Decoupling Regime and Mechanistic Heterogeneity via Hierarchical Mixture-of-Experts for Global Carbon Flux Prediction

Chengjie Fang
Hangzhou Institute for Advanced Study, University of Chinese Academy of Sciences
Hangzhou, Zhejiang, China
fangchengjie24@mails.ucas.ac.cn

Enyuan Zhao*
Hangzhou Institute for Advanced Study, University of Chinese Academy of Sciences
Hangzhou, Zhejiang, China
zhaoenyuan@ucas.ac.cn

Hongming Yao
Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences
Shenzhen, Guangdong, China
hm.yao@siat.ac.cn

Yujia Yang
Hangzhou Institute for Advanced Study, University of Chinese Academy of Sciences
Hangzhou, Zhejiang, China
yangyujia24@mails.ucas.ac.cn

Di Gao
Hangzhou Institute for Advanced Study, University of Chinese Academy of Sciences
Hangzhou, Zhejiang, China
gaodi@ucas.ac.cn

Yanbiao Li
Computer Network Information Center, Chinese Academy of Sciences
Beijing, China
Hangzhou Institute for Advanced Study, University of Chinese Academy of Sciences
Hangzhou, Zhejiang, China
lybmath@cnic.cn

Abstract

The integration of physical laws and machine learning has emerged as a compelling paradigm for modeling complex real-world systems. Compared to traditional numerical solvers, this approach significantly improves computational efficiency while preserving physical consistency. However, the source–sink dynamics of global carbon fluxes are driven by a complex coupling of diverse physical and biogeochemical mechanisms, resulting in pronounced spatiotemporal heterogeneity. Existing physics-informed methods typically embed governing equations directly into model architectures or loss functions. Inherently, this imposes strong assumptions of regime uniformity and mechanistic homogeneity, neglecting the intricate interactions across different regimes and physical processes. Consequently, while these methods excel on idealized partial differential equations (PDEs), they often exhibit limited generalizability when applied to complex real-world systems. To bridge this gap, we propose a hierarchical Mixture-of-Experts (MoE) framework, termed **Re**gime and **M**echanism Decoupled **M**ixture-**o**f-**E**xperts (**REM-MoE**), that explicitly encodes regimes and evolutionary mechanisms as two independent inductive biases during prediction. The core innovation lies in structurally decoupling complex systems: first-level experts model regime-specific heterogeneity, while second-level experts capture the diversity of underlying mechanisms. By dynamically combining these two levels, REM-MoE adaptively represents evolutionary dynamics under varying "regime–mechanism" conditions, thereby modeling state-dependent changes in regime–mechanism interactions. We validate our approach on a global carbon flux prediction task, where observational signals are jointly driven by multiple mechanisms across diverse regimes. Experimental results show that the proposed method consistently outperforms strong baselines in prediction accuracy while learning regime-dependent mechanism representations that are aligned with physical reference patterns. Furthermore, REM-MoE maintains strong performance over longer forecasting horizons and yields physically interpretable routing patterns. The code is publicly available at: https://github.com/Son4ta/REM-MoE.

CCS Concepts

• **Computing methodologies** → *Machine learning approaches*; • **Applied computing** → **Environmental sciences**.

Keywords

Physics-informed machine learning, ocean carbon flux, mixture-of-experts, multi-variable processing

ACM Reference Format:
Chengjie Fang, Enyuan Zhao, Hongming Yao, Yujia Yang, Di Gao, and Yanbiao Li. 2026. Decoupling Regime and Mechanistic Heterogeneity via Hierarchical Mixture-of-Experts for Global Carbon Flux Prediction. In *Proceedings of the 34th ACM International Conference on Multimedia (MM '26), November 10–14, 2026, Rio de Janeiro, Brazil.* ACM, New York, NY, USA, 9 pages.

*Corresponding author



This work is licensed under a Creative Commons Attribution 4.0 International License.
MM '26, Rio de Janeiro, Brazil
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2213-4/2026/11

1 Introduction

Against the backdrop of global warming and intensifying human activities, the precise quantification and dynamic prediction of the carbon cycle have emerged as core scientific challenges in Earth system

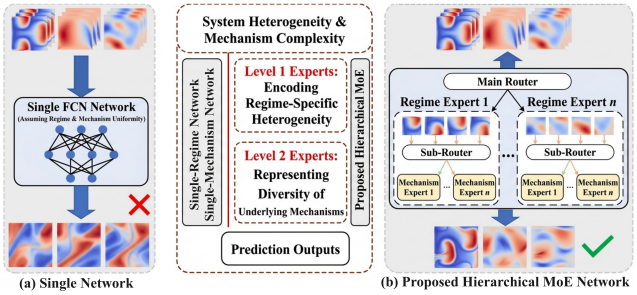


Figure 1: Comparison of modeling paradigms. (a) Traditional single-network architecture, which implicitly assumes regime uniformity and mechanistic homogeneity, leading to entangled representations of heterogeneous physical processes; (b) the proposed hierarchical MoE architecture, where first-level experts model regime-specific heterogeneity, and second-level experts capture diverse underlying mechanisms, enabling adaptive representation of state-dependent regime–mechanism interactions.

science, marine chemistry, and climate change research [11, 19, 24]. However, the source–sink dynamics of global carbon fluxes are driven by a complex coupling of diverse physical and biogeochemical mechanisms, exhibiting strong nonlinearities and pronounced spatiotemporal heterogeneity [35, 38]. These properties make it inherently challenging to reliably infer the underlying drivers from limited observations and to deliver robust forecasts across different spatial regimes and evolving system states [6, 12, 40].

In recent years, the integration of physical laws and machine learning has emerged as a compelling paradigm for modeling complex real-world systems. Compared to traditional numerical solvers, this approach significantly improves computational efficiency while preserving physical consistency [3]. A common strategy in existing physics-informed methods is to embed governing equations directly into model architectures or loss functions. Nevertheless, many of these approaches overlook a fundamental difficulty: the assumption of spatiotemporal uniformity. By applying global physical constraints, these methods inherently impose strong assumptions of regime uniformity and mechanistic homogeneity. Consequently, they neglect the intricate interactions across different regimes and diverse physical processes. As a result, while these methods excel on idealized partial differential equations (PDEs), they often fall short when applied to complex real-world systems characterized by pronounced spatiotemporal heterogeneity.

Physics-informed learning improves modeling efficiency by incorporating physical constraints, but global constraints often assume regime and mechanistic uniformity. For global carbon fluxes, where dominant processes vary with region and system state, such assumptions can entangle heterogeneous drivers and limit state-dependent modeling. We therefore propose a hierarchical MoE that encodes regimes and mechanisms as separate inductive biases: first-level experts model regime-specific heterogeneity, while second-level experts capture process diversity (Fig. 1).

Furthermore, training MoE architectures often encounters a severe cold-start problem, which frequently leads to routing collapse and expert homogenization. To overcome this, we inject explicit geospatial (regime) and mechanistic priors into the network. These priors provide strong initial guidance, enabling the model to rapidly navigate the difficult cold-start phase. This prior guidance mitigates routing collapse and expert homogenization during early training, promoting spatially and physically meaningful expert specialization.

