ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-6874-1/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICASSP49660.2025.10888694

Climate Downscaling Using Neural Operator: Spatiotemporal Multimodal Fusion Operator with State-Query Coupled Kernel

Haodi Zhang[1], Yichi Wang[1], Yifan Jian[1], Jiahui Jiang[1], Zhaohai Bai[2], Lin Ma[3,*]
[1]*College of Computer Science and Software Engineering, Shenzhen University*
[2]*Center for Agricultural Resources Research, Institute of Genetic and Developmental Biology, Chinese Academy of Sciences*
[3]*State Key Laboratory of Pollution Control and Resource Reuse, School of the Environment, Nanjing University*

***Abstract*—Climate downscaling is crucial for detailed small-scale analysis and for acquiring climate data in regions without weather stations. Operator learning has proven potential for this task. However, several challenges remain in operator learning, such as multimodal fusion, spatiotemporal fusion and input state and query adaptation. To address these challenges, we propose a Spatiotemporal Multimodal Fusion Operator with a State-Query Coupled Kernel (SMCK). This framework includes a latent space fusion encoder that encodes climate variables using position-wise multihead attention for multimodal fusion and integrates historical information to generate robust and precise representation. Additionally, we introduce a state-query coupled kernel that combines radial basis functions and discrete fourier encoding to enhance query location representation, while also adapting to the state to obtain the coupled kernel. Extensive experiments demonstrate that our method achieves state-of-the-art performance and provides strong support for climate downscaling and the planning of climate-related strategies.**

***Index Terms*—Multimodal fusion, Climate downscaling, Deep Learning , Operator learning, Time series forecasting**

I. Introduction

Climate change poses an increasingly severe threat to humanity, causing substantial loss of life and property worldwide through unprecedented extreme events such as wildfires, droughts, floods, and heatwaves [4]. To better understand and prepare for these events, accurate climate information at regional and local scales is crucial. While Global Climate Models (GCMs) such as ClimODE [29], FengWu [9], and Pangu [5] are valuable for large-scale climate simulations, they have limitations at smaller scales due to their coarse spatial resolutions. Additionally, it is challenging to obtain climate information in regions where establishing weather stations is difficult. To address these issues, the concept of climate downscaling has drawn significant attention [20], [24], [30], as it aims to obtain higher-resolution information. However, existing downscaling methods operate at fixed scales, which will lead to error propagation when trying to obtain information at arbitrary scales. In contrast, operator learning has been theoretically proven to be able to arbitrarily approximate any nonlinear operator [2], [10]. Meanwhile, they have been successfully applied in partial differential equation tasks such as fluid mechanics and aerodynamics [14], [22]. However, there is still no application in the climate downscaling task that complies with climate dynamics [29]. We make the first attempt to apply it to this task, which is of great significance in the field of climate downscaling.

Previous work such as DeepONet [22], utilizes a branch-trunk network architecture to separately handle input function and query location. LOCA [15] builds upon this by leveraging the Wavelet Scattering Network [25] and introducing coupled attention mechanisms to enhance performance. Despite recent advancements, operator learning for climate downscaling remains highly challenging, yielding unsatisfactory results. Key obstacles include *multimodal fusion*, *spatiotemporal fusion* and *input state and query adaptation*. First, climate observations typically involve multiple variables or modalities, such as temperature, humidity, and wind speed, which vary across different hectopascals. This complexity necessitates capturing diverse modalities to fully leverage the data for accurate predictions. Second, incomplete observations hinder the capture of all possible variables, rendering the observation process approximate markovian and highlighting the need to leverage time series information. Third, like attention mechanism, varying input states affect contributions to query locations. However, methods such as DeepONet and LOCA, which treat input functions and query locations separately, lack robust mechanisms for adaptation, underscoring the need to couple query locations with the input state for more effective learning.

