NeuralOM: Neural Ocean Model for Subseasonal-to-Seasonal Simulation

Yuan Gao[1*], Hao Wu[1*†], Fan Xu[2], Yanfei Xiang[1], Ruijian Gou[3], Ruiqi Shu[1], Qingsong Wen[4], Xian Wu[5‡], Kun Wang[6‡], Xiaomeng Huang[1‡]

[1]Tsinghua University,
[2]Shenzhen Loop Area Institute,
[3]Ocean University of China,
[4]Squirrel Ai Learning,
[5]Tencent,
[6]Nanyang Technological University
{yuangao24, wu-h25, xiangyf22, srq24}@mails.tsinghua.edu.cn; fanxu@slai.edu.cn; grj@stu.ouc.edu.cn; qingsongwen@squirrelai.com; kevinxwu@tencent.com; wang.kun@ntu.edu.sg; hxm@tsinghua.edu.cn

Abstract

Long-term, high-fidelity simulation of slow-changing physical systems, such as the ocean and climate, presents a fundamental challenge in scientific computing. Traditional autoregressive machine learning models often fail in these tasks as minor errors accumulate and lead to rapid forecast degradation. To address this problem, we propose **NeuralOM**, a general neural operator framework designed for simulating complex, slow-changing dynamics. NeuralOM's core consists of two key innovations: (1) a ***Progressive Residual Correction Framework*** that decomposes the forecasting task into a series of fine-grained refinement steps, effectively suppressing long-term error accumulation; and (2) a ***Physics-Guided Graph Network*** whose built-in **adaptive messaging mechanism** explicitly models multi-scale physical interactions, such as gradient-driven flows and multiplicative couplings, thereby enhancing physical consistency while maintaining computational efficiency. We validate NeuralOM on the challenging task of global Subseasonal-to-Seasonal (S2S) ocean simulation. Extensive experiments demonstrate that NeuralOM not only surpasses state-of-the-art models in forecast accuracy and long-term stability, but also excels in simulating extreme events. For instance, at a 60-day lead time, NeuralOM achieves a 13.3% lower RMSE compared to the best-performing baseline, offering a stable, efficient, and physically-aware paradigm for data-driven scientific computing.

Code — https://github.com/YuanGao-YG/NeuralOM

Introduction

The long-term simulation of complex physical systems (Wang et al. 2024b; Wu et al. 2025c), such as the global ocean and climate, is a cornerstone of modern science yet remains a great challenge. While machine learning (ML) models offer a computationally efficient path to complement traditional numerical solvers, they face a fundamental bottleneck in such tasks: ***long-term stability***. This issue is especially severe in ***slow-changing systems***, where the state evolves in a subtle manner. In standard autoregressive frameworks, minor prediction errors at each step inevitably accumulate, causing the simulation to diverge catastrophically from the true physical dynamics, a well-known problem of **compounding error**. This error cascade often leads data-driven models to fail in long-term rollouts. As shown in Figure 1 (bottom left), a state-of-the-art baseline model generates substantial, physically unrealistic artifacts in its long-range forecast.

To address this challenge, existing machine learning approaches have been explored, but they are limited in two key aspects. ❶. ***First***, they often struggle to capture the

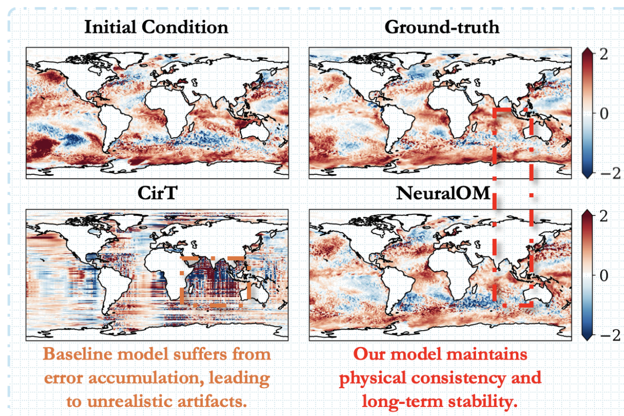

Figure 1: NeuralOM overcomes the long-term instability challenge in simulating slow-changing physical systems. Shown is a 60-day Subseasonal-to-Seasonal (S2S) simulation for sea surface temperature anomaly. (***Bottom Left***) A state-of-the-art baseline (CirT) collapses due to compounding errors, producing physically unrealistic artifacts and failing to capture the ocean state. (***Bottom Right***) In contrast, NeuralOM maintains high fidelity and long-term stability, accurately reproducing the complex patterns of the ground-truth, as highlighted in the detailed view (red dashed lines).

[*]These authors contributed equally.
[†]Project lead and technical guidance.
[‡]Corresponding author.

fine-grained dynamics inherent to slow-changing systems. Many models are designed to predict sharp, significant shifts, making them ill-suited for learning the small, incremental updates that define ocean dynamics. This typically results in either over-smoothing the predictions or failing to capture critical physical anomalies. ❷. ***Second***, the internal mechanisms of these models are often physics-agnostic. For instance, standard message passing in Graph Neural Networks (GNNs) (Keisler 2022) treats all node interactions uniformly, unable to distinguish between fundamentally different physical processes like gradient-driven flows and multiplicative couplings between variables. Furthermore, they lack the ability to adaptively model phenomena across different spatio-temporal scales, from large-scale, slow-moving currents to small-scale, energetic eddies.

To systematically address these issues, we introduce **NEURALOM**, a neural operator framework designed for stable, high-fidelity simulation of slow-changing dynamics. NEURALOM's core is built on two key innovations. First, we propose a **Progressive Residual Correction Framework**. This framework decomposes the complex forecasting task into a cascade of refinement steps, where each step focuses on learning and correcting the residual error from the previous one. This strategy of "small, iterative corrections" enables the model to capture extremely fine-grained dynamic changes, thereby effectively suppressing long-term error accumulation. Second, we design a novel **Physics-Guided Graph Network** with a built-in **adaptive messaging mechanism**. It incorporates physical priors into the network by explicitly modeling gradient-driven flows and multiplicative couplings between variables. Crucially, its aggregation scheme is not fixed but dynamically adjusts, allowing it to inherently handle multi-scale dynamics.