Experiments on global ocean carbon-flux prediction demonstrate that the proposed REM-MoE framework achieves state-of-the-art predictive performance while producing mechanism-routing patterns consistent with state-dependent physical variations. By modeling regime-specific spatial heterogeneity and mechanistic diversity separately, the framework maintains robust forecasting performance across different horizons and heterogeneous ocean regimes.

Our main contributions are threefold:

(1) **Hierarchical decoupling of spatiotemporal heterogeneity.** We propose a novel dual-layer MoE architecture that explicitly models regimes and evolutionary mechanisms as independent inductive biases. This structural design relaxes the regime-uniformity and mechanistic-homogeneity assumptions commonly adopted by globally constrained physics-informed models.
(2) **Prior-guided expert training.** We introduce a strategy to inject explicit geospatial and mechanistic priors into the routing process to address the MoE cold-start problem. This guidance mitigates routing collapse and expert homogenization, encouraging spatially and physically distinct expert specialization.
(3) **Robust performance and physical interpretability.** REM-MoE consistently outperforms strong baselines in prediction accuracy. Furthermore, the learned "regime–mechanism" routing provides physically meaningful indications of how mechanism dominance varies across ocean regimes and seasons.

2 Related Work

Marine carbon-cycle modeling requires handling nonlinear, multi-scale dynamics while balancing physical consistency and computational efficiency. This has motivated learning-based approaches that combine data-driven flexibility with physical inductive biases.

Neural Operator Methods. Neural operator methods leverage the strong nonlinear approximation capability of deep neural networks to directly learn operator mappings between functional spaces, thereby bypassing explicit numerical solvers. Representative approaches include DeepXDE [18], the Fourier Neural Operator (FNO) [16], and their extensions such as U-FNO [30], LSM [31]and Geo-FNO [15]. These methods have demonstrated favorable efficiency in high-dimensional spatiotemporal prediction. However, neural operators typically treat the system as a black-box mapping without explicitly modeling latent physical quantities or underlying dynamical mechanisms. When applied to global carbon flux prediction, they struggle to account for the state-dependent transitions across different regimes, resulting in limited physical interpretability and poor extrapolation in out-of-distribution scenarios.

Physics-Informed Learning. Physics-informed learning incorporates governing equations or physical updates into the loss

function or network architecture to improve consistency and interpretability. Beyond standard Physics-Informed Neural Networks (PINNs) [22], recent studies have moved toward explicitly modeling latent physical quantities—such as velocity or fluxes—to provide more robust guidance. For instance, PICABU [10] employs causal structures for interpretability, while PhyMPGN [37] combines graph neural networks with numerical integrators. Other methods like PINP [2] and DeltaPhi [36] attempt to model multiple coupled variables or residuals to achieve structured learning. Nevertheless, these approaches typically impose physical constraints globally, inherently assuming mechanistic homogeneity and spatial uniformity. This makes them less effective for complex systems like global carbon fluxes, where interactions vary significantly across different regimes, leading to signal aliasing and a failure to capture local mechanism shifts.

Mixture-of-Experts. The Mixture-of-Experts paradigm offers a principled mechanism for expanding model capacity without a proportional increase in computational cost, with its core idea being to selectively activate a relevant subset of specialized expert networks for each input through a learned gating mechanism [13]. Building on this foundation, MoE architectures have been widely adopted in natural language processing and computer vision, where sparse gating strategies have demonstrated remarkable scalability [14]. Beyond these classical domains, MoE has been broadly extended to specialized vertical fields, leveraging its capacity to decompose complex tasks into simpler sub-problems handled by domain-specific experts [39]. In the physical sciences, variants have been proposed to handle non-stationary dynamics, multi-modal distributions, and regime-switching behavior [26].

3 Method

3.1 Overview of the Framework

We propose REM-MoE, a hierarchical two-level Mixture-of-Experts framework for global monthly surface ΔpCO_2 prediction (Figure 2). Multivariate historical ocean observations are first encoded into contextualized spatiotemporal tokens. A regime-level router then softly partitions the ocean into large-scale dynamical regimes, capturing geographical and climatic heterogeneity. Conditioned on this regime context, a process-level router activates thermodynamic, biogeochemical, and mixing experts to model region- and season-dependent mechanisms. A Global Regime Dictionary (GRD) further provides physically motivated spatial prototypes and process priors through annealed guidance, gradually shifting from prior-informed routing to data-driven specialization. The fused expert outputs finally produce the global ΔpCO_2 prediction. By explicitly modeling both regime heterogeneity and mechanistic diversity, REM-MoE reduces feature entanglement and better captures state-dependent ocean carbon dynamics.

3.2 Spatio-Temporal Encoding

To model the spatio-temporal evolution of ocean states, we first transform raw gridded multivariate observations into channel-isolated, context-aware representations. Given a sliding window of T monthly snapshots on a global 0.25° grid, the input contains $C = 5$ ocean variables—mixed layer depth, sea surface temperature, surface pCO_2, dissolved inorganic carbon, and chlorophyll-a—where B denotes the batch size and T the input sequence length, yielding $X_{\text{raw}} \in \mathbb{R}^{B\times T\times C\times H_{\text{grid}}\times W_{\text{grid}}}$. The grid is partitioned into non-overlapping $P\times P$ patches ($P = 8$), and land-dominated patches are removed, yielding N valid ocean patches. This produces a patch sequence $X_p \in \mathbb{R}^{B\times T\times N\times (C\cdot P^2)}$.

To reduce spurious coupling across physical mechanisms, we adopt a slotted embedding strategy. The variables are assigned to four feature slots: a shared slot containing all variables, and three specialized slots corresponding to thermal, biogeochemical, and mixing processes. Each slot is projected independently into its own subspace:

$$Z_s = \text{Linear}_s(X_p^{(s)}), \quad s \in \{\text{shared, thermal, bio, mixed}\} \tag{1}$$

The resulting slot features are concatenated to form the initial embedding $X_{embed} \in \mathbb{R}^{B\times T\times N\times H}$, where $H = \sum_s D_s$.

We then inject spatial and seasonal context by adding geographic positional encodings and applying month-conditioned FiLM modulation:

$$X_{all} = (X_{embed} + E_{pos}) \odot (1 + \gamma_t) + \beta_t. \tag{2}$$

where γ_t and β_t are generated from the calendar month.

Finally, the latest frame is locally smoothed through spatial message passing and used as a query to attend over the historical frames, yielding a temporally compressed representation. This representation is further processed by a Perceiver IO encoder with 512 learnable latent vectors to capture basin-scale interactions. The output, $\mathbf{h}_{shared} \in \mathbb{R}^{B\times N\times H}$, preserves channel isolation while providing the contextual input for the downstream hierarchical MoE.

3.3 Hierarchical MoE Architecture

Conventional single-network architectures attempt to fit multiple physical processes across all global regions and seasons using a globally shared parameter space, often resulting in extreme local physical signals being smoothed out by global averages. To address this, REM-MoE utilizes a hierarchically decoupled design: macroscopic spatial partitioning followed by microscopic process-level routing.

Regime-Level Routing. The primary objective of Level-1 routing is to identify the ocean climatic regime of the current spatiotemporal node. Unlike purely data-driven linear classifiers, our router employs explicit spatial prototype matching. We define a learnable prototype matrix $\mathbf{M} \in \mathbb{R}^{K\times H}$ within the GRD, physically initialized based on standard oceanographic partitions (e.g., RECCAP-2), where $K = 11$ is the number of RECCAP-2-derived regimes. Given $\mathbf{h}_{shared}$, we compute $\mathbf{h}_c = \mathbf{h}_{shared} - \bar{\mathbf{h}}$, where $\bar{\mathbf{h}}$ denotes the training-period global mean of $\mathbf{h}_{shared}$ accumulated over batch elements and valid ocean patches. The raw routing score $\mathbf{S}$ is calculated via the centered cosine similarity against the L2-normalized prototype matrix $\mathbf{M}_{eff}$:

$$\mathbf{S} = \frac{\mathbf{h}_c}{\|\mathbf{h}_c\|_2} \cdot \left(\frac{\mathbf{M}_{eff}}{\|\mathbf{M}_{eff}\|_2}\right)^{\top} \in \mathbb{R}^{B\times N\times K}, \tag{3}$$

A temperature-scaled softmax yields discrete regime assignment probabilities $\mathbf{P}_{regime}$:

$$\mathbf{P}_{regime} = \text{softmax}(\mathbf{S}/\tau). \tag{4}$$

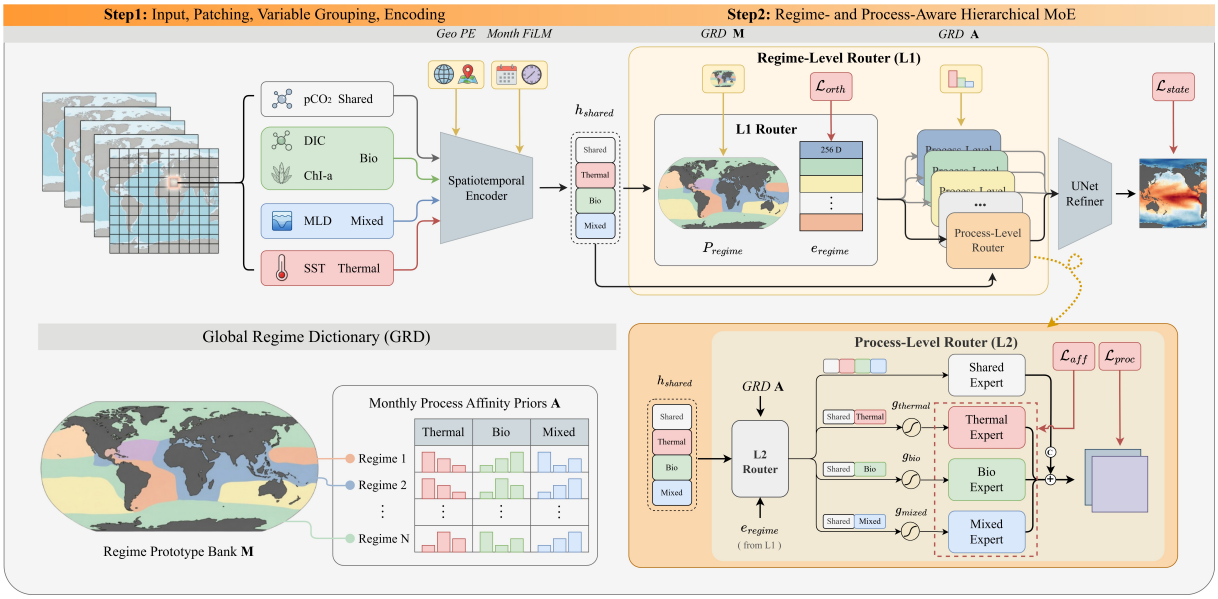


Figure 2: Schematic architecture of the proposed Regime and Mechanism Decoupled Mixture-of-Experts (REM-MoE).

To condition the downstream process experts, $\mathbf{P}_{regime}$ is projected via an orthogonally initialized embedding matrix $\mathbf{R} \in \mathbb{R}^{K\times E}$ into a continuous macro-context vector $\mathbf{e}_{regime} \in \mathbb{R}^{B\times N\times E}$, which encodes the geographical and dynamical background.

Process-Level Routing. Conditioned on the macro-background, the Level-2 router evaluates the concurrent contributions of thermodynamic, biogeochemical, and physical mixing processes. Using a non-mutually exclusive sigmoid gating mechanism, multiple experts can activate simultaneously. The gating weights $\mathbf{g} = [g_{th}, g_{bio}, g_{mix}] \in (0,1)^{B\times N\times 3}$ are generated by concatenating $\mathbf{h}_{shared}$ and $\mathbf{e}_{regime}$:

$$\mathbf{h}_{gate} = \mathrm{GELU}([\mathbf{h}_{shared}; \mathbf{e}_{regime}]\mathbf{W}_{g1} + \mathbf{b}_{g1}), \tag{5}$$

$$\mathbf{g} = \sigma(\mathbf{h}_{gate}\mathbf{W}_{g2} + \mathbf{b}_{g2}), \tag{6}$$

To enforce structural separation of process-specific inputs, each expert employs a "Slotted Reading" mechanism. Expert inputs are constructed by concatenating the generic shared slot $\mathbf{h}_{sh}$ with their strictly corresponding physical mechanism slot $(\mathbf{h}_{th}, \mathbf{h}_{bio}, \mathbf{h}_{mix})$. The inference of each process expert $j \in \{th, bio, mix\}$ is defined as:

$$\hat{y}_j = g_j \odot \mathrm{ExpertMLP}_j([\mathbf{h}_{sh}; \mathbf{h}_j]). \tag{7}$$

A Shared Expert also predicts a global residual without sigmoid attenuation.

Spatial Fusion. The 1D sequence outputs of the three process experts are element-wise summed to yield a unified physical process prediction $\hat{y}_{proc} = \sum_j \hat{y}_j$. Both $\hat{y}_{proc}$ and the shared expert output $\hat{y}_{shared}$ are scattered back to 2D feature maps. They are concatenated with an ocean-land mask and processed by a lightweight UNet to produce the final monthly ΔpCO_2 field.

3.4 Integration of Global Regime Dictionary (GRD)

To prevent routing collapse and cold-start ambiguity inherent to purely data-driven MoE architectures, the GRD injects physical priors through annealed guidance and maintains long-term spatial memory.

Regime Prototype Bank. To anchor early training physically, the prototype matrix $\mathbf{M}$ is initialized using the RECCAP-2 climatology. It incorporates learnable offsets $\Delta\mathbf{M}$ and undergoes Exponential Moving Average (EMA) updates based on batch feature residuals to sustain geographic memory. The effective matrix used for routing is defined as:

$$\mathbf{M}_{\text{eff}} = \text{L2-normalize}(\mathbf{M} + \Delta\mathbf{M}). \tag{8}$$

Monthly Process Affinity Priors. To guide the Level-2 router, we establish statistical affinity priors $\mathbf{A}_{prior,m} \in \mathbb{R}^{12\times K\times 3}$ for the three processes across months and regimes, derived via ridge regression on historical data. A target direction vector $\mathbf{a}_{dir,n} \in \mathbb{R}^3$ for token n in month m is computed using $\mathbf{P}_{regime}$ and a learnable offset $\Delta\mathbf{A}$:

$$\mathbf{a}_{dir,n} = \sum_{k=1}^{K} P_{regime}^{(n,k)} \cdot \left(\mathbf{A}_{prior,m}[k] + \Delta\mathbf{A}[k]\right). \tag{9}$$

Annealed Prior Guidance. The priors act via an annealed strategy: heavily weighted initially to construct a robust physical framework, and linearly annealed to zero as training progresses, fully activating $\Delta\mathbf{M}$ and $\Delta\mathbf{A}$ to allow data-driven degrees of freedom.

3.5 Objectives and Loss Functions

The overall optimization objective comprises a main task loss and customized auxiliary losses designed to enforce physical constraints.

Embedding Orthogonality Loss. To ensure diversity in the macroscopic regime representations, we penalize the off-diagonal

elements of the Gram matrix of the L2-normalized embedding vectors $\hat{\mathbf{r}}$:

$$\mathcal{L}_{orth} = \frac{1}{K(K-1)} \sum_{i=1}^{K} \sum_{j \neq i} (\hat{\mathbf{r}}_i \cdot \hat{\mathbf{r}}_j)^2, \tag{10}$$

Physical Process Consistency Loss. To guarantee that the joint contribution of the physical experts consistently explains the overall signal without any single expert dominating or all being suppressed, we apply a Huber loss on the summed 2D output $Y^{(proc)}$ against the normalized ground truth y:

$$\mathcal{L}_{proc} = \frac{1}{|\mathcal{M}_{2D}|} \sum_{i \in \mathcal{M}_{2D}} \text{Huber}(Y_i^{(proc)}, y_i), \tag{11}$$

where $\mathcal{M}_{2D}$ is the set of valid spatial pixels.

Expert Affinity Alignment Loss. To align the experts with regional statistical priors during the cold-start phase, we apply a magnitude-weighted cosine alignment loss:

$$\mathcal{L}_{aff} = \frac{\sum_{n \in \mathcal{M}} \|\mathbf{c}_n\|_2 \cdot (1 - \cos(\mathbf{c}_n, \mathbf{a}_{dir,n}))}{\sum_{n \in \mathcal{M}} \|\mathbf{c}_n\|_2 + \varepsilon}, \tag{12}$$

where $\mathbf{c}_n = [\hat{y}_{th,n}, \hat{y}_{bio,n}, \hat{y}_{mix,n}]$ is the concatenated expert output vector, and $\|\mathbf{c}_n\|_2$ acts as a stop-gradient magnitude weight to focus penalization on dynamically active regions.

Main Task Loss. The primary objective minimizes the prediction error of the final surface ΔpCO_2 variations using a robust Huber loss ($\delta = 1.0$) to mitigate the influence of oceanic anomalies:

$$\mathcal{L}_{state} = \frac{1}{|\mathcal{M}|} \sum_{i \in \mathcal{M}} \text{Huber}(\hat{y}_i, y_i), \tag{13}$$

Total Loss. The complete loss function integrates the state prediction and physical constraints:

$$\mathcal{L}_{total} = \mathcal{L}_{state} + \lambda_{proc}\mathcal{L}_{proc} + \beta\mathcal{L}_{orth} + \gamma_{aff}\mathcal{L}_{aff}. \tag{14}$$

where λ_{proc}, β, and γ_{aff} are hyperparameters scaling the auxiliary objectives.

4 Experiments

We evaluate REM-MoE along three complementary axes:

(Q1) Forecasting accuracy: can hierarchical regime–mechanism decoupling improve forecasting performance over strong data-driven and physics-informed baselines, especially under long-horizon rollouts where error accumulation is severe?

(Q2) Structural validity: are the decoupled latent design and explicit physical constraints *necessary* for stable optimization and faithful intermediate states, rather than merely helpful engineering choices?

(Q3) Physical consistency: do the learned routing patterns and expert outputs exhibit physically meaningful correspondence with established physical reference patterns?

4.1 Experimental Setup

Data protocol and splits. We construct the dataset from January 1985 to December 2024 and use a strictly chronological split: 1985–2018 for training, 2019–2021 for validation, and 2022–2024 for testing. The dataset uses CMEMS multi-source observation reconstructions and reanalysis products[1].

[1] https://data.marine.copernicus.eu/product/MULTIOBS_GLO_BIO_CARBON_SURFACE_MYNRT_015_008

Preprocessing. All input variables are Z-score normalized to account for scale differences across physical and biogeochemical fields. The normalization statistics are computed only on the training set and reused for validation and testing. For land-masked regions and missing values, we fill inputs with the corresponding climatological mean in anomaly space, preserving the zero-mean property while avoiding artificial spatial signals.

Tasks and metrics. We evaluate three multi-step forecasting settings: $3 \rightarrow 3$, $6 \rightarrow 6$, and $12 \rightarrow 12$ months (historical inputs $\rightarrow$ predicted future window), all under the same spatial resolution and ocean mask. Performance is assessed using MSE, MAE, PSNR, and PCC.

Baselines. We compare REM-MoE with representative spatiotemporal forecasting models (ConvLSTM, SimVP, and Earthformer) and operator/physics-informed approaches (FNO and PINP). All methods use the same data splits, input variables, temporal windows, spatial resolution, and ocean mask; baseline hyperparameters are selected on the validation set for fair comparison.

4.2 Global Spatiotemporal Forecasting Performance

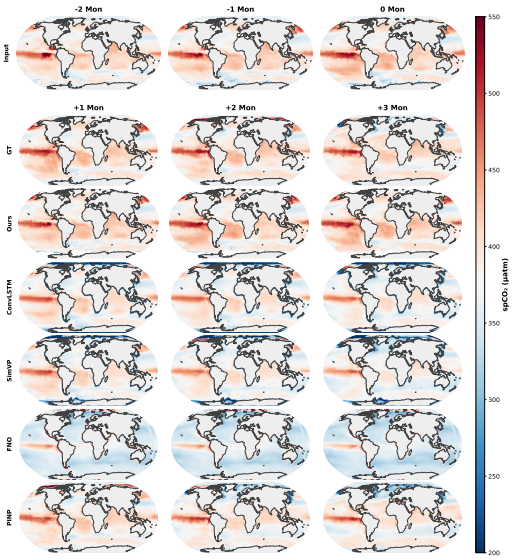


Figure 3: Multi-step forecasting in the $3 \rightarrow 3$ setting. The panels visualize absolute surface $spCO_2$ states; the model output is reconstructed from predicted monthly ΔpCO_2 increments and the preceding state.

Qualitative multi-step forecasts. As illustrated in Fig. 3, traditional single-network architectures (whether purely data-driven or globally physics-constrained) are highly susceptible to high-frequency signal attenuation and fine-scale structure loss during autoregressive rollouts. As discussed in Section 1, this degradation is particularly pronounced in regions characterized by intense dynamical variability (e.g., the equatorial Pacific and western boundary currents). By attempting to fit intrinsically distinct physical processes with a homogeneous set of rules, these baselines suffer from the entangled representations of heterogeneous drivers,

Table 1: Comparison on three forecasting tasks. Metrics with arrows indicate improvement direction (↓: lower is better, ↑: higher is better). The best results are in bold and the second-best are underlined.

Category	Model	Short-term (3→3)				Medium-term (6→6)				Long-term (12→12)			
		MSE↓	MAE↓	PSNR↑	PCC↑	MSE↓	MAE↓	PSNR↑	PCC↑	MSE↓	MAE↓	PSNR↑	PCC↑
Neural Ops.	FNO (2020)	0.4532	0.4915	21.185	0.8345	0.6475	0.6015	20.455	0.7681	0.6892	0.5991	20.035	0.7812
	U-NO (2023)	0.5241	0.5204	20.565	0.8142	0.7185	0.6265	19.995	0.7438	0.7725	0.6291	19.535	0.7525
	U-FNO (2022)	0.4371	0.4819	21.321	0.8402	0.6362	0.5945	20.515	0.7715	0.6671	0.5845	20.125	0.7935
	F-FNO (2021)	0.4425	0.4831	21.294	0.8396	0.6385	0.5981	20.495	0.7708	0.6795	0.5882	20.112	0.7895
	Transolver (2024a)	0.8082	0.6501	18.652	0.6881	0.9635	0.7092	18.705	0.6405	1.1612	0.7505	17.745	0.5285
	LSM (2023)	0.4431	0.4835	21.285	0.8392	0.6371	0.5942	20.525	0.7735	0.6785	0.5852	20.085	0.7915
	GNOT (2023)	1.1095	0.8171	17.285	0.4902	1.2145	0.7665	17.545	0.5115	1.3065	0.8885	17.385	0.4135
Deep Learning	ConvLSTM (2015)	0.1761	0.2802	26.291	0.9175	0.1835	0.2861	26.435	0.9131	0.1881	0.2882	26.595	0.9098
	PhyDNet (2020)	0.6152	0.5471	20.688	0.7795	0.6221	0.5521	20.875	0.7738	0.6235	0.5531	21.095	0.7735
	MotionRNN (2021)	0.2035	0.3028	25.515	0.9015	0.2058	0.3042	25.725	0.9018	0.2125	0.3135	25.825	0.8981
	U-Net (2015)	0.4682	0.4985	21.041	0.8301	0.6625	0.6085	20.335	0.7621	0.7125	0.6005	19.865	0.7775
	SmaAtUnet (2021)	0.1368	0.2479	27.195	0.9348	0.1465	0.2575	27.185	0.9291	0.1835	0.2835	26.425	0.9125
	SimVP (2022b)	0.1451	0.2508	27.215	0.9325	0.1478	0.2555	27.575	0.9305	0.1805	0.2862	26.485	0.9155
	TAU (2023)	0.1305	0.2325	27.482	0.9385	0.1315	0.2325	27.655	0.9389	0.1382	0.2355	27.685	0.9351
	Earthformer (2022a)	0.2232	0.3241	25.068	0.8898	0.2192	0.3195	25.415	0.8928	0.2212	0.3205	25.615	0.8928
	Galerkin (2021)	0.8205	0.6572	18.615	0.6812	0.9551	0.7101	18.765	0.6478	1.1625	0.7551	17.765	0.5225
	MWT (2021)	0.5485	0.5391	20.368	0.7971	0.7442	0.6391	19.845	0.7285	0.7481	0.6315	19.682	0.7585
	PastNet (2024b)	0.1728	0.2912	25.945	0.9135	0.1738	0.2921	26.185	0.9145	0.1758	0.2971	26.365	0.9131
	WaST (2024)	0.1681	0.2665	26.352	0.9241	0.1995	0.2925	25.875	0.9108	0.2265	0.3125	25.575	0.8965
Physics	PINP (2025)	0.0858	0.1585	30.185	0.9411	0.0975	0.1772	29.175	0.9331	0.1112	0.1965	28.485	0.9368
	PICABU (2025)	0.1205	0.2182	27.882	0.9391	0.1465	0.2465	26.835	0.9242	0.1905	0.3068	25.882	0.9038
	Alphapre (2025)	0.2115	0.3119	25.378	0.9012	0.2182	0.3175	25.435	0.8958	0.2275	0.3228	25.505	0.8875
Ours		**0.0376**	**0.0954**	**34.582**	**0.9821**	**0.0468**	**0.1135**	**33.412**	**0.9752**	**0.0685**	**0.1625**	**29.854**	**0.9562**

often yielding spatial artifacts. In stark contrast, REM-MoE maintains sharper physical gradients and more faithful spatial structures throughout the forecast, consistent with the explicit spatial routing and process-specific expert design.

Consistent gains across horizons. Quantitative evaluations are summarized in Table 1. Across the $3 \rightarrow 3$, $6 \rightarrow 6$, and $12 \rightarrow 12$ month settings, REM-MoE consistently outperforms all compared baselines, with a larger advantage in the 12-month setting. This trend is consistent with our motivation that state-aware regime–mechanism modeling can provide a useful structural inductive bias for long-horizon forecasting.

4.3 Ablation Study

To investigate the contributions of REM-MoE's main architectural and training components, we conduct a series of ablation experiments. As shown in Table 2, we ablated the full model across three dimensions: model architecture (Model), physical constraints (Constraint), and training strategy (Strategy). Furthermore, to comprehensively quantify the performance degradation of each variant in terms of statistical fidelity and dynamical stability, the Taylor diagram in Fig. 4 (Right) illustrates the evaluation results concerning correlation coefficient, standard deviation, and centered RMSE.

Model Ablation. To verify the effectiveness of the dual-layer decoupling architecture, we compared the model's performance when different components were removed. When the second-level physical process experts (M3: No L2 Experts) were removed, the prediction error increased significantly (ΔpCO_2 MSE increased to 0.125). This indicates that characterizing regional regimes solely through the first layer cannot adequately represent the superimposed evolution of multiple physical and biogeochemical mechanisms. If the first-level spatial routing (M4: w/o L1 Router) is removed, relying directly on global features to allocate physical experts, the error also increases notably (MSE 0.117). This confirms that process routing without spatial background anchoring leads

Table 2: Ablation study of REM-MoE. REM-MoE (Full Model) represents the baseline with all modules and training strategies enabled. Lower is better (↓).

Ablation Setting	Target (ΔpCO_2)	
	MSE ↓	MAE ↓
REM-MoE (Full Model)	**0.0685**	**0.1625**
Constraint Ablation		
C2. w/o Aff. Prior	0.0823	0.1794
C3. w/o Anch. GRD	0.1035	0.2049
C4. w/o Reg. Orth.	0.0733	0.1681
Model Ablation		
M2. No L1 Expert	0.0872	0.1853
M3. No L2 Experts	0.1252	0.2298
M4. w/o L1 Router	0.1174	0.2216
M5. w/o Geo-Season	0.0755	0.1712
Strategy Ablation		
S2. End-to-End Rt.	0.0768	0.1736
S3. Entangled Experts	0.1129	0.2165

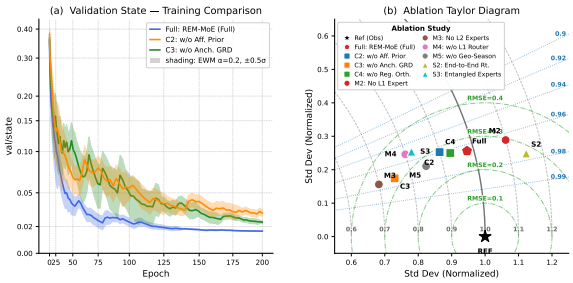


Figure 4: Validation loss curves and Taylor diagram. (a) Comparison of validation state loss curves between the main experiment and variants removing prior constraints during training; (b) Taylor diagram evaluation of all ablation variants.

to the inability of experts to effectively divide labor for specific regions. Additionally, the performance degradation of the M2 variant (removing the first-level shared expert, MSE 0.087) indicates the existence of global background signals in the real system that cannot be fully decomposed. The shared expert helps maintain the overall stability of the model.

Constraint Ablation. During the optimization process, the absence of explicit constraints may cause the model to reduce training loss by memorizing surface patterns rather than extracting physical features. The C2 variant (w/o Aff. Prior) removed the affinity prior component (A) in the GRD module, resulting in performance degradation. This suggests that the affinity prior facilitates process-specific specialization during training. The C3 variant (w/o Anch. GRD) removed the prototype anchoring component (M) in the GRD module. Fig. 5 displays the spatial routing diagnostic results for this variant. After removing the prototype anchoring, the Shannon entropy map in Fig. 5(b) shows increased routing uncertainty across the global ocean. The bar chart in Fig. 5(c) reveals routing collapse, where regional allocation is concentrated into a few categories, indicating a reduced capacity to represent spatial heterogeneity. The validation loss curves in Fig. 4 (Left) show that models lacking M anchoring (C3) or A prior (C2) exhibited substantial fluctuations in the early training stages, indicating that physical constraints help smooth the optimization space.

Strategy Ablation. Finally, we study the training strategy. In S3 (Entangled Experts), removing slot isolation and allowing all experts to access the full latent representation degrades performance (MSE = 0.112), suggesting that expert specialization weakens without structural isolation. The S2 variant, which removes the annealing strategy, also performs worse than the full model, indicating that annealed training helps avoid poor early routing and improves optimization stability.

4.4 Physical Interpretability Analysis

We analyze the latent routing patterns and expert outputs of REM-MoE to examine whether the hierarchical design yields physically meaningful, process-related representations.

L1 spatial routing. The first-level experts are designed to capture large-scale spatial heterogeneity. As shown in Fig. 6(a), the learned regimes form distinct clusters with clear boundaries in the SST–Chl-a phase space, while their monthly centroid trajectories

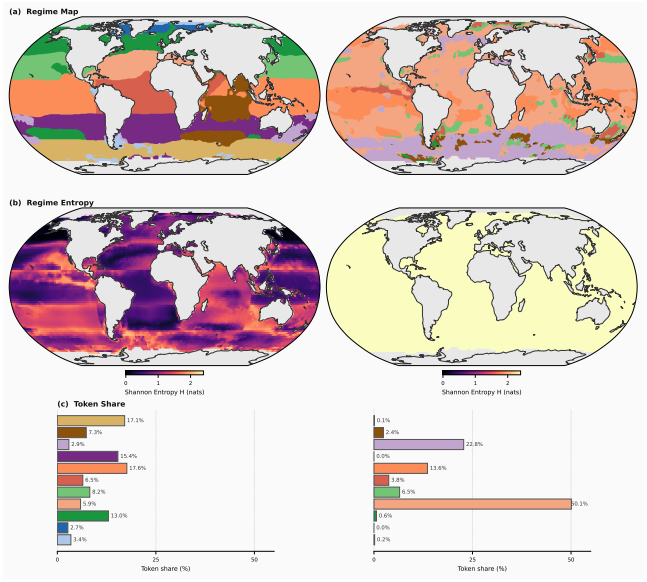


Figure 5: Spatial routing diagnostics comparison. (a) Gaussian-smoothed regime assignment maps; (b) Shannon entropy maps of regime assignments, where brighter colors indicate higher uncertainty; (c) Token proportions across regimes, highlighting the routing collapse in C3.

trace closed loops, reflecting annual oceanic cycles. The polar plots in Fig. 6(b)–(e) further reveal regime-dependent seasonal rhythms in environmental variables. For example, high-latitude and equatorial regimes exhibit clear differences in the phase and amplitude of MLD and DIC. These results suggest that L1 routing yields regionally meaningful partitions consistent with marine biogeochemical structure.

L2 mechanism routing. The second-level router determines which physical mechanisms are emphasized under different ocean states. To assess its physical correspondence, we use an empirical physical trend reference (Formula-A) in Fig. 7, constructed from established empirical relationships following [12]. In the six representative regimes shown in Fig. 7(a), the learned routing distributions align well with the formula-based density contours in ternary phase space. The seasonal centroid trajectories in Fig. 7(b)–(g) further exhibit trend-consistent shifts relative to Formula-A. This suggests that the learned routing preserves physically meaningful seasonal structure in the mapping from ocean state to process allocation.

L2 expert channel isolation. Expert outputs exhibit broad-scale spatial patterns and zonal profiles corresponding to their physical reference proxies (Fig. 8), complementing the gate-space comparison in Fig. 7. The maps show whether broad geographic structures are retained by each expert channel, while the zonal profiles test whether these similarities persist after latitudinal aggregation. Local and amplitude differences remain, so the agreement does not imply pixel-level identity. Because the process outputs are summed and combined with a shared residual for forecasting, this comparison supports process-related channel specialization rather than a unique additive or causal decomposition of the observed carbon-flux signal.

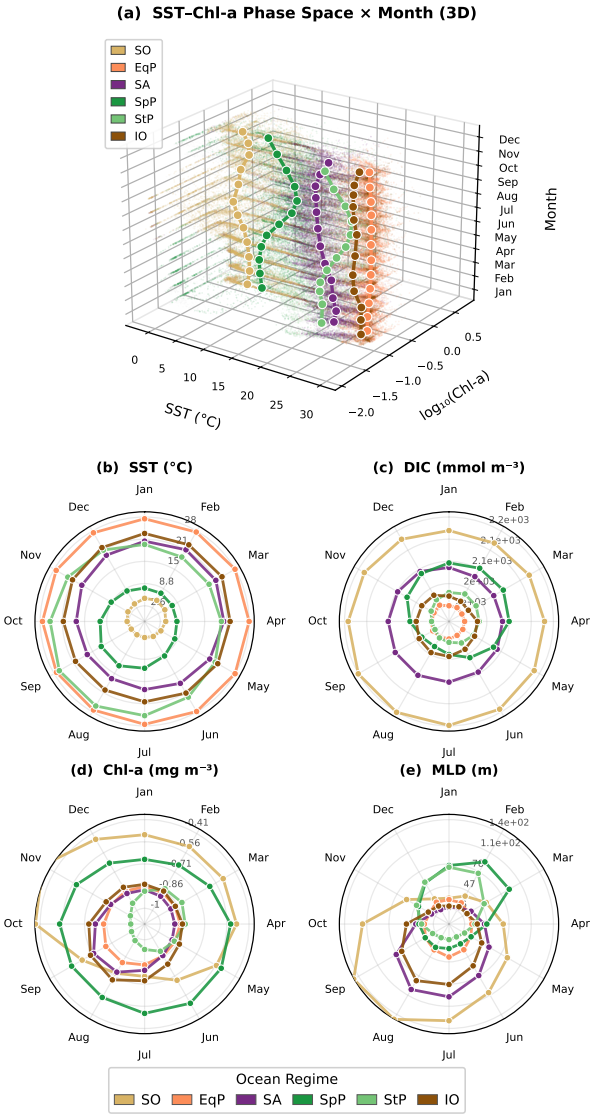


Figure 6: Ocean regime phase space and seasonal cycles. (a) Ocean regimes and monthly trajectories in the SST–Chl-a phase space. (b)–(e) Polar plots of seasonal cycles for SST, DIC, Chl-a, and MLD within each regime.

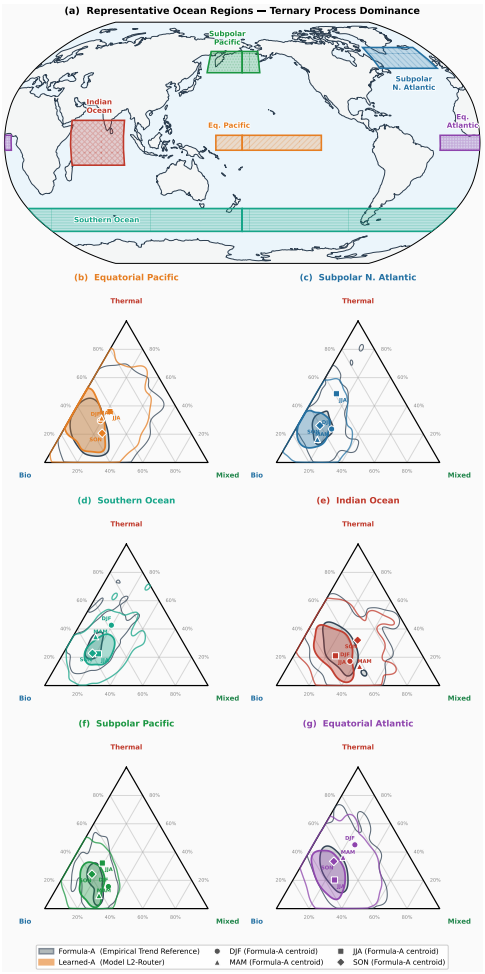


Figure 7: Alignment of physical dominance in ternary phase space. (a) Geographic distribution of representative regimes. (b)–(g) KDE comparison between the empirical trend reference (Formula-A) and the learned routing distribution (Learned-A) for thermodynamic, biological, and mixed processes.

5 Conclusion

In this work, we present REM-MoE, a hierarchical Mixture-of-Experts framework for modeling the spatiotemporal heterogeneity of global ocean carbon dynamics. By separating regime-level and process-level routing and incorporating prior-guided optimization, REM-MoE promotes expert specialization and mitigates routing collapse. Experiments demonstrate state-of-the-art long-horizon forecasting performance together with physically interpretable, trend-consistent routing patterns, supporting hierarchical regime–mechanism modeling for global ocean carbon-flux forecasting.

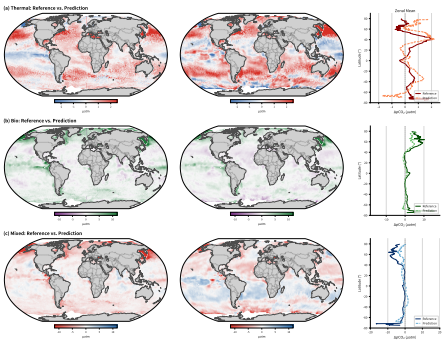


Figure 8: Expert outputs versus physical reference proxies. (a) Thermal process. (b) Biological process. (c) Mixed process.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant No. 42206035), the Science and Technology Program of Zhejiang Province (Grant No. 2025C02038), the Hangzhou Qianjiang Distinguished Expert Program (Grant No. HZ20251217834), and the Shenzhen Science and Technology Program (Grant No. JCYJ20250604183032041).

References

[1] Shuhao Cao. 2021. Choose a transformer: Fourier or galerkin. *Advances in neural information processing systems* 34 (2021), 24924–24940.

[2] Huaguan Chen, Yang Liu, and Hao Sun. 2025. PINP: Physics-Informed Neural Predictor with latent estimation of fluid flows. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net. https://openreview.net/forum?id=vAuodZOQEZ

[3] Siyuan Duan, Wenyuan Wu, Peng Hu, Zhenwen Ren, Dezhong Peng, and Yuan Sun. 2025. CoPINN: Cognitive Physics-Informed Neural Networks. In *Forty-second International Conference on Machine Learning*.

[4] Zhihan Gao, Xingjian Shi, Hao Wang, Yi Zhu, Yuyang Bernie Wang, Mu Li, and Dit-Yan Yeung. 2022. Earthformer: Exploring space-time transformers for earth system forecasting. *Advances in Neural Information Processing Systems* 35 (2022), 25390–25403.

[5] Zhangyang Gao, Cheng Tan, Lirong Wu, and Stan Z Li. 2022. Simvp: Simpler yet better video prediction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 3170–3180.

[6] Nicolas Gruber, Dorothee CE Bakker, Tim DeVries, Luke Gregor, Judith Hauck, Peter Landschützer, Galen A McKinley, and Jens Daniel Müller. 2023. Trends and variability in the ocean carbon sink. *Nature Reviews Earth & Environment* 4, 2 (2023), 119–134.

[7] Vincent Le Guen and Nicolas Thome. 2020. Disentangling physical dynamics from unknown factors for unsupervised video prediction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 11474–11484.

[8] Gaurav Gupta, Xiongye Xiao, and Paul Bogdan. 2021. Multiwavelet-based operator learning for differential equations. *Advances in neural information processing systems* 34 (2021), 24048–24062.

[9] Zhongkai Hao, Zhengyi Wang, Hang Su, Chengyang Ying, Yinpeng Dong, Songming Liu, Ze Cheng, Jian Song, and Jun Zhu. 2023. GNOT: A General Neural Operator Transformer for Operator Learning. *arXiv preprint arXiv:2302.14376* (2023).

[10] Sebastian Hickman, Ilija Trajkovic, Julia Kaltenborn, Francis Pelletier, Alex Archibald, Yaniv Gurwicz, Peer Nowack, David Rolnick, and Julien Boussard. 2025. Causal Climate Emulation with Bayesian Filtering. *The Thirty-ninth Annual Conference on Neural Information Processing Systems* (2025).

[11] Wolfgang Koeve, Angela Landolfi, Andreas Oschlies, and Ivy Frenger. 2024. Marine carbon sink dominated by biological pump after temperature overshoot. *Nature Geoscience* 17, 11 (2024), 1093–1099.

[12] Peter Landschützer, Nicolas Gruber, Dorothee CE Bakker, Irene Stemmler, and Katharina D Six. 2018. Strengthening seasonal marine CO2 variations due to increasing atmospheric CO2. *Nature Climate Change* 8, 2 (2018), 146–150.

[13] Lujun Li, Qiyuan Zhu, Jiacheng Wang, Xiaoyu Qin, Wei Li, Hao Gu, Sirui Han, and Yike Guo. 2026. Sub-moe: Efficient mixture-of-expert llms compression via subspace expert merging. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 40. 22994–23002.

[14] Yunxin Li, Shenyuan Jiang, Baotian Hu, Longyue Wang, Wanqi Zhong, Wenhan Luo, Lin Ma, and Min Zhang. 2025. Uni-moe: Scaling unified multimodal llms with mixture of experts. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2025).

[15] Zongyi Li, Daniel Zhengyu Huang, Burigede Liu, and Anima Anandkumar. 2023. Fourier neural operator with learned deformations for pdes on general geometries. *Journal of Machine Learning Research* 24, 388 (2023), 1–26.

[16] Zongyi Li, Nikola Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. 2020. Fourier Neural Operator for Parametric Partial Differential Equations. arXiv:2010.08895 [cs.LG]

[17] Kenghong Lin, Baoquan Zhang, Demin Yu, Wenzhi Feng, Shidong Chen, Feifan Gao, Xutao Li, and Yunming Ye. 2025. AlphaPre: Amplitude-Phase Disentanglement Model for Precipitation Nowcasting. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 17841–17850.

[18] Lu Lu, Xuhui Meng, Zhiping Mao, and George Em Karniadakis. 2021. DeepXDE: A deep learning library for solving differential equations. *SIAM review* 63, 1 (2021), 208–228.

[19] Moritz Mathis, Fabrice Lacroix, Stefan Hagemann, David Marcolino Nielsen, Tatiana Ilyina, and Corinna Schrum. 2024. Enhanced CO2 uptake of the coastal ocean is dominated by biological carbon fixation. *Nature Climate Change* 14, 4 (2024), 373–379.

[20] Xuesong Nie, Yunfeng Yan, Siyuan Li, Cheng Tan, Xi Chen, Haoyuan Jin, Zhihang Zhu, Stan Z Li, and Donglian Qi. 2024. Wavelet-driven spatiotemporal predictive learning: Bridging frequency and time variations. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 38. 4334–4342.

[21] Md Ashiqur Rahman, Zachary E. Ross, and Kamyar Azizzadenesheli. 2023. U-NO: U-shaped Neural Operators. *Transactions on Machine Learning Research* (2023). arXiv:2204.11127.

[22] M. Raissi, P. Perdikaris, and G. E. Karniadakis. 2019. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *J. Comput. Phys.* 378 (2019), 686–707. doi:10.1016/j.jcp.2018.10.045

[23] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*. 234–241.

[24] Nadine Schubert, Fernando Tuya, Viviana Peña, Paulo A Horta, Vinícius W Salazar, Pedro Neves, Cláudia Ribeiro, Francisco Otero-Ferrer, Fernando Espino, Kathryn Schoenrock, et al. 2024. "Pink power"—the importance of coralline algal beds in the oceanic carbon cycle. *Nature Communications* 15, 1 (2024), 8282.

[25] Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. 2015. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. *Advances in neural information processing systems* 28 (2015).

[26] Dengdi Sun, Xiaoya Zhou, Xiao Wang, Hao Si, Wanli Lyu, Jin Tang, and Bin Luo. 2026. NESTOR: A Nested MOE-based Neural Operator for Large-Scale PDE Pre-Training. *arXiv preprint arXiv:2602.22059* (2026).

[27] Cheng Tan, Zhangyang Gao, Lirong Wu, Yongjie Xu, Jun Xia, Siyuan Li, and Stan Z Li. 2023. Temporal attention unit: Towards efficient spatiotemporal predictive learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 18770–18782.

[28] Alasdair Tran, Alexander Mathews, Lexing Xie, and Cheng Soon Ong. 2021. Factorized Fourier Neural Operators. In *NeurIPS Workshop on Machine Learning for Physical Sciences*. arXiv:2111.13802.

[29] Kevin Trebing, Tomasz Stańczyk, and Siamak Mehrkanoon. 2021. SmaAt-UNet: Precipitation nowcasting using a small attention-UNet architecture. *Pattern Recognition Letters* 145 (2021), 178–186.

[30] Gege Wen, Zongyi Li, Kamyar Azizzadenesheli, Anima Anandkumar, and Sally M Benson. 2022. U-FNO—An enhanced Fourier neural operator-based deep-learning model for multiphase flow. *Advances in Water Resources* 163 (2022), 104180.

[31] Haixu Wu, Tengge Hu, Huakun Luo, Jianmin Wang, and Mingsheng Long. 2023. Solving High-Dimensional PDEs with Latent Spectral Models. In *International Conference on Machine Learning*.

[32] Haixu Wu, Huakun Luo, Haowen Wang, Jianmin Wang, and Mingsheng Long. 2024. Transolver: A Fast Transformer Solver for PDEs on General Geometries. In *International Conference on Machine Learning*.

[33] Hao Wu, Fan Xu, Chong Chen, Xian-Sheng Hua, Xiao Luo, and Haixin Wang. 2024. PastNet: Introducing Physical Inductive Biases for Spatio-temporal Video Prediction. In *Proceedings of the 32nd ACM International Conference on Multimedia* (Melbourne VIC, Australia) *(MM '24)*. Association for Computing Machinery, New York, NY, USA, 2917–2926. doi:10.1145/3664647.3681489

[34] Haixu Wu, Zhiyu Yao, Jianmin Wang, and Mingsheng Long. 2021. MotionRNN: A flexible model for video prediction with spacetime-varying motions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 15435–15444.

[35] Demin Yu, Xutao Li, Yunming Ye, Baoquan Zhang, Chuyao Luo, Kuai Dai, Rui Wang, and Xunlai Chen. 2024. Diffcast: A unified framework via residual diffusion for precipitation nowcasting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 27758–27767.

[36] Xihang Yue, Yi Yang, and Linchao Zhu. 2025. DeltaPhi: Physical States Residual Learning for Neural Operators in Data-Limited PDE Solving. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

[37] Bocheng Zeng, Qi Wang, Mengtao Yan, Yang Liu, Ruizhi Chengze, Yi Zhang, Hongsheng Liu, Zidong Wang, and Hao Sun. 2025. Phympgn: Physics-encoded message passing graph network for spatiotemporal pde systems. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24–28, 2025*. OpenReview.net. https://openreview.net/forum?id=fU8H4lzkIm

[38] Yuchen Zhang, Mingsheng Long, Kaiyuan Chen, Lanxiang Xing, Ronghua Jin, Michael I Jordan, and Jianmin Wang. 2023. Skilful nowcasting of extreme precipitation with NowcastNet. *Nature* 619, 7970 (2023), 526–532.

[39] Ze Zhang, Enyuan Zhao, Yi Jiang, Jie Nie, and Xinyue Liang. 2025. Challenging Dataset and Multi-Modal Gated Mixture of Experts Model for Remote Sensing Copy-Move Forgery Understanding. In *2025 IEEE International Conference on Multimedia and Expo (ICME)*. 1–6. doi:10.1109/ICME59968.2025.11210097

[40] Hao Zheng, Yuting Zheng, Hanbo Huang, Chaofan Sun, Enhui Liao, Lin Liu, Yi Han, Hao Zhou, and Shiyu Liang. 2025. CO2-Net: A Physics-Informed Spatio-Temporal Model for Global Surface CO2 Reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 6220–6230.