Recent approaches attempt to leverage the powerful capabilities of transformers [28] to tackle these challenges. For instance, OFormer [17] introduces an encoder-decoder structure with Galerkin-type attention [8] and incorporates coordinate information. However, it does not deal with multimodal fusion nor align different modalities. GNOT [12] uses a simple multimodal fusion method by encoding with MLP [21] and employs a mixture-of-experts approach to independently generate weights for query locations, but it does not address spatiotemporal information. Moreover, their incorporation of coordinates is relatively simplistic. For example, DeepONet and GNOT merely encode the query location without exploiting the relationship between the query location and the input function resolution. To our knowledge, no existing work simultaneously tackles all these challenges, thus limiting the practical application of neural operators.

[*] Corresponding author email: malin1979@nju.edu.cn

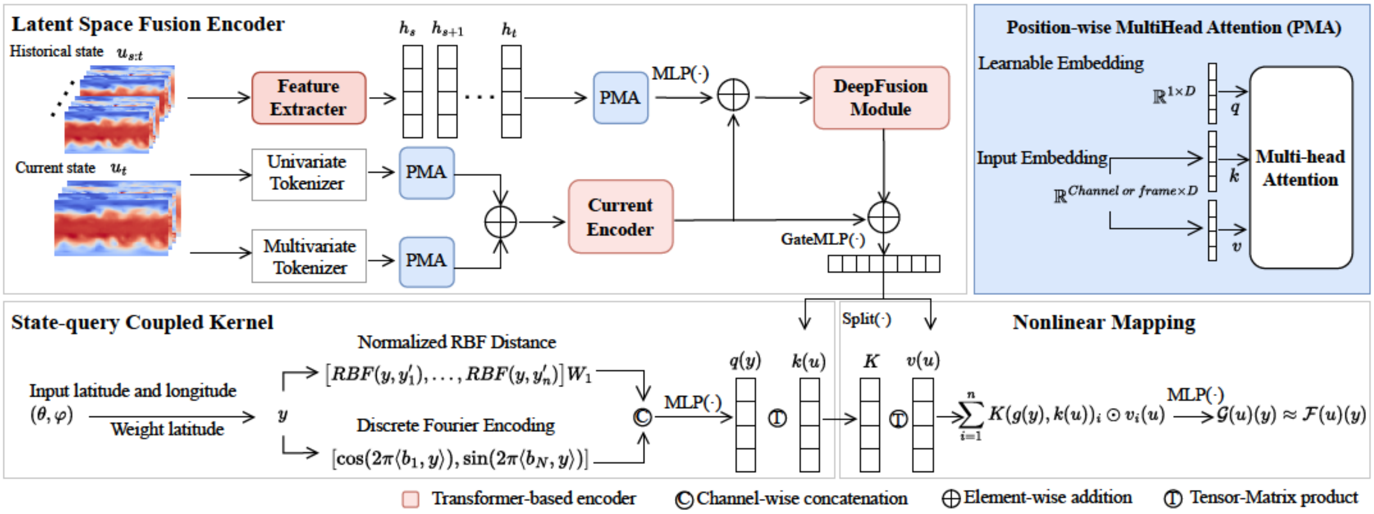
Latent Space Fusion Encoder
Historical state $u_{s:t}$
Feature Extracter
h_s
h_{s+1}
h_t
PMA
MLP(·)
DeepFusion Module
Current state u_t
Univariate Tokenizer
PMA
Multivariate Tokenizer
PMA
Current Encoder
GateMLP(·)
Position-wise MultiHead Attention (PMA)
Learnable Embedding
$\mathbb{R}^{1\times D}$
q
Input Embedding
$\mathbb{R}^{Channel\ or\ frame\times D}$
k
v
Multi-head Attention
State-query Coupled Kernel
Input latitude and longitude
(θ,φ)
Weight latitude
y
Normalized RBF Distance
$[RBF(y,y_1'),\dots,RBF(y,y_n')]W_1$
Discrete Fourier Encoding
$[\cos(2\pi\langle b_1,y\rangle),\sin(2\pi\langle b_N,y\rangle)]$
MLP(·)
$q(y)$
$k(u)$
Split(·)
Nonlinear Mapping
K
$v(u)$
$\sum_{i=1}^{n} K(g(y),k(u))_i \odot v_i(u)$
MLP(·)
$\mathcal{G}(u)(y)\approx\mathcal{F}(u)(y)$
Transformer-based encoder
Channel-wise concatenation
Element-wise addition
Tensor-Matrix product

Fig. 1. Our framework is composed of a LSF encoder, a state-query coupling kernel, and a nonlinear mapping. In the LSF encoder, the lower part is the current state encoder and the upper part is the historical state fusion. The outputs of query location encoding and GateMLP are used to construct the state-query coupled kernel. Then it is processed through the nonlinear mapping to predict the state of the query location.