We validate NEURALOM on the challenging benchmark of global Subseasonal-to-Seasonal (S2S) ocean simulation and forecasting. Extensive experiments provide strong evidence of its superior performance:

Experimental Observations. ❶ ***State-of-the-art accuracy in long-term simulation*:** As shown in Table 1, in 60-day simulation, NEURALOM achieves a normalized RMSE of just 0.7014, a 13.3**%** reduction compared to the best-performing baseline, WenHai (0.8091). It also improves the Anomaly Correlation Coefficient (ACC) by 12.0**%**, indicating more accurate tracking of the physical field's evolution. ❷ ***Superior visual fidelity and physical consistency*:** As illustrated in Figure 1 and Figure 3, unlike baseline models like CirT that collapse due to error explosion, NEURALOM produces physically consistent forecasts that remain highly aligned with the ground truth even at 60 days, successfully preserving critical fine-scale structures. ❸ ***Validated effectiveness of our designs through ablation studies*:** As detailed in Table 2 and Table 3, removing any of our core components, either the Progressive Residual Correction Framework or the Physics-Guided Adaptive Graph Messaging, leads to a significant drop in performance. In particular, incorporating climatology priors improves the ACC metric by more than 5**-fold**, highlighting that every design choice is crucial to achieve accurate and stable simulations.

In summary, our main contributions are threefold: (1) a progressive residual correction framework designed for slow-changing systems; (2) a physics-guided graph network with an adaptive, multi-scale messaging mechanism; and (3) state-of-the-art performance on the S2S ocean simulation benchmark, establishing a new paradigm for data-driven scientific computing.

Related Work

In recent years, previous works in spatio-temporal mining have achieved stupendous achievements (Pfaff et al. 2020; Li et al. 2020; Wu et al. 2023, 2024d; Wang et al. 2024a; Wu et al. 2024b,a, 2025b,a,d; Jia et al. 2025). However, many of these methods are only applicable in idealized or relative simple scenario, whereas ocean system is a complex dynamical system in the real world, requiring the design of specialized methods for accurate prediction (Wu et al. 2024c). Research in data-driven ocean simulation is primarily divided into two streams: task-specific predictive models and general-purpose global foundation models.

Task-Specific ML Models. These models aim to simulate specific oceanic or atmospheric phenomena and have achieved impressive results in their respective domains. Examples include work on El Nio-Southern Oscillation (ENSO) prediction (Ham, Kim, and Luo 2019; Hu et al. 2021; Chang et al. 2023; Lyu et al. 2024; Chen et al. 2025), as well as models for the Madden-Julian Oscillation (MJO) (Kim et al. 2021; Delaunay and Christensen 2022; Yang et al. 2023; Shin et al. 2024a,b) and Marine Heatwaves (MHWs) (Jacox et al. 2022; Sun, Jing, and Liu 2024; Parasyris et al. 2025). The success of these models validates the immense potential of machine learning for complex geoscience problems. However, their limitation is also apparent: they are typically highly optimized for a single objective and often overlook the complex interactions between different physical variables and across multiple scales. This makes them difficult to generalize into a foundation model capable of capturing the full dynamics of the ocean system.

Global Ocean System Foundation Models. Inspired by the success of foundation models in atmospheric science (Pathak et al. 2022; Bi et al. 2023; Chen et al. 2023a,b; Lam et al. 2023; Oskarsson et al. 2024; Nguyen et al. 2024; Kochkov et al. 2024; Gao et al. 2025), a series of global ocean foundation models have recently emerged, such as AI-GOMS (Xiong et al. 2023), Xihe (Wang et al. 2024c), LangYa (Yang et al. 2024), and WenHai (Cui et al. 2025). These models aim to build a unified digital twin of the ocean and have already surpassed traditional numerical methods in computational efficiency and, in some cases, accuracy (Hao et al. 2025). However, they still face the two core challenges we identified in the introduction. First, they generally adopt standard autoregressive frameworks and have not fundamentally solved the problem of error accumulation in long-term simulations, leading to a decline in forecast fidelity over time. Second, their designs often lack sufficient physical constraints. For example, while models like WenHai (Cui et al. 2025) incorporate domain knowledge into a Swin Transformer (Liu et al. 2021), their feature interaction mechanisms remain relatively physics-agnostic, failing

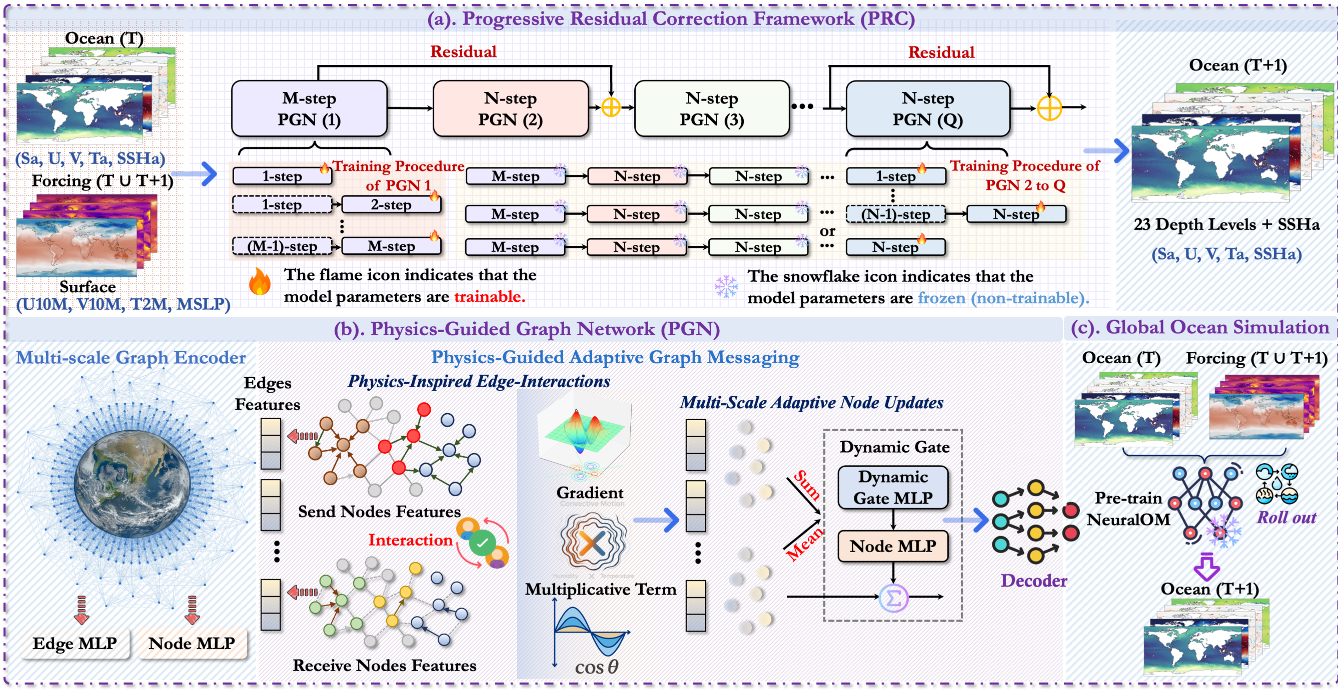

Figure 2: Overview of Our NEURALOM. **(a)** The overall architecture of the progressive residual correction framework, input variables (subtracting climatology for periodic variables), progressive residual correction stage, and visualization of simulation ocean variables; **(b)** The proposed physics-guided graph network; **(c)** The global ocean simulation module uses a rollout approach to generate future results.

to explicitly model physical processes like gradient changes or variable couplings, nor can they adaptively handle multi-scale dynamics.

