Bias represents the systematic offset between predictions and observations, while ACT is defined as the standard deviation of the predicted anomalies and reflects the model's ability to reproduce the observed variability amplitude given a measure of the smoothness of the forecast.

C.1. Anomaly Correlation Coefficient (ACC)

The ACC measures the correlation between the predicted and observed anomalies, evaluating the model's ability to predict deviations from the climatology. It is calculated as:

$$\mathrm{ACC} = \frac{\overline{(a_f - \overline{a_f})(a_o - \overline{a_o})}}{\sqrt{\overline{(a_f - \overline{a_f})^2}}\sqrt{\overline{(a_o - \overline{a_o})^2}}}, \tag{C.1}$$

where $a_f = \hat{x}_i^t - c_i^t$ denotes the predicted anomaly (prediction deviation from climatology c_i^t) and $a_o = x_i^t - c_i^t$ the observed anomaly (ground truth deviation) for a given time and location. $\overline{(\cdot)} = \frac{1}{|\mathbb{G}|}\sum_{v_i\in\mathbb{G}} w_\phi(\cdot)$ represents the latitudinal weighted spatial average, where w_ϕ is the latitude-weighting factor based on the cosine of the latitude expressed in radians (Geer, 2016; Rasp et al., 2020; Lam et al., 2023). This weighting scheme ensures that regions near the equator, where the longitudinal distance between grid points is larger, are not underrepresented in the score calculation. Note that while the model training (Appendix A.3) utilized a simplified binary mask due to the limited regional domain, this rigorous weighting is applied during evaluation to ensure strict adherence to standard verification protocols (e.g., WMO, ECMWF).

The ACC ranges between −1 and 1, where values close to 1 indicate a high positive correlation between the predicted and observed anomalies. This metric is widely used in forecast verification studies (ECMWF).

C.2. Root Mean Square Error (RMSE)

The RMSE quantifies the average magnitude of the difference between the predicted and observed values, providing a measure of the model's overall accuracy. It is defined as:

$$\mathrm{RMSE} = \sqrt{\overline{\left(\hat{x}_i^t - x_i^t\right)^2}}, \tag{C.2}$$

where $\hat{x}_i^t$ is the predicted value and x_i^t is the ground truth value at a given time and grid cell.

A lower RMSE indicates better predictive accuracy, making it one of the most common metrics for forecast verification.

C.3. Activity (ACT)

The Activity assesses the model's ability to reproduce the observed variability by measuring the standard deviation of the predicted and observed anomaly fields:

$$\mathrm{ACT}_f = \sqrt{\overline{(a_f - \overline{a_f})^2}}, \qquad \mathrm{ACT}_o = \sqrt{\overline{(a_o - \overline{a_o})^2}}. \tag{C.3}$$

A similar ACT between predictions and observations suggests that the model accurately captures the spatial variability of the ocean field. This score was originally proposed by Thorpe et al. (2013) and later adopted by Bouallègue et al. (2024) to assess the smoothness of the forecast.

The relative activity (RA) of a model is defined as the ratio between the forecast ACT and the observed ACT (Bechtold et al., 2008). Analyzing RA as a function of forecast lead time reveals whether the model underestimates variability (producing smoother forecasts) or overestimates it (producing noisier forecasts):

$$\mathrm{RA} = \frac{\mathrm{ACT}_f}{\mathrm{ACT}_o}. \tag{C.4}$$

C.4. Bias

The Bias quantifies the systematic difference between the predicted and observed values, indicating whether the model tends to overestimate or underestimate the observations. It is calculated as:

$$\mathrm{Bias} = \overline{\left(\hat{x}_i^t - x_i^t\right)}. \tag{C.5}$$

A Bias close to zero implies that the model does not present systematic overestimation or underestimation of the observations. This metric is commonly used in the verification of ocean and atmospheric models.

These metrics provide a comprehensive assessment of the models' performance, allowing us to quantify their predictive skill across different aspects of the forecast quality.

Data availability

The data and code used in this paper are available at the following URLs:

• The European North West Shelf/Iberia Biscay Irish Seas - High Resolution L4 Sea Surface Temperature Reprocessed is available at https://doi.org/10.48670/moi-00153.

• EasyCORA is available at https://doi.org/10.17882/46219.

• GLORYS is available at https://doi.org/10.48670/moi-00021.

• The source code of our models is publicly available at https://github.com/gacuervol/regional-graphcast-sst.git under the MIT licence.

References

Adcroft, A., Hallberg, R., Dunne, J.P., Samuels, B.L., Galt, J.A., Barker, C.H., Payton, D., 2010. Simulations of underwater plumes of dissolved oil in the Gulf of Mexico. Geophys. Res. Lett. 37 (18), http://dx.doi.org/10.1029/2010GL044689.

Arístegui, J., Barton, E.D., Álvarez-Salgado, X.A., Santos, A.M.P., Figueiras, F.G., Kifani, S., Hernández-León, S., Mason, E., Machú, E., Demarcq, H., 2009. Sub-regional ecosystem variability in the Canary Current upwelling. Prog. Oceanogr. 83 (1), 33–48. http://dx.doi.org/10.1016/j.pocean.2009.07.031, Eastern Boundary Upwelling Ecosystems: Integrative and Comparative Approaches.

Arístegui, J., Sangrá, P., Hernández-León, S., Cantón, M., Hernández-Guerra, A., Kerling, J., 1994. Island-induced eddies in the Canary Islands. Deep. Sea Res. Part I: Oceanogr. Res. Pap. 41 (10), 1509–1525. http://dx.doi.org/10.1016/0967-0637(94)90058-2.

Barton, E., Arístegui, J., Tett, P., Cantón, M., García-Braun, J., Hernández-León, S., Nykjaer, L., Almeida, C., Almunia, J., Ballesteros, S., Basterretxea, G., Escánez, J., García-Weill, L., Hernández-Guerra, A., López-Laatzen, F., Molina, R., Montero, M., Navarro-Pérez, E., Rodríguez, J., van Lenning, K., Vélez, H., Wild, K., 1998. The transition zone of the Canary Current upwelling region. Prog. Oceanogr. 41 (4), 455–504. http://dx.doi.org/10.1016/S0079-6611(98)00023-8.

Battaglia, P.W., Hamrick, J.B., Bapst, V., Sanchez-Gonzalez, A., Zambaldi, V., Malinowski, M., Tacchetti, A., Raposo, D., Santoro, A., Faulkner, R., Gulcehre, C., Song, F., Ballard, A., Gilmer, J., Dahl, G., Vaswani, A., Allen, K., Nash, C., Langston, V., Dyer, C., Heess, N., Wierstra, D., Kohli, P., Botvinick, M., Vinyals, O., Li, Y., Pascanu, R., 2018. Relational inductive biases, deep learning, and graph networks. http://dx.doi.org/10.48550/arXiv.1806.01261, ArXiv E-Prints arXiv:1806.01261.