To tackle these challenges, we propose a spatiotemporal multimodal fusion operator with state-query coupled kernel (SMCK), which consists of three main components. First, we propose a Latent Space Fusion Encoder (LSF) that encodes climate variables at the current time by categorizing them into distinct sets and applying position-wise multihead attention (PMA) across modalities for various geographic locations. Additionally, we utilize a feature extractor to extract historical information, and then fuse it using the PMA, enabling us to effectively handle spatiotemporal multimodal information. Second, inspired by multihead attention (MHA) mechanisms [3], we introduce a state-query coupled kernel that uses an encoding function based on radial basis functions [23] and discrete fourier encoding [18] for query location, enhancing its representation. This coupled kernel allows the query location to adapt to the current state rather than being constrained solely to the encoding of the query location. Finally, we employ a nonlinear mapping to predict the state at the query location. Extensive experiments demonstrate that our method achieves state-of-the-art performance.

II. Preliminaries

Kernel integral operations are typically implemented by integrating input values weighted by kernel values, representing pairwise interactions between elements of the input function domain and the output function domain. These operations align well with the process in neural operators, where global transformations capture interactions between input points. Recently, it has been shown that the kernel integral operation can be approximated by the attention mechanism [8], [16]. In its continuous limit, with suitable mesh weights, it becomes a Riemann sum approximation of the kernel integral transform:

$$z_i = \sum_{j=1}^{N_v} h(q_i, k_j) v_j \mu_j \approx \int_{\mathcal{A}} \kappa(\xi, x_i) v(\xi)\, d\xi, \qquad (1)$$

where $q_i, k_j, v_j \in \mathbb{R}^d$, $h : \mathbb{R}^d \times \mathbb{R}^d \mapsto \mathbb{R}$ is a weight function, μ_j represents the mesh-based measure for grid point x_j, and parameterizes the kernel function κ, $\mathcal{A}$ is continuous space.

III. Method

A. Problem Formulation

Given N pairs of climate input-output data as functions [26] $\{u^{\ell}_{s:t}, s^{\ell}(y)\}_{\ell=1}^{N}$, generated by a unknown ground truth operator $\mathcal{F} : \mathcal{C}(\mathbb{R}^2, \mathbb{R}^{d_u}) \rightarrow \mathcal{C}(\mathbb{R}^2, \mathbb{R}^{d_s})$, where $\mathcal{C}(\mathcal{X}, \mathcal{Y})$ denotes the set of continuous functions from $\mathcal{X}$ to $\mathcal{Y}$, d_u and d_s are numbers of climate variables. $u^{\ell} \in \mathcal{C}(\mathbb{R}^2, \mathbb{R}^{d_u})$, the $u^{\ell}_{s:t}$ is a time series $\{u^{\ell}_s, u^{\ell}_{s+1}, \dots, u^{\ell}_t\}$, $s^{\ell} \in \mathcal{C}(\mathbb{R}^2, \mathbb{R}^{d_s})$, and $y \in \mathbb{R}^2$ is a query location. Ω is the query set. Our goal is to learn an operator $\mathcal{G}$ that approximates the operator $\mathcal{F}$.

$$\begin{aligned}
\mathcal{L}_{\text{L2}} &= \frac{1}{N}\sum_{\ell=1}^{N}\frac{1}{|\Omega|}\sum_{y\in\Omega} L(y)\left((\mathcal{G}(u^{\ell}_{s:t})(y) - s^{\ell}(y)\right)^2 \\
\mathcal{L}_{\text{phy}} &= \frac{1}{N}\sum_{\ell=1}^{N}\frac{1}{|\Omega|}\sum_{y\in\Omega} L(y)\left(\nabla\mathcal{G}(u^{\ell}_{s:t})(y) - \nabla s^{\ell}(y)\right)^2
\end{aligned} \qquad (2)$$

We also incorporate a physics-informed loss $\mathcal{L}_{\text{phy}}$. The total loss $\mathcal{L} = \mathcal{L}_{\text{L2}} + \lambda_1\mathcal{L}_{\text{phy}}$. ∇ represents the gradient operator, λ_1 is a hyperparameter. $L(y)$ is the weighting factor [27].