NEURALOM is designed precisely to address these limitations. Unlike task-specific models, it is a general-purpose global model. Compared to existing global foundation models, NEURALOM introduces two fundamental breakthroughs: (1) Its **Progressive Residual Correction Framework** is a novel architecture specifically designed to suppress error accumulation in slow-changing systems and ensure long-term stability. (2) Its **Physics-Guided Graph Network**, with its adaptive messaging mechanism, explicitly integrates physical interactions (e.g., gradients and couplings) into the model, addressing the physics-agnostic problem of existing models.

Methodology

Preliminaries. We formulate the ocean system simulation as an autoregressive forecasting problem. Given the system state $\mathbf{O}_t$ at the current time step t (such as ocean temperature and salinity) and the external forcing $\mathbf{F}_t$, $\mathbf{F}_{t+1}$ (e.g., atmospheric fields), our goal is to predict the system state at the next time step, $\hat{\mathbf{O}}_{t+1}$. This process is represented by a neural network $\mathcal{M}$:

$$\hat{\mathbf{O}}_{t+1} = \mathcal{M}(\mathbf{O}_t, \mathbf{F}_t, \mathbf{F}_{t+1}; \Theta), \tag{1}$$

where Θ represents the learnable parameters of the model. During inference, the model performs long-range forecasting by recursively feeding its own prediction from the previous step, $\hat{\mathbf{O}}_t$, as input for the next step, starting from an initial state $\mathbf{O}_0$. Further details on the specific variables and data preprocessing used in our experiments are provided in the Appendix.

Progressive Correction for Stability

To overcome the challenges of error accumulation and the physics-agnostic nature of existing models, we propose NEURALOM, a stable and efficient simulation framework designed for slow-changing physical systems. Its core design comprises a ***Progressive Residual Correction Framework*** and a ***Physics-Guided Graph Network*** that serves as its core engine.

Progressive Residual Correction When simulating slow-changing systems, directly predicting the entire next state $\mathbf{O}_{t+1}$ is difficult, as the true change $(\mathbf{O}_{t+1} - \mathbf{O}_t)$ is remarkably small. Most models struggle to capture this subtle signal accurately, leading to rapid error accumulation.

To address this, we design a **Progressive Residual Correction** framework. It decomposes the complex single-step prediction task into a series of simpler, cascaded correction steps. First, a base model makes an initial, coarse prediction of the next state. Subsequently, a series of residual models are invoked, with each model focusing on learning and correcting the prediction error from the previous stage. This process can be abstractly formulated as:

$$\hat{\mathbf{O}}_{t+1}^{(1)} = \mathcal{M}_{\text{base}}(\mathbf{X}_t), \tag{2}$$

$$\hat{\mathbf{O}}_{t+1}^{(q)} = \hat{\mathbf{O}}_{t+1}^{(q-1)} + \mathcal{M}_{\text{residual}}^{(q)}(\hat{\mathbf{O}}_{t+1}^{(q-1)}), \quad \text{for } q = 2, \ldots, Q, \tag{3}$$

where $\mathbf{X}_t$ is the input of model, $\hat{\mathbf{O}}_{t+1}^{(1)}$ is the initial prediction, and $\hat{\mathbf{O}}_{t+1}^{(q)}$ is the refined prediction after $q-1$ correction stages. The final output is $\hat{\mathbf{O}}_{t+1} = \hat{\mathbf{O}}_{t+1}^{(Q)}$. This approach shifts the model's focus from predicting a large state vector to capturing fine-grained, physically meaningful anomalies, thereby significantly enhancing long-term stability and fidelity. Additionally, for variables with strong periodicity (e.g., sea salinity, sea temperature, and sea surface height), we first subtract their climatological mean before feeding them into the model to further help it focus on learning small anomaly changes.

Physics-Guided Adaptive Graph Messaging. Each model within this framework ($\mathcal{M}_{\text{base}}, \mathcal{M}_{\text{residual}}$) is realized by our novel **Physics-Guided Graph Network**. Unlike standard GNNs, our model deeply integrates physical priors into the graph's message passing and node update procedures through its core ***Physics-guided Adaptive Graph Messaging*** module.

Physics-Inspired Edge-Interactions. Traditional GNNs are often *physics-agnostic*, typically just concatenating adjacent node features and passing them through an MLP. To more realistically emulate the dynamic relationships between physical variables, we design physics-inspired edge interactions (PEI). The message between two nodes (i.e., the edge feature) is constructed not blindly, but through a set of explicit physical operators, including: (1) **feature differencing** to model **gradient-driven flows**; (2) **multiplicative coupling** to capture the joint effects of variables (like temperature and salinity on density); and (3) **cosine similarity** to identify similar water masses. These operators capture core relationships in ocean dynamics and provide a strong inductive bias. This process can be abstractly represented as:

$$\mathbf{e}_{uv} = f_{\text{PEI}}(h_u, h_v) = \phi([\underbrace{h_u - h_v}_{\text{Gradient}}, \underbrace{h_u \odot h_v}_{\text{Coupling}}, \underbrace{\text{sim}(h_u, h_v)}_{\text{Similarity}}]), \tag{4}$$

where h_u, h_v are node features and ϕ is a neural network that fuses these physical features.

Multi-scale Adaptive Node Aggregation. Physical systems often contain motions at different scales. To capture both large-scale, slow-changing processes and small-scale, energetic events simultaneously, we introduce a **dynamic gating mechanism** for aggregating neighbor information. This mechanism adaptively arbitrates between **sum** and **mean** aggregation based on local dynamics. For localized, high-energy events (like eddies), the gate favors sum aggregation to preserve their intensity. For large-scale, smooth flows, it favors mean aggregation to obtain a more robust, macroscopic representation.

$$h'_v = (1-\gamma)\cdot \text{Aggregate}_{\text{mean}}(\{\mathbf{e}_{uv}\}) + \gamma \cdot \text{Aggregate}_{\text{sum}}(\{\mathbf{e}_{uv}\}), \tag{5}$$

where the gating coefficient γ is dynamically learned by the network. This adaptive aggregation allows NEURALOM to efficiently handle complex multi-scale dynamics within a unified framework. For details on graph construction, encoder, and decoder, please refer to the Appendix.