Battaglia, P.W., Pascanu, R., Lai, M., Rezende, D., Kavukcuoglu, K., 2016. Interaction Networks for Learning about Objects, Relations and Physics. Technical Report, Cornell University, http://dx.doi.org/10.48550/arXiv.1612.00222, arXiv:1612.00222.

Bechtold, P., Köhler, M., Jung, T., Leutbecher, M., Rodwell, M., Vitart, F., 2008. Advances in simulating atmospheric variability with IFS cycle 32r3. In: ECMWF Newsletter No. 114 - Winter 2007/08. Vol. 114, pp. 29–38, URL: https://www.ecmwf.int/en/elibrary/78219-newsletter-no-114-winter-200708.

Bell, M.J., Lefèbvre, M., Traon, P.-Y.L., Smith, N., Wilmer-Becker, K., 2009. GODAE: the global ocean data assimilation experiment. Oceanography 22 (3), 14–21. http://dx.doi.org/10.5670/oceanog.2009.62.

Bengio, Y., Simard, P., Frasconi, P., 1994. Learning long-term dependencies with gradient descent is difficult. IEEE Trans. Neural Netw. 5 (2), 157–166. http://dx.doi.org/10.1109/72.279181.

Bi, K., Xie, L., Zhang, H., Chen, X., Gu, X., Tian, Q., 2023. Accurate medium-range global weather forecasting with 3D neural networks. Nature 619 (7970), 533–538. http://dx.doi.org/10.1038/s41586-023-06185-3.

Bindoff, N.L., Cheung, W.W.L., Kairo, J.G., Arístegui, J., Guinder, V.A., Hallberg, R., Hilmi, N., Jiao, N., Karim, M.S., Levin, L., O'Donoghue, S., Cuicapusa, S.R.P., Rinkevich, B., Suga, T., Tagliabue, A., Williamson, P., 2019. Changing ocean, marine ecosystems, and dependent communities. In: Pörtner, H.-O., Roberts, D.C., Masson-Delmotte, V., Zhai, P., Tignor, M., Poloczanska, E., Mintenbeck, K., Alegría, A., Nicolai, M., Okem, A., Petzold, J., Rama, B., Weyer, N.M. (Eds.), IPCC Special Report on the Ocean and Cryosphere in a Changing Climate. Cambridge University Press, Cambridge, UK and New York, NY, USA, pp. 447–587. http://dx.doi.org/10.1017/9781009157964.007.

Bodnar, C., Bruinsma, W.P., Lucic, A., Stanley, M., Brandstetter, J., Garvan, P., Riechert, M., Weyn, J., Dong, H., Vaughan, A., Gupta, J.K., Tambiratnam, K., Archibald, A., Heider, E., Welling, M., Turner, R.E., Perdikaris, P., 2024. A Foundation Model of the Atmosphere. Technical Report, Cornell University, http://dx.doi.org/10.48550/arXiv.2405.13063, arXiv:2405.13063.

Bouallègue, Z.B., Clare, M.C.A., Magnusson, L., Gascón, E., Maier-Gerber, M., Janoušek, M., Rodwell, M., Pinault, F., Dramsch, J.S., Lang, S.T.K., Raoult, B., Rabier, F., Chevallier, M., Sandu, I., Dueben, P., Chantry, M., Pappenberger, F., 2024. The rise of data-driven weather forecasting: A first statistical assessment of machine learning–based weather forecasts in an operational-like context. Bull. Am. Meteorol. Soc. 105 (6), E864 – E883. http://dx.doi.org/10.1175/BAMS-D-23-0162.1.

Brasseur, P., Verron, J., 2006. The SEEK filter method for data assimilation in oceanography: a synthesis. Ocean Dyn. 56, 650–661. http://dx.doi.org/10.1007/s10236-006-0080-3.

Cabanes, C., Grouazel, A., von Schuckmann, K., Hamon, M., Turpin, V., Coatanoan, C., Paris, F., Guinehut, S., Boone, C., Ferry, N., de Boyer Montégut, C., Carval, T., Reverdin, G., Pouliquen, S., Le Traon, P.-Y., 2013. The CORA dataset: validation and diagnostics of in-situ ocean temperature and salinity measurements. Ocean Sci. 9 (1), 1–18. http://dx.doi.org/10.5194/os-9-1-2013.

Chambers, D.P., Cazenave, A., Champollion, N., Dieng, H., Llovel, W., Forsberg, R., von Schuckmann, K., Wada, Y., 2017. Evaluation of the global mean sea level budget between 1993 and 2014. In: Integrative Study of the Mean Sea Level and Its Components. Springer, pp. 315–333.

Chattopadhyay, A., Gray, M., Wu, T., Lowe, A.B., He, R., 2024. OceanNet: a principled neural operator-based digital twin for regional oceans. Sci. Rep. 14 (1), 21181. http://dx.doi.org/10.1038/s41598-024-72145-0.

Chattopadhyay, A., Mustafa, M., Hassanzadeh, P., Kashinath, K., 2021. Deep spatial transformers for autoregressive data-driven forecasting of geophysical turbulence. In: Proceedings of the 10th International Conference on Climate Informatics. In: CI2020, Association for Computing Machinery, New York, NY, USA, pp. 106–112. http://dx.doi.org/10.1145/3429309.3429325.

Chattopadhyay, A., Sun, Y.Q., Hassanzadeh, P., 2023. Challenges of learning multi-scale dynamics with AI weather models: Implications for stability and one solution. http://dx.doi.org/10.48550/arXiv.2304.07029, ArXiv E-Prints, arXiv:2304.07029.

CMEMS, 2024. European North West Shelf/Iberia biscay irish seas - High resolution L4 sea surface temperature reprocessed. http://dx.doi.org/10.48670/moi-00153.

CMEMS, 2025. Global ocean- CORA- in-situ observations yearly delivery in delayed mode (INSITU_GLO_PHY_TS_DISCRETE_MY_013_001). E.U. Copernicus Marine Service Information (CMEMS). Marine Data Store (MDS). http://dx.doi.org/10.17882/46219, Dataset. (Accessed 08 January 2026).

Couairon, G., Lessig, C., Charantonis, A., Monteleoni, C., 2024. Archesweather: An Efficient AI Weather Forecasting Model at 1.5° Resolution. Technical Report 2405.14527, Cornell University, http://dx.doi.org/10.48550/arXiv.2405.14527.

Crocker, R., Maksymczuk, J., Mittermaier, M., Tonani, M., Pequignet, C., 2020. An approach to the verification of high-resolution ocean models using spatial methods. Ocean Sci. 16 (4), 831–845. http://dx.doi.org/10.5194/os-16-831-2020.

Cropper, T.E., Hanna, E., Bigg, G.R., 2014. Spatial and temporal seasonal trends in coastal upwelling off Northwest Africa, 1981–2012. Deep. Sea Res. Part I: Oceanogr. Res. Pap. 86, 94–111. http://dx.doi.org/10.1016/j.dsr.2014.01.007.

Cushman-Roisin, B., Tang, B., Chassignet, E.P., 1990. Westward motion of mesoscale eddies. J. Phys. Oceanogr. 20 (5), 758–768.

Dee, D.P., Uppala, S.M., Simmons, A.J., Berrisford, P., Poli, P., Kobayashi, S., Andrae, U., Balmaseda, M.A., Balsamo, G., Bauer, P., Bechtold, P., Beljaars, A.C.M., van de Berg, L., Bidlot, J., Bormann, N., Delsol, C., Dragani, R., Fuentes, M., Geer, A.J., Haimberger, L., Healy, S.B., Hersbach, H., Hólm, E.V., Isaksen, L., Kållberg, P., Köhler, M., Matricardi, M., McNally, A.P., Monge-Sanz, B.M., Morcrette, J.-J., Park, B.-K., Peubey, C., de Rosnay, P., Tavolato, C., Thépaut, J.-N., Vitart, F., 2011. The ERA-Interim reanalysis: configuration and performance of the data assimilation system. Q. J. R. Meteorol. Soc. 137 (656), 553–597. http://dx.doi.org/10.1002/qj.828.

Docquier, D., Grist, J.P., Roberts, M.J., Roberts, C.D., Semmler, T., Ponsoni, L., Massonnet, F., Sidorenko, D., Sein, D.V., Iovino, D., et al., 2019. Impact of model resolution on Arctic sea ice and North Atlantic Ocean heat transport. Clim. Dyn. 53, 4989–5017.

Dueben, P.D., Bauer, P., 2018. Challenges and design choices for global weather and climate models based on machine learning. Geosci. Model. Dev. 11 (10), 3999–4009. http://dx.doi.org/10.5194/gmd-11-3999-2018.

Ebert, E.E., 2008. Fuzzy verification of high-resolution gridded forecasts: a review and proposed framework. Meteorol. Appl. 15 (1), 51–64. http://dx.doi.org/10.1002/met.25.

El Aouni, A., Gaudel, Q., Johnson, J.E., Charly, R., Le Sommer, J., van Gennip, S., fablet, r., Drevillon, M., DRILLET, Y., Le Traon, P., 2025. OceanBench: A benchmark for data-driven global ocean forecasting systems. In: Belgrave, D., Zhang, C., Lin, H., Pascanu, R., Koniusz, P., Ghassemi, M., Chen, N. (Eds.), Advances in Neural Information Processing Systems. Vol. 38, Curran Associates, Inc.

Estrada-Allis, S.N., Rodríguez-Santana, Á., Naveira-Garabato, A.C., García-Weil, L., Arcos-Pulido, M., Emelianov, M., 2023. Enhancement of turbulence and nutrient fluxes within an Eastern Boundary Upwelling Filament: a diapycnal entrainment approach. Front. Mar. Sci. Volume 10 - 2023, http://dx.doi.org/10.3389/fmars.2023.1113879.

Ezraty, R., Girard-Ardhuin, F., Piollé, J.-F., Kaleschke, L., Heygster, G., 2007. Arctic and Antarctic Sea Ice Concentration and Arctic Sea Ice Drift Estimated from Special Sensor Microwave Data, vol. 2, Département D'Océanographie Physique et Spatiale, IFREMER, Brest, France and University of Bremen Germany.

Falkowski, P.G., Ziemann, D., Kolber, Z., Bienfang, P.K., 1991. Role of eddy pumping in enhancing primary production in the ocean. Nature 352 (6330), 55–58. http://dx.doi.org/10.1038/352055a0.

Galloudec, O.L., Chune, S.L., Nouel, L., Fernandez, E., Derval, C., Tressol, M., Dussurget, R., Biardeau, A., Tonani, M., 2024. Product User Manual for Global Ocean Physics Analysis and Forecasting Product. Copernicus Marine Environment Monitoring Service (CMEMS), URL: https://documentation.marine.copernicus.eu/PUM/CMEMS-GLO-PUM-001-024.pdf, CMEMS-GLO-PUM-001-024, Issue 2.1.

Geer, A.J., 2016. Significance of changes in medium-range forecast scores. Tellus A: Dyn. Meteorol. Oceanogr. 68 (1), 30229. http://dx.doi.org/10.3402/tellusa.v68.30229.

Gers, F., Schmidhuber, J., 2000. Recurrent nets that time and count. In: Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks. IJCNN 2000. Neural Computing: New Challenges and Perspectives for the New Millennium. Vol. 3, pp. 189–194. http://dx.doi.org/10.1109/IJCNN.2000.861302.

Gers, F., Schmidhuber, J., Cummins, F., 1999. Learning to forget: continual prediction with LSTM. In: 1999 Ninth International Conference on Artificial Neural Networks ICANN 99. (Conf. Publ. No. 470). Vol. 2, pp. 850–855. http://dx.doi.org/10.1049/cp:19991218.

Greff, K., Srivastava, R.K., Koutník, J., Steunebrink, B.R., Schmidhuber, J., 2017. LSTM: A search space odyssey. IEEE Trans. Neural Netw. Learn. Syst. 28 (10), 2222–2232. http://dx.doi.org/10.1109/TNNLS.2016.2582924.

Hagen, E., Zulicke, C., Feistel, R., 1996. Near-surface structures in the Cape Ghir filament off Morocco. Oceanol. Acta 19 (6), 577–598, URL: https://archimer.ifremer.fr/doc/00096/20728/.

Hausmann, U., Czaja, A., 2012. The observed signature of mesoscale eddies in sea surface temperature and the associated heat transport. Deep. Sea Res. Part I: Oceanogr. Res. Pap. 70, 60–72. http://dx.doi.org/10.1016/j.dsr.2012.08.005.

Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. Neural Comput. 9 (8), 1735–1780. http://dx.doi.org/10.1162/neco.1997.9.8.1735.

Holmberg, D., Clementi, E., Roos, T., 2024. Regional ocean forecasting with hierarchical graph neural networks. http://dx.doi.org/10.48550/arXiv.2410.11807, ArXiv E-Prints arXiv:2410.11807.

Hoyer, S., Hamman, J., 2017. Xarray: N-D labeled arrays and datasets in Python. J. Open Res. Softw. 5 (1), http://dx.doi.org/10.5334/jors.148.

Hyun, S., Mishra, A., Follett, C.L., Jonsson, B., Kulk, G., Forget, G., Racault, M.-F., Jackson, T., Dutkiewicz, S., Müller, C.L., Bien, J., 2022. Ocean mover's distance: using optimal transport for analysing oceanographic data. Proc. R. Soc. A: Math. Phys. Eng. Sci. 478 (2262), 20210875. http://dx.doi.org/10.1098/rspa.2021.0875.

Jean-Michel, L., Eric, G., Romain, B.-B., Gilles, G., Angélique, M., Marie, D., Clément, B., Mathieu, H., Olivier, L.G., Charly, R., Tony, C., Charles-Emmanuel, T., Florent, G., Giovanni, R., Mounir, B., Yann, D., Pierre-Yves, L.T., 2021. The copernicus global 1/12° oceanic and sea ice GLORYS12 reanalysis. Front. Earth Sci. 9, http://dx.doi.org/10.3389/feart.2021.698876.

Keisler, R., 2022. Forecasting global weather with graph neural networks. http://dx.doi.org/10.48550/arXiv.2202.07575, ArXiv E-Prints arXiv:2202.07575.

Kochkov, D., Yuval, J., Langmore, I., Norgaard, P.C., Smith, J.A., Mooers, G., Klöwer, M., Lottes, J., Rasp, S., Düben, P.D., Hatfield, S., Battaglia, P.W., Sanchez-Gonzalez, A., Willson, M., Brenner, M.P., Hoyer, S., 2023. Neural general circulation models for weather and climate. Nature 632, 1060–1066. http://dx.doi.org/10.1038/s41586-024-07744-y.

Lam, R., Sánchez-González, A., Willson, M., Wirnsberger, P., Fortunato, M., Alet, F., Ravuri, S., Ewalds, T., Eaton-Rosen, Z., Hu, W., Merose, A., Hoyer, S., Holland, G., Vinyals, O., Stott, J., Pritzel, A., Mohamed, S., Battaglia, P., 2023. Learning skillful medium-range global weather forecasting. Science 382 (6677), 1416–1421. http://dx.doi.org/10.1126/science.adi2336.

Lang, S., Alexe, M., Chantry, M., Dramsch, J., Pinault, F., Raoult, B., Clare, M.C.A., Lessig, C., Maier-Gerber, M., Magnusson, L., Ben Bouallègue, Z., Prieto Nemesio, A., Dueben, P.D., Brown, A., Pappenberger, F., Rabier, F., 2024. AIFS – ECMWF's data-driven forecasting system. http://dx.doi.org/10.48550/arXiv.2406.01465, ArXiv E-Prints arXiv:2406.01465.

Lellouche, J.-M., Greiner, E., Le Galloudec, O., Garric, G., Regnier, C., Drevillon, M., Benkiran, M., Testut, C.-E., Bourdalle-Badie, R., Gasparin, F., Hernandez, O., Levier, B., Drillet, Y., Remy, E., Le Traon, P.-Y., 2018. Recent updates to the copernicus marine service global ocean monitoring and forecasting real-time 1/12° high-resolution system. Ocean Sci. 14 (5), 1093–1126. http://dx.doi.org/10.5194/os-14-1093-2018.

Lellouche, J.-M., Le Galloudec, O., Drévillon, M., Régnier, C., Greiner, E., Garric, G., Ferry, N., Desportes, C., Testut, C.-E., Bricaud, C., Bourdallé-Badie, R., Tranchant, B., Benkiran, M., Drillet, Y., Daudin, A., De Nicola, C., 2013. Evaluation of global monitoring and forecasting systems at Mercator Océan. Ocean Sci. 9 (1), 57–81. http://dx.doi.org/10.5194/os-9-57-2013.

Li, S., Zhao, Y., Varma, R., Salpekar, O., Noordhuis, P., Li, T., Paszke, A., Smith, J., Vaughan, B., Damania, P., Chintala, S., 2020. PyTorch distributed: Experiences on accelerating data parallel training. http://dx.doi.org/10.48550/arXiv.2006.15704, ArXiv E-Prints arXiv:2006.15704.

Liu, Y., MacFadyen, A., Ji, Z.-G., Weisberg, R.H., 2013. Monitoring and Modeling the Deepwater Horizon Oil Spill: A Record Breaking Enterprise. John Wiley & Sons, http://dx.doi.org/10.1029/GM195.

Madec, G., Bell, M., Benshila, R., Blaker, A., Boudrallé-Badie, R., Bricaud, C., Bruciaferri, D., Carneiro, D., Castrillo, M., Calvert, D., Chanut, J., Clementi, E., Coward, A., de Lavergne, C., Dobricic, S., Epicoco, I., Éthé, C., Fiedler, E., Ford, D., Furner, R., Ganderton, J., Graham, T., Harle, J., Hutchinson, K., Iovino, D., King, R., Lea, D., Levy, C., Lovato, T., Maisonnave, E., Mak, J., Sanchez, J.M.C., Martin, M., Martin, N., Martins, D., Masson, S., Mathiot, P., Mele, F., Mocavero, S., Moulin, A., Müller, S., Nurser, G., Oddo, P., Paronuzzi, S., Paul, J., Peltier, M., Person, R., Rousset, C., Rynders, S., Samson, G., Schroeder, D., Storkey, D., Storto, A., Téchené, S., Vancoppenolle, M., Wilson, C., 2024. NEMO ocean engine reference manual. http://dx.doi.org/10.5281/zenodo.14515373.

Madec, G., et al., 2008. NEMO Ocean Engine. Note Du Pôle De Modélisation. Vol. 27, Institut Pierre-Simon Laplace (IPSL), France, pp. 1288–1619.

Mourre, B., Aguiar, E., Juzà, M., Hernández-Lasheras, J., Reyes, E., Heslop, E., Escudier, R., Cutolo, E., Ruiz, S., Mason, E., Pascual, A., Tintoré, J., 2018. Assessment of high-resolution regional ocean prediction systems using multi-platform observations: Illustrations in the Western Mediterranean Sea. New Front. Oper. Oceanogr. http://dx.doi.org/10.17125/gov2018.ch24.

Mu, B., Peng, C., Yuan, S., Chen, L., 2019. ENSO forecasting over multiple time horizons using ConvLSTM network and rolling mechanism. In: 2019 International Joint Conference on Neural Networks. IJCNN, pp. 1–8. http://dx.doi.org/10.1109/IJCNN.2019.8851967.

Mu, B., Qin, B., Yuan, S., 2021. ENSO-ASC 1.0.0: ENSO deep learning forecast model with a multivariate air–sea coupler. Geosci. Model. Dev. 14 (11), 6977–6999. http://dx.doi.org/10.5194/gmd-14-6977-2021.

Oskarsson, J., Landelius, T., Lindsten, F., 2023. Graph-based neural weather prediction for limited area modeling. http://dx.doi.org/10.48550/arXiv.2309.17370, ArXiv E-Prints arXiv:2309.17370.

Pelegrí, J., Arístegui, J., Cana, L., González-Dávila, M., Hernández-Guerra, A., Hernández-León, S., Marrero-Díaz, A., Montero, M., Sangrà, P., Santana-Casiano, M., 2005. Coupling between the open ocean and the coastal upwelling region off northwest Africa: water recirculation and offshore pumping of organic matter. J. Mar. Syst. 54 (1), 3–37. http://dx.doi.org/10.1016/j.jmarsys.2004.07.003, A general study of the Spanish North Atlantic boundaries: an interdisciplinary approach.

Pfaff, T., Fortunato, M., Sanchez-Gonzalez, A., Battaglia, P.W., 2020. Learning mesh-based simulation with graph networks. http://dx.doi.org/10.48550/arXiv.2010.03409, ArXiv E-Prints arXiv:2010.03409.

Piollé, J.-F., Autret, E., 2023. QUID for SST TAC Products SST_GLO_SST_L3S_NRT_OBSERVATIONS_010_010. Quality Information Document 2.1, E.U. Copernicus Marine Service Information (CMEMS), http://dx.doi.org/10.48670/moi-00164.

Price, I., Sanchez-Gonzalez, A., Alet, F., Andersson, T.R., El-Kadi, A., Masters, D., Ewalds, T., Stott, J., Mohamed, S., Battaglia, P.W., Lam, R., Willson, M., 2024. Probabilistic weather forecasting with machine learning. Nature 637, 84–90. http://dx.doi.org/10.1038/s41586-024-08252-9.

Pujol, M.-I., Faugère, Y., Taburet, G., Dupuy, S., Pelloquin, C., Ablain, M., Picot, N., 2016. DUACS DT2014: the new multi-mission altimeter data set reprocessed over 20 years. Ocean Sci. 12 (5), 1067–1090. http://dx.doi.org/10.5194/os-12-1067-2016.

Rasp, S., Dueben, P.D., Scher, S., Weyn, J.A., Mouatadid, S., Thuerey, N., 2020. WeatherBench: a benchmark data set for data-driven weather forecasting. J. Adv. Model. Earth Syst. 12 (11), e2020MS002203. http://dx.doi.org/10.1029/2020MS002203.

Rasp, S., Hoyer, S., Merose, A., Langmore, I., Battaglia, P., Russel, T., Sanchez-Gonzalez, A., Yang, V., Carver, R., Agrawal, S., Chantry, M., Bouallegue, Z.B., Dueben, P., Bromberg, C., Sisk, J., Barrington, L., Bell, A., Sha, F., 2024. WeatherBench 2: A Benchmark for the Next Generation of Data-Driven Global Weather Models. Technical Report 2308.15560, Cornell University, http://dx.doi.org/10.48550/arXiv.2308.15560.

Rio, M.H., Guinehut, S., Larnicol, G., 2011. New CNES-CLS09 global mean dynamic topography computed from the combination of GRACE data, altimetry, and in situ measurements. J. Geophys. Res.: Oceans 116 (C7), http://dx.doi.org/10.1029/2010JC006505.

Roberts, M., Vidale, P., Senior, C., Hewitt, H., Bates, C., Berthou, S., Chang, P., Christensen, H., Danilov, S., Demory, M.-E., et al., 2018. The benefits of global high resolution for climate simulation: process understanding and the enabling of stakeholder decisions at the regional scale. Bull. Am. Meteorol. Soc. 99 (11), 2341–2359.

Sangrà, P., Pascual, A., Rodríguez-Santana, Á., Machín, F., Mason, E., McWilliams, J.C., Pelegrí, J.L., Dong, C., Rubio, A., Arístegui, J., Marrero-Díaz, Á., Hernández-Guerra, A., Martínez-Marrero, A., Auladell, M., 2009. The canary eddy corridor: A major pathway for long-lived eddies in the subtropical North Atlantic. Deep. Sea Res. Part I: Oceanogr. Res. Pap. 56 (12), 2100–2114. http://dx.doi.org/10.1016/j.dsr.2009.08.008.

Sangrà, P., Troupin, C., Barreiro-González, B., Desmond Barton, E., Orbi, A., Arístegui, J., 2015. The Cape Ghir filament system in August 2009 (NW Africa). J. Geophys. Res.: Oceans 120 (6), 4516–4533. http://dx.doi.org/10.1002/2014JC010514.

Scher, S., 2018. Toward data-driven weather and climate forecasting: Approximating a simple general circulation model with deep learning. Geophys. Res. Lett. 45 (22), 12,616–12,622. http://dx.doi.org/10.1029/2018GL080704.

Shi, X., Chen, Z., Wang, H., Yeung, D.-Y., Wong, W.-k., Woo, W.-c., 2015. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. http://dx.doi.org/10.48550/arXiv.1506.04214, ArXiv E-Prints arXiv:1506.04214.

Smith, G.C., Fortin, A.-S., 2022. Verification of eddy properties in operational oceanographic analysis systems. Ocean Model. 172, 101982. http://dx.doi.org/10.1016/j.ocemod.2022.101982.

Sommer, J.L., Chassignet, E.P., Wallcraft, A.J., 2018. Ocean circulation modeling for operational oceanography: Current status and future challenges. New Front. Oper. Oceanogr. http://dx.doi.org/10.17125/gov2018.ch12.

Szekely, T., Gourrion, J., Pouliquen, S., Reverdin, G., Merceur, F., 2019. CORA, Coriolis Ocean Dataset for Reanalysis. Sea Scientific Open Data Publication, http://dx.doi.org/10.17882/46219.

Taylor, K.E., 2001. Summarizing multiple aspects of model performance in a single diagram. J. Geophys. Res.: Atmospheres 106 (D7), 7183–7192. http://dx.doi.org/10.1029/2000JD900719.

Thompson, B., Sanchez, C., Heng, B.C.P., Kumar, R., Liu, J., Huang, X.-Y., Tkalich, P., 2021. Development of a MetUM (v 11.1) and NEMO (v 3.6) coupled operational forecast model for the maritime continent – Part 1: Evaluation of ocean forecasts. Geosci. Model. Dev. 14 (2), 1081–1100. http://dx.doi.org/10.5194/gmd-14-1081-2021.

Thorpe, A., Bauer, P., Magnusson, L., Richardson, D., 2013. An evaluation of recent performance of ECMWF's forecasts. In: ECMWF Newsletter No. 137 - Autumn 2013. pp. 15–18, URL: https://www.ecmwf.int/en/elibrary/78198-newsletter-no-137-autumn-2013.

Treguier, A.-M., Chassignet, E.p., Le Boyer, A., Pinardi, N., 2017. Modeling and forecasting the "weather of the ocean" at the mesoscale. J. Mar. Res. 75 (3), 301–329. http://dx.doi.org/10.1357/002224017821836842.

Troupin, C., Mason, E., Beckers, J.-M., Sangrà, P., 2012. Generation of the Cape Ghir upwelling filament: A numerical study. Ocean Model. 41, 1–15. http://dx.doi.org/10.1016/j.ocemod.2011.09.001.

Vissio, G., Lembo, V., Lucarini, V., Ghil, M., 2020. Evaluating the performance of climate models based on wasserstein distance. Geophys. Res. Lett. 47 (21), e2020GL089385. http://dx.doi.org/10.1029/2020GL089385.

Wang, X., Wang, R., Hu, N., Wang, P., Huo, P., Wang, G., Wang, H., Wang, S., Zhu, J., Xu, J., Yin, J., Bao, S., Luo, C., Zu, Z., Han, Y., Zhang, W., Ren, K., Deng, K., Song, J., 2024. XiHe: A Data-Driven Model for Global Ocean Eddy-Resolving Forecasting. Technical Report, Cornell University, http://dx.doi.org/10.48550/arXiv.2402.02995, arXiv:2402.02995.

Watters, N., Tacchetti, A., Weber, T., Pascanu, R., Battaglia, P., Zoran, D., 2017. Visual interaction networks. http://dx.doi.org/10.48550/arXiv.1706.01433, ArXiv E-Prints arXiv:1706.01433.

Wooster, W.S., Bakun, A., McLain, D.R., 1976. The seasonal upwelling cycle along the eastern boundary of the North Atlantic. J. Mar. Res. 34 (2), 131–141.

Yang, X., Zhang, F., Sun, P., Li, X., Du, Z., Liu, R., 2022. A spatio-temporal graph-guided convolutional LSTM for tropical cyclones precipitation nowcasting. Appl. Soft Comput. 124, 109003. http://dx.doi.org/10.1016/j.asoc.2022.109003.

Yang, J., Zhang, T., Zhang, J., Lin, X., Wang, H., Feng, T., 2024. A ConvLSTM nearshore water level prediction model with integrated attention mechanism. Front. Mar. Sci. 11, http://dx.doi.org/10.3389/fmars.2024.1470320.

Zeng, X., Li, Y., He, R., 2015. Predictability of the loop current variation and eddy shedding process in the Gulf of Mexico using an artificial neural network approach. J. Atmospheric Ocean. Technol. 32 (5), 1098–1111. http://dx.doi.org/10.1175/JTECH-D-14-00176.1.

EGU26-10343, updated on 16 Jul 2026
https://doi.org/10.5194/egusphere-egu26-10343
EGU General Assembly 2026

Lagrangian evaluation of surface transport around the Canary Islands using drifter observations and OpenDrift simulations

Jacob S. Torres-Ojeda and Ángel Rodríguez Santana
ECOAQUA, Universidad de Las Palmas de Gran Canaria, Edificio de Ciencias Básicas, 35017, Las Palmas de Gran Canaria, SPAIN

The prediction of ocean surface trajectories remains a key challenge in coastal and island-influenced regions, were strong spatial variability limits model skill. Previous Lagrangian studies have shown the usefulness of drifter observations to assess trajectory predictability and to compare different sources of surface currents (e.g. Dagestad and Röhrs, 2019). In this context, Lagrangian approaches provide a direct and observation-based framework to evaluate surface transport.
This study assesses surface transport predictability around the Canary Islands using trajectories from two surface drifters (CODE/Davis type, drogued at 1 m depth) and numerical simulations performed with the OpenDrift framework (Dagestad et al., 2018). Simulations are forced with surface currents from the Iberia–Biscay–Ireland (IBI) regional ocean model distributed by the Copernicus Marine Environment Monitoring Service (CMEMS), and, where available, from the high-resolution coastal forecasting system SAMOA (Sotillo et al., 2019), operationally implemented for Spanish ports. Wind forcing is provided by ERA5 atmospheric fields, and wave-induced Stokes drift is included using IBI wave products from CMEMS.
From each observed drifter position, short-term forward simulations are performed to predict the subsequent drifter location. Model performance is quantified through the separation distance between simulated and observed positions, allowing a direct comparison of transport skill between different current products and forcing configurations.
The oceanic and atmospheric datasets used in this study correspond to operational or near-real-time products rather than fully consolidated reanalysis, reflecting realistic conditions for trajectory forecasting applications. The results reveal pronounced spatial and temporal variability in the separation between modeled and observed positions, with the relative performance of SAMOA and IBI depending on location and conditions, and neither consistently outperforming the other. While further improvements in transport predictability are expected once consolidated reanalysis products become available, the present results already provide a robust assessment of Lagrangian model skill under operational conditions.

Acknowledgments:
This work was supported by the projects SIRENA and SIRENA 2, funded by the collaboration of the Biodiversity Foundation of the Ministry for the Ecological Transition and the Demographic Challenge, through the Pleamar Program, and are co-financed by the European Union through the

EMFAF (European Maritime, Fisheries and Aquaculture Fund).

References:

Dagestad, K.-F., Röhrs, J., Breivik, Ø., & Ådlandsvik, B. (2018): OpenDrift v1.0: a generic framework for trajectory modelling, Geoscientific Model Development, 11, 1405–1420, https://doi.org/10.5194/gmd-11-1405-2018

Dagestad, K.-F., & Röhrs, J. (2011): Prediction of ocean surface trajectories using satellite derived vs. modeled ocean currents, Ocean Modelling. https://doi.org/10.1016/j.rse.2019.01.001

Sotillo, M. G., Cerralbo, P., Lorente, P., Grifoll, M., Espino, M., Sanchez-Arcilla, A., & Álvarez-Fanjul, E. (2019): Coastal ocean forecasting in Spanish ports: the SAMOA operational service, Journal of Operational Oceanography, 13, 37–54, https://doi.org/10.1080/1755876X.2019.1606765

Copernicus Marine Environment Monitoring Service (CMEMS): IBI Ocean Currents Product, https://doi.org/10.48670/moi-00027

Copernicus Marine Environment Monitoring Service (CMEMS): IBI Stokes Drift Product, https://doi.org/10.48670/moi-00025

Hersbach, H. et al. (2020): ERA5 global reanalysis, Copernicus Climate Change Service (C3S), https://doi.org/10.24381/cds.adbb2d47

EGU26

EGU26-15146, updated on 16 Jul 2026
https://doi.org/10.5194/egusphere-egu26-15146
EGU General Assembly 2026

Mitigating Voronoi-induced artifacts in GNN-based sea surface temperature forecasting using bathymetry-aware adaptive meshes

Giovanny Alejandro Cuervo Londoño[1], Ángel Rodríguez Santana[1], and Javier Sánchez[2]
[1]Instituto Universitario de Investigación en Acuicultura Sostenible y Ecosistemas Marinos (ECOAQUA), Universidad de Las Palmas de Gran Canaria, 35017 Las Palmas de Gran Canaria, Spain
[2]Centro de Tecnologías de la Imagen (CTIM), Instituto Universitario de Cibernética, Empresas y Sociedad (IUCES), Universidad de Las Palmas de Gran Canaria, 35017 Las Palmas de Gran Canaria, Spain.

Accurately forecasting Sea Surface Temperature (SST) is critical for understanding ocean dynamics, climate change impacts, and marine ecosystem management (Brito-Morales et al., 2020; Gattuso et al., 2018). In recent years, Graph Neural Networks (GNNs) have emerged as a powerful tool for spatiotemporal oceanographic forecasting, offering advantages over traditional Euclidean deep learning models by operating on unstructured grids (Liang et al., 2023; Zhang et al., 2025). However, the transition from structured satellite-derived data to mesh-based representations often introduces numerical artifacts, particularly due to the grid-to-mesh coupling mechanisms (Cuervo-Londoño et al., 2026; Cuervo-Londoño, Sánchez, et al., 2025).

This study investigates the origin of "Voronoi-induced artifacts" in GNN architectures applied to SST forecasting in the Northwest African region and the Canary Islands. We demonstrate that the grid-to-mesh association is algebraically equivalent to an order-k Voronoi partition (Cuervo-Londoño, Reyes, et al., 2025; Okabe et al., 2000), implying that the way nodes are distributed and how they associate with the underlying data grid significantly influences the quality of the predictions. To address these issues, we propose and evaluate four different mesh configurations: structured quadrangular meshes (Holmberg et al., 2024; Lam et al., 2023) and three unstructured approaches, including novel bathymetry-aware meshes.

Our findings reveal that connectivity plays a decisive role in mitigating artifact formation. Specifically, using approximately four connections per node under optimized grid-to-mesh association rules significantly reduces errors. Furthermore, the results show that densifying the node distribution according to the seabed's topography (bathymetry) not only reduce spatial artifacts but also increases forecast accuracy. The bathymetry-based meshes with optimized connectivity (3-4 connections) achieved a 30% improvement in performance compared to traditional structured mesh baselines. These insights suggest that incorporating geographical and topological priors into GNN design is essential for developing robust and reliable machine-learning surrogates for physical oceanography (Reichstein et al., 2019).

Acknowledgments: This work was supported by the projects SIRENA and SIRENA 2, funded by the collaboration of the Biodiversity Foundation of the Ministry for the Ecological Transition and the

Demographic Challenge, through the Pleamar Program, and are co-financed by the European Union through the EMFAF (European Maritime, Fisheries and Aquaculture Fund).

References

Cuervo-Londoño, G. A., Reyes, J. G., Rodríguez-Santana, Á., & Sánchez, J. (2025). Voronoi-Induced Artifacts from Grid-to-Mesh Coupling and Bathymetry-Aware Meshes in Graph Neural Networks for Sea Surface Temperature Forecasting. Electronics, 14(24), 4841. https://doi.org/10.3390/electronics14244841

Cuervo-Londoño, G. A., Sánchez, J., & Rodríguez-Santana, Á. (2025). Deep Learning Weather Models for Subregional Ocean Forecasting: A Case Study on the Canary Current Upwelling System (No. arXiv:2505.24429). arXiv.https://doi.org/10.48550/arXiv.2505.24429

Cuervo-Londoño, G. A., Sánchez, J., & Rodríguez-Santana, Á. (2026). Forecasting Sea Surface Temperature from Satellite Images with Graph Neural Networks. In M. Castrillón-Santana, C. M. Travieso-González, O. Deniz Suarez, D. Freire-Obregón, D. Hernández-Sosa, J. Lorenzo-Navarro, & O. J. Santana (Eds.), Computer Analysis of Images and Patterns (pp. 329–339). Springer Nature Switzerland. https://doi.org/10.1007/978-3-032-05060-1_28

Sostenibilidad y Cambio Climático Póster P125

PROYECTO SIRENA: MONITORIZACIÓN Y PREDICCIÓN DE RIESGOS EN ACUICULTURA OFFSHORE

L. Yubero-Lozoya 1 3, A.M. Mancho 1, A. Rodríguez-Santana 2 , A.J. González-Ramos 2 , A. García 2 , J.S. Torres 2, A. Marrero-Díaz 2
1 CSIC, Instituto de Ciencias Matemáticas, C/ Nicolás Cabrera 13-15, Campus Cantoblanco, 28049, Madrid, ESPAÑA. 2 ECOAQUA, Universidad de Las Palmas de Gran Canaria, Edificio de Ciencias Básicas, 35017, Las Palmas de Gran Canaria, ESPAÑA. 3 Escuela Técnica Superior de Ingenieros de Telecomunicación, Universidad Politécnica de Madrid, Av. Complutense, 30, 28040 Madrid, ESPAÑA

Abstract
Offshore aquaculture is carried out in open sea conditions, where control over environmental risks is inherently limited. These include anthropogenic events, such as discharges associated with maritime traffic, as well as natural phenomena such as algal blooms, which may affect aquaculture facilities. In this context, continuous monitoring is essential to anticipate and mitigate their impacts. The SIRENA project develops to this end advanced monitoring tools based on remote sensing, in situ observations, and mathematical modelling. Remote sensing enables the early detection of threats, while high-resolution models predict their evolution and potential impact, with results validated through in situ observations. Tested in the Canary Islands, this scalable system contributes to improving the resilience and sustainability of the aquaculture sector in the face of marine environmental challenges.

Resumen
La acuicultura offshore se desarrolla en mar abierto, donde el control sobre los riesgos ambientales es limitado. Estos incluyen eventos de origen antrópico, como vertidos asociados al tráfico marítimo, y fenómenos naturales, como proliferaciones de algas, que pueden afectar a las instalaciones acuícolas. En este contexto, la monitorización continua resulta esencial para anticipar y mitigar sus efectos. El proyecto SIRENA desarrolla con este fin herramientas avanzadas basadas en teledetección, observaciones in situ y modelización matemática. La teledetección permite la identificación temprana de amenazas, mientras que modelos de alta resolución predicen su evolución e impacto, siendo validados mediante observaciones in situ. Testado en Canarias, este sistema escalable contribuye a mejorar la resiliencia y sostenibilidad del sector acuícola frente a los desafíos del medio marino

Introducción
La acuicultura offshore requiere una monitorización continua frente a múltiples riesgos ambientales, ya que se desarrolla en un entorno altamente dinámico donde la predicción de las condiciones oceanográficas y de la evolución de posibles eventos nocivos sigue siendo un reto significativo. En esta presentación se abordan dos casos de uso implementados durante la ejecución del proyecto SIRENA que ilustran esta problemática. En el primer caso, el proyecto SIRENA ha constatado que la teledetección satelital ofrece una excelente cobertura espacial para detectar anomalías en la superficie del océano. En efecto, resultó clave, por ejemplo, en la identificación de las anomalías asociadas al vertido del buque Akhisar durante su repostaje en el Puerto de La Luz (Gran Canaria) en septiembre de 2024. Las limitaciones derivadas del tiempo de revisita de los satélites se complementaron mediante la modelización de la evolución del vertido, lo que permitió predecir y evaluar la posible afección a granjas acuícolas en la zona. El segundo caso evalúa la predicción mediante modelos a escala local, dado que sigue siendo un desafío. Modelos operacionales como el del Servicio Marino de Copernicus o, en España, el de Puertos del Estado carecen de resolución suficiente para capturar, por ejemplo, variaciones termohalinas críticas para los cultivos. Por ello, en SIRENA se ha implementado un modelo hidrodinámico tridimensional de muy alta resolución (45 m) para un enclave acuícola del sur de Gran Canaria. Las observaciones in situ han aportado mediciones en la columna de agua que han permitido calibrar y validar este modelo, mejorando su fiabilidad. No obstante, su elevado coste logístico, la dependencia de las condiciones meteorológicas y su limitado alcance espacial restringen su uso. Estas limitaciones pueden mitigarse mediante su complementariedad con la modelización, un aspecto clave que también ha sido abordado en SIRENA. El modelo interpola los vacíos espacio-temporales y transforma observaciones aisladas en una herramienta robusta, continua y predictiva, indispensable para anticipar el impacto de las amenazas sobre las instalaciones acuícolas.

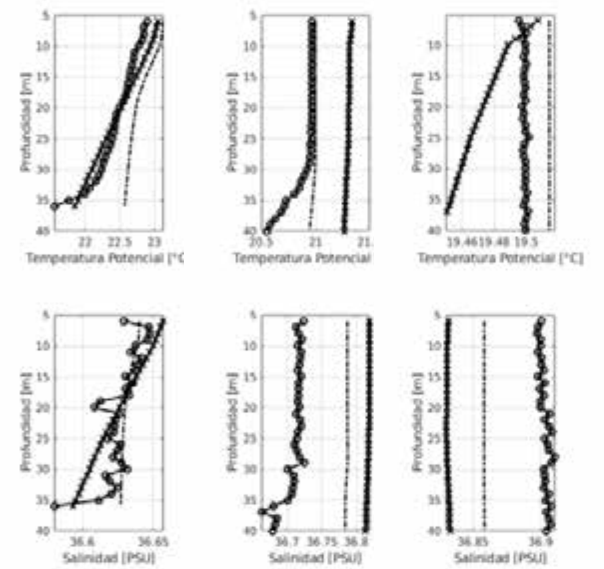
Material y métodos
En SIRENA se han utilizado imágenes de teledetección obtenidas por satélite y proporcionadas de manera gratuita a través del programa Copernicus. En particular, para la detección de anomalías en superficie asociadas a vertidos, las imágenes radar proporcionadas por Sentinel-1 resultan especialmente adecuadas, ya que no se ven afectadas ni por la presencia de nubes ni por la falta de iluminación solar.
Para la obtención de medidas in situ se ha empleado instrumentación oceanográfica diversa, incluyendo sondas multiparamétricas, perfiladores de turbulencia y drifters tipo CODE-Davis.
Para evaluar el transporte de sustancias por las corrientes marinas se han utilizado datos proporcionados por Puertos del Estado. A partir de estos datos se han realizado simulaciones mediante partículas pasivas superficiales (García-

Garrido et al., 2016; García-Sánchez et al., 2021). Asimismo, estos datos se han empleado para inicializar y forzar en las fronteras el modelo hidrodinámico de muy alta resolución desarrollado en ROMS, al que adicionalmente, se han incorporado forzamientos atmosféricos del viento, la radiación solar y la precipitación. Tras su ajuste con observaciones in situ, el modelo proporciona perfiles verticales horarios de temperatura, salinidad y corrientes muy útiles para la monitorización de las granjas acuícolas.

Resultados y discusión

Fig. 1

Fig 2



En la Figura 1 se observa la concordancia entre las simulaciones (token en amarillo) y los puntos afectados por el vertido (en rojo), identificados a partir de una imagen radar. En la Figura 2 se comparan los resultados del modelo oceánico de alta resolución con el modelo de Puertos del Estado (PE) y con observaciones in situ de temperatura y salinidad. Se muestran los perfiles a un mismo lugar y hora en las fechas 28 de octubre de 2025, 9 de diciembre de 2025 y 22 de enero de 2026. Estos muestran una clara correspondencia entre el modelo desarrollado y las observaciones in situ.

Bibliografía

Abdel-Tawwab, M.et al. 2025. Dietary acidifiers blend modulates growth, antioxidant, immunity, cytokine performances, and the disease control in Nile tilapia fingerlings fed on high-soybean feeds. Fish Physiol Biochem 51, 194.

Álvarez, E., M. et al. 2016. SAMOA: la oceanografía operacional al servicio de las Autoridades Portuarias. En: Libro de Resúmenes. Valle C. et al. (Eds,). IV Encuentro Oceanografía Física Española. Universidad de Alicante, Alicante. 143 pp. ISBN: 978-84-16724-17-8.

Garcia-Garrido, V.J., et al. 2016. A dynamical systems perspective for a real-time response to a marine oil spill. Marine Pollution Bulletin 112, 201-210.

García-Sánchez, G., et al. 2021. Very High Resolution Tools for the Monitoring and Assessment of Environmental Hazards in Coastal Areas. Frontiers in Marine Science, 7:605804.

Pereiro, D., et al. 2025. Operational oceanographic services in support of aquaculture and biodiversity in Galway Bay, Ireland. EGU General Assembly 2025, EGU25-4517.

Sotillo, M. G., et al. 2015. The MyOcean IBI Ocean Forecast and Reanalysis Systems: operational products and roadmap to the future Copernicus Service. Journal of Operational Oceanography. 8, 63–79.

Sotillo, M. G.,et al. 2020. Coastal ocean forecasting in spanish ports: the SAMOA operational service. Journal of Operational Oceanography, 13(1), 37-54.

Agradecimientos

Este trabajo se ha llevado a cabo en el marco de los proyectos SIRENA y SIRENA 2. Los proyectos SIRENA y SIRENA 2 se desarrollan con la colaboración de la Fundación Biodiversidad, del Ministerio para la Transición Ecológica y el Reto Demográfico, a través del Programa Pleamar, y se cofinancian por la Unión Europea por el FEMPA (Fondo Europeo Marítimo, de Pesca y de Acuicultura).