FORMA: FEEDBACK-CONDITIONED SPATIAL MEMORY FOR FIELD FORECASTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Long-horizon geophysical forecasting requires a predictive state that can remember persistent seasonal structure while rapidly adapting to moving fronts, changing spatial interactions, and unreliable observations. Existing forecasting architectures provide powerful temporal or spatial modeling mechanisms, but coupling long-term memory with observation-dependent spatial adaptation remains difficult. We introduce FORMA, a forecasting architecture centered on **Field Memory**, a field-valued predictive state whose evolution is regulated by feedback from the observed geophysical field. Seasonal phase, observation reliability, and interface dynamics determine what information is retained, what new evidence is written, and how memory propagates across space. FORMA implements this idea with an RWKV-style temporal recurrence together with efficient global and local spatial interaction, preserving spatial tokens throughout the historical scan while maintaining linear complexity in history length and grid size. On 90-day sea-surface-temperature forecasting, FORMA reduces area RMSE by **19.1%** over a matched RWKV recurrence and by **7.9%** over the strongest baseline in our comparison, with consistent gains in interface and basin-scale accuracy. The same architecture also transfers to Arctic sea-ice concentration forecasting, reducing concentration RMSE by **4.3%** and sea-ice-extent error by **25.4%**. These results demonstrate that explicitly modeling a feedback-regulated spatial predictive state is a promising alternative for long-horizon forecasting of evolving geophysical fields.

1 INTRODUCTION

Forecasting sea-surface temperature and sea-ice concentration requires modeling the evolution of spatial fields over time. These fields contain both broad seasonal patterns and localized structures, including thermal fronts and ice margins. Their evolution depends on information at different spatial and temporal scales. Seasonal context can remain useful across many observations, while a moving boundary requires frequent updates to its local features. Observation quality also varies across dates, affecting how much information a model can extract from the available history.

Convolutional recurrent models such as ConvLSTM and PredRNN maintain spatially structured hidden states over time (Shi et al., 2015; Wang et al., 2017), whereas feed-forward models such as SimVP directly map historical to future latent fields (Gao et al., 2022a). Neural operators enable efficient long-range interaction over spatial fields (Li et al., 2021; Tran et al., 2023). Recent Earth-system forecasting systems span graph-based, Transformer-based, and hybrid learned-physics designs, extending the scale and scope of learned forecasting (Lam et al., 2023; Bi et al., 2023; Gao et al., 2022b; Kochkov et al., 2024; Nguyen et al., 2024). Meanwhile, state-space and recurrent architectures such as Mamba and RWKV offer linear-time alternatives for long-sequence modeling (Gu & Dao, 2024; Peng et al., 2023), with recent spatiotemporal variants extending these ideas to long-range SST prediction (Jiang et al., 2025).

For long-horizon field forecasting, however, history does not contribute uniformly. Seasonal information may remain useful for months, while fronts and boundaries can change over much shorter intervals; observation quality also varies across dates. The model should therefore vary how strongly past information is retained, new observations are incorporated, and the predictive state is propagated

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

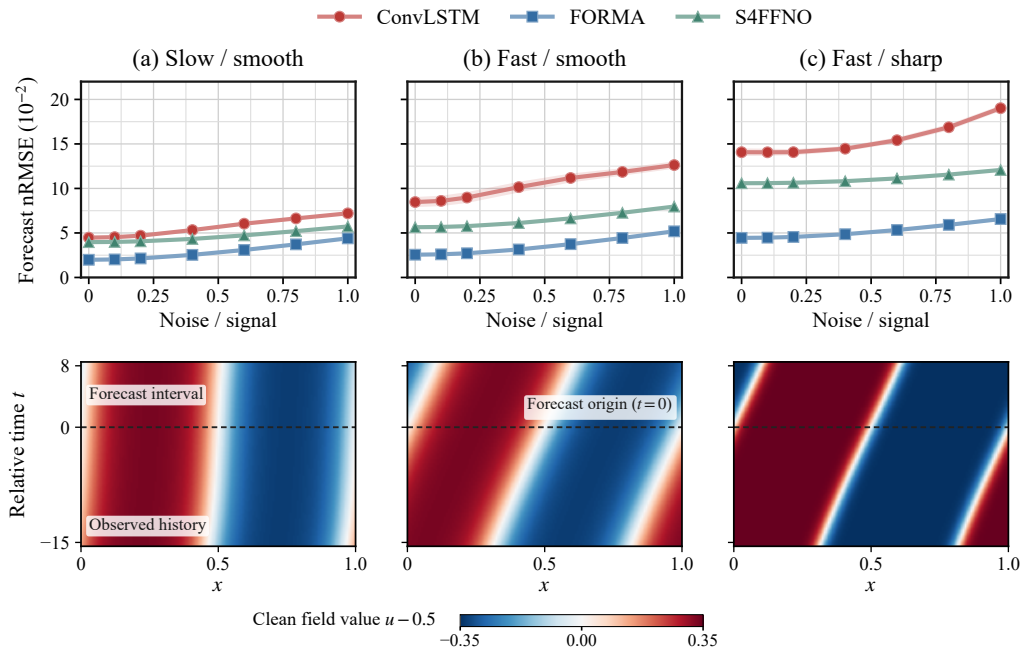


Figure 1: Controlled field-forecasting diagnostic under varying dynamics and input noise. The analytically generated fields contain a sheared translating front. Columns show slow/smooth, fast/smooth, and fast/sharp regimes. Top: forecast normalized RMSE as the relative noise-to-signal level increases; bottom: representative spatiotemporal fields. This descriptive study isolates the effects of temporal dynamics, interface sharpness, and observation quality.

across space. This suggests representing the observed history not only as a feature sequence, but as a spatial state whose update depends on the observed field.

Based on this view, we introduce FORMA, a spatiotemporal forecasting architecture built around Field Memory, a feedback-conditioned field-valued predictive state that evolves throughout the observed history. Field Memory uses three feedback signals—seasonal phase, input reliability, and interface features—to regulate state retention, the incorporation of new evidence, and spatial propagation. Its temporal component follows an RWKV-style recurrence, while a low-rank global route and an interface-conditioned local route provide complementary spatial interactions. The spatial state is preserved throughout the historical scan, allowing long-range temporal aggregation and spatial communication with linear complexity in history length and grid size at fixed feature width and route rank.

We evaluate FORMA on 90-day SST forecasting and apply the same architecture to Arctic SIC. The SST experiments compare forecasting accuracy, interface errors, and basin-mean error with convolutional, recurrent, spectral, Transformer, state-space, and ocean-specific models. Component ablations and input perturbations examine the contribution of feedback conditioning. Additional studies evaluate forecast lead time, history length, and model capacity.

To make this design setting concrete, Figure 1 reports a controlled diagnostic on analytically generated translating fronts rather than a separate benchmark. The diagnostic varies the apparent temporal dynamics (slow versus fast), the width of the field transition (smooth versus sharp), and the quality of the observed history through controlled Gaussian corruption. It therefore isolates the conditions under which a forecaster must retain long-range context while updating a moving interface; the protocol and model settings are summarized in Appendix D.1.

Diagnostic interpretation. Across the three regimes, FORMA retains the lowest error as corruption increases, with the clearest separation in the fast and sharp cases. The lower panels make the corresponding forecasting challenge explicit: the predictive state must carry a moving, spatially localized front across the forecast origin.

Our main contributions are:

(1) We propose FORMA, a spatiotemporal forecasting architecture that models evolving Earth-system fields with a feedback-regulated spatial predictive state. Seasonal phase, input reliability, and interface features condition how the state retains historical information, accepts new evidence, and exchanges information across locations. The architecture combines these updates with a spatial encoder and a multi-horizon decoder, allowing seasonal context, changing observation quality, and moving boundaries to guide the features accumulated for forecasting.

(2) We develop Field Memory, a spatial memory module that combines RWKV-style temporal aggregation with global and local propagation. A low-rank global route exchanges information between distant locations, and an interface-conditioned local route adapts neighborhood interactions to changing boundary structure. Feedback-conditioned retention and write gates regulate the state carried forward and the evidence added at each observed date. The recurrent scan and factorized spatial routes give linear complexity in history length and grid size at fixed feature width and route rank.

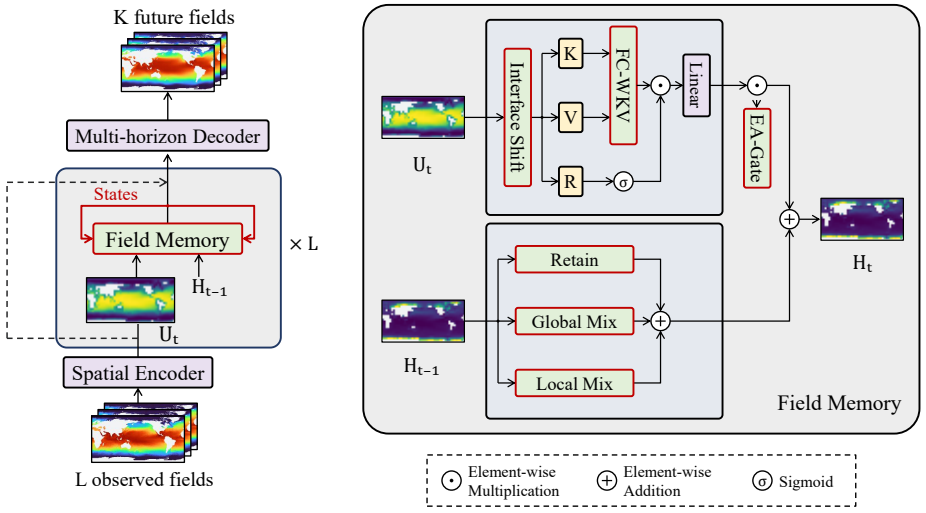
(3) We evaluate FORMA on 90-day SST forecasting using field and interface accuracy, spatial-pattern correlation, and basin-mean error. The full model reduces area RMSE by **19.1%** relative to the matched RWKV recurrence, from 1.1600 to **0.9390**°C, and achieves the lowest basin-mean error of **0.0552**°C among the compared models. Feedback ablations and input-quality perturbations examine the contribution of the memory update, while capacity and history-length studies characterize the model under different configurations. We also apply the architecture to independently trained Arctic SIC forecasting with a physically defined ice-edge target.