Optimization and Inference

Objective. We train each component of NEURALOM by minimizing the relative L2-norm between the predicted and ground-truth states. The loss function $\mathcal{L}$ is defined as:

$$\mathcal{L}(\Theta) = \mathbb{E}_{t,k}\left[\frac{\left\|\hat{O}_{i,j,k}^{t+1} - O_{i,j,k}^{t+1}\right\|_2}{\left\|O_{i,j,k}^{t+1}\right\|_2}\right], \tag{6}$$

where $\hat{O}$ and O are the predicted and ground-truth values, respectively, indexed over time t, spatial locations (i, j), and variable channels k. $\|\cdot\|_2$ denotes the Euclidean norm over the spatial grid (i, j) for each channel k. $E_{t,k}$ denotes the empirical expectation taken over training time steps t and variable channels k, i.e., averaging the channel-wise relative errors over training samples.

Inference Strategy. For long-range simulation and forecasting, we employ a standard autoregressive (or "rollout") strategy. Starting from an initial state $\mathbf{X}_0$, the model continuously feeds its own previous prediction as input for the next step to generate a continuous trajectory spanning several weeks or months. For idealized *simulation* experiments, we use ground-truth atmospheric data as forcing. This approach eliminates interference from atmospheric forecast errors, allowing for a fairer assessment of the ocean models' intrinsic performance. For more realistic *forecasting* experiments, we drive NEURALOM with forecast fields generated by a state-of-the-art atmospheric forecasting model, OneForecast (Gao et al. 2025).

Algorithm 1: NeuralOM Forward Pass

Require: Initial state $\mathbf{O}_0$, forcing sequence $\{\mathbf{F}_t\}_{t=0}^{T}$, total steps $T+1$
Ensure: Predicted trajectory $\{\hat{\mathbf{O}}_t\}_{t=1}^{T}$

```
O^prev ← O_0
for t = 0, 1, ..., T − 1 do
    X_t ← Concat(O^prev, F_t, F_{t+1})    ▷ Combine current inputs
    # Stage 1: Progressive Residual Correction Framework
    Ô^(1)_{t+1} ← PGN_base(X_t)    ▷ Initial prediction by the base model
    Ô_intermediate ← Ô^(1)_{t+1}
    for q = 2, ..., Q do    ▷ Perform Q-1 rounds of residual correction
        Residual ← PGN^(q)_residual(Ô_intermediate)    ▷ Residual model predicts the correction
        Ô_intermediate ← Ô_intermediate + Residual    ▷ Apply the residual correction
    end for
    Ô_{t+1} ← Ô_intermediate    ▷ Get the final single-step prediction
    O^prev ← Ô_{t+1}    ▷ Update state for the next autoregressive step
end for
return {Ô_t}_{t=1}^T
```

Model	Simulate Horizon							
	40-day		45-day		50-day		60-day	
	RMSE	ACC	RMSE	ACC	RMSE	ACC	RMSE	ACC
FourCastNet (Pathak et al. 2022) arXiv 2022	5.6295	0.0782	> 10	0.0190	> 10	0.0046	> 10	0.0007
CirT (Liu et al. 2025) ICLR 2025	2.7510	0.0143	2.8686	0.0115	2.9514	0.0099	3.0631	0.0081
WenHai (Cui et al. 2025) Nature Communications 2025	<u>0.7199</u>	<u>0.5065</u>	<u>0.7467</u>	<u>0.4703</u>	<u>0.7700</u>	<u>0.4385</u>	<u>0.8091</u>	<u>0.3851</u>
NeuralOM (ours)	**0.6509**	**0.5297**	**0.6665**	**0.5009**	**0.6795**	**0.4754**	**0.7014**	**0.4314**
NeuralOM Improvement	9.59%	4.58%	10.74%	6.49%	11.74%	8.42%	13.32%	12.01%

Table 1: Performance comparison of NeuralOM against baselines on global ocean simulation task. Average RMSE (normalized, ↓) and ACC (↑) across all 93 ocean variables are shown. Best results in **bold**, second best <u>underlined</u>.

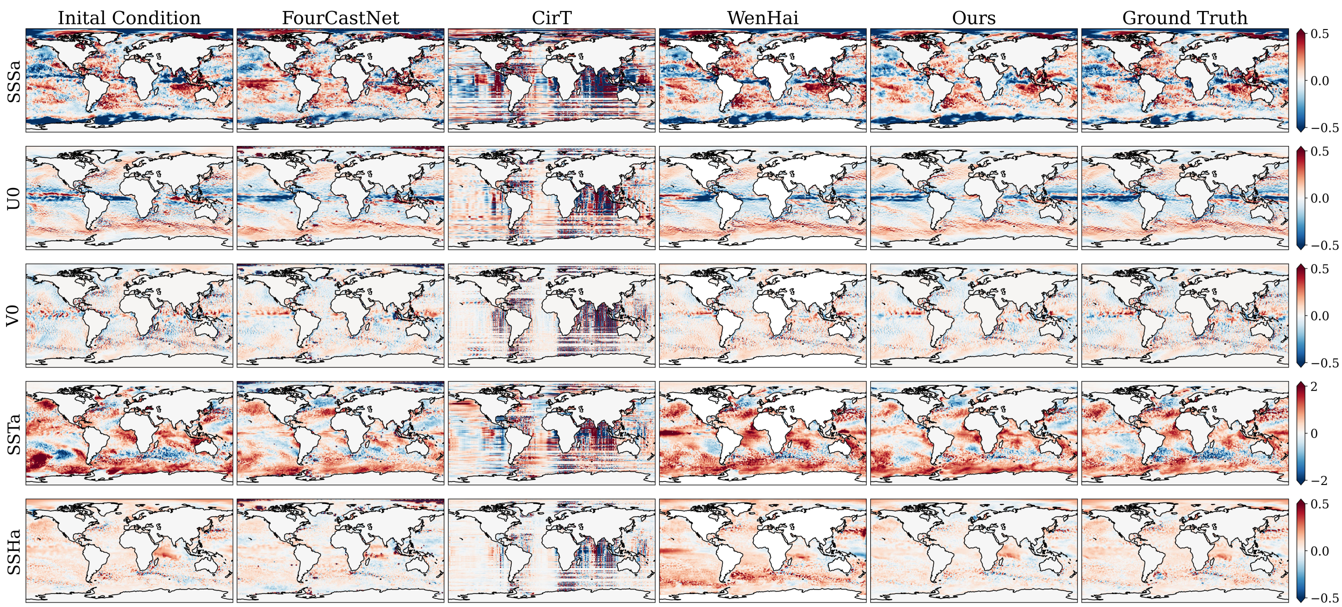

Figure 3: 60-day simulation results of different models.