B. Framework Overview

As shown in Fig. 1, SMCK consists of three main components: LSF Encoder, State-query Coupled Kernel, and a nonlinear mapping function. The LSF Encoder processes input climate function $u_{s:t}$ using two submodules: the *Current State Encoder* f_{CSE} and *Historical State Fusion* f_{HSR}. The f_{CSE} synthesizes multimodal climate functions into a unified representation. f_{HSR} integrates historical states $u_{s:t}$ encoded by *Feature Extractor* f_{FE} individually and integrates in latent space, and then refining the f_{CSE} output through the

TABLE I

EVALUATION OF MODELS FOR MSE, RMSE, ACC. BEST PERFORMANCES ARE HIGHLIGHTED IN RED AND UNDERLINED; SECOND-BEST IN BLUE.

Metric	Variable	LOCA	Oformer	GNOT	DeepONet	WaveONet	SMCK
MSE(↓)	Zonal wind speed at 10 meters (U10)	0.1371	0.1387	0.1527	0.1524	0.1518	**0.1361**
	Meridional wind speed at 10 meters (V10)	**0.1720**	0.1759	0.1926	0.1912	0.1901	0.1723
	Temperature at 2 meters (T2m)	0.0056	0.0050	0.0109	0.0108	0.0107	**0.0042**
	Temperature at 850 hPa (T850)	0.0069	0.0068	0.0074	0.0073	0.0072	**0.0064**
	Geopotential at 500 hPa (Z500)	0.0009	0.0010	0.0010	0.0010	0.0009	**0.0009**
	Total Variable Average	0.0645	0.0655	0.0729	0.0725	0.0722	**0.06399**
RMSE(↓)	Zonal wind speed at 10 meters (U10)	2.0424	2.0544	2.1556	2.1535	2.1494	**2.0348**
	Meridional wind speed at 10 meters (V10)	**1.9698**	1.9920	2.0850	2.0770	2.0714	1.9714
	Temperature at 2 meters (T2m)	1.5910	1.5033	2.2181	2.2087	2.2015	**1.3832**
	Temperature at 850 hPa (T850)	1.2993	1.2892	1.3461	1.3401	1.3342	**1.2578**
	Geopotential at 500 hPa (Z500)	107.1335	107.4628	107.4355	107.2406	106.4886	**104.8935**
	Total Variable Average	22.8072	22.8603	23.0481	23.0040	22.8490	**22.3081**
ACC(↑)	Zonal wind speed at 10 meters (U10)	0.8755	0.8740	0.8620	0.8622	0.8627	**0.8761**
	Meridional wind speed at 10 meters (V10)	0.8843	0.8817	0.8711	0.8720	0.8726	**0.8840**
	Temperature at 2 meters (T2m)	0.9388	0.9453	0.8888	0.8899	0.8903	**0.9533**
	Temperature at 850 hPa (T850)	0.9567	0.9574	0.9538	0.9542	0.9546	**0.9593**
	Geopotential at 500 hPa (Z500)	0.9939	0.9939	0.9939	0.9939	0.9940	**0.9941**
	Total Variable Average	0.9298	0.9304	0.9139	0.9145	0.9148	**0.9334**

DeepFusion Module f_{DFM}. The state-query coupled kernel utilizes normalized RBF distance and discrete fourier encoding to encode query location y, followed by a parameterized kernel to fuse these with the state. The final climate state for a query location is obtained through a nonlinear projection. Our model is summarized below, excluding residual way:

$$\mathcal{G}(u_{s:t})(y) = f\left(\sum_{i=1}^{n} K(q(y), k(u_{s:t}))_i \odot v_i(u_{s:t})\right) \tag{3}$$

Where f is the nonlinear mapping, K is the state-query coupled kernel, q is the location encoding function, and k and v are state encoding functions follow the LSF Encoder.

