scientific reports

Check for updates

OPEN

Accurate Mediterranean Sea forecasting via graph-based deep learning

Daniel Holmberg[1,2]✉, Emanuela Clementi[2], Italo Epicoco[2,3] & Teemu Roos[1]

Accurate ocean forecasting systems are essential for understanding marine dynamics, which play a crucial role in sectors such as shipping, aquaculture, environmental monitoring, and coastal risk management. Traditional numerical solvers, while effective, are computationally expensive and time-consuming. Recent advancements in machine learning have revolutionized weather forecasting, offering fast and energy-efficient alternatives. Building on these advancements, we introduce SeaCast, a neural network designed for high-resolution regional ocean forecasting. SeaCast employs a graph-based framework to effectively handle the complex geometry of ocean grids and integrates external forcing data tailored to the regional ocean context. Our approach is validated through experiments at a high horizontal resolution using the operational numerical forecasting system of the Mediterranean Sea, along with both numerical and data-driven atmospheric forcings. Results demonstrate that SeaCast consistently outperforms the operational model over the conventional 10-day forecast window and further extends skillful predictions to 15 days, marking a significant advancement in regional ocean prediction.

Keywords Regional ocean forecasting, Graph neural networks, Learned simulation

Predicting sea dynamics is a formidable scientific challenge, driven by the need to anticipate changes in ocean conditions that affect weather systems, marine ecosystems, and a wide range of maritime activities[1]. While the need for improved ocean and coastal data is global, actionable decisions in sectors such as shipping, marine resource management, and coastal planning often depend on regional, high-resolution models that can deliver accurate forecasts tailored to local conditions[2–6].

Here, we focus on the Mediterranean Sea, a region characterized by intense mesoscale activity, complex coastlines, and dense coastal populations. Regional ocean forecasts in this area are important for managing marine operations and mitigating coastal risks. Accurate and timely predictions empower local agencies to address a wide range of challenges, from water quality monitoring[7] and oil spill containment[8] to maritime route optimization[9]. Currently, the *Mediterranean Forecasting System* (MedFS)[10] delivered through the Copernicus Marine Service (CMEMS) provides medium-range (10 days) forecasts using a physics-based two-way coupled wave-current modeling system, and serves as the operational reference system for the region.

Despite their accuracy, numerical models such as MedFS are computationally intensive and time-consuming to run. Recent advancements in machine learning-based weather prediction (MLWP) have demonstrated the potential for fast and performant alternatives. Several autoregressive machine learning models now rival or surpass state-of-the-art physics-based systems used by meteorological agencies worldwide. Notable developments include architectures based on Transformers[11,12], neural operators[13], and graph neural networks (GNNs)[14–16].

While these advances have been transformative in the atmospheric domain, its application to ocean forecasting remains relatively nascent. Existing work has focused mostly on forecasting the global ocean, at climate[17], seasonal[18,19], and medium-range[20–22] timescales. However, in contrast to global approaches, operational regional forecasting requires realistic lateral boundary treatment and higher horizontal resolution that remain largely unexplored in current data-driven ocean frameworks. Regional studies to date have primarily targeted the forecasting of a small set of surface variables, such as sea ice[23], sea surface height (SSH)[24], or surface currents and temperature[25]. However, regional data-driven ocean forecasting systems for depth-resolved prediction in high resolution operational settings are still lacking.

To address this gap for the Mediterranean Sea, we introduce *SeaCast*, an autoregressive machine learning model designed for regional ocean forecasting, extending the methodology from weather forecasting using

[1]Department of Computer Science, University of Helsinki, Helsinki, Finland. [2]CMCC Foundation-Euro-Mediterranean Center on Climate Change, Lecce, Italy. [3]Department of Engineering for Innovation, University of Salento, Lecce, Italy. ✉email: daniel.holmberg@helsinki.fi

hierarchical GNNs[16]. Our approach involves several key features that enable accurate prediction of ocean states: (1) we adapt the graph creation, training, and evaluation processes to accommodate the irregular geometry of ocean grids; (2) the model includes relevant atmospheric forcing near the sea surface; and (3) lateral boundary forcing is applied to account for water inflow and outflow, ensuring compatibility with the ocean at large. SeaCast is evaluated in an operational setting alongside MedFS, including validation against observational data. In addition, we conduct targeted experiments to assess the influence of atmospheric forcing components and the effects of training period length.

Results

SeaCast: data-driven regional ocean forecasting

SeaCast is a data-driven forecasting model for the Mediterranean Sea that delivers 15-day forecasts across 18 depth levels on a $1/24^\circ$ (approximately 4 km) horizontal grid, matching the resolution of the operational MedFS system (see Supplementary A). The model predicts key physical ocean variables, including depth-resolved zonal and meridional currents, salinity, and temperature, along with SSH, resulting in 73 predicted fields in total. A major advantage of SeaCast is its computational efficiency: it produces a full 15-day forecast in just 20 seconds on a single GPU, compared to approximately 70 minutes required by MedFS to generate a 10-day forecast on 89 CPU cores, using a 120-second timestep and producing outputs at 141 depth levels. Although the two systems differ in how they operate, the data-driven approach represents a significant speedup in producing upper ocean forecasts compared to what was previously possible.

The model architecture follows an encode–process–decode framework[26] operating on a hierarchical graph mesh tailored to the Mediterranean basin (see Supplementary B). As illustrated in Fig. 1, the input sea state and atmospheric forcing are first encoded onto a coarser multi-resolution mesh representation. The latent features are then processed by GNN layers in a hierarchical fashion, enabling the model to capture both short- and long-range ocean interactions. The processed output is subsequently decoded back onto the original high-resolution grid. Rather than directly predicting the next sea state, the model learns the tendency, i.e. the expected change over a one-day interval, which is added to the current state to obtain the forecast. Dynamic boundary conditions are then incorporated to generate the complete next sea state. This predicted state is then fed back into the model as input for the next step, enabling the model to produce forecasts at different lead times by repeating the application of the same cycle in an autoregressive manner. Unlike multiscale models such as GraphCast[15], which connect nodes on a single mesh level, the hierarchical approach we employ separates the domain into multiple,

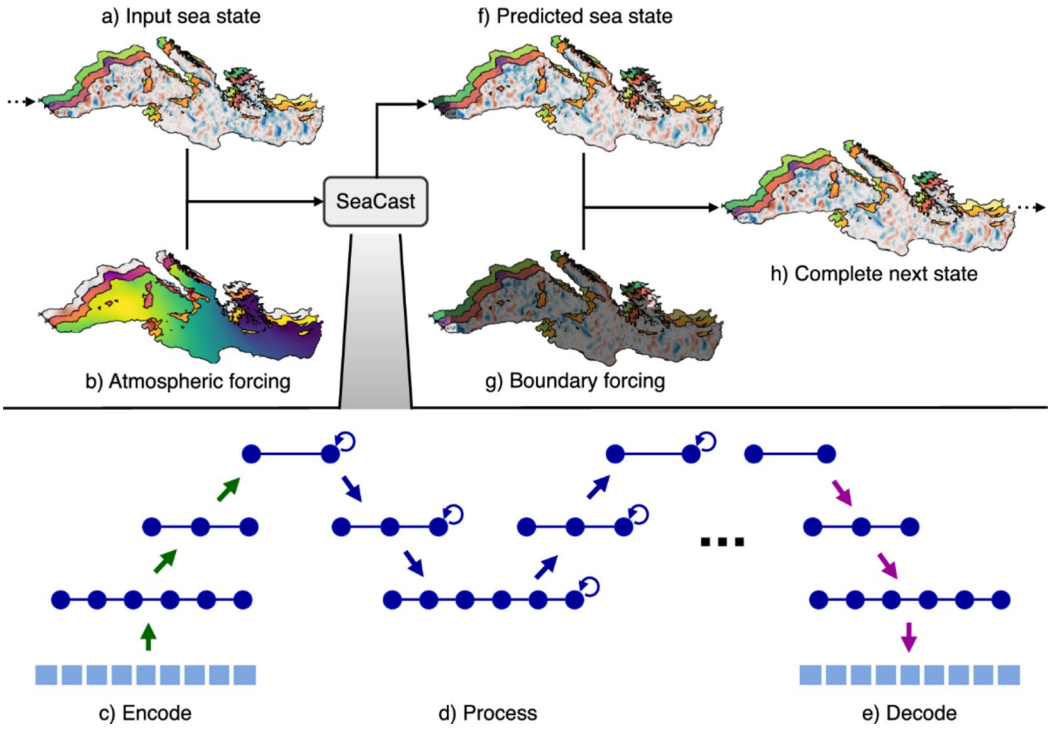


Fig. 1. SeaCast performs autoregressive ocean forecasting using a graph neural network (GNN). (**a**) The input sea state and (**b**) atmospheric forcing are projected onto a coarser mesh representation through the (**c**) encoder. (**d**) GNN layers process this latent representation on a hierarchical mesh. The output is then (**e**) decoded back onto the original grid to form (**f**) a predicted sea state. (**g**) Boundary conditions (here exaggerated in size) are incorporated to produce (**h**) the complete next state. This forecast is then fed back into the system as a new input (as shown by the arrow looping from h to a), enabling multi-step forecasting through repeated application of the encode–process–decode cycle.

distinct mesh levels. This design results in more uniform connectivity from the mesh to the grid, helping to mitigate artifacts associated with nodes having variable neighborhood sizes[16].

The atmospheric forcing used in SeaCast includes 10-meter wind stress components, 2-meter temperature, and mean sea level pressure (MSLP), combined with the sine and cosine of the day of year as seasonal indicators. Open boundary conditions are handled by overwriting the predicted sea state with the true state in boundary regions located at the Strait of Gibraltar and the Dardanelles Strait during training, and with MedFS data during evaluation. This approach is necessary to ensure realistic representation of inflow and outflow dynamics for the region considered[27].

SeaCast is trained on 35 years (1987–2021) of Mediterranean reanalysis daily mean data[28] and fine-tuned using two additional years (2022–2023) of daily operational analysis[10]. Fine-tuning serves multiple purposes: it exposes the model to more recent sea states, enables the use of analysis fields as initial conditions in an operational setting, and allows the model to adapt to updates introduced in the operational MedFS system that are absent in the reanalysis. The atmospheric forcings driving the surface of the model are taken from the European Centre for Medium-Range Weather Forecasts (ECMWF) ERA5 reanalysis data[29] during the training phase, and during testing either from the ECMWF *ensemble control forecast* (ENS)[30] or the recent *artificial intelligence forecasting system* (AIFS)[31]. SeaCast is evaluated on a daily test set with initializations spanning from the beginning of July 2024 to the end of December 2024, each producing a 15-day forecast. Performance is assessed using both model-based reference fields and satellite observations, and benchmarked against the operational MedFS. SeaCast uses the same initial conditions available to MedFS making for a fair comparison. Further details on the method are provided in "Methods".

Comparing SeaCast to the operational MedFS

SeaCast is evaluated against the operational MedFS, the primary regional ocean forecasting system for the Mediterranean Sea. MedFS provides deterministic forecasts up to 10-days lead, adhering to the 10-day standard of CMEMS ocean products, and also the extent of ECMWF's legacy high-resolution (HRES) atmospheric forecasts used as surface forcing. Recent changes to the ECMWF Integrated Forecasting System (IFS), extending medium-range predictions from 10 to 15 days in Cycle 48r1[32], was taken into account when developing SeaCast. By default, the model uses AIFS as atmospheric forcing, and in "Effect of atmospheric forcing", also ENS for comparison. By using these products, and persisting the lateral boundary conditions the last 5 days, SeaCast is capable of producing 15-day forecasts. This extended forecast horizon represents an important step forward, enabling potential earlier warnings of marine extremes and enhancing medium-range planning capabilities.

To benchmark performance, we evaluate the forecast skill of both SeaCast and MedFS across six key ocean variables: zonal and meridional currents, salinity, temperature, sea surface temperature (SST), and sea level anomaly (SLA). Subsurface variables are validated against daily mean MedFS analysis fields. These gridded analysis fields, which assimilate quality-controlled observations into the physics model, provide a stable reference for subsurface conditions. SST and SLA are compared to Level-3 (L3) satellite observations. Specifically, SST predictions from the model's uppermost layer are evaluated against merged multi-sensor satellite data at $1/16^\circ$ resolution[33], while SSH forecasts are converted to SLA and compared to along-track 5 Hz altimeter measurements from multiple satellite missions (see "Satellite data" for further details).

A persistence baseline is included as a naive benchmark, generated by repeating the initial conditions over all lead times. This baseline provides a conservative lower bound for model performance. As shown in Fig. 2, both SeaCast and MedFS significantly outperform persistence across all variables. Notably, SeaCast demonstrates improved forecast skill relative to MedFS, and the gap between the two tends to increase at extended lead times.

Figure 3 presents a depth-resolved comparison of SeaCast and MedFS, showing the normalized RMSE difference, defined as $(\mathrm{RMSE}_{\mathrm{SeaCast}} - \mathrm{RMSE}_{\mathrm{MedFS}})/\mathrm{RMSE}_{\mathrm{MedFS}}$, across all vertical levels and lead times. Negative values indicate improved performance by SeaCast. The improvements are generally larger at longer forecast lead times, which is in line with results for the global ocean[22]. For temperature and current components, the relative gains are strongest near the surface, while for salinity the largest improvements are found at greater depths. SeaCast generally does not outperform MedFS that much at the lowest level included here (192 m depth), likely because influence from below this level are unaccounted for and affect the conditions there.

To further assess regional performance, Fig. 4 shows the spatial distribution of normalized RMSE differences in SST between SeaCast and MedFS, for forecast lead times of 1, 4, 7, and 10 days. Across the majority of the Mediterranean basin, SeaCast exhibits increasing skill relative to MedFS as the lead time grows. These improvements are particularly notable in the Western basin, especially in the Alboran Sea, and in the Aegean Sea, areas characterized by intense mesoscale processes. Meanwhile, MedFS demonstrates enhanced skill in the Adriatic and Ligurian Seas. This heterogeneous spatial pattern of skill may be linked to differences in the atmospheric forcing used by the two systems.

Additional evaluation metrics and results are presented in Supplementary C and D. These include validation against independent in-situ observations (Supplementary D.1), depth-resolved RMSE relative to the analysis fields (Supplementary D.2), and vertical error profiles (Supplementary D.3). When compared against the in-situ dataset, SeaCast shows improved skill relative to MedFS for both currents and temperature across all lead times. For salinity, SeaCast also achieves lower RMSE; however, due to the sparsity of observations, the 50% confidence intervals overlap for SeaCast, MedFS, and even the persistence baseline, making this comparison less conclusive.

Beyond basin averaged RMSE-based diagnostics, visual examples of forecasted fields are provided in Supplementary E. These allow us to assess whether the model preserves spatially coherent ocean structures, including mesoscale eddies and frontal features. The predicted fields retain coherent vortices, sharp frontal gradients, and realistic filamentary structures at the correct spatial scales over the 15-day forecast horizon. In particular maps of currents and sea level (Figs. S17, S18, S21) indicate that SeaCast manages to capture key dynamical patterns of the Mediterranean circulation such as the Alboran gyres, the Gulf of Lion gyre, the Middle

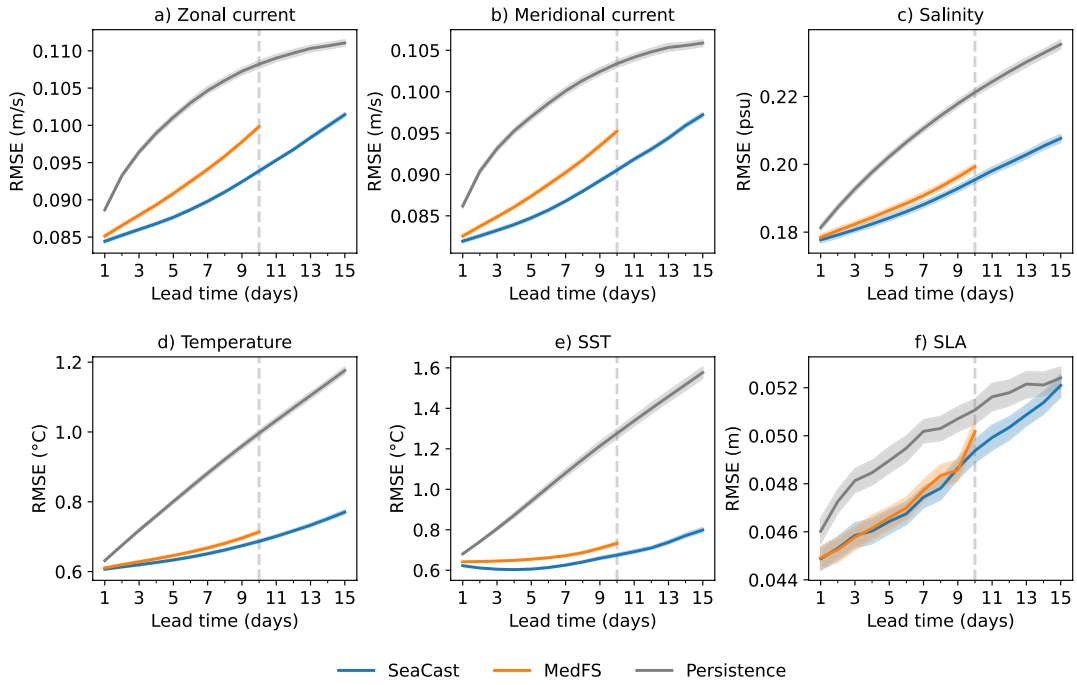


Fig. 2. RMSE as a function of lead time for SeaCast, MedFS, and a persistence baseline. Sea state variables with several vertical levels (zonal and meridional currents, salinity, and temperature) are evaluated against analysis fields where the error is an average over all depths. SST and SLA are validated against L3 satellite observations, using the uppermost modeled temperature level for the SST comparison. The shading corresponds to the 50% confidence intervals estimated via bootstrapping.

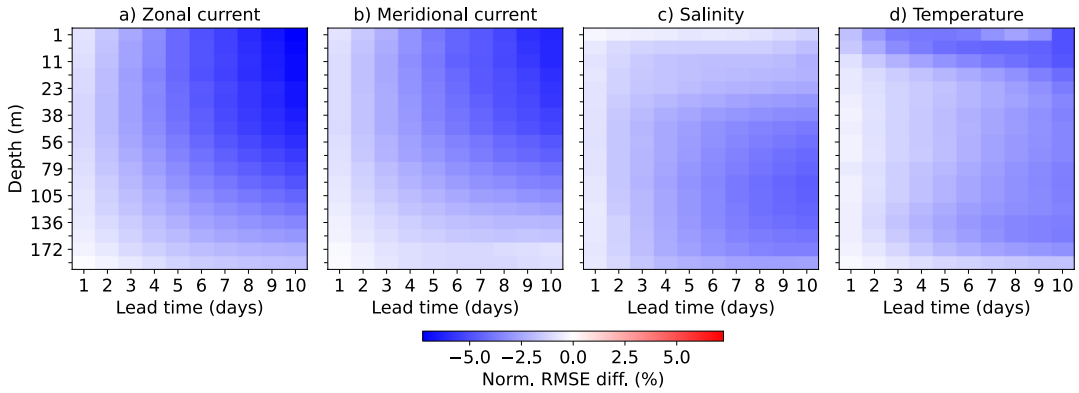


Fig. 3. Normalized RMSE difference between SeaCast and MedFS across all depth levels, evaluated against analysis fields. Negative values indicate better performance by SeaCast. Improvements are most pronounced near the surface for temperature and the current components, and at greater depths for salinity.

and South Adriatic gyres, the Mersa Matruh Gyre System as well as the Alegrian current, the Liguro-Provenal-Catalan current, the Adriatic coastal currents and the Levantine currents. The horizontal maps also confirm that the quality of the forecast degrades with the increased forecast lead time as expected.

Detecting high temperatures

Machine learning–based forecasting methods that optimize the mean squared error (MSE) such as we do are trained to predict the expected value of the forecast distribution. As a result, they could potentially struggle with accurately capturing extreme events. In the marine context, one such extreme is unusually warm SST.

To evaluate SeaCast's ability to predict temperature extremes, we draw inspiration from the marine heatwave (MHW) definition proposed by Hobday et al.[34], where an event is defined as SST exceeding the 90th percentile of the daily climatology for at least five consecutive days, based on an 11-day rolling mean. In our case, we adopt a simplified criterion by identifying individual days when the daily SST exceeds the 90th percentile threshold, allowing us to produce high temperature detections rates at each of the 15 lead times. The SST climatology is

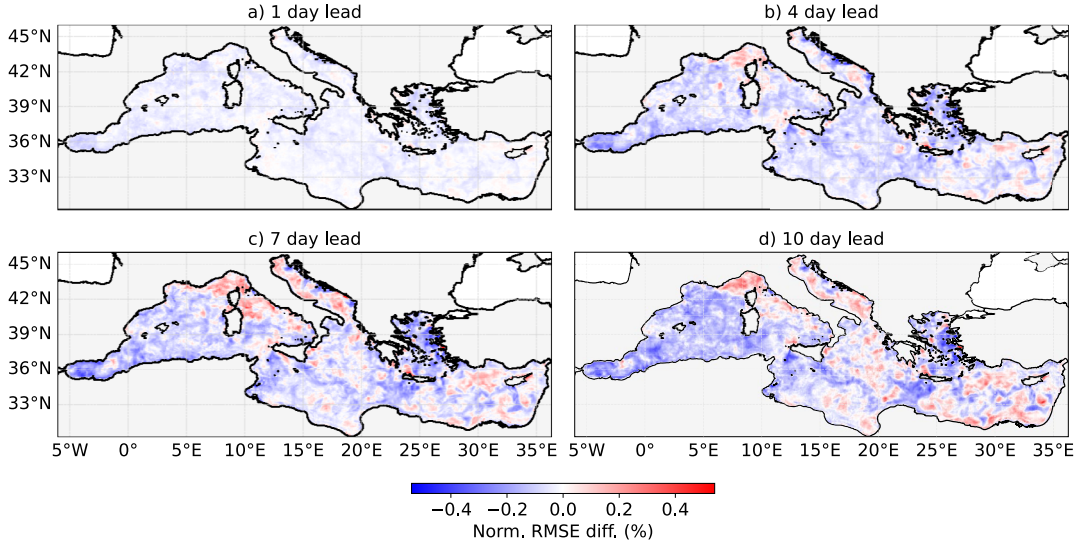


Fig. 4. Spatial distribution of normalized RMSE difference in SST between SeaCast and MedFS, evaluated against L3 satellite data. Results are shown for lead times of 1, 4, 7, and 10 days. Negative values indicate less error by SeaCast. Relative skill increases with lead time across most of the Mediterranean.

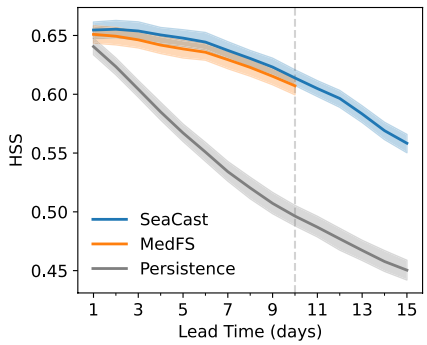


Fig. 5. Heidke Skill Score (HSS) for detecting SST anomalies above the 90th percentile, relative to climatology. Results are shown for SeaCast, MedFS, and the persistence baseline. The shading corresponds to the 50% confidence intervals estimated via bootstrapping.

computed from L3 satellite observations spanning every day in the years 2008 to 2023, using an 11-day rolling mean. For each calendar day, the 90th percentile is calculated to define the threshold for extreme temperature events.

We then evaluate each model's performance in detecting high temperatures using the Heidke Skill Score (HSS) at each forecast lead time, and estimate 50% confidence intervals via bootstrapping. Higher HSS values indicate more accurate classification of events relative to random chance. As shown in Fig. 5, both SeaCast and MedFS significantly outperform the persistence baseline. Furthermore, SeaCast slightly outperforms MedFS also on temperature extremes. An additional advantage of SeaCast is its ability to provide forecasts up to 15 days, compared to the 10-day horizon of MedFS, offering potential for earlier warnings of MHWs.

Effect of atmospheric forcing

Atmospheric forcing plays a critical role in driving ocean dynamics, particularly near the sea surface. To assess the sensitivity of SeaCast to different atmospheric inputs, we perform a controlled ablation experiment by randomly permuting each forcing variable across the spatial grid dimension during inference. This preserves the statistical distribution of the variable but removes its spatial coherence, effectively making it uninformative while leaving all other inputs intact. The resulting performance degradation reveals the relative importance of each atmospheric variable for ocean forecasting.

Figure 6 presents the normalized RMSE difference per lead time for each variable, evaluated against the original, unperturbed SeaCast model. The results indicate that wind stress is one of the most critical drivers across all ocean state variables. By transferring momentum from the atmosphere to the ocean, wind stress generates horizontal currents and drives vertical transport. Thus, the results show that for zonal and meridional currents in particular, wind stress is the sole contributor to predictive skill among the atmospheric forcing components.

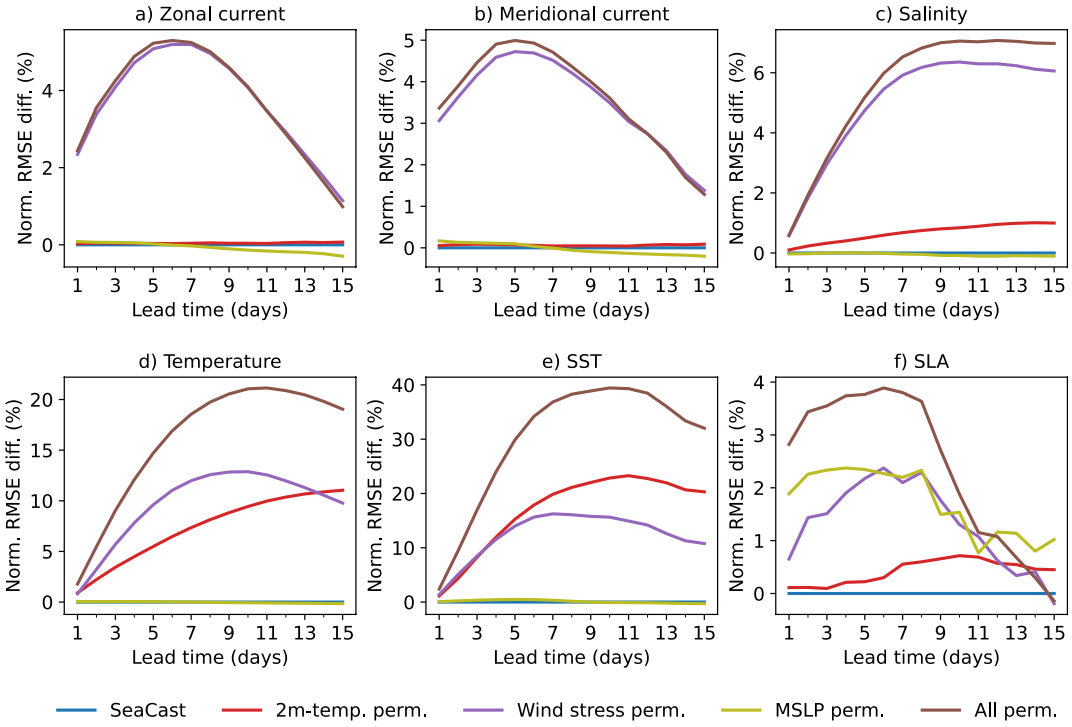


Fig. 6. Impact of permuted atmospheric forcings on forecast skill. Normalized RMSE difference per lead time is shown relative to the original SeaCast configuration with unmodified atmospheric inputs included as a reference line in blue. Each panel shows a different predicted ocean variable, while the curves correspond to the permuted atmospheric forcing variables. Larger positive values indicate greater importance of the corresponding variable.

Salinity forecasts are also strongly influenced by wind stress, an intuitive result given its central role in driving surface mixing and the lateral advection of freshwater. The second most influential variable for salinity is the 2-meter air temperature, which indirectly affects surface buoyancy and stratification triggering convective overturning and mixing fresh or salty surface water into the subsurface[36]. For temperature profiles, wind stress again dominates across depth. Near the surface, however, 2-meter temperature becomes increasingly important due to its direct influence on air–sea heat flux, particularly through modulation of the sensible heat flux and its impact on vertical mixing in the subsurface[37].

For SLA, the dominant atmospheric driver is MSLP, which directly affects the sea level through the inverted barometer effect[38] and large-scale ocean response. Moreover, MSLP gradients give rise to surface wind stress through atmospheric geostrophic balance[39], making wind stress the second most significant contributor to SLA.

The patterns we see here confirm known ocean–atmosphere coupling mechanisms and support the robustness of SeaCast's sensitivity to physically meaningful drivers. A depth-resolved analysis of the effect of atmospheric forcing on individual variables is further provided in Supplementary D.2.

Additionally, we assess SeaCast's performance using two different atmospheric forcing products: AIFS and ENS. Both forcings include the same set of variables, providing a consistent basis for comparison. Figure 7 presents scorecards of normalized RMSE differences across depth levels. Results indicate that the AIFS forcing leads to improved forecast skill for currents and temperature, particularly at longer lead times, with the most notable gains observed in SST. This is consistent with the evaluation of AIFS itself, which demonstrates lower forecast errors compared to its numerical counterpart beyond day one[31]. However, for salinity, SeaCast with ENS forcing exhibits slightly better performance in the upper layers here.

Effect of training period

To evaluate the impact of training data duration and fine-tuning on model performance, we compare SeaCast variants trained on different historical timespans and examine the effect of omitting fine-tuning. Figure 8 presents the normalized RMSE difference relative to a persistence baseline across forecast lead times, with MedFS included for reference. The full SeaCast model is trained on 35 years of reanalysis data (January 1987 to December 2021), followed by fine-tuning on 2 years of more recent analysis data (January 2022 to December 2023). We compare this to a version trained solely on the 35-year reanalysis dataset, without fine-tuning. Additionally, we include a shorter-term model variant trained on 8 years of reanalysis data (January 2014 to December 2021), both with and without fine-tuning on the same 2-year analysis period. This shorter-term model is denoted as SeaCast (10y). The models are trained using the same number of epochs and follow identical epoch-wise learning rate schedules. As a result, SeaCast (10y, w/o finetuning) undergoes fewer optimization steps during the pre-training phase compared to the full model, and requires a smaller computational budget (6 h vs. 20.5 h on 64 GPUs).

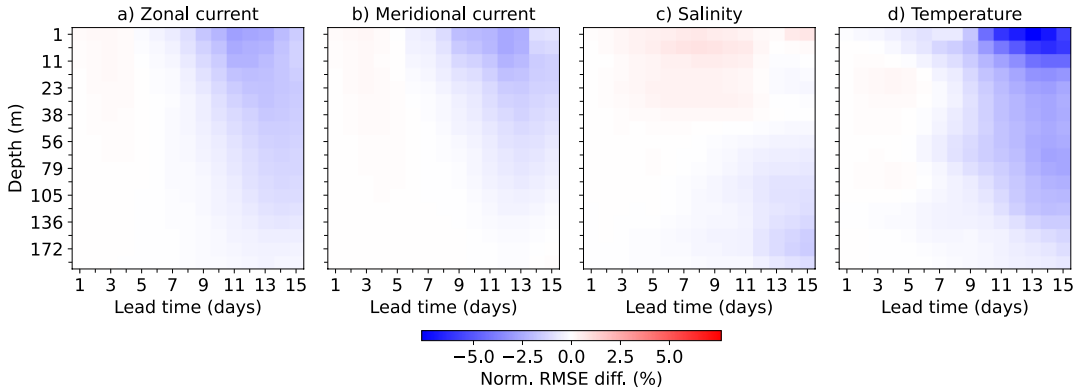


Fig. 7. Normalized RMSE difference scorecard comparing SeaCast performance when forced with AIFS versus ENS atmospheric inputs. Evaluations are performed against analysis fields across all depth levels. Negative values indicate less error when using AIFS forcing.

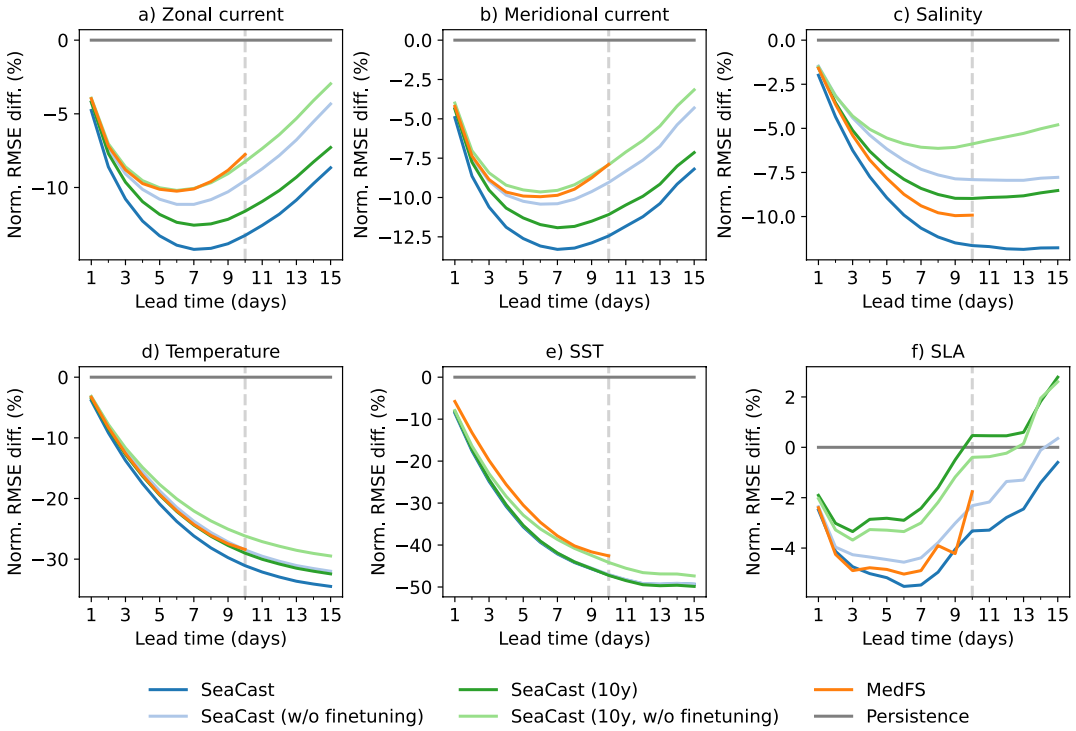


Fig. 8. Normalized RMSE difference per lead time for SeaCast variants and MedFS compared to a persistence baseline. SeaCast is trained on 35 years of reanalysis (1987–2021) with 2-year fine-tuning (2022–2023). The 10y variant uses only 8 years of reanalysis (2014–2021), with or without the same fine-tuning. Results show the impact of training duration and fine-tuning on forecast performance.

The results show that the 10-year variants of SeaCast generally perform on par with or slightly better than MedFS for zonal and meridional currents, temperature profiles, and sea surface temperature. However, for salinity and SLA, only the full SeaCast model, trained on 35 years of reanalysis data and fine-tuned on analysis data, outperforms MedFS. The SLA error for the fine-tuned 10-year variant is slightly higher than that of the base version, which may seem counterintuitive. This can be partly explained by the fact that SLA evaluation is performed along satellite tracks, whereas training is done on regularly gridded simulation outputs. Minimizing MSE on simulated fields does not guarantee lower error on satellite tracks, even if SLA is assimilated into the physics-based model. Moreover, the satellite tracks yield a sparser and more variable target compared to other fields, producing a wider confidence interval, as observed in Fig. 2. On other fields, the fine-tuned model consistently outperforms the base version. These findings are encouraging for regional settings with access to limited historical data, suggesting that even 10 years of reanalysis data and a more modest compute budget can produce forecast skill competitive with numerical forecast models.

Fine-tuning on analysis data yields consistent performance improvements across variables, which is expected given that several evaluation targets are derived from the same analysis system. These analysis states are more up-to-date than the reanalysis and incorporate system enhancements not present in the older data, including open boundary conditions at the Dardanelles Strait, coupling with a wave model[40], and tidal forcing[41]. By fine-tuning on this updated dataset, SeaCast adapts to the characteristics of the current operational system, leading to better alignment with the evaluation targets and improved overall forecast accuracy.

Discussion

In this study, we introduced SeaCast, a graph-based machine learning model designed for high-resolution regional ocean forecasting. SeaCast learns directly from historical reanalysis and analysis data to predict key ocean variables in the Mediterranean Sea. The model demonstrates improved skill for all modeled variables across all depth levels compared to the operational MedFS when considering both analysis fields and satellite observations as references. Another benefit of SeaCast is its speed: once trained, it delivers a full 15-day, 18-level forecast on a $1/24^\circ$ grid in just 20 s on a single GPU, substantially faster than the physics-based model on a CPU cluster. This performance can enable new capabilities for Mediterranean forecasting such as rapid scenario testing and ensemble forecasting with many members.

Our sensitivity experiments underscore the central role of atmospheric forcing, where wind stress emerges as a dominant driver of predictive skill across all ocean variables, with surface temperature and MSLP exerting secondary influence motivated by the air–sea heat-flux and inverted-barometer effects. These results suggest that physically meaningful coupling mechanisms are captured by the data-driven framework. We further demonstrate that the length and recency of the training period materially affect forecast accuracy. The 10-year variant of SeaCast achieves performance on par with MedFS for currents, temperature, and SST, while relying on very attainable historical data archives and computational resources.

A key conceptual distinction between SeaCast and conventional dynamical models lies in how physical consistency is handled. Like other recent data-driven ocean forecasting systems[20–22], SeaCast does not explicitly enforce conservation of mass, energy, or momentum. Instead, it learns to generate physically plausible forecasts by training on a large number of simulation states with assimilated observations, thereby capturing ocean dynamics statistically from the training data. An alternative approach to achieve stronger physical guarantees could be to combine a differentiable ocean solver with machine-learning components, similar to neural general circulation models for weather[35], which embed the governing equations directly into the architecture while learning the effect of unresolved processes such as cloud formation and subgrid-scale dynamics.

Despite the promising outcomes, several avenues remain for future improvement. First, enhancements to the underlying reanalysis systems, in line with recent progress on MedFS[10], could further increase the fidelity of data-driven forecasting models. Additionally, increasing the temporal resolution of archived datasets, e.g. from daily to 6-hourly fields, would provide richer training data and enable models to capture diurnal variability. Furthermore, recent work has demonstrated that machine learning can be used to produce continuous-time forecasts[42,43], a concept that could potentially be adapted to ocean modeling.

While SeaCast uses MedFS fields to impose lateral boundary conditions, leveraging its existing regridded connection to the global ocean, this setup introduces a degree of dependence on the operational system it is compared against. As a result, the sea dynamics inside the lateral boundaries are the same as MedFS, whereas the interior evolution is governed entirely by SeaCast. This approach ensures realistic boundary conditions and the comparison remains meaningful for assessing predictive skill within the regional domain. However, to create a truly standalone data-driven regional forecasting model one could take inspiration from early studies in limited area weather modeling with machine learning, where a globally stretched grid[44] or learned mappings from surrounding forecasts[45] have been used to handle boundary conditions. Integrating both data-driven global ocean and atmospheric forecasts, which can be rolled out arbitrarily far, into machine learning-based regional ocean models would enable similarly long lead-time predictions for these models as well. This opens up the question of how far into the future such forecasts can remain skillful.

The interdependence of physics, biogeochemistry, and wave dynamics in the MedFS system[10] suggests a natural extension for machine learning frameworks: a coupled neural architecture that jointly predicts sea state, biochemical tracers, and wave variables, enabling truly holistic forecasts. While these particular components are usually represented on compatible grids, recent advances in foundation models for weather and climate have shown that even heterogeneous spatial and temporal datasets, can be effectively used for training, resulting in models that generalize across diverse prediction tasks[46,47]. Adopting this paradigm for ocean forecasting could enable the development of transferable representations, supporting scalable and adaptable forecasting systems across regions and variable types.

Another direction of future work is developing probabilistic ocean forecasting systems. While SeaCast currently produces deterministic forecasts akin to traditional operational ocean forecasting systems, emerging generative machine learning methods for weather forecasting emphasize uncertainty quantification through ensemble generation[48,49]. These approaches enable better decision-making by accounting for forecast spread and producing potentially more physically consistent ensemble members, which could be very useful for risk-sensitive applications such as early warning systems and coastal hazard planning.

In conclusion, SeaCast demonstrates the viability of machine learning for regional ocean forecasting by outperforming a physics-based system. Beyond the Mediterranean, the same framework could be adapted to other regional settings such as the Bay of Biscay or the Baltic Sea. This work also presents novel sensitivity experiments in data-driven ocean forecasting, revealing the roles of atmospheric forcing components and training period characteristics. Further improvements in data fidelity, model coupling, and probabilistic prediction will be key to realizing the full potential of machine learning–based regional ocean forecasting.

Methods

Problem definition

The forecasting problem is characterized by mapping a sequence of initial states $X^{-p:0} = (X^{-p}, ..., X^0)$, where p is the length of the past window, to a sequence of future states $X^{1:T} = (X^1, \ldots, X^T)$, where T is the length of the forecast. Each state $X^t \in \mathbb{R}^{N \times d_x}$, contains d_x variables at N locations represented on a grid. Variables include both those at multiple vertical levels and single-level surface measures. In addition to this, forcing inputs $F^{1:T} = (F^1, ..., F^T)$ encompassing known dynamic factors relevant to the forecasting problem, are also available.

SeaCast

Graph-based neural forecasting

Initial states typically cover the two preceding time steps $X^{-1:0}$, enabling the capture of first-order dynamics[15]. Concatenating a single-step forecast $\hat{X}^t = f(X^{t-2:t-1}, F^t)$ to the initial state and repeating the process with t incremented by one allows autoregressive forecasts $\hat{X}^{1:T}$ of arbitrary length T. An integral part of this framework is mapping from N surface grid points onto a mesh graph $\mathcal{G}_M = (\mathcal{V}_M, \mathcal{E}_M)$ coarser than the simulation domain[50]. The function f is implemented as a sequence of GNN layers following an encode-process-decode architecture[26] where: (1) grid inputs are encoded onto the mesh representation; (2) a number of GNN layers process this latent representation; (3) the processed data is mapped back onto the original grid. SeaCast predicts the next step as a residual update to the most recent input state, making it an easier learning task than predicting the next state directly. The mappings between grid and mesh nodes occur through bipartite grid-to-mesh $\mathcal{G}_{G2M}$ and mesh-to-grid $\mathcal{G}_{M2G}$ graphs. Upwards node and edge updates are facilitated through GNN layers in the form of *propagation networks*[16], and the remaining updates through *interaction networks*[51], with all layers mapping to a latent dimensionality d_z. This design leverages the inherent inductive bias of interaction networks to retain information, whereas propagation networks are more effective at forwarding new information through the graph. Multilayer perceptrons within GNN layers consists of a single hidden layer using the Swish activation function[52], followed by layer normalization[53].

Hierarchical graph

The hierarchical mesh structure described in[16] consists of multiple graph levels $\mathcal{G}_1, \ldots, \mathcal{G}_H$, where each level $\mathcal{G}_h = (\mathcal{V}_h, \mathcal{E}_h)$ gets progressively coarser. The first level connects directly to the grid, forming $\mathcal{G}_{G2M} = (\mathcal{V}_G \cup \mathcal{V}_1, \mathcal{E}_{G2M})$ and $\mathcal{G}_{M2G} = (\mathcal{V}_G \cup \mathcal{V}_1, \mathcal{E}_{M2G})$. A processing step on the mesh is defined as a complete sweep through the hierarchy $\mathcal{G}_{h,h+1} = (\mathcal{V}_h \cup \mathcal{V}_{h+1}, \mathcal{E}_{h,h+1})$ where the graph sequences are $\mathcal{G}_{1,2}, ..., \mathcal{G}_{H-1,H}$ and $\mathcal{G}_{H,H-1}, ..., \mathcal{G}_{2,1}$, respectively.

Mesh construction

Regional oceans can take very irregular shapes, calling for a customized mesh for the modeled geographical area. The foundation for our mesh is a quadrilateral construction used for graph-based limited area weather modeling[16]. It is initialized by selecting only the nodes corresponding to the ocean surface grid. All nodes are connected with bidirectional edges to their neighbors horizontally, vertically and diagonally. Nodes on higher resolution levels are positioned in the center of a 3×3 square on the level below. Upward edges are created by connecting each node at level h to the closest nodes at level $h + 1$, and the downward edges mirror these. Edges crossing land areas with a threshold of 8 grid points are excluded, both for inter- and intra-level graphs. This procedure results in a mesh that conforms to the shape of the regional ocean.

Rollout masking

We want to ensure (1) that the learning task of the model is exclusively for grid nodes inside the regional ocean at each depth level, and (2) that the predictions are aligned with the influence of the global ocean. To address the first point we only propagate predictions part of the internal depth-wise ocean grid $\mathbb{G}$ in the autoregressive rollout. In response to the second point, boundary forcing is applied at each time step by replacing predictions inside the boundary region $\mathbb{B}$ with the ground truth forecast X_t. We update the row for each node v as:

$$\hat{X}^t_{v,i} \leftarrow \left(\mathbb{I}_{\{v \in \mathbb{G}_{l,i}\}} - \mathbb{I}_{\{v \in \mathbb{B}_{l,i}\}}\right) \hat{X}^t_{v,i} + \mathbb{I}_{\{v \in \mathbb{B}_{l,i}\}} X^t_{v,i} \; \forall l \in \mathbb{L}_i \tag{1}$$

Here, $\mathbb{I}_{\{\cdot\}}$ denotes the indicator function, equal to 1 if the specified condition is true and 0 otherwise. $\mathbb{G}_{l,i}$ represents the set of oceanic grid nodes at depth level l associated with variable i, with $\mathbb{B}_{l,i}$ defined analogously for the boundary region. Finally, $\mathbb{L}_i$ denotes the set of vertical levels associated with variable i.

Training objective

The model is trained to minimize the mean squared error (MSE) between the predictions and the ground truth over a rolled-out sequence of states. The loss function we use is similar to what is commonly applied in MLWP, with the distinction that we account for the bathymetry, or the ocean grid structure at each depth level. The complete loss function is defined as:

$$\mathcal{L} = \frac{1}{T_{\text{rollout}}} \sum_{t=1}^{T_{\text{rollout}}} \sum_{i=1}^{C} \lambda_i \sum_{l=1}^{L_i} \frac{w_l}{N_l} \sum_{v=1}^{N_l} a_v \left(\hat{X}^t_{v,i} - X^t_{v,i}\right)^2 \tag{2}$$

where T_{rollout} is the number of steps in the rollout, C is the number of ocean variables in the tensor, L_i is the number of depth levels for variable i, N_l is the number of ocean grid nodes at depth level l, a_v is the latitude-

longitude area of grid cell v normalized to unit mean, w_l is the loss weight for depth level l, and λ_i is the inverse variance of time differences for variable i.

The area factor a_v accounts for variations in the size of grid cells due to the spherical geometry of the Earth, assigning lower weights to cells near the poles and higher weights near the equator, and is normalized to have unit mean across the domain. The depth-specific weight w_l is designed to prioritize prediction accuracy in the upper ocean, reflecting the greater societal relevance of shallower depths. For depth-resolved variables, w_l is proportional to $(200 - d)$, where d is the depth in meters, and the weights are normalized to have unit mean across all depths. The single-level SSH variable is assigned a weight of one. Further normalizing by the magnitude of the dynamics for each feature, signified by the λ_i factor, ensures that the model evaluates errors across vertical levels on a physically meaningful scale[14].

Model and training

We train SeaCast with 3 mesh levels and 3 processor layers with $d_z = 256$ hidden units, totaling 17.7 M trainable parameters. Training initiates with a warm-up phase of five epochs, starting from a learning rate of 10^{-5} and incrementing epoch-wise to a base rate of 10^{-3}. Following the warm-up, we employ a cosine decay schedule. For optimization, we use AdamW[54], configured with $\beta_1 = 0.9$, $\beta_2 = 0.95$, and a weight decay of 0.1. SeaCast is trained on reanalysis data for 200 epochs using a local batch size of 1. The model is further finetuned on analysis data for 30 epochs, with the same learning rate schedule, but the initial learning rate set to 10^{-7} and the base rate as 10^{-5}. Furthermore, the number of rollout steps is progressively increased to 2 at epoch 10 and 3 at epoch 20. Additional rollout steps can be added to the loss function, although this comes at an increased training cost, and have been observed to yield diminishing returns in MLWP[14]. The pretraining on 35 years of daily reanalysis data took 20.5 h on 64 AMD MI250x GPUs (1312 GPU hours) in a data-parallel configuration, and finetuning on 2 years of analysis data took 3.5 h on 8 GPUs (28 GPU hours).

Computational complexity

SeaCast can produce a complete 15-day forecast in 20 seconds on a single AMD MI250x GPU, which is roughly equivalent to 1.3 s per simulated day (~ 178 SYPD). The SeaCast forecast includes 18 depth levels, and outputs predictions at a daily temporal resolution. In contrast, the computational time required for the MedFS is approximately 70 minutes to run a 10-day forecast (~ 0.6 SYPD), using 89 CPU cores on Lenovo SD650 V3 computing nodes. This includes the time for generating and writing outputs for all 141 vertical levels. Both models produce outputs at the same $1/24^\circ$ spatial resolution. In terms of energy use, a single 15-day SeaCast forecast consumes roughly 2.8 Wh, whereas a 10-day MedFS forecast requires around 650 Wh. These numbers are based on the nominal power ratings of the respective GPU and CPU hardware. Model training incurs a one-time cost of approximately 656 kWh for pretraining and 14 kWh for fine-tuning. Further replacing the computationally expensive wave component not considered here with a learned surrogate could reduce energy consumption at forecast time even more.

Dataset

Ocean state

The Mediterranean Sea physics analysis and forecasting system leverages a two-way coupled hydrodynamic–wave modeling framework composed of the Nucleus for European Modelling of the Ocean (NEMO) v4.2[55], which includes an explicit representation of tides, and WAVEWATCH III v6.07[56]. Model solutions are corrected using the 3D variational data assimilation scheme OceanVar[57], which integrates quality-controlled in situ and satellite observations to improve ocean dynamics. The Mediterranean reanalysis system is based on NEMO v3.6 (without tides) and OceanVar, assimilating reprocessed observational data. The model is implemented at a high horizontal grid resolution of $1/24^\circ$, utilizing a total of 141 vertical levels with uneven spacing. By choosing every other depth until 200 m the list of selected depths included in this study becomes: 1.02, 5.46, 10.5, 16.3, 22.7, 29.9, 37.9, 46.7, 56.3, 66.9, 78.6, 91.2, 105, 120, 136, 153, 172, and 192 meters. The topography is based on an interpolation of the General Bathymetric Chart of the Oceans (GEBCO) 30 arcsecond grid[58]. SeaCast is trained on 37 years of daily reanalysis/analysis data, totaling 13 508 samples, and each epoch the model is validated on analysis data from January to June of 2024, comprising 177 samples. The test data consist of daily forecasts issued between July 3rd and December 31st, 2024 comprising 182 daily initializations, and forecast skill is assessed up to January 14th, 2025 to account for the 15-day SeaCast forecast horizon.

Atmospheric forcing

Atmospheric forcing plays an important role in data-driven modeling of the ocean's response to atmospheric conditions[25], especially in driving marine dynamics near the sea surface. We incorporate four key atmospheric variables: the 2-m temperature, MSLP, and the 10-m zonal and meridional wind stress components derived from wind components using the Hellerman and Rosenstein formulation[59] for the drag coefficient. The atmospheric data are sourced from daily mean aggregates of 6-hourly ERA5 reanalysis data[29]. For testing, we compare the 6-hourly daily means of the numerical ENS and the data-driven AIFS forecasts. All atmospheric forcing variables are bi-linearly interpolated from their native $1/4^\circ$ resolution to the $1/24^\circ$ resolution of the sea grid. Additionally, the sine and cosine of the day of the year, normalized between 0 and 1, are included as forcing features to account for seasonal variations. The atmospheric forcing used in our model is windowed over three consecutive time steps. This means that each forcing input F_t includes data from times $t - 1$, t, and $t + 1$.

Lateral boundary forcing

We define the lateral boundary as the grid nodes west of longitude -5.2° to the edge of the grid, covering the Strait of Gibraltar, as well as the Dardanelles Strait, which spans latitudes 39.9° to 40.4° and longitudes

25.9° to 26.4°. Note that we use a boundary region that lies inside the propagated data grid, allowing us to use Mediterrenaen forecast data as boundary forcing. Current ocean forecasts at CMEMS are available 10 days in the future for the most part, following the length of HRES atmospheric forcing. However, we use the ENS/AIFS standard of 15-day forecasts. Hence we have to increase the length of the boundary forcing, and we do so by persisting the last forecast state in the boundary region five times at the end.

Satellite data

Satellite observations of SST and SLA over the Mediterranean Sea are used for evaluation. For SST, we use the super-collated Level-3S (L3S) multi-sensor merged product[33] delivered through CMEMS, which provides daily SST fields at a spatial resolution of $1/16^\circ$. These observations are representative of nighttime conditions, thereby minimizing the effects of the diurnal heating cycle. The L3S dataset is constructed by merging only the highest quality Level-2P (L2P) input data from several satellite sensors, selected within a strict local nighttime window to reduce cloud contamination and sensor inconsistencies. The main contributing sensors include SLSTR from Sentinel-3A and -3B, VIIRS from NOAA-20 and Suomi-NPP, AVHRR from Metop-B and -C, and SEVIRI from Meteosat. Model forecasts are bilinearly interpolated to the L3S SST grid during evaluation.

For SLA, we use the CMEMS along-track L3 near-real-time product. This product contains high-resolution measurements, where we use 5 Hz measurements from satellite altimeters HaiYang-2B, Jason-3 (interleaved orbit), Sentinel-3A and -3B, Sentinel-6A, and the SWOT mission (nadir track). The filtered version of this dataset is used to minimize the influence of measurement noise. Model SSH is converted to SLA and bilinearly interpolated onto the satellite track coordinates during evaluation. The process of aligning model and observation data is detailed in Supplementary A.3.

Evaluation metrics

We assess the performance of SeaCast using several variations of RMSE, which quantifies the average magnitude of prediction errors relative to the reference observations. RMSE is computed per lead time and per variable, and includes area weighting to account for the changing size of ocean grid cells with latitude, consistent with the loss function used during training. For the 3D variables temperature, salinity, and currents, we report depth-averaged RMSE to provide a vertically integrated summary of forecast skill at each lead time. For SST, we additionally compute per-location RMSE by comparing model outputs interpolated onto the L3S SST grid to the observations.

To evaluate the model's ability to detect temperature extremes, we use HSS, which measures categorical forecast skill relative to random chance. HSS is computed based on a binary classification of observed threshold exceedances, defined as SST values above the 90th percentile of the daily climatology, using an 11-day rolling mean. An HSS of 1 indicates a perfect forecast, 0 indicates no skill beyond chance, and negative values indicate performance worse than a random forecast.

Additional details on the verification metrics and their mathematical definitions are provided in Supplementary C.

Data availability

The model weights and preprocessed data to reproduce the results for this study are stored at https://doi.org/10.57967/hf/5342. The original data is publicly available via CMEMS for the ocean data and ECMWF for the atmospheric data. The Mediterranean reanalysis and analysis data can be downloaded from https://doi.org/10.25423/cmcc/medsea_multiyear_phy_006_004_e3r1 and https://doi.org/10.25423/cmcc/medsea_analysisforecast_phy_006_013_eas8. The atmospheric forcing is obtained from https://doi.org/10.21957/open-data. The observed SST, SLA, and in-situ data used for evaluation are available via https://doi.org/10.48670/moi-00171, https://doi.org/10.48670/moi-00140, and https://doi.org/10.48670/moi-00044, respectively.

Code availability

The source code for SeaCast is available at https://doi.org/10.5281/zenodo.15367632. The repository contains scripts for data downloading and preprocessing, as well as code for training and evaluating the models presented.

Received: 5 July 2025; Accepted: 1 December 2025
Published online: 06 December 2025

References

1. Le Traon, P. Y. et al. From observation to information and users: The Copernicus Marine Service perspective. *Front. Mar. Sci.* **6**, 234 (2019).
2. Sun, R. et al. SKRIPS v1.0: A regional coupled ocean-atmosphere modeling framework (MITgcm-WRF) using ESMF/NUOPC, description and preliminary results for the Red Sea. *Geosci. Model Dev.* **12**, 4221–4244 (2019).
3. Sakamoto, K. et al. Development of a 2-km resolution ocean model covering the coastal seas around Japan for operational application. *Ocean Dyn.* **69**, 1181–1202 (2019).
4. Ciliberti, S. A. et al. Monitoring and forecasting the ocean state and biogeochemical processes in the Black Sea: Recent developments in the Copernicus Marine Service. *J. Mar. Sci. Eng.* **9**, 1146 (2021).
5. Kärnä, T. et al. Operational marine forecast model for the Baltic Sea. Nemo-Nordic 2.0. *Geosci. Model Dev.* **14**, 5731–5749 (2021).
6. Zhu, X. et al. Improvements in the regional South China Sea operational oceanography forecasting system (SCSOFSv2). *Geosci. Model Dev.* **15**, 995–1015 (2022).
7. Bruschi, A. et al. Indexes for the assessment of bacterial pollution in bathing waters from point sources: The northern Adriatic Sea CADEAU service. *J. Environ. Manag.* **293**, 112878 (2021).
8. Liubartseva, S. et al. Modeling chronic oil pollution from ships. *Mar. Pollut. Bull.* (2023).

9. Mannarini, G., Salinas, M. L., Carelli, L., Petacco, N. & Orović, J. VISIR-2: Ship weather routing in Python. *Geosci. Model Dev.* **17**, 4355–4382 (2024).
10. Coppini, G. et al. The Mediterranean forecasting system-Part 1: Evolution and performance. *Ocean Sci.* **19**, 1483–1516 (2023).
11. Bi, K. et al. Accurate medium-range global weather forecasting with 3D neural networks. *Nature* **619**, 533–538 (2023).
12. Nguyen, T. et al. Scaling transformer neural networks for skillful and reliable medium-range weather forecasting. *Adv. Neural Inf. Process. Syst.* **37**, 68740–68771 (2024).
13. Pathak, J. et al. FourCastNet: A global data-driven high-resolution weather model using adaptive Fourier neural operators. arXiv preprint arXiv:2202.11214 (2022).
14. Keisler, R. Forecasting global weather with graph neural networks. arXiv preprint arXiv:2202.07575 (2022).
15. Lam, R. et al. Learning skillful medium-range global weather forecasting. *Science* **382**, 1416–1421 (2023).
16. Oskarsson, J., Landelius, T., Deisenroth, M. & Lindsten, F. Probabilistic weather forecasting with hierarchical graph neural networks. *Adv. Neural Inf. Process. Syst.* **37**, 41577–41648 (2024).
17. Dheeshjith, S. et al. Samudra: An AI global ocean emulator for climate. *Geophys. Res. Lett.* **52**(10), e2024GL114318 (2025).
18. Wang, C. et al. Coupled ocean-atmosphere dynamics in a machine learning Earth system model. arXiv preprint arXiv:2406.08632 (2024).
19. Guo, Z. et al. Data-driven global ocean modeling for seasonal to decadal prediction. arXiv preprint arXiv:2405.15412 (2024).
20. Wang, X. et al. XiHe: A data-driven model for global ocean eddy-resolving forecasting. arXiv preprint arXiv:2402.02995 (2024).
21. Aouni, A. E. et al. GLONET: Mercator's end-to-end neural forecasting system. arXiv preprint arXiv:2412.05454 (2024).
22. Cui, Y. et al. Forecasting the eddying ocean with a deep neural network. *Nat. Commun.* **16**, 2268 (2025).
23. Andersson, T. R. et al. Seasonal Arctic sea ice forecasting with probabilistic deep learning. *Nat. Commun.* **12**, 5124 (2021).
24. Chattopadhyay, A., Gray, M., Wu, T., Lowe, A. B. & He, R. OceanNet: A principled neural operator-based digital twin for regional oceans. *Sci. Rep.* **14**, 21181 (2024).
25. Subel, A. & Zanna, L. Building ocean climate emulators. arXiv preprint arXiv:2402.04342 (2024).
26. Sanchez-Gonzalez, A. et al. Learning to simulate complex physics with graph networks. In *International Conference on Machine Learning*. 8459–8468 (2020).
27. Marchesiello, P., McWilliams, J. C. & Shchepetkin, A. Open boundary conditions for long-term integration of regional oceanic models. *Ocean Model.* **3**, 1–20 (2001).
28. Escudier, R. et al. A high resolution reanalysis for the Mediterranean Sea. *Front. Earth Sci.* **9**, 702285 (2021).
29. Hersbach, H. et al. ERA5 monthly averaged data on single levels from 1940 to present. In *Copernicus Climate Change Service (C3S) Climate Data Store (CDS)* (2023).
30. Molteni, F., Buizza, R., Palmer, T. N. & Petroliagis, T. The ECMWF ensemble prediction system: Methodology and validation. *Q. J. R. Meteorol. Soc.* **122**, 73–119 (1996).
31. Lang, S. *et al.* AIFS - ECMWF's data-driven forecasting system. arXiv preprint arXiv:2406.01465 (2024).
32. ECMWF. Plans for high resolution forecast (HRES) and ensemble forecast (ENS). *In focus* (2024). https://www.ecmwf.int/en/about/media-centre/focus/2024/plans-high-resolution-forecast-hres-and-ensemble-forecast-ens.
33. Nardelli, B. B., Tronconi, C., Pisano, A. & Santoleri, R. High and ultra-high resolution processing of satellite sea surface temperature data over southern European seas in the framework of MyOcean project. *Remote Sens. Environ.* **129**, 1–16 (2013).
34. Hobday, A. J. et al. A hierarchical approach to defining marine heatwaves. *Prog. Oceanogr.* **141**, 227–238 (2016).
35. Kochkov, D. et al. Neural general circulation models for weather and climate. *Nature* **632**, 1060–1066 (2024).
36. Marshall, J. & Schott, F. Open-ocean convection: Observations, theory, and models. *Rev. Geophys.* **37**, 1–64 (1999).
37. Large, W. G., McWilliams, J. C. & Doney, S. C. Oceanic vertical mixing: A review and a model with a nonlocal boundary layer parameterization. *Rev. Geophys.* **32**, 363–403 (1994).
38. Wunsch, C. & Stammer, D. Atmospheric loading and the oceanic inverted barometer effect. *Rev. Geophys.* **35**, 79–107 (1997).
39. Gill, A. E. *Atmosphere–Ocean Dynamics* (Academic Press, 1982).
40. Clementi, E. et al. Coupling hydrodynamic and wave models: first step and sensitivity experiments in the Mediterranean Sea. *Ocean Dyn.* **67**, 1293–1312 (2017).
41. McDonagh, B., Clementi, E., Goglio, A. C. & Pinardi, N. The characteristics of tides and their effects on the general circulation of the Mediterranean Sea. *Ocean Sci.* **20**, 1051–1066 (2024).
42. Rühling Cachay, S., Zhao, B., Joren, H. & Yu, R. Dyffusion: A dynamics-informed diffusion model for spatiotemporal forecasting. *Adv. Neural Inf. Process. Syst.* **36**, 45259–45287 (2023).
43. Andrae, M., Landelius, T., Oskarsson, J. & Lindsten, F. Continuous ensemble weather forecasting with diffusion models. In *International Conference on Learning Representations* (2025).
44. Nipen, T. N. et al. Regional data-driven weather modeling with a global stretched-grid. arXiv preprint arXiv:2409.02891 (2024).
45. Adamov, S. et al. Building machine learning limited area models: Kilometer-scale weather forecasting in realistic settings. arXiv preprint arXiv:2504.09340 (2025).
46. Nguyen, T., Brandstetter, J., Kapoor, A., Gupta, J. K. & Grover, A. ClimaX: A foundation model for weather and climate. In *International Conference on Machine Learning* (2023).
47. Bodnar, C. et al. A foundation model for the Earth system. *Nature* 1–8 (2025).
48. Price, I. et al. Probabilistic weather forecasting with machine learning. *Nature* **637**, 84–90 (2025).
49. Larsson, E., Oskarsson, J., Landelius, T. & Lindsten, F. Diffusion-LAM: Probabilistic limited area weather forecasting with diffusion. arXiv preprint arXiv:2502.07532 (2025).
50. Fortunato, M., Pfaff, T., Wirnsberger, P., Pritzel, A. & Battaglia, P. Multiscale meshgraphnets. arXiv preprint arXiv:2210.00612 (2022).
51. Battaglia, P., Pascanu, R., Lai, M., Jimenez Rezende, D. et al. Interaction networks for learning about objects, relations and physics. *Adv. Neural Inf. Process. Syst.* **29** (2016).
52. Ramachandran, P., Zoph, B. & Le, Q. V. Searching for activation functions. arXiv preprint arXiv:1710.05941 (2017).
53. Ba, J., Kiros, J. & Hinton, G. Layer normalization. arXiv preprint arXiv:1607.06450 (2016).
54. Loshchilov, I. & Hutter, F. Decoupled weight decay regularization. In *International Conference on Learning Representations* (2019).
55. Madec, G. et al. NEMO ocean engine. In *Scientific Notes of Climate Modelling Center*. Vol. 27 (2017).
56. WAVEWATCH III Development Group (WW3DG). User Manual and System Documentation of WAVEWATCH III Version 6.07. In Technical Note 333, NOAA/NWS/NCEP/MMAB, College Park, MD, USA (2019).
57. Dobricic, S. & Pinardi, N. An oceanographic three-dimensional variational data assimilation scheme. *Ocean Model.* **22**, 89–105 (2008).
58. Weatherall, P. et al. A new digital bathymetric model of the world's oceans. *Earth Sp. Sci.* **2**, 331–345 (2015).
59. Hellerman, S. & Rosenstein, M. Normal monthly wind stress over the world ocean with error estimates. *J. Phys. Oceanogr.* **13**, 1093–1104 (1983).

Acknowledgements

Open access funded by Helsinki University Library. This work was financially supported by the Research Council of Finland under the FAISER project (grant no. 361902). Computing resources were provided by the LUMI supercomputer, owned by the EuroHPC Joint Undertaking and hosted by CSC–IT Center for Science.

Author contributions

D.H. conceptualized the project and conducted the experiments. E.C. and I.E. assisted with the assessment of regional ocean forecasting systems and provided guidance on interpreting the results. T.R. supervised the project. All authors provided feedback on results at various stages of the project and contributed to the writing of the manuscript. All authors have reviewed the manuscript and consent to its publication.

Funding

This project was supported by the Research Council of Finland.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-31177-w.

Correspondence and requests for materials should be addressed to D.H.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2025