Summary. In conclusion, the proposed NeuralOM framework forms an end-to-end solution for slow-changing physical systems, bridging macroscopic correction strategies with microscopic physical modeling. To transparently illustrate its computational flow, we outline the complete forward pass of NeuralOM in Algorithm 1. In the following section, we will show its practical effectiveness through a series of rigorous benchmarks and ablation studies.

Experiments

To comprehensively evaluate the performance of NeuralOM and validate its effectiveness in addressing the challenges of long-term stability and physical consistency, we design a series of rigorous experiments. We first benchmark NeuralOM against several state-of-the-art models on an idealized global ocean **simulation** task. We then apply it to a more challenging real-world **forecasting** scenario. Finally, we dissect the contributions of our model's innovative components through extensive **ablation studies**. All experiments were conducted on 64 NVIDIA A100 GPUs.

Experimental Setup

Dataset and Task. Our experiments are based on the **GLORYS12** global ocean reanalysis dataset, which provides daily mean ocean states since 1993. We split the data into a training set (1993-2017), a validation set (2018-2019), and a test set (2020). The core task is autoregressive forecasting at the **Subseasonal-to-Seasonal (S2S)** scale, with rollout horizons ranging from several days up to 60 days. The input consists of 93 ocean state variables (including salinity, velocity, temperature at various depths, and sea surface height) and 8 atmospheric variables as external forcing. All spatial data is downsampled to a 24-hour temporal resolution (daily mean for ocean and 12:00 UTC for forcings) and 0.5° spatial resolution. Further details on data processing and variables are available in the Appendix.

Baselines. We compare NeuralOM against three representative models: (1) **FourCastNet** (Pathak et al. 2022), a widely recognized atmospheric science foundation model based on Fourier Neural Operators. (2) **CirT** (Liu et al.

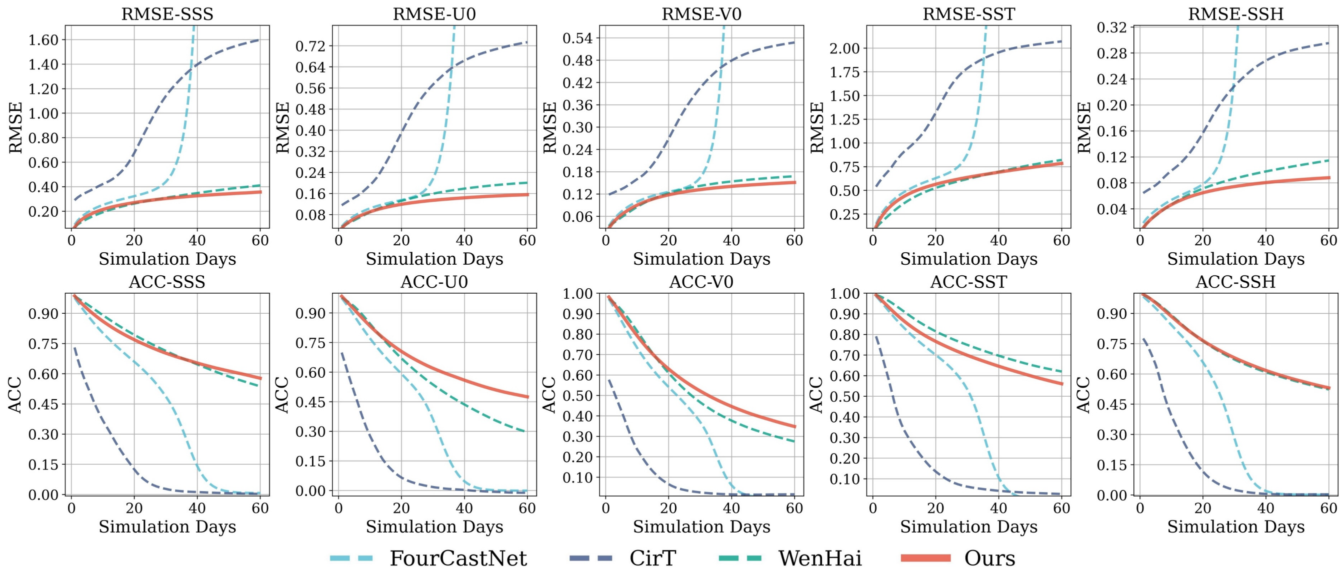

Figure 4: The latitude-weighted RMSE and ACC results of several important ocean surface variables.

2025), a state-of-the-art atmospheric forecasting model designed for the S2S scale. (3) **WenHai** (Cui et al. 2025), a state-of-the-art ML-based ocean model that has demonstrated superior performance over traditional numerical models and ML-based ocean foundation models. To ensure a fair comparison, all models, except for the officially released WenHai results, are retrained under our unified framework.

Evaluation Metrics. We use four metrics to assess performance: (1) **Root Mean Square Error (RMSE)** to measure the average prediction error, and (2) **Anomaly Correlation Coefficient (ACC)** to evaluate the model's ability to capture deviations from the climatological mean (i.e., anomalies). Additionally, we use (3) **Critical Success Index (CSI)**, and (4) **Symmetric Extremal Dependence Index (SEDI)** to evaluate the model's performance on extreme events.

Main Results and Analysis

NeuralOM sets a state-of-the-art in long-term simulation. We first evaluate NeuralOM in an idealized simulation setting using ground truth atmospheric forcings. As shown in **Table 1**, NeuralOM consistently outperforms all baselines. At a 60-day lead time, it reduces RMSE by 13.3**%** and improves ACC by 12.0**%** over the strongest baseline, WenHai. This superior performance is visually confirmed in **Figure 3**, where NeuralOM is the only model to maintain physical consistency and avoid the error cascades that cause other models to collapse. The significantly lower error growth rate, detailed in **Figure 4**, further underscores its long-term stability. The poor performance of powerful atmospheric models like CirT (60-day ACC of 0.0081) also highlights the unique challenges of ocean simulation and validates our specialized design.