C. LSF Encoder

1) Position-wise Multi-Head Attention: We introduce a learnable embedding $q \in \mathbb{R}^D$ as the query. The key k and value v are the same input embedding x whose dimension is $\mathbb{R}^{L\times(Channel\ or\ Frame)\times D}$, L is the location patch series.

$$\text{PMA}(x) := \text{MHA}(q, x, x) \tag{4}$$

2) Current State Encoder: We first categorize the climate variables into two sets: the univariate set US and the multivariate set MS. In US, each element is in $\mathbb{R}^{1\times H\times W}$, and in MS, each element is in $\mathbb{R}^{C'\times H\times W}$, where C' represents the number of different observation heights (800 hPa and 900 hPa), H, W are resolution locations. Then, we use tokenizers [11] to tokenize every element in US and MS, and concatenate them individually. $[\cdot]$ denotes vector concatenation.

$$v_i = \text{Tokenizer}(s_i, Channel = 1), \quad s_i \in US \tag{5}$$

$$w_j = \text{Tokenizer}(m_j, Channel = C'), \quad m_j \in MS \tag{6}$$

$$a_{us} = [v_0, \ldots, v_n], \quad a_{ms} = [w_0, \ldots, w_m] \tag{7}$$

$$z_{CSE} = f_{CE}(\text{PMA}(a_{us}) + \lambda_2 \text{PMA}(a_{ms})) \tag{8}$$

Secondly, we employ PMA in a_{us} and a_{ms}. The λ_2 is a hyperparameter. Then we use the *current encoder* f_{CE} (which is transformer-based) to get the final current representation.

3) Historical State Fusion: We apply self-supervised learning to f_{FE} using JEPA [1] and keep it frozen during downscaling task. The f_{FE} encodes $u_{s:t}$ individually to generate features $h_{s:t}$, which are also integrated by PMA. f_{FE}, f_{DFM} are also transformer-based:

$$h = [f_{FE}(u_s), f_{FE}(u_i), \ldots, f_{FE}(u_t)], \quad u_i \in u_{s:t} \tag{9}$$

$$z_{LSA} = f_{MLP}(\text{PMA}(h)) \tag{10}$$

$$z = f_{DFM}(z_{CSE} + z_{LSA}) + z_{CSE} \tag{11}$$

D. State-query Coupled Kernel

We introduce the State-query Coupled Kernel to address the limitation of where query location fail to adapt effectively to the input function.

1) Query location Encoding: We employ a normalized RBF distance κ and discrete fourier encoding to encode y.

$$\kappa(y, y') = \frac{\exp\left(-\|y - y'\|^2\right)}{\int_{\mathcal{Y}} \exp\left(-\|y - y'\|^2\right)\, dy'} \tag{12}$$

$$a_{rbf} = [\kappa(y, y'_0), \ldots, \kappa(y, y'_n)]\, W_1, y'_i \in \mathcal{Y} \tag{13}$$

$$\begin{aligned} a_{cos} &= \left[\cos(2\pi b_1^T y), \ldots, \cos(2\pi b_N^T y)\right] \\ a_{sin} &= \left[\sin(2\pi b_1^T y), \ldots, \sin(2\pi b_N^T y)\right] \end{aligned} \tag{14}$$

$$q(y) := f_{MLP}([a_{rbf}, a_{cos}, a_{sin}]) \tag{15}$$

Where the W_1 is a linear mappings, b_i is learnable vector that $i \in [1, N]$. Each of these operations is applied to the query location y, $\mathcal{Y}$ is the low resolution location set. f_{MLP} is implemented by MLP.