2 Related Work

Spatiotemporal and Earth-system forecasting. Recent Earth-system forecasting now spans graph, Transformer, neural-operator, and probabilistic systems (Lam et al., 2023; Bi et al., 2023; Gao et al., 2022b; Kochkov et al., 2024; Price et al., 2025; Nguyen et al., 2024; 2023; Zhang et al., 2023; Rasp et al., 2024). Earlier convolutional recurrent models such as ConvLSTM, PredRNN, and PredRNNv2 maintain spatial hidden states (Shi et al., 2015; Wang et al., 2017; 2023), while SimVP separates spatial encoding from temporal translation (Gao et al., 2022a). Large-scale systems increasingly combine learned propagation with physical components, probabilistic refinement, or observation updates; examples include graph-based forecasting, learned atmospheric dynamics, and generative ensembles (Lam et al., 2023; Kochkov et al., 2024; Price et al., 2025; Chen et al., 2023; Singh et al., 2024). Ocean and sea-ice forecasting adds seasonal nonstationarity, incomplete observations, and moving fronts or ice edges. Recent studies address these properties with multi-source convolutional prediction, Mamba-based spatiotemporal decomposition, probabilistic or generative modeling, multi-scale fusion, case adaptation, and ocean-specific dynamics (Xie et al., 2024; Jiang et al., 2025; Andersson et al., 2021; Xu et al., 2025a;b; Wang et al., 2026; 2025; Gao et al., 2026; Jiang et al., 2026; Bushuk et al., 2024; Lu et al., 2026). We study how seasonal, reliability, and interface information can jointly guide the update of a spatial memory.

Neural operators. DeepONet and FNO learn mappings between input and output functions (Lu et al., 2021; Li et al., 2021; Kovachki et al., 2023). Factorized and physics-informed operators improve efficiency or incorporate differential structure (Tran et al., 2023; Li et al., 2022; 2023; Brandstetter et al., 2022; Takamoto et al., 2022; Lippe et al., 2022). Geometry-aware, spherical, and equivariant operators extend these ideas to irregular domains and global grids (Tang et al., 2026; Shumaylov et al., 2025). Learned simulators and observation-driven models combine state propagation with new measurements (Pfaff et al., 2021; Singh et al., 2024). Field Memory uses a low-rank global operator and a sparse local operator within a recurrent update, with weights that depend on the current encoded field and feedback features.

Efficient recurrent and state-space models. Global interaction is commonly implemented with Transformer attention (Vaswani et al., 2017), token-mixing architectures (Tolstikhin et al., 2021; Liu et al., 2021; 2022), or IO-aware attention kernels (Dao et al., 2022; Dao, 2024). S4, Mamba, Mamba-2, RWKV, Hyena, and related recurrent models instead target long sequences with structured or linear-time updates (Gu et al., 2022; Gu & Dao, 2024; Dao & Gu, 2024; Peng et al., 2023; Poli et al.,



(a) FORMA Architecture

(b) Field Memory

Figure 2: Overall architecture of FORMA. (a) The forecasting architecture. (b) Field Memory. Interface Shift denotes interface-conditioned neighbor mixing; FC-WKV denotes feedback-conditioned weighted key–value aggregation; EA-Gate controls incoming evidence. $\mathbf{U}_t$ denotes encoded observation features, and $\mathbf{H}_{t-1}$ denotes the preceding internal memory state.

2023; Orvieto et al., 2023); multidimensional and visual variants include Mamba-ND, VMamba, VideoMamba, xLSTM, and V-RWKV (Li et al., 2024b; Liu et al., 2024; Li et al., 2024a; Beck et al., 2024; Duan et al., 2025). Latent dynamical models further emphasize scalable state evolution and long-horizon prediction (Hansen et al., 2024). FORMA uses an RWKV-style recurrence along the time dimension at each spatial location, then adds feedback-conditioned retention, writing, and spatial propagation to organize the accumulated evidence.

3 FORMA

3.1 Overall Architecture

FORMA maps L observed fields to K future fields. As shown in Figure 2, the model contains a shared spatial encoder, feedback feature construction, Field Memory, and a multi-horizon decoder. The encoder produces a latent grid for each observation. Field Memory scans these grids in temporal order and updates a spatial predictive state. The decoder combines the final state with features from the latest observation.

Let $\mathbf{X}_t \in [0,1]^{H\times W\times C}$ contain the normalized field and its validity channel, and let $\mathbf{A}_t$ contain auxiliary information available at date t. The forecasting task is

$$\widehat{\mathbf{Y}}_{t+1:t+K} = f_\theta(\mathbf{X}_{t-L+1:t}, \mathbf{A}_{t-L+1:t}). \tag{1}$$

Within each input window, we index observations by $1,\dots,L$. The encoder maps each observation to $\mathbf{U}_t \in \mathbb{R}^{N\times D}$, where N is the number of latent spatial locations and D is the feature width. Field Memory maintains a state $\mathbf{H}_t \in \mathbb{R}^{N\times D}$ on the same grid. This state stores the features accumulated during the historical scan, and the decoder predicts all K future fields jointly from its terminal output.

3.2 Feedback Features

We derive three causal feedback signals from the calendar and the observations available at date t: seasonal phase $\mathbf{p}_t$, input reliability q_t, and interface features $\mathbf{B}_t$.

Seasonal phase. The phase code uses annual and semi-annual sine/cosine features, including a shifted annual component, and conditions temporal retention, writing, and spatial aggregation.

Input reliability. The date-level score $q_t \in [0, 1]$ is a causal observation-quality descriptor. Broadcast over the latent grid, it conditions retention, writing, and global aggregation; the write gate uses it to scale newly added evidence.

Interface features. A learned head extracts interface features from the current encoded field and its temporal change:

$$\mathbf{B}_t = \chi_\theta([\mathbf{U}_t, \mathbf{U}_t - \mathbf{U}_{t-1}]). \tag{2}$$

The first difference in a window is zero. Historical supervision uses a gradient band for SST and a band around the physical 15% SIC threshold; the resulting features condition local aggregation and writing.

Before temporal aggregation, we mix each encoded feature with its four spatial neighbors:

$$\widetilde{\mathbf{U}}_{t,n} = \mathbf{U}_{t,n} + \eta \sum_{\delta \in \mathcal{N}(n)} g_{t,n,\delta} \mathbf{U}_{t,\delta}, \tag{3}$$

where $g_{t,n,\delta}$ is obtained by applying a softmax to a projection of $(\mathbf{B}_{t,n}, \mathbf{p}_t)$. This shift supplies local context to the temporal recurrence. Longitude is periodic, and missing neighbors at the latitude boundaries contribute zero.

3.3 Field Memory

Field Memory first aggregates historical features at each spatial location, then uses the resulting evidence to update the spatial memory. The update combines retention, new writes, global propagation, and local propagation.

Temporal aggregation. Pointwise projections of $\widetilde{\mathbf{U}}_t$ produce keys $\mathbf{K}_t$, values $\mathbf{V}_t$, and sigmoid receptance gates $\mathbf{R}_t$. Feedback-conditioned weighted key–value aggregation (FC-WKV) retains the normalized RWKV recurrence and conditions its decay on the observed field and feedback signals. It maintains a numerator $\mathbf{S}_t$ and denominator $\mathbf{N}_t$:

$$\mathbf{S}_t = \boldsymbol{\lambda}_t \odot \mathbf{S}_{t-1} + \exp(\mathbf{K}_t) \odot \mathbf{V}_t, \tag{4}$$

$$\mathbf{N}_t = \boldsymbol{\lambda}_t \odot \mathbf{N}_{t-1} + \exp(\mathbf{K}_t), \tag{5}$$

$$\mathbf{E}_t = W_o\left(\mathbf{R}_t \odot \frac{\mathbf{S}_t}{\mathbf{N}_t + \epsilon}\right). \tag{6}$$

The decay gate $\boldsymbol{\lambda}_t$ depends on the spatially pooled encoded field, seasonal phase, and reliability. The recurrence produces evidence $\mathbf{E}_t$ at every location while retaining a fixed-size pair of temporal states.