Robust performance is maintained in ocean forecasting. To assess practical applicability, we drive the ocean models with atmospheric *forecasts* from OneForecast, which introduces real-world uncertainty. As common atmospheric forecast models don't provide all forcings required by WenHai at the S2S scale, we can't assess its results. Even in this more challenging scenario, **Table 4** shows that NeuralOM retains its lead, achieving significantly better RMSE (0.68) and ACC (0.48) at a 35-day lead time. This demonstrates the robustness of NeuralOM for real-world applications, further supported by the high-fidelity visuals in **Figure 5(a)**.

NeuralOM excels at capturing high-impact extreme events. Beyond average metrics, a model's skill in predicting extreme events is critical. We use CSI and SEDI to evaluate predictions of extreme surface currents. As shown in **Figure 5(b)**, NeuralOM achieves higher scores than all baselines at both 30 and 60-day lead times. This enhanced capability to capture potentially hazardous dynamics demonstrates its value for critical applications like risk assessment.

Variants	RMSE	ACC
NeuralOM w/o PRC	0.6692	0.5502
NeuralOM w/o PEI	0.6865	0.5166
NeuralOM w/o MANA Sum	0.7037	0.5359
NeuralOM w/o MANA Mean	0.6978	0.5160
NeuralOM	**0.6009**	**0.6058**

Table 2: Ablation Studies on model design, the best results are in **bold**.

Ablation Studies

To understand the driving factors behind NeuralOM's superior performance and to validate the effectiveness of each of our design choices, we conduct detailed ablation studies. We primarily investigate two aspects: the importance of the input representation (i.e., subtracting climatology) and the contribution of each component in our core architecture.

Subtracting climatology priors is critical for performance. As mentioned in our methodology, we argue that focusing the model on learning "anomalies" rather than the full state is crucial for capturing slow-changing dynamics. To verify this, we train a variant of NeuralOM without

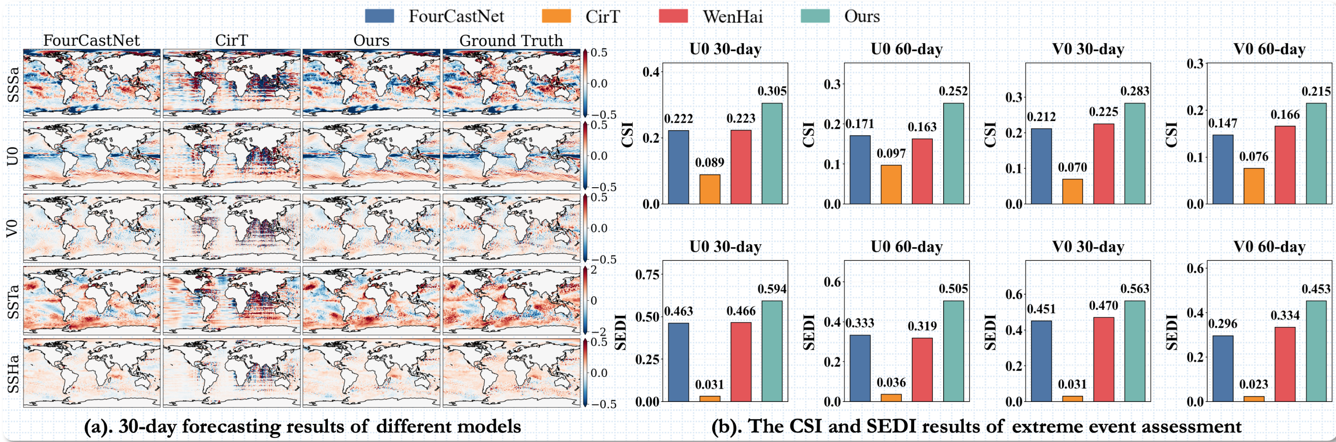

Figure 5: Performance on forecasting and extreme event assessment. **(a)** Visualizations of 30-day forecasts. Our model's outputs align closely with the ground truth, while baseline models exhibit significant artifacts. **(b)** CSI and SEDI scores for extreme surface current events. NEURALOM achieves higher scores than baselines at both 30 and 60-day lead times.

VARIANTS	23 LAYERS S		23 LAYERS T		SSH	
	RMSE	ACC	RMSE	ACC	RMSE	ACC
NEURALOM W/O SUBTRACT CLIMATOLOGY	0.0515	0.1066	0.1518	0.4532	0.2427	0.5737
NEURALOM W SUBTRACT CLIMATOLOGY	**0.0105**	**0.6412**	**0.0985**	**0.6507**	**0.1375**	**0.6690**

Table 3: Ablation Studies on whether subtracting the climatology from periodicity variables will improve the simulation performance, the best results are in **bold**.

MODELS	METRICS					
	15-DAY		25-DAY		35-DAY	
	RMSE	ACC	RMSE	ACC	RMSE	ACC
FOURCASTNET	0.55	0.68	0.68	0.47	1.30	0.19
CIRT	1.36	0.14	2.08	0.05	2.61	0.02
OURS	**0.52**	**0.73**	**0.62**	**0.59**	**0.68**	**0.48**

Table 4: In the real-world global ocean forecasting task, we compare the performance of our NEURALOM with 2 baselines. The average results for all 93 ocean variables of RMSE and ACC are recorded. The best results are in **bold**.

subtracting the climatological mean. The results, shown in **Table 3**, are striking. For sea salinity (S) across 23 depth layers, the model with climatology priors (W Subtract Climatology) achieves an ACC of 0.6412, whereas the model without them (W/O Subtract Climatology) only reaches 0.1066—a performance gap of over **5-fold**. This result provides strong evidence that incorporating physical priors like climatology into the input representation is a **prerequisite** for enabling the model to effectively learn fine-grained physical signals and achieve high-fidelity simulations.

Each core architectural component is indispensable. We evaluate the individual contributions of NEURALOM's core components by removing them one by one. The results, summarized in **Table 2**, show that removing any single component leads to a significant degradation in performance:

➠ ***Without Progressive Residual Correction*** **(w/o PRC):** Replacing the multi-stage PRC framework with a single model increases RMSE from 0.6009 to 0.6692, confirming that decomposing the complex prediction task into a series of simpler residual correction steps is vital for suppressing error accumulation and ensuring long-term stability.

➠ ***Without Physics-inspired Edge-Interactions*** **(w/o PEI):** Removing PEI, thereby reverting to a traditional, physics-agnostic message passing scheme, caused the ACC to drop from 0.6058 to 0.5166. This confirms that explicitly modeling physical processes like gradients and couplings in message passing improves the model's physical consistency.

➠ ***Multi-scale Adaptive Node Aggregation*** **(w/o MANA Sum/Mean):** Removing either the sum or the mean aggregation path degrades performance. This confirms the necessity of our adaptive aggregation mechanism, which allows the model to flexibly handle ocean dynamics at different scales.

In summary, these ablation studies confirms that the success of NEURALOM is not coincidental but is the result of a synergistic combination of its carefully designed components from the input representation and macroscopic framework to the microscopic interaction mechanisms.

Conclusion

In this paper, we introduce NEURALOM, a neural operator framework designed to solve the problem of long-term stability by simulating slow-changing physical systems. Its core innovations, a Progressive Residual Correction Framework and a Physics-Guided Graph Network, synergistically improve long-term stability and physical consistency. In the challenging task of global S2S ocean simulation, experiments demonstrate that NEURALOM significantly outperforms state-of-the-art models in forecast accuracy, long-term stability, and the ability to capture extreme events. Future work could explore integrating strict physical conservation laws into the loss function and reducing computational costs.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (42125503, 42430602).

References

Bi, K.; Xie, L.; Zhang, H.; Chen, X.; Gu, X.; and Tian, Q. 2023. Accurate medium-range global weather forecasting with 3D neural networks. *Nature*, 619(7970): 533–538.

Chang, X.; Wang, J.; Zhang, G.; Yang, M.; Xi, Y.; Xi, C.; Chen, G.; Nie, X.; Meng, B.; and Quan, X. 2023. Predicting colorectal cancer microsatellite instability with a self-attention-enabled convolutional neural network. *Cell Reports Medicine*, 4(2).

Chen, K.; Han, T.; Gong, J.; Bai, L.; Ling, F.; Luo, J.-J.; Chen, X.; Ma, L.; Zhang, T.; Su, R.; et al. 2023a. Fengwu: Pushing the skillful global medium-range weather forecast beyond 10 days lead. *arXiv preprint arXiv:2304.02948*.

Chen, L.; Zhong, X.; Zhang, F.; Cheng, Y.; Xu, Y.; Qi, Y.; and Li, H. 2023b. FuXi: a cascade machine learning forecasting system for 15-day global weather forecast. *npj climate and atmospheric science*, 6(1): 190.

Chen, Y.; Jin, Y.; Liu, Z.; Shen, X.; Chen, X.; Lin, X.; Zhang, R.-H.; Luo, J.-J.; Zhang, W.; Duan, W.; et al. 2025. Combined dynamical-deep learning ENSO forecasts. *Nature Communications*, 16(1): 3845.

Cui, Y.; Wu, R.; Zhang, X.; Zhu, Z.; Liu, B.; Shi, J.; Chen, J.; Liu, H.; Zhou, S.; Su, L.; et al. 2025. Forecasting the eddying ocean with a deep neural network. *Nature Communications*, 16(1): 2268.

Delaunay, A.; and Christensen, H. M. 2022. Interpretable deep learning for probabilistic MJO prediction. *Geophysical Research Letters*, 49(16): e2022GL098566.

Gao, Y.; Wu, H.; Shu, R.; Dong, H.; Xu, F.; Chen, R.; Yan, Y.; Wen, Q.; Hu, X.; Wang, K.; et al. 2025. OneForecast: A Universal Framework for Global and Regional Weather Forecasting. *arXiv preprint arXiv:2502.00338*.

Ham, Y.-G.; Kim, J.-H.; and Luo, J.-J. 2019. Deep learning for multi-year ENSO forecasts. *Nature*, 573(7775): 568–572.

Hao, R.; Zhao, Y.; Zhang, S.; and Deng, X. 2025. Deep Learning for Ocean Forecasting: A Comprehensive Review of Methods, Applications, and Datasets. *IEEE Transactions on Cybernetics*.

Hu, J.; Weng, B.; Huang, T.; Gao, J.; Ye, F.; and You, L. 2021. Deep residual convolutional neural network combining dropout and transfer learning for ENSO forecasting. *Geophysical Research Letters*, 48(24): e2021GL093531.

Jacox, M. G.; Alexander, M. A.; Amaya, D.; Becker, E.; Bograd, S. J.; Brodie, S.; Hazen, E. L.; Pozo Buil, M.; and Tommasi, D. 2022. Global seasonal forecasts of marine heatwaves. *Nature*, 604(7906): 486–490.

Jia, H.; Zhao, P.; Wu, H.; Gao, Y.; Tao, Y.; and Cui, B. 2025. Learning from History: A Retrieval-Augmented Framework for Spatiotemporal Prediction. *arXiv preprint arXiv:2510.24049*.

Keisler, R. 2022. Forecasting global weather with graph neural networks. *arXiv preprint arXiv:2202.07575*.

Kim, H.; Ham, Y.; Joo, Y.; and Son, S. 2021. Deep learning for bias correction of MJO prediction. *Nature Communications*, 12(1): 3087.

Kochkov, D.; Yuval, J.; Langmore, I.; Norgaard, P.; Smith, J.; Mooers, G.; Klöwer, M.; Lottes, J.; Rasp, S.; Düben, P.; et al. 2024. Neural general circulation models for weather and climate. *Nature*, 632(8027): 1060–1066.

Lam, R.; Sanchez-Gonzalez, A.; Willson, M.; Wirnsberger, P.; Fortunato, M.; Alet, F.; Ravuri, S.; Ewalds, T.; Eaton-Rosen, Z.; Hu, W.; et al. 2023. Learning skillful medium-range global weather forecasting. *Science*, 382(6677): 1416–1421.

Li, Z.; Kovachki, N.; Azizzadenesheli, K.; Liu, B.; Bhattacharya, K.; Stuart, A.; and Anandkumar, A. 2020. Fourier neural operator for parametric partial differential equations. *arXiv preprint arXiv:2010.08895*.

Liu, Y.; Zheng, Z.; Cheng, J.; Tsung, F.; Zhao, D.; Rong, Y.; and Li, J. 2025. Cirt: Global subseasonal-to-seasonal forecasting with geometry-inspired transformer. In *The Thirteenth International Conference on Learning Representations*.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, 10012–10022.

Lyu, P.; Tang, T.; Ling, F.; Luo, J.-J.; Boers, N.; Ouyang, W.; and Bai, L. 2024. ResoNet: Robust and explainable ENSO forecasts with hybrid convolution and transformer networks. *Advances in Atmospheric Sciences*, 41(7): 1289–1298.

Nguyen, T.; Shah, R.; Bansal, H.; Arcomano, T.; Maulik, R.; Kotamarthi, R.; Foster, I.; Madireddy, S.; and Grover, A. 2024. Scaling transformer neural networks for skillful and reliable medium-range weather forecasting. *Advances in Neural Information Processing Systems*, 37: 68740–68771.

Oskarsson, J.; Landelius, T.; Deisenroth, M.; and Lindsten, F. 2024. Probabilistic weather forecasting with hierarchical graph neural networks. *Advances in Neural Information Processing Systems*, 37: 41577–41648.

Parasyris, A.; Metheniti, V.; Kampanis, N.; and Darmaraki, S. 2025. Marine Heatwaves in the Mediterranean Sea: A Convolutional Neural Network study for extreme event prediction. *Ocean Science*, 21(3): 897–912.

Pathak, J.; Subramanian, S.; Harrington, P.; Raja, S.; Chattopadhyay, A.; Mardani, M.; Kurth, T.; Hall, D.; Li, Z.; Azizzadenesheli, K.; et al. 2022. Fourcastnet: A global data-driven high-resolution weather model using adaptive fourier neural operators. *arXiv preprint arXiv:2202.11214*.

Pfaff, T.; Fortunato, M.; Sanchez-Gonzalez, A.; and Battaglia, P. 2020. Learning mesh-based simulation with graph networks. In *International conference on learning representations*.

Shin, N.-Y.; Kang, D.; Kim, D.; Lee, J.-Y.; and Kug, J.-S. 2024a. Data-driven investigation on the boreal summer MJO predictability. *npj Climate and Atmospheric Science*, 7(1): 248.

Shin, N.-Y.; Kim, D.; Kang, D.; Kim, H.; and Kug, J.-S. 2024b. Deep learning reveals moisture as the primary predictability source of MJO. *npj Climate and Atmospheric Science*, 7(1): 11.

Sun, D.; Jing, Z.; and Liu, H. 2024. Deep learning improves sub-seasonal marine heatwave forecast. *Environmental Research Letters*, 19(6): 064035.

Wang, K.; Wu, H.; Duan, Y.; Zhang, G.; Wang, K.; Peng, X.; Zheng, Y.; Liang, Y.; and Wang, Y. 2024a. Nuwadynamics: Discovering and updating in causal spatio-temporal modeling. In *The Twelfth International Conference on Learning Representations*.

Wang, K.; Wu, H.; Zhang, G.; Fang, J.; Liang, Y.; Wu, Y.; Zimmermann, R.; and Wang, Y. 2024b. Modeling spatio-temporal dynamical systems with neural discrete learning and levels-of-experts. *IEEE Transactions on Knowledge and Data Engineering*.

Wang, X.; Wang, R.; Hu, N.; Wang, P.; Huo, P.; Wang, G.; Wang, H.; Wang, S.; Zhu, J.; Xu, J.; et al. 2024c. Xihe: A data-driven model for global ocean eddy-resolving forecasting. *arXiv preprint arXiv:2402.02995*.

Wu, H.; Gao, Y.; Shi, X.; Li, S.; Xu, F.; Zhang, F.; Zhu, Z.; Wang, W.; Luo, X.; Wang, K.; et al. 2025a. Spatiotemporal Forecasting as Planning: A Model-Based Reinforcement Learning Approach with Generative World Models. *arXiv preprint arXiv:2510.04020*.

Wu, H.; Gao, Y.; Shu, R.; Han, Z.; Xu, F.; Zhu, Z.; Wen, Q.; Wu, X.; Wang, K.; and Huang, X. 2025b. Turb-L1: Achieving Long-term Turbulence Tracing By Tackling Spectral Bias. *arXiv preprint arXiv:2505.19038*.

Wu, H.; Gao, Y.; Shu, R.; Wang, K.; Gou, R.; Wu, C.; Liu, X.; He, J.; Cao, S.; Fang, J.; et al. 2025c. Advanced long-term earth system forecasting by learning the small-scale nature. *arXiv preprint arXiv:2505.19432*.

Wu, H.; Gao, Y.; Xu, F.; Zhang, F.; Wen, Q.; Wang, K.; Huang, X.; and Wu, X. 2025d. Differential-Integral Neural Operator for Long-Term Turbulence Forecasting. *arXiv preprint arXiv:2509.21196*.

Wu, H.; Wang, C.; Xu, F.; Xue, J.; Chen, C.; Hua, X.-S.; and Luo, X. 2024a. Pure: Prompt evolution with graph ode for out-of-distribution fluid dynamics modeling. *Advances in Neural Information Processing Systems*, 37: 104965–104994.

Wu, H.; Wang, H.; Wang, K.; Wang, W.; Ye, C.; Tao, Y.; Chen, C.; Hua, X.-S.; and Luo, X. 2024b. Prometheus: Out-of-distribution Fluid Dynamics Modeling with Disentangled Graph ODE. In *Proceedings of the 41st International Conference on Machine Learning*, PMLR 235. Vienna, Austria: PMLR.

Wu, H.; Wang, S.; Liang, Y.; Zhou, Z.; Huang, W.; Xiong, W.; and Wang, K. 2023. Earthfarseer: Versatile Spatio-Temporal Dynamical Systems Modeling in One Model. *AAAI2024*.

Wu, H.; Weng, K.; Zhou, S.; Huang, X.; and Xiong, W. 2024c. Neural Manifold Operators for Learning the Evolution of Physical Dynamics. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 3356–3366.

Wu, H.; Xu, F.; Chen, C.; Hua, X.-S.; Luo, X.; and Wang, H. 2024d. Pastnet: Introducing physical inductive biases for spatio-temporal video prediction. In *Proceedings of the 32nd ACM international conference on multimedia*, 2917–2926.

Xiong, W.; Xiang, Y.; Wu, H.; Zhou, S.; Sun, Y.; Ma, M.; and Huang, X. 2023. Ai-goms: Large ai-driven global ocean modeling system. *arXiv preprint arXiv:2308.03152*.

Yang, N.; Wang, C.; Zhao, M.; Zhao, Z.; Zheng, H.; Zhang, B.; Wang, J.; and Li, X. 2024. LangYa: Revolutionizing Cross-Spatiotemporal Ocean Forecasting. *arXiv preprint arXiv:2412.18097*.

Yang, Y.-M.; Kim, J.-H.; Park, J.-H.; Ham, Y.-G.; An, S.-I.; Lee, J.-Y.; and Wang, B. 2023. Exploring dominant processes for multi-month predictability of western Pacific precipitation using deep learning. *Npj Climate and Atmospheric Science*, 6(1): 157.