2) State-location Attention: We apply GateMLP [19] $f_{GateMLP}$ in z from (11) lifting to a $\mathbb{R}^{2D}$ embedding, which then be splited into k and v, each of dimension $\mathbb{R}^D$:

$$k, v = \text{Split}(f_{GateMLP}(z), D) \tag{16}$$

We then use an attention kernel to fuse the state feature and query location. The final output is obtained by MHA in a residual way, then through use a nonlinear mapping $f(\cdot)$ implemented by MLP:

$$\mathcal{G}(u_{s:t})(y) = f(q(y) + \mathrm{MHA}(q(y), k, v)) \tag{17}$$

IV. Experiments

We evaluate our model on multiple variables and challenging metrics, also conduct vary ablation and hyperparameter studies to demonstrate its effectiveness.

A. Datasets

The ERA5 dataset [13] provides comprehensive atmospheric reanalysis dataset. Integrating state-of-the-art forecasting models from the ECMWF's Integrated Forecasting System (IFS) [6]. In our setting, we utilize 6 year of data (2010-2015) for training, comprising 4,800 samples; 1 year of data (2016) for validation, comprising 800 samples; and 2 years of data (2017-2018) for testing, comprising 1,600 samples. The comparison variables include U10, V10, T2m, Z500, and T850. Full variables names can be found in Table I.

B. Metrics

We evaluate benchmarks using three commonly used metrics: latitude-weighted RMSE, latitude-weighted MSE, and Anomaly Correlation Coefficient (ACC) after denormalizing the predictions. The $L(y)$ and MSE are discussed before, the MSE is the $\mathcal{L}_{\mathrm{L2}}$ in (2). RMSE is below:

$$\frac{1}{N}\sum_{\ell=1}^{N}\sqrt{\frac{1}{|\Omega|}\sum_{y\in\Omega}L(y)\left(\mathcal{G}(u_{s:t}^{\ell})(y) - s^{\ell}(y)\right)^{2}} \tag{18}$$

Where $\mathcal{G}(u_{s:t})$ represents the prediction, s^{ℓ} is ground truth and y denote the location, Ω is the total locations set.

Anomaly correlation coefficient (ACC) is the spatial correlation between prediction value $\mathcal{G}(u_{s:t}^{\ell})$ relative to climatology and ground truth s^{ℓ} relative to climatology:

$$\frac{1}{N}\sum_{\ell=1}^{N}\frac{\sum_{y\in\Omega}L(y)\mathcal{G}(u_{s:t}^{\ell})(y)'s^{\ell}(y)'}{\sqrt{\sum_{y\in\Omega}L(y)\mathcal{G}(u_{s:t}^{\ell})(y)'^{2}\sum_{y\in\Omega}L(y)s^{\ell}(y)'^{2}}} \tag{19}$$

The $s^{\ell}(y)'$ and $\mathcal{G}(u_{s:t}^{\ell})(y)'$ denote the centered versions of $s^{\ell}(y)$ and $\mathcal{G}(u_{s:t}^{\ell})(y)$. The ACC gauges a model's capacity to predict extreme weather or climate events. It can be utilized to determine a model's ability to capture abnormal weather or climate phenomena, while RMSE and MSE account for area variations across grid cells at different latitudes to assess errors in climate state.

TABLE II
COMPARISON OF THE ABLATION PERFORMANCE

RMSE	w/o RBF	w/o DFE	w/o SQCK	Ours
U10	2.0402	2.0896	2.0352	**2.0348**
V10	1.9768	2.0271	1.9720	**1.9713**
T2m	1.4066	1.7237	**1.3736**	1.3832
T850	1.2626	1.2956	1.2582	**1.2578**
Z	105.3823	105.9473	105.2457	**104.8935**
Total Average	22.4137	22.6167	22.3769	**22.3081**

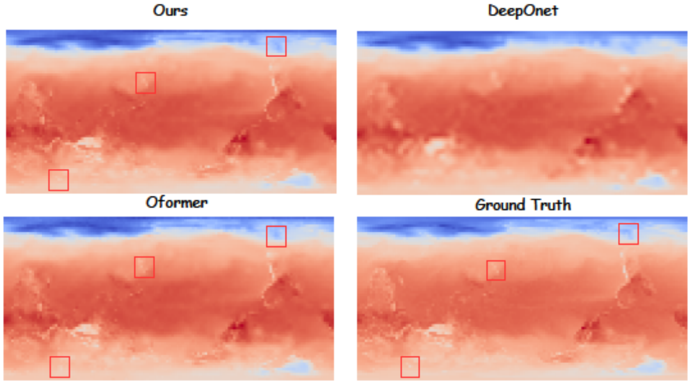

Fig. 2. Visualization climate downscaling of Temperature at 2 meters.