Spatial memory update. The memory state is updated by

$$\mathbf{H}_t = \mathbf{F}_t \mathbf{H}_{t-1} + \boldsymbol{\Gamma}_t \odot \mathbf{E}_t, \qquad \mathbf{F}_t = (1 - \alpha)\mathbf{D}_t + \alpha(\omega_G \mathbf{P}_t^G + \omega_E \mathbf{P}_t^E), \tag{7}$$

where $\mathbf{D}_t$ applies location- and channel-wise retention, $\boldsymbol{\Gamma}_t$ is the write gate, and $\mathbf{P}_t^G$ and $\mathbf{P}_t^E$ are the global and local spatial operators. The feedback-conditioned heads generate these factors:

$$(\mathbf{D}_t, \boldsymbol{\Gamma}_t, \mathbf{P}_t^G, \mathbf{P}_t^E) = \Xi_\theta(\widetilde{\mathbf{U}}_t, \mathbf{p}_t, q_t, \mathbf{B}_t). \tag{8}$$

Retention uses a sigmoid gate. Writing is additionally modulated by the causal observation-quality descriptor. In the reference configuration, $\alpha = 0.1$ and the two active routes have equal weights $\omega_G = \omega_E = 1/2$.

Global propagation. The global route exchanges information across the latent grid through a rank-r factorization:

$$\mathbf{P}_t^G = \mathbf{A}_t^G (\mathbf{C}_t^G)^\top, \qquad \mathbf{A}_t^G, \mathbf{C}_t^G \in \mathbb{R}^{N \times r}. \tag{9}$$

The factors are predicted from the encoded field, phase, and reliability. A softmax normalizes $\mathbf{C}_t^G$ over spatial locations and $\mathbf{A}_t^G$ over the rank dimension. Computation proceeds in two steps: $(\mathbf{C}_t^G)^\top \mathbf{H}_{t-1}$ forms r global summaries, and $\mathbf{A}_t^G$ distributes these summaries back to the grid. This gives each location access to the full spatial state at a cost of $O(NrD)$.

Local propagation. The local route aggregates memory from four adjacent locations:

$$\mathbf{P}_t^E = \mathbf{A}_{\mathcal{G}} \odot \rho_\theta(\mathbf{B}_t, \mathbf{p}_t), \tag{10}$$

where $\mathbf{A}_{\mathcal{G}}$ specifies grid adjacency and ρ_θ produces normalized neighbor weights. These weights adapt local communication to the interface features and seasonal phase. The global route provides long-range interaction, while the local route supplies spatially varying neighborhood updates at a cost of $O(ND)$.

Equation 7 combines location-wise retention with normalized global and local propagation. The temporal states and spatial memory are initialized at the start of each input window; iterating the update over L observations produces the final memory $\mathbf{H}_L$.

State-space interpretation. Conditioned on the observed history, the feedback heads specify a sequence of transition operators, so Equation 7 is a non-autonomous affine state-space system with input $\mathbf{W}_t = \mathbf{\Gamma}_t \odot \mathbf{E}_t$. Defining $\mathbf{\Phi}_{T\leftarrow s} = \mathbf{F}_T \cdots \mathbf{F}_{s+1}$ gives

$$\mathbf{H}_T = \mathbf{\Phi}_{T\leftarrow 0}\mathbf{H}_0 + \sum_{s=1}^{T} \mathbf{\Phi}_{T\leftarrow s}\mathbf{W}_s. \tag{11}$$

Thus the terminal memory is an explicit sum of dated evidence writes, each propagated by the subsequent observation-dependent dynamics. The softmax normalizations make the global route row-stochastic and the local route row-substochastic. Consequently, if $\bar{d}_t = \max_{n,d} D_{t,n,d} < 1$, then $\|\mathbf{F}_t\|_\infty \le (1-\alpha)\bar{d}_t + \alpha < 1$: with the input-derived factors fixed, perturbations to the retained state contract rather than amplify. Feedback therefore changes the propagation kernel, while normalized spatial mixing controls its stability. Appendix C gives the full multi-step bound and the corresponding space–time propagation kernel.

3.4 Decoder and Training

Multi-horizon decoder. We reshape $\mathbf{H}_L$ and $\mathbf{U}_L$ to the latent grid and add a 1×1 projection of the latest encoded observation to the memory. This path supplies recent features directly to the prediction head. The combined feature map is resized and decoded into K future fields and K interface-logit maps. A sigmoid maps field predictions to the normalized output range.

Training objective. We train the encoder, Field Memory, and decoder end to end with

$$\mathcal{L} = \mathcal{L}_{\text{field}} + \beta_e \mathcal{L}_{\text{interface}} + \beta_h \mathcal{L}_{\text{hist-interface}} + \beta_a \mathcal{L}_{\text{region}} + \beta_t \mathcal{L}_{\text{temp}}. \tag{12}$$

The field term is a masked Huber loss. Future and historical interface losses supervise the predicted boundaries and the interface features used during the scan. The regional term penalizes differences in area-weighted spatial means, and the temporal term matches changes between adjacent forecast dates. The complete definitions and weights are given in Appendix B.

Computation. Input-dependent projections are batched over the observed history before the recurrent scan. At fixed feature width and route rank, the scan has linear complexity in both history length and the number of spatial tokens. Including pointwise projections, its cost is $O(LND^2 + LNrD + L|\mathcal{G}|D)$. The implementation uses low-rank matrix products for global propagation and optional Triton kernels for neighbor aggregation and elementwise state updates.

4 Experiments

We evaluate FORMA on SST and Arctic SIC forecasting with 90 observed daily fields and a 90-day prediction horizon. The experiments cover field and interface accuracy, spatial-pattern agreement, feedback conditioning, and model capacity.

4.1 Experimental Setup

Data and splits. We use NOAA OISSTv2.1 SST (Reynolds et al., 2007; Banzon et al., 2016) and NOAA/NSIDC CDR v6 SIC (National Snow and Ice Data Center, 2026), both regridded to $1°$. Complete forecast intervals determine the chronological train (1982–2018), validation (2019–2021), and test (2022–2024) splits. Statistics and SST interface thresholds are fitted on the training set, and invalid cells are excluded from losses and metrics. SIC uses valid Arctic-ocean cells north of $30°$N; native-resolution SST is evaluated separately at $0.25°$.

Optimization. Unless noted otherwise, every trainable configuration in the SST and SIC experiments (including baselines, ablations, resolution, capacity, and history/lead-time studies) is trained on Nvidia A100 GPUs for 20 epochs with batch size 16. FORMA configurations use AdamW with a learning rate of 3×10^{-4} and weight decay of 1×10^{-4}. We use mixed precision and gradient clipping; checkpoints are selected by validation area RMSE.

Baselines. The comparison covers convolutional (3D U-Net (Xie et al., 2024)), feed-forward (SimVP (Gao et al., 2022a)), recurrent (PredRNNv2 (Wang et al., 2023)), spectral (FFNO (Tran et al., 2023)), Transformer (Earthformer (Gao et al., 2022b)), state-space (STDMamba (Jiang et al., 2025)), and ocean-specific (NeuralOM (Gao et al., 2026); SSTODE (Jiang et al., 2026)) architectures.

All models use the common split and dates; PredRNNv2 retains its recurrent decoder, while other baselines use task-specific multi-horizon heads. Public sources and adaptations are summarized in Appendix D.2. The matched RWKV control shares FORMA's encoder, decoder, recurrence, and latest-observation path.

Metrics. SST evaluation uses cosine-latitude area RMSE, interface-weighted RMSE, correlation deficit $D_{\mathrm{corr}} = 10^3(1 - \mathrm{ACC})$, and basin-mean error. SIC evaluation uses concentration RMSE, D_{corr}, sea-ice-extent (SIE) error, and integrated ice-edge error (IIEE). Definitions are given in Appendix B.2.

4.2 Main SST Results

Table 1: SST forecasting with a 90-day history and a 90-day forecast horizon. Temperature-error columns are displayed in 10^{-2} °C. The correlation deficit is $D_{\mathrm{corr}} = 10^3(1 - \mathrm{ACC})$, where ACC is the cosine-latitude area-weighted spatial correlation. Lower values are better for all metrics. All models use the common data split and validation-selected checkpoints.

Method	Area RMSE↓	Interface RMSE↓	D_{corr} ↓	Basin error↓
3D U-Net (Xie et al., 2024)	102.0	118.7	4.5	21.6
SimVP (Gao et al., 2022a)	115.6	131.9	5.3	25.6
PredRNNv2 (Wang et al., 2023)	133.0	147.0	7.2	42.5
FFNO (Tran et al., 2023)	105.6	122.2	4.8	21.1
Earthformer (Gao et al., 2022b)	103.4	122.2	4.9	6.1
STDMamba (Jiang et al., 2025)	117.0	138.6	6.1	9.6
NeuralOM (Gao et al., 2026)	141.4	164.4	9.2	14.1
SSTODE (Jiang et al., 2026)	146.6	168.7	9.9	7.2
FORMA	**93.9**	**109.6**	**4.1**	**5.5**

FORMA is best in every column of Table 1, with area RMSE 7.9% below the strongest competing baseline, 3D U-Net, and 9.2% below Earthformer.

Native-resolution SST. Direct $0.25°$ prediction improves area RMSE, interface RMSE, and D_{corr} from 0.9681, 1.0174, and 4.4 to **0.8609**, **0.9015**, and **3.4**, respectively; the gain remains after area-aware aggregation (Table 2).

Table 2: Native-resolution SST evaluation. Temperature-error columns are displayed in 10^{-2} °C; D_{corr} is defined in Table 1. *Aggregated* evaluates the native-grid prediction after area-aware aggregation to $1°$. The upsampled configuration interpolates the output of a $1°$ model.

Configuration	Grid	Area RMSE↓	Interface RMSE↓	D_{corr} ↓	Aggregated RMSE↓
FORMA reference	$1°$	93.9	109.6	4.1	N/A
FORMA native	$0.25°$	**86.1**	**90.1**	**3.4**	**85.8**
$1°$ forecast, upsampled	$0.25°$	96.8	101.7	4.4	97.1

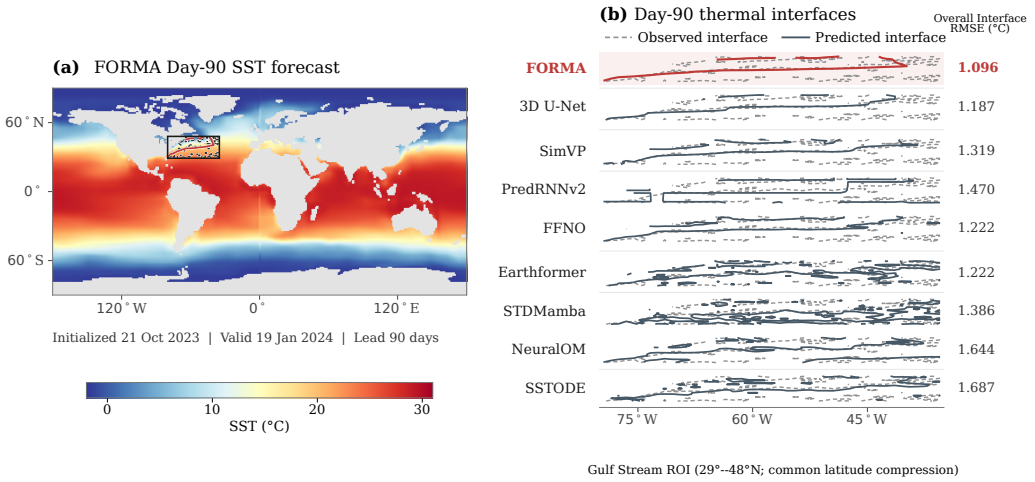


Figure 3: Day-90 SST forecast and thermal interfaces. (a) FORMA prediction for the representative test origin; the box marks the Gulf Stream region shown in (b). (b) Observed and predicted interfaces for the same target. The origin is closest to FORMA's median Day-90 interface RMSE. Values at right reproduce the test-set averages in Table 1, rather than regional errors; latitude is compressed identically across rows.

Figure 3 shows the displacement, fragmentation, and smoothing of the predicted interface; FORMA best preserves the principal observed front.

4.3 Cross-Field Validation

Settings. We train the same architecture independently on Arctic SIC, using the ocean mask, concentrations in $[0, 1]$, and a soft band around the 15% ice threshold. Table 3 compares external baselines and a control without interface-related Field Memory components; metric definitions are in Appendix B.2.

Table 3: Arctic SIC forecasting with a 90-day history and a 90-day horizon. SIC RMSE is displayed in 10^{-2} concentration units, while SIE error and IIEE are displayed in 10^{-2} million km^2. The correlation deficit is $D_{\text{corr}} = 10^3(1 - \text{ACC})$. All trainable models use 20 epochs and batch size 16.

Method	SIC RMSE↓	D_{corr} ↓	SIE error↓	IIEE↓
SimVP (Gao et al., 2022a)	12.2	48.9	41.3	112.3
3D U-Net (Xie et al., 2024)	12.7	53.0	44.3	124.5
PredRNNv2 (Wang et al., 2023)	20.6	126.0	173.0	268.2
FFNO (Tran et al., 2023)	12.0	46.7	48.1	112.0
IceMamba (Wang et al., 2025)	11.7	44.8	39.0	104.6
Field Memory without interface route	11.4	43.7	36.2	108.2
FORMA	**11.2**	**42.0**	**29.1**	**104.2**

FORMA ranks first on all four SIC metrics, reducing concentration RMSE by **4.3%** and SIE error by **25.4%**; the same architecture transfers without modification.

Cross-field interpretation. The SIC results test the same state formulation when the interface is defined by a concentration threshold rather than a temperature gradient. The gains in SIE and IIEE indicate that the improvement extends beyond concentration values to the location and total extent of the ice edge.

Table 4: SST mechanism ablations and input-reliability tests. Temperature errors are in 10^{-2} °C; all forecasts use a 90-day history and horizon. Input perturbations affect the history only.

(a) Mechanism ablation.

Variant	Area RMSE↓	Interface RMSE↓	D_{corr} ↓	Basin error↓
Static transition gates	157.2	180.0	11.4	8.3
No phase conditioning	153.9	177.9	10.8	21.1
No trust conditioning	120.7	141.2	6.6	16.9
No interface route	112.8	132.5	5.9	10.6
Matched RWKV recurrence	116.0	133.0	6.1	6.6
Full FORMA	**93.9**	**109.6**	**4.1**	**5.5**

(b) Input-reliability stress test.

	Full FORMA			No trust conditioning		
	Clean history	25% contiguous gap	25% temporal outliers	Clean history	25% contiguous gap	25% temporal outliers
Area RMSE	**93.9**	**307.5**	**396.8**	120.7	382.0	413.7
Interface RMSE	**109.6**	**311.7**	**473.8**	141.2	345.9	488.8

4.4 Ablations, Reliability, and Capacity

Feedback conditioning. With other components fixed, static gates, no phase, and no reliability yield area RMSEs of 1.5721, 1.5392, and 1.2069°C, versus **0.9390**°C for the full model (Table 4(a)).

Spatial memory. Removing the interface route and its associated conditioning gives 1.1281°C, while the matched RWKV control gives 1.1600°C. Appendix B specifies the variants.

Spatial diagnostics. In the median-improvement case, FORMA reduces Day-90 interface RMSE from 1.5518 to **1.1691**°C (**24.7%**); Figure 4 localizes this gain around observed fronts.

Model capacity. Scaling the model from $D = 48$ to 192 reduces area RMSE from 1.1120 to **0.8772**°C; at $D = 96$, rank 8 performs best among ranks 2, 8, and 16. Full results are in Appendix D.4.

Input reliability. Under identical contiguous gaps and temporal outliers in 25% of the history, reliability conditioning lowers both area and interface RMSE, with the larger margin under missing observations (Table 4(b)).

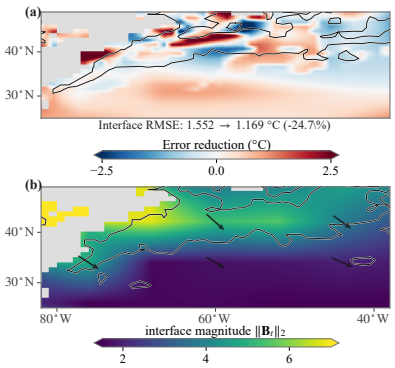


Figure 4: Median-improvement SST case. (a) Day-90 absolute-error reduction over matched RWKV; positive values favor FORMA. (b) Interface magnitude and local routing at the last observation. Contours show observed fronts; arrows show preferred source neighbors, not physical flow.

Forecast lead time and history length. FORMA has the lowest area RMSE at days 30, 60, and 90 (Appendix D); at day 90 it records **1.1163**°C, versus 1.4369°C for Earthformer and 1.6825°C for STDMamba. Using 90 rather than 14 history days reduces average area RMSE from 1.0108 to **0.9390**°C.

5 Conclusion

We introduced FORMA, a forecasting model built around feedback-conditioned spatial memory. Field Memory uses seasonal phase, input reliability, and interface features to control retention, writing, and spatial propagation, combining an RWKV-style recurrence with global and local routes. On 90-day SST forecasting, FORMA achieves the lowest errors across all compared model families, reducing area RMSE by 7.9% relative to the strongest competing model in our comparison and by 19.1% relative to the matched RWKV recurrence. The same architecture transfers to independently trained Arctic SIC forecasting and ranks first on every reported metric. Future work will extend the design to coupled fields and probabilistic forecasting.

AI USE STATEMENT

Generative AI tools were used to assist with literature search and synthesis, research ideation, formulation and refinement of the conceptual framework and mathematical presentation, methodology and experiment design, software implementation, reference formatting, and the creation and revision of scientific figures and other research artifacts. The authors manually reviewed the cited literature and all AI-assisted text, equations, code, and figures. The authors take responsibility for the final content of this work, including all claims and artifacts produced with AI assistance.

REFERENCES

T. R. Andersson, J. Scott Hosking, M. Pérez-Ortiz, et al. Seasonal arctic sea ice forecasting with probabilistic deep learning. *Nature Communications*, 12:5124, 2021. doi: 10.1038/s41467-021-25257-4.

V. Banzon, T. M. Smith, T. M. Chin, C. Liu, and W. Hankins. A long-term record of blended satellite and in situ sea-surface temperature data for climate research. *Earth System Science Data*, 8:165–176, 2016. doi: 10.5194/essd-8-165-2016.

Maximilian Beck, Korbinian Poppel, Markus Spanjer, et al. xLSTM: Extended long short-term memory. In *Advances in Neural Information Processing Systems*, 2024.

Kaifeng Bi, Lingxi Xie, Heng Zhang, Xin Chen, Xiaotao Gu, and Qi Tian. Accurate medium-range global weather forecasting with 3d neural networks. *Nature*, 619:533–538, 2023. doi: 10.1038/s41586-023-06185-3.

Johannes Brandstetter, Rob Hesselink, Elise van der Wijk, et al. Message passing neural PDE solvers. In *International Conference on Machine Learning*, 2022.

Mitchell Bushuk et al. Predicting september arctic sea ice: A multi-model seasonal skill comparison. *Bulletin of the American Meteorological Society*, 2024. doi: 10.1175/BAMS-D-23-0163.1.

Kang Chen et al. Fuxi: A cascade machine learning forecasting system for 15-day global weather forecast. *npj Climate and Atmospheric Science*, 6:190, 2023. doi: 10.1038/s41612-023-00512-1.

Tri Dao. Flashattention-2: Faster attention with better parallelism and work partitioning. In *International Conference on Learning Representations*, 2024.

Tri Dao and Albert Gu. Transformers are SSMs: Generalized models and efficient algorithms through structured state space duality. In *International Conference on Machine Learning*, 2024.

Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Re. Flashattention: Fast and memory-efficient exact attention with io-awareness. In *Advances in Neural Information Processing Systems*, 2022.

Yuchen Duan, Weiyun Wang, Zhe Chen, Xizhou Zhu, Lewei Lu, Tong Lu, Yu Qiao, Hongsheng Li, Jifeng Dai, and Wenhai Wang. Vision-rwkv: Efficient and scalable visual perception with rwkv-like architectures. In *International Conference on Learning Representations*, 2025.

Yuan Gao, Hao Wu, Fan Xu, Yanfei Xiang, Ruijian Gou, Ruiqi Shu, Qingsong Wen, Xian Wu, Kun Wang, and Xiaomeng Huang. NeuralOM: Neural ocean model for subseasonal-to-seasonal simulation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 14756–14764, 2026. doi: 10.1609/aaai.v40i17.38495.

Zhangyang Gao, Cheng Tan, Lirong Wu, and Stan Z. Li. SimVP: Simpler yet better video prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022a.

Zhihan Gao, Xingjian Shi, Hao Wang, Yi Zhu, Yuyang Wang, Mu Li, and Dit-Yan Yeung. Earthformer: Exploring space-time transformers for earth system forecasting. In *Advances in Neural Information Processing Systems*, volume 35, pp. 25390–25403, 2022b. doi: 10.52202/068431-1841. URL https://github.com/amazon-science/earth-forecasting-transformer.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *Transactions on Machine Learning Research*, 2024. arXiv preprint arXiv:2312.00752.

Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. In *International Conference on Learning Representations*, 2022.

Nicklas Hansen, Hao Su, and Xiaolong Wang. TD-MPC2: Scalable, robust world models for continuous control. In *International Conference on Learning Representations*, 2024.

Xudong Jiang, Shuyu Wang, Wengen Li, Hanchen Yang, Jihong Guan, Yichao Zhang, and Shuigeng Zhou. STDMamba: Spatiotemporal decomposition mamba for long-term fine-grained SST prediction. *IEEE Transactions on Geoscience and Remote Sensing*, 63:1–16, 2025. doi: 10.1109/TGRS.2025.3624051. URL https://github.com/XDjiang25/STDMamba.

Zheng Jiang, Wei Wang, Gaowei Zhang, and Yi Wang. SSTODE: Ocean–atmosphere physics-informed neural ODEs for sea surface temperature prediction. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(45):38661–38669, 2026. doi: 10.1609/aaai.v40i45.41209. URL https://ojs.aaai.org/index.php/AAAI/article/view/41209.

Dmitrii Kochkov, J"urgen Yuval, Ian Langmore, et al. Neural general circulation models for weather and climate. *Nature*, 635:965–973, 2024. doi: 10.1038/s41586-024-07744-y.

Nikola B. Kovachki, Zongyi Li, Kamyar Azizzadenesheli, et al. Neural operators: Learning maps between function spaces with applications to PDEs. *Journal of Machine Learning Research*, 24: 1–97, 2023.

Remi Lam, Alvaro Sanchez-Gonzalez, Matthew Willson, Peter Wirnsberger, Meire Fortunato, Ferran Alet, Suman Ravuri, Timo Ewalds, Zach Eaton-Rosen, Weihua Hu, Alexander Merose, Stephan Hoyer, George Holland, Oriol Vinyals, Jacklynn Stott, Alexander Pritzel, Shakir Mohamed, and Peter Battaglia. Learning skillful medium-range global weather forecasting. *Science*, 382(6677):1416–1421, 2023. doi: 10.1126/science.adi2336.

Kunchang Li, Xinhao Li, Yi Wang, et al. Videomamba: State space model for efficient video understanding. In *European Conference on Computer Vision*, 2024a.

Shufan Li, Harkanwar Singh, and Aditya Grover. Mamba-nd: Selective state space modeling for multi-dimensional data. In *European Conference on Computer Vision*, 2024b. doi: 10.1007/978-3-031-73414-4_5.

Zongyi Li, Nikola Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Fourier neural operator for parametric partial differential equations. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=c8P9NQVtmnO.

Zongyi Li, Daniel Huang, Burigede Liu, et al. Physics-informed neural operator for learning partial differential equations. In *International Conference on Learning Representations*, 2022.

Zongyi Li, Nikola B. Kovachki, Chris Choy, Boyi Li, Jean Kossaifi, Shourya Otta, Mohammad Amin Nabian, Maximilian Stadler, Christian Hundt, Kamyar Azizzadenesheli, and Animashree Anandkumar. Geometry-informed neural operator for large-scale 3d PDEs. In *Advances in Neural Information Processing Systems*, pp. 35836–35854, 2023. doi: 10.52202/075280-1556.

Phillip Lippe, Bastiaan S. Veeling, Paris Perdikaris, and Johannes Brandstetter. PDE-refiner: An iterative PDE-solver for forecasting the evolution of physical systems. In *Advances in Neural Information Processing Systems*, 2022.

Yue Liu, Yunjie Tian, Yuzhong Zhao, et al. VMamba: Visual state space model. In *Advances in Neural Information Processing Systems*, 2024.

Ze Liu, Yutong Lin, Yue Cao, et al. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, et al. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Bin Lu, Jingjing Shen, Ze Zhao, Jianping Zhou, Haonan Qi, Luyu Han, Xiaoying Gan, Meng Jin, Lei Zhou, Luoyi Fu, Xinbing Wang, and Chenghu Zhou. Oceanverse: Evaluable 4d ocean element reconstruction dataset under realistic sparsity. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 9474–9485, 2026. doi: 10.1145/3770855.3817534.

Lu Lu, Pengzhan Jin, George Pang, Zhiping Zhang, and George Em Karniadakis. Learning nonlinear operators via DeepONet based on the universal approximation theorem of operators. *Nature Machine Intelligence*, 3:218–229, 2021.

National Snow and Ice Data Center. Noaa/nsidc climate data record of passive microwave sea ice concentration, version 6, 2026. URL https://nsidc.org/data/g02202/versions/6.

Tung Nguyen, Johannes Brandstetter, Ashish Kapoor, Jayesh K. Gupta, and Aditya Grover. ClimaX: A foundation model for weather and climate. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 25904–25938. PMLR, 2023. URL https://proceedings.mlr.press/v202/nguyen23a.html.

Tung Nguyen, Rohan Shah, Hritik Bansal, Troy Arcomano, Romit Maulik, Veerabhadra Kotamarthi, Ian Foster, Sandeep Madireddy, and Aditya Grover. Scaling transformer neural networks for skillful and reliable medium-range weather forecasting. In *Advances in Neural Information Processing Systems*, volume 37, 2024. arXiv:2312.03876.

Antonio Orvieto, Samuel L. Smith, Albert Gu, et al. Resurrecting recurrent neural networks for long sequences. In *Advances in Neural Information Processing Systems*, 2023.

Bo Peng, Eric Alcaide, Quentin Anthony, Alon Albalak, Stella Biderman, Huanqi Cao, Xin Cheng, Michael Chung, Leon Derczynski, Xingjian Du, Matteo Grella, Kranthi Gv, Xuzheng He, Haowen Hou, Przemyslaw Kazienko, Jan Kocon, Jiaming Kong, Bartłomiej Koptyra, Hayden Lau, Jiaju Lin, Krishna Sri Ipsit Mantri, Ferdinand Mom, Atsushi Saito, Guangyu Song, Xiangru Tang, Johan Wind, Stanislaw Wozniak, Zhenyuan Zhang, Qinghua Zhou, Jian Zhu, and Rui-Jie Zhu. Rwkv: Reinventing rnns for the transformer era. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 14048–14077, 2023. doi: 10.18653/v1/2023.findings-emnlp.936.

Tobias Pfaff, Meire Fortunato, Alvaro Sanchez-Gonzalez, and Peter W. Battaglia. Learning mesh-based simulation with graph networks. In *International Conference on Learning Representations*, 2021.

Michael Poli, Stefano Massaroli, Eric Nguyen, et al. Hyena hierarchy: Towards larger convolutional language models. In *Advances in Neural Information Processing Systems*, 2023.

Ilan Price, Alvaro Sanchez-Gonzalez, Ferran Alet, Torsten R. Andersson, Zied Ben Bouallegue, Amina El-Kadi, Dylan Masters, Timo Ewalds, Stephan Hoyer, Remi Lam, Ian Langmore, Michael Levi, Daniele Maglietta, Ria Nuttall, Alexander Pritzel, Suman Ravuri, John Shaman, Jeroen Sprinkhuizen, Niamh Thain, Matthew Willson, et al. Probabilistic weather forecasting with machine learning. *Nature*, 637:84–90, 2025. doi: 10.1038/s41586-024-08252-9.

Stephan Rasp, Stephan Hoyer, Alexander Merose, Ian Langmore, Peter Battaglia, Tyler Russell, Alvaro Sanchez-Gonzalez, Vivian Yang, Rob Carver, Shreya Agrawal, Matthew Chantry, Zied Ben Bouallegue, Peter Dueben, Carla Bromberg, Jared Sisk, Luke Barrington, Aaron Bell, and Fei Sha. Weatherbench 2: A benchmark for the next generation of data-driven global weather models. *Journal of Advances in Modeling Earth Systems*, 16(6):e2023MS004019, 2024. doi: 10.1029/2023MS004019.

Richard W. Reynolds, Thomas M. Smith, Chun-yi Liu, Dudley B. Chelton, Kenneth S. Casey, and Michael G. Schlax. Daily high-resolution-blended analyses for sea surface temperature. *Journal of Climate*, 20(22):5473–5496, 2007. doi: 10.1175/2007JCLI1824.1.

Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-Chun Woo. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. In *Advances in Neural Information Processing Systems*, 2015.

Zakhar Shumaylov, Peter Zaika, James Rowbottom, Ferdia Sherry, Melanie Weber, and Carola-Bibiane Schönlieb. Lie algebra canonicalization: Equivariant neural operators under arbitrary lie groups. In *International Conference on Learning Representations*, 2025. Published as a conference paper at ICLR 2025; arXiv:2410.02698.

Ashutosh Singh, Ricardo Borsoi, Deniz Erdogmus, and Tales Imbiriba. Learning semilinear neural operators: A unified recursive framework for prediction and data assimilation. In *International Conference on Learning Representations*, 2024. Published as a conference paper at ICLR 2024; arXiv:2402.15656.

Makoto Takamoto, Tanuj Praditia, Raphael Leiteritz, et al. PDEBench: An extensive benchmark for scientific machine learning. In *Advances in Neural Information Processing Systems*, 2022.

Hao Tang, Hao Chen, and Chao Li. Generalized spherical neural operators: Green's function formulation. In *International Conference on Learning Representations*, 2026. ICLR 2026 Poster; arXiv:2512.10723.

Ilya O. Tolstikhin, Neil Houlsby, Alexander Kolesnikov, et al. MLP-Mixer: An all-mlp architecture for vision. In *Advances in Neural Information Processing Systems*, 2021.

Alasdair Tran, Alexander Mathews, Lexing Xie, and Cheng Soon Ong. Factorized fourier neural operators. In *International Conference on Learning Representations*, 2023. arXiv:2111.13802; code: https://github.com/alasdairtran/fourierflow.

Ashish Vaswani, Noam Shazeer, Niki Parmar, et al. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017.

Wei Wang, Weidong Yang, Lei Wang, Guihua Wang, and Ruibo Lei. Seasonal forecasting of pan-arctic sea ice with state space model. *npj Climate and Atmospheric Science*, 8:172, 2025. doi: 10.1038/s41612-025-01058-0.

Wei Wang, Weidong Yang, Xing Liu, Xing Xing, and Wenfang Cheng. Icecbr: Climate-shift-aware case adaptation for non-stationary arctic sea-ice forecasting. In *Case-Based Reasoning Research and Development*, Lecture Notes in Computer Science, 2026. doi: 10.1007/978-3-032-33865-5_27.

Yunbo Wang, Mingsheng Long, Jianmin Wang, Zhifeng Gao, and Philip S. Yu. PredRNN: Recurrent neural networks for predictive learning using spatiotemporal LSTMs. In *Advances in Neural Information Processing Systems*, 2017.

Yunbo Wang, Haixu Wu, Jianjin Zhang, Zhifeng Gao, Jianmin Wang, Philip S. Yu, and Mingsheng Long. PredRNN: A recurrent neural network for spatiotemporal predictive learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2):2208–2225, 2023. doi: 10.1109/TPAMI.2022.3165153.

Bowen Xie, Jifeng Qi, Shuguo Yang, Guimin Sun, Zhongkun Feng, Baoshu Yin, and Wenwu Wang. Sea surface temperature and marine heat wave predictions in the south china sea: A 3d u-net deep learning model integrating multi-source data. *Atmosphere*, 15(1):86, 2024. doi: 10.3390/atmos15010086.

Jingyi Xu, Siwei Tu, Weidong Yang, Ben Fei, Shuhao Li, Keyi Liu, Yeqi Luo, Lipeng Ma, and Lei Bai. Icediff: High resolution and high-quality arctic sea ice forecasting with generative diffusion prior. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10567–10576, 2025a. doi: 10.1109/CVPR52734.2025.00988.

Jingyi Xu, Shengnan Wang, Weidong Yang, Keyi Liu, Yeqi Luo, Ben Fei, and Lei Bai. Sifusion: A unified fusion framework for multi-granularity arctic sea ice forecasting. In *Advances in Neural Information Processing Systems*, volume 38, pp. 153855–153882, 2025b. doi: 10.52202/085713-4630.

Yuchen Zhang, Mingsheng Long, Kaiyuan Chen, Lanxiang Xing, Ronghua Jin, Michael I. Jordan, and Jianmin Wang. Skilful nowcasting of extreme precipitation with nowcastnet. *Nature*, 619 (7970):526–532, 2023. doi: 10.1038/s41586-023-06184-4.

A Scope and Limitations

FORMA uses a finite-width spatial memory and a deterministic multi-horizon decoder. Its current applications use independently trained SST and SIC models with a fixed 90-day forecast horizon. Causal observation-quality descriptors are derived from the available historical inputs, and their calibration depends on the observation product. Interface supervision depends on gradient bands for SST and concentration thresholds for SIC. Missing observations and ambiguous boundaries can affect these targets. Coupled ocean–atmosphere dynamics, uncertainty estimation, and autoregressive deployment remain directions for further study.

B Implementation Details

B.1 Default Hyperparameters

The reference configuration uses an 8×8 spatial patch, latent width $D = 96$, interface width $D_b = 24$, global rank $r = 8$, and decoder width 64. The forecast horizon is 90 days. The history-length study uses $L \in \{14, 90\}$, and lead-day results are taken at days 30, 60, and 90 from the same forecast output. SST is normalized using the training-set 0.5th and 99.5th percentiles; SIC remains in $[0, 1]$. We train with AdamW (learning rate 3×10^{-4}, weight decay 1×10^{-4}), mixed precision, and gradient clipping. Checkpoints are selected by validation area RMSE.

B.2 Evaluation Metrics

Let $t = 1, \ldots, T$ index all forecast maps across test origins and lead dates. Let V_t denote the valid cells at forecast date t, a_i their areas, and $e_{t,i} = \widehat{Y}_{t,i} - Y_{t,i}$ the field error. For SST, our area RMSE is

$$\mathrm{AreaRMSE} = \left[\frac{1}{T} \sum_{t=1}^{T} \frac{\sum_{i \in V_t} a_i e_{t,i}^2}{\sum_{i \in V_t} a_i} \right]^{1/2}.$$

The area weights are proportional to cosine latitude. We define the interface RMSE by replacing the area weights with the soft interface target $I_{t,i}$ (defined in Appendix B.3):

$$\mathrm{InterfaceRMSE} = \left[\frac{1}{T} \sum_{t=1}^{T} \frac{\sum_{i \in V_t} a_i I_{t,i} e_{t,i}^2}{\sum_{i \in V_t} a_i I_{t,i}} \right]^{1/2}.$$

Thus, Interface RMSE is an explicitly front-focused diagnostic introduced for this study; it assigns greater weight to errors near the target interface while retaining the same area measure. The SST basin error is

$$\mathrm{BasinErr} = \frac{1}{T} \sum_{t=1}^{T} \left| \frac{\sum_{i \in V_t} a_i \widehat{Y}_{t,i}}{\sum_{i \in V_t} a_i} - \frac{\sum_{i \in V_t} a_i Y_{t,i}}{\sum_{i \in V_t} a_i} \right|.$$

For SIC, concentration RMSE uses the same area-weighted formula as Area RMSE, with Y denoting concentration. Define the observed ice mask $M_t = \mathbb{1}[Y_t \geq 0.15]$ and predicted ice mask $\widehat{M}_t = \mathbb{1}[\widehat{Y}_t \geq 0.15]$. With cell areas a_i, SIE error is $\left| \sum_{i \in V_t} a_i (\widehat{M}_{t,i} - M_{t,i}) \right|$, and IIEE is $\sum_{i \in V_t} a_i |\widehat{M}_{t,i} - M_{t,i}|$. Both are evaluated on valid Arctic-ocean cells, averaged over forecast dates and test origins, and reported in million km^2. For ACC, let $\bar{\widehat{Y}}_t = \left(\sum_{i \in V_t} a_i \widehat{Y}_{t,i} \right) / \left(\sum_{i \in V_t} a_i \right)$ and $\bar{Y}_t = \left(\sum_{i \in V_t} a_i Y_{t,i} \right) / \left(\sum_{i \in V_t} a_i \right)$ be the spatial means over valid cells V_t. The area-weighted spatial-pattern correlation is

$$\mathrm{ACC}_t = \frac{\sum_{i \in V_t} a_i (\widehat{Y}_{t,i} - \bar{\widehat{Y}}_t)(Y_{t,i} - \bar{Y}_t)}{\sqrt{\sum_{i \in V_t} a_i (\widehat{Y}_{t,i} - \bar{\widehat{Y}}_t)^2} \sqrt{\sum_{i \in V_t} a_i (Y_{t,i} - \bar{Y}_t)^2}}, \qquad \mathrm{ACC} = \frac{1}{T} \sum_{t=1}^{T} \mathrm{ACC}_t.$$

Each predicted and observed map is centered by its own spatial mean. The same definition is used for SST and SIC, averaging the per-map correlations over all test origins and forecast dates. Tables and plots report the correlation deficit $D_{\mathrm{corr}} = 10^3(1 - \mathrm{ACC})$, where zero denotes perfect correlation and lower is better. This presentation preserves the underlying correlation definition and averaging, and is calculated from the unrounded ACC before display rounding.

B.3 Feedback Feature Construction

The phase code is

$$\mathbf{p}_t = \left[\sin\frac{2\pi d_t}{365.25}, \cos\frac{2\pi d_t}{365.25}, \sin\frac{4\pi d_t}{365.25}, \cos\frac{4\pi d_t}{365.25}, \phi_t^{\mathrm{tc}}\right]. \tag{13}$$

Here, d_t is the day index within the year and $\phi_t^{\mathrm{tc}} = \sin(2\pi(d_t - 90)/365.25)$ is the shifted annual feature. The pipeline also provides a scalar causal observation-quality descriptor derived from the valid coverage, finite-value status, temporal consistency, and product metadata available at each date. The descriptor is normalized to $[0, 1]$, broadcast over the latent grid, and recomputed from the modified histories in the perturbation experiments. No future target information is used in its construction.

For SST, the interface target is $I_t = \sigma((\|\nabla X_t\| - \tau_g)/\sigma_e)$, with τ_g set to the 85th percentile of training-set normalized gradients and $\sigma_e = 0.03$. For SIC, the target is $I_t = \exp[-(X_t - 0.15)^2/(2\sigma_e^2)]$ with $\sigma_e = 0.05$. Both are evaluated on valid task cells.

B.4 Complete Training Objective

For validity mask V, let $\langle A\rangle_V = \sum VA/\max(1, \sum V)$. The five terms in Equation 12 are

$$\mathcal{L}_{\mathrm{field}} = \langle \mathrm{Huber}_{0.05}(\widehat{Y} - Y)\rangle_V, \tag{14}$$

$$\mathcal{L}_{\mathrm{interface}} = \langle \mathrm{BCEWithLogits}(\widehat{\ell}^f, I(Y))\rangle_V, \tag{15}$$

$$\mathcal{L}_{\text{hist-interface}} = \langle \mathrm{BCEWithLogits}(\widehat{\ell}^h, \mathcal{D}(I(X_{1:L})))\rangle_{V^h}, \tag{16}$$

$$\mathcal{L}_{\mathrm{region}} = K^{-1}\sum_k |\bar{\widehat{Y}}_k - \bar{Y}_k|, \tag{17}$$

$$\mathcal{L}_{\mathrm{temp}} = \langle \mathrm{SL1}(\Delta\widehat{Y} - \Delta Y)\rangle_{V^\Delta}. \tag{18}$$

Regional means use cosine-latitude weights; $V^\Delta_{kij} = V_{k+1,ij}V_{k,ij}$. We set $(\beta_e, \beta_h, \beta_a, \beta_t) = (0.20, 0.10, 0.10, 0.02)$.

B.5 Efficient Execution

The implementation batches input-dependent projections before the temporal scan. The optional Triton backend provides kernels for four-neighbor aggregation and elementwise retention/write updates. Neighbor aggregation uses periodic longitude and non-periodic latitude boundaries. Pointwise operations can also use `torch.compile`. A vectorized PyTorch backend supports the same recurrence. The native-resolution configuration can decode on the latent grid before resizing the predicted fields, reducing the cost of the full-resolution decoder.

B.6 Input-only robustness perturbations

The contiguous-gap test replaces one randomly positioned block covering 25% of the historical dates with the normalized missing value and clears its validity channel. The temporal-outlier test adds a fixed normalized offset to 25% of the dates and clips the result to the valid range. The compared models receive identical perturbed histories, while targets and model parameters remain fixed during evaluation.

B.7 Memory Diagnostics

A temporal persistence summary can be computed from the mean retention gate $\bar{r}_t$ and mean write gate $\bar{w}_t$. Averaging over spatial locations and channels gives the gate-based influence proxy

$$I_{T,s} = \bar{w}_s \prod_{j=s+1}^{T} \bar{r}_j \tag{19}$$

for a historical write at date s and a later memory state at date T. The relative contribution of the spatial routes can also be summarized by the global branch's mean absolute output divided by the sum of the global and local mean absolute outputs.

B.8 PSEUDOCODE

The forward and training procedure for a history $\{\mathbf{X}_t\}_{t=1}^{L}$ is:

1. Initialize the spatial memory $\mathbf{H}_0$ and temporal accumulators $\mathbf{S}_0, \mathbf{N}_0$ to zero.
2. Encode the historical fields into $\mathbf{U}_t$, construct phase and reliability features, and compute $\mathbf{B}_t$ from consecutive encoded fields.
3. Apply the interface-conditioned neighbor shift and compute the temporal projections and memory-update factors.
4. For each observed date, update $\mathbf{S}_t, \mathbf{N}_t$, form temporal evidence $\mathbf{E}_t$, and update $\mathbf{H}_t$ using Equation 7.
5. Add the projected latest observation to $\mathbf{H}_L$ and decode the K future fields and interface logits.
6. During training, optimize Equation 12 using the target fields and interface supervision.

B.9 ABLATION CONFIGURATIONS

The no-phase and no-trust variants remove the corresponding signal from all conditioning heads. The no-interface variant disables the local memory route, the input neighbor shift, and interface conditioning; its training configuration also disables interface supervision. The static-gate variant uses constant retention, write, and neighbor weights, together with global mean aggregation. The matched RWKV recurrence retains temporal aggregation and the latest-observation path while disabling phase, reliability, interface conditioning, and spatial propagation.

C STATE PROPAGATION

C.1 EXPANSION OF THE MEMORY UPDATE

For a fixed sequence of input-derived transition factors, define the ordered propagation operator

$$\mathbf{\Phi}_{T\leftarrow s} = \mathbf{F}_T \mathbf{F}_{T-1} \cdots \mathbf{F}_{s+1}, \qquad \mathbf{\Phi}_{s\leftarrow s} = \mathbf{I}. \tag{20}$$

Expanding Equation 7 over the observed history gives

$$\mathbf{H}_T = \mathbf{\Phi}_{T\leftarrow 0}\mathbf{H}_0 + \sum_{s=1}^{T} \mathbf{\Phi}_{T\leftarrow s}(\mathbf{\Gamma}_s \odot \mathbf{E}_s). \tag{21}$$

Each term is a historical write followed by the retention and spatial propagation operations at later dates. The product preserves the temporal order of these operations.

Let $\mathbf{G}_s$ be the diagonal operator corresponding to the elementwise write gate $\mathbf{\Gamma}_s$. The block

$$\mathbf{K}^{\text{state}}_{(T,n),(s,m)} = [\mathbf{\Phi}_{T\leftarrow s}\mathbf{G}_s]_{n,m} \tag{22}$$

maps evidence at source location m and date s to memory location n at the end of the scan. It is conditional on the realized feedback sequence. The recurrence computes this propagation through the factorized updates in Equation 7.

C.2 PROPAGATION BOUND

Consider two memory states with identical feedback factors and evidence writes. Their difference evolves as

$$\Delta\mathbf{H}_t = \mathbf{F}_t \Delta\mathbf{H}_{t-1}. \tag{23}$$

The normalized global route has row sum one, and the local route has row sum at most one, including boundary locations. If $D_{t,n,d} \le \bar{d}_t < 1$ and $0 \le \alpha < 1$, then

$$\|\mathbf{F}_t\|_\infty \le (1-\alpha)\bar{d}_t + \alpha < 1. \tag{24}$$

Applying the bound over successive dates yields

$$\|\Delta \mathbf{H}_T\|_\infty \le \prod_{j=s+1}^{T} \big((1-\alpha)\bar{d}_j + \alpha\big) \|\Delta \mathbf{H}_s\|_\infty. \tag{25}$$

This bounds the propagation of a memory perturbation with the input-derived transition factors and evidence writes held fixed.

D Additional Experimental Details

D.1 Controlled field diagnostic

Figure 1 uses a small, fully controlled field generator to expose the interaction between temporal evolution, interface structure, and observation noise. On a normalized 64×64 periodic-longitude grid, each clean field is a sheared translating front,

$$\begin{aligned} f_t(x,y) &= \frac{1}{2} + a \tanh\left(\frac{\sin\big(2\pi[x-u_t(y)]\big) + \kappa \sin\big(4\pi[x-u_t(y)] + \phi_2\big) + \eta(y)}{w}\right), \\ u_t(y) &= v_0[0.8 + 0.2\cos(2\pi y)]\left[t + \frac{\beta}{\omega}\{\sin(\phi+\omega t) - \sin\phi\}\right]. \end{aligned} \tag{26}$$

where w controls the transition width, v_0, β, and $\omega = 2\pi/96$ control the displacement, and the remaining phase and shape terms add stationary spatial heterogeneity. The slow and fast columns use the two prescribed temporal phases of the generator, while smooth and sharp use $w = 0.55$ and $w = 0.18$, respectively; thus "fast" denotes the regime with the larger apparent front displacement over the forecast window, not a change in grid resolution.

For each history, independent Gaussian perturbations are added to the observed frames. The plotted noise-to-signal level ϵ scales the perturbation standard deviation by the training-signal standard deviation, so the horizontal axis is dimensionless and comparable across regimes. The generated dataset contains 4,096 training sequences and 512 validation sequences, with a 64-sequence validation sweep per regime; the pilot runs use a fixed 2,048-sequence training subset, 16 observed frames, and an 8-frame forecast horizon. ConvLSTM, S4FFNO, and FORMA receive identical corrupted histories and forecast horizons; their diagnostic configurations use width 64 for ConvLSTM and S4FFNO (four Fourier layers for S4FFNO) and $(D, D_b, r) = (96, 24, 8)$ for FORMA. We report the fixed-protocol validation sweeps as descriptive evidence for the field-memory design, not as an additional leaderboard claim.

D.2 Published-Baseline Adaptation Protocol

Table B.1 summarizes the public baseline implementations and the task-level adaptations used for the comparison. The adaptations map the field and validity inputs to each backbone and return forecasts on the common target grid. The convolutional, spectral, Transformer, and state-space models use task-specific multi-horizon heads. PredRNNv2 retains its recurrent decoder and training-time reverse scheduled sampling. The complete source code, including FORMA and the baseline adaptations, will be released upon publication.

Table B.1. Public baseline implementations and task-level adaptations. The SST comparison uses the first eight models. The SIC comparison evaluates 3D U-Net, SimVP, PredRNNv2, FFNO, and IceMamba.

Method	Official repository	Task-level adaptation
3D U-Net	task adapter for the cited SST implementation	$3\times4\times4$ average-pooled time–latitude–longitude volume; genuine Conv3d encoder–decoder; direct 90-field output restored to the common grid
SimVP	OpenSTL/SimVP	target/validity input embedding; direct multi-horizon decoder
PredRNNv2	official PredRNNv2 implementation	target/validity input embedding; recurrent 90-field decoder with reverse scheduled sampling
FFNO	alasdairtran/fourierflow	factorized spectral layers; direct 90-field decoder
Earthformer	amazon-science/earth-forecasting-transformer	Cuboid Transformer backbone; field/validity input adapter; direct 90-field decoder
STDMamba	XDjiang25/STDMamba	spatiotemporal decomposition Mamba backbone; field/validity input adapter; direct 90-field decoder
NeuralOM	YuanGao-YG/NeuralOM	single-field OISST input; direct 90-field decoder; no external forcing
SSTODE	nicezheng/SSTODE-code	latent velocity and advection–diffusion/energy-exchange blocks; OISST-only input; direct 90-field decoder; no ERA5 forcing
IceMamba	official IceMamba implementation (Wang et al., 2025)	SIC/validity input embedding; direct 90-field decoder; sea-ice mask and 15% edge-compatible output head

Each baseline is trained from scratch on the chronological split used for FORMA. The models receive the same field and validity history and are evaluated at the same forecast dates. Table B.2 organizes the parameter and computational measurements.

Table B.2. Computational comparison on the SST task. Measurements use A100 inference with inference batch size 16, FP16 mixed precision, and the PyTorch backend; training configurations use 20 epochs and batch size 16. FLOPs count convolution and matrix-multiplication arithmetic per forecast sample, with one multiply–accumulate counted as two operations; FFT and elementwise operations are excluded.

Method	Params (M)	FLOPs (G)	Peak memory (GiB)	Samples/s
3D U-Net (Xie et al., 2024)	0.025	1.315	0.143	498.99
SimVP (Gao et al., 2022a)	6.473	74.673	1.052	95.77
PredRNNv2 (Wang et al., 2023)	0.522	193.600	0.535	5.81
FFNO (Tran et al., 2023)	0.084	1.253	0.100	650.24
Earthformer (Gao et al., 2022b)	4.095	1161.063	5.313	12.37
STDMamba (Jiang et al., 2025)	0.719	84.889	0.445	151.70
NeuralOM (Gao et al., 2026)	0.057	11.714	0.114	53.88
FORMA	0.184	26.785	0.527	23.83

D.3 Forecast Lead Time and History Length

Table B.3. Lead-day comparison from the same direct 90-day checkpoints. Temperature-error columns are displayed in 10^{-2} °C and the correlation-deficit columns in 10^{-3} units; units are stated in the caption rather than repeated in the headers.

	Lead day 30		Lead day 60		Lead day 90	
Method	Area RMSE↓	D_{corr} ↓	Area RMSE↓	D_{corr} ↓	Area RMSE↓	D_{corr} ↓
SimVP (Gao et al., 2022a)	107.3	4.3	119.3	5.7	135.9	7.9
Earthformer (Gao et al., 2022b)	87.6	3.6	114.2	5.9	143.7	9.5
STDMamba (Jiang et al., 2025)	92.0	3.8	133.9	8.0	168.3	12.3
Matched RWKV recurrence	96.8	4.3	127.4	7.4	157.0	11.3
FORMA	**87.5**	**3.6**	**99.1**	**4.6**	**111.6**	**5.8**

Table B.4. History-window sensitivity at a fixed 90-day forecast horizon. Each row is a separately trained model whose input window matches the history shown in the column group: the 14-day-history model uses 14 preceding daily fields, and the 90-day-history model uses 90 preceding daily fields; both predict the same 90 future days. Temperature-error columns are displayed in 10^{-2} °C and correlation-deficit columns in 10^{-3} units; units are stated in the caption rather than repeated in the headers. For the 14-day FORMA configuration, overall RMSE is 1.0014°C and area RMSE is 1.0108°C; MAE is 0.6752°C, ACC is 0.99526 (equivalently $D_{\mathrm{corr}} = 4.745$), and day-30/60/90 area RMSE is $0.9338/1.0699/1.2184$°C.

	14-day history		90-day history	
Method	Area RMSE↓	D_{corr} ↓	Area RMSE↓	D_{corr} ↓
Matched RWKV recurrence	144.2	9.5	116.0	6.1
FORMA	**101.08**	**4.745**	**93.9**	**4.1**

Table B.5. Efficiency measurements for the history-window study using the inference protocol of Table B.2. Area RMSE is displayed in 10^{-2} °C and averaged over the 90-day forecast.

Method	Context	Params	Peak mem. (GiB)	Samples/s	Area RMSE↓
Matched RWKV recurrence	14	183998	0.086	298.87	144.2
FORMA	14	183998	0.093	166.58	101.08
Matched RWKV recurrence	90	183998	0.238	61.37	116.0
FORMA	90	183998	0.527	23.83	93.9

D.4 Capacity-scaling configurations

All capacity configurations use the 1° SST data, a 90-day history, and a 90-day forecast horizon. Models are trained for 20 epochs with batch size 16, mixed precision, and the Triton backend. Checkpoints are selected by validation area RMSE. The joint-capacity study changes D, D_b, r, and decoder width together. The global-rank study fixes $D = 96$, $D_b = 24$, and decoder width 64 while varying $r \in \{2, 8, 16\}$.

Table B.6. Model configurations and SST test results. Error columns are displayed in 10^{-2} units. The base and rank-8 entries use the Main reference checkpoint. Parameter counts are computed from the model configurations.

Sweep	Variant	D	D_b	r	Decoder	Params	Area RMSE↓	D_{corr} ↓
Joint capacity	compact	48	12	4	32	53706	111.2	5.8
Joint capacity	base	96	24	8	64	183998	93.9	4.1
Joint capacity	expanded	192	48	16	128	674982	87.7	3.6
Global rank	rank2	96	24	2	64	182756	99.2	4.6
Global rank	rank8	96	24	8	64	183998	93.9	4.1
Global rank	rank16	96	24	16	64	185654	98.4	4.5

Increasing rank from 2 to 8 reduces area RMSE from 0.9918 to 0.9390°C and reduces the correlation deficit from 4.6 to 4.1. Rank 16 gives 0.9844°C and a deficit of 4.5. The larger joint-capacity configuration improves both metrics, while increasing global rank alone beyond 8 gives similar results in this study.

D.5 Additional 90-day Metrics

Table B.7. Additional metrics for the 90-day forecast at fixed 90-day history. Error columns are displayed in 10^{-2} units.

Method	MAE↓	D_{corr} ↓	Interface RMSE↓	Basin error↓
3D U-Net	69.3	4.5	118.7	21.6
SimVP	83.6	5.3	131.9	25.6
PredRNNv2	91.1	7.2	146.9	42.5
FFNO	70.5	4.8	122.2	21.1
Earthformer	63.9	4.9	122.2	6.1
STDMamba	69.2	6.1	138.6	9.6
NeuralOM	85.4	9.2	164.4	14.1
SSTODE	85.4	9.9	168.7	7.2
FORMA	**62.2**	**4.1**	**109.6**	**5.5**