C. Main results and case Study

As demonstrated in Table I, SMCK is comprehensively evaluated against current baseline methods, including DeepONet [22], LOCA [15], GNOT [12], Oformer [17], and WaveONet, which utilizes a wavelet scattering network [7] as the encoder in the DeepONet framework to augment encoding capabilities. Our model consistently outperforms these established approaches over various climate variables and multiple challenging metrics, highlighting the superiority and robustness of our model. We also visualize prediction compared with some baseline methods in Fig. 2. The areas highlighted in the red boxes demonstrate the superiority of our approach.

D. Hyperparameters and ablation study

In our hyperparameter experiments, we tested time frames of 0, 4, 8, 12, where each frame represents a 1 hour interval, 0 frame is the ablation in LSF encoder (**w/o LSF**). The average RMSE results are 22.352, 22.349, **22.308**, and 22.356 respectively, We find that frame of 8 performed best, aligning with our hypothesis of the observation process's approximate Markov property. The ablation study Table II using a time frame of 8 evaluated different query location encodings (**w/o RBF** and **w/o DFE**) and the state-query coupled kernel (**w/o SQCK**). Each component enhance model outcome , confirm the efficacy of our approach.

V. Conclusion

In this paper, we introduce a novel operator SMCK for climate downscaling. Building upon the approximate Markov property of climate processes and the multimodal nature of climate processes, we propose a spatiotemporal multimodal fusion encoder, which efficiently and robustly integrates information. Additionally, we present an improved query location encoding method that combines radial basis function and discrete fourier encoding. Finally, we introduce the state-query coupling kernel, which leverages the attention mechanism to couple query locations with the input state for more effective learning. Extensive experiments demonstrate the effectiveness and robustness of SMCK, thus making contributions to the domain of climate downscaling and analysis research.

REFERENCES

[1] M. Assran, Q. Duval, I. Misra, P. Bojanowski, P. Vincent, M. Rabbat, Y. LeCun, and N. Ballas, "Self-supervised learning from images with a joint-embedding predictive architecture," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 15 619–15 629.

[2] K. Azizzadenesheli, N. Kovachki, Z. Li, M. Liu-Schiaffini, J. Kossaifi, and A. Anandkumar, "Neural operators for accelerating scientific simulations and design," *Nature Reviews Physics*, pp. 1–9, 2024.

[3] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *CoRR*, vol. abs/1409.0473, 2014. [Online]. Available: https://api.semanticscholar.org/CorpusID:11212020

[4] P. Bauer, A. Thorpe, and G. Brunet, "The quiet revolution of numerical weather prediction," *Nature*, p. 47–55, Sep 2015. [Online]. Available: http://dx.doi.org/10.1038/nature14956

[5] K. Bi, L. Xie, H. Zhang, X. Chen, X. Gu, and Q. Tian, "Accurate medium-range global weather forecasting with 3d neural networks," *Nature*, vol. 619, no. 7970, pp. 533–538, 2023.

[6] P. Bougeault *et al.*, "The thorpex interactive grand global ensemble," *Bulletin of the American Meteorological Society*, vol. 91, no. 8, pp. 1059–1072, 2010. [Online]. Available: https://doi.org/10.1175/2010BAMS2853.1

[7] J. Bruna and S. Mallat, "Invariant scattering convolution networks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 8, pp. 1872–1886, 2013.

[8] S. Cao, "Choose a transformer: Fourier or galerkin," *Advances in neural information processing systems*, vol. 34, pp. 24 924–24 940, 2021.

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

[10] T. Chen and H. Chen, "Universal approximation to nonlinear operators by neural networks with arbitrary activation functions and its application to dynamical systems," *IEEE transactions on neural networks*, vol. 6, no. 4, pp. 911–917, 1995.

[11] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[12] Z. Hao, Z. Wang, H. Su, C. Ying, Y. Dong, S. Liu, Z. Cheng, J. Song, and J. Zhu, "Gnot: A general neural operator transformer for operator learning," in *International Conference on Machine Learning*. PMLR, 2023, pp. 12 556–12 569.

[13] H. Hersbach, B. Bell, P. Berrisford, S. Hirahara, A. Horányi, J. Muñoz-Sabater, J. Nicolas, C. Peubey, R. Radu, D. Schepers, A. Simmons, C. Soci, S. Abdalla, X. Abellan, G. Balsamo, P. Bechtold, G. Biavati, J. Bidlot, M. Bonavita, G. Chiara, P. Dahlgren, D. Dee, M. Diamantakis, R. Dragani, J. Flemming, R. Forbes, M. Fuentes, A. Geer, L. Haimberger, S. Healy, R. J. Hogan, E. Hólm, M. Janisková, S. Keeley, P. Laloyaux, P. Lopez, C. Lupu, G. Radnoti, P. Rosnay, I. Rozum, F. Vamborg, S. Villaume, and J. Thépaut, "The era5 global reanalysis," *Quarterly Journal of the Royal Meteorological Society*, p. 1999–2049, Jul 2020. [Online]. Available: http://dx.doi.org/10.1002/qj.3803

[14] P. Jin, S. Meng, and L. Lu, "Mionet: Learning multiple-input operators via tensor product," *SIAM Journal on Scientific Computing*, vol. 44, no. 6, pp. A3490–A3514, 2022.

[15] G. Kissas, J. H. Seidman, L. F. Guilhoto, V. M. Preciado, G. J. Pappas, and P. Perdikaris, "Learning operators with coupled attention," *Journal of Machine Learning Research*, vol. 23, no. 215, pp. 1–63, 2022.

[16] N. Kovachki, Z. Li, B. Liu, K. Azizzadenesheli, K. Bhattacharya, A. Stuart, and A. Anandkumar, "Neural operator: Learning maps between function spaces with applications to pdes," *Journal of Machine Learning Research*, vol. 24, no. 89, pp. 1–97, 2023.

[17] Z. Li, K. Meidani, and A. B. Farimani, "Transformer for partial differential equations' operator learning," *Trans. Mach. Learn. Res.*, vol. 2023, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:249152256

[18] Z. Li, N. B. Kovachki, K. Azizzadenesheli, K. Bhattacharya, A. Stuart, A. Anandkumar *et al.*, "Fourier neural operator for parametric partial differential equations," in *International Conference on Learning Representations*, 2021.

[19] H. Liu, Z. Dai, D. So, and Q. V. Le, "Pay attention to mlps," *Advances in neural information processing systems*, vol. 34, pp. 9204–9215, 2021.

[20] Y. Liu, A. R. Ganguly, and J. Dy, "Climate downscaling using ynet: A deep convolutional network with skip connections and fusion," in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2020, pp. 3145–3153.

[21] I. D. Longstaff and J. F. Cross, "A pattern recognition approach to understanding the multi-layer perception," *Pattern Recognition Letters*, vol. 5, no. 5, pp. 315–319, 1987.

[22] L. Lu, P. Jin, G. Pang, Z. Zhang, and G. E. Karniadakis, "Learning nonlinear operators via deeponet based on the universal approximation theorem of operators," *Nature Machine Intelligence*, vol. 3, pp. 218 – 229, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:233822586

[23] C. Micchelli and M. Pontil, "Kernels for multi–task learning," in *Advances in Neural Information Processing Systems*, L. Saul, Y. Weiss, and L. Bottou, Eds., vol. 17. MIT Press, 2004. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2004/file/c4f796afbc6267501964b46427b3f6ba-Paper.pdf

[24] T. Nguyen, J. Jewik, H. Bansal, P. Sharma, and A. Grover, "Climatelearn: Benchmarking machine learning for weather and climate modeling," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[25] E. Oyallon, E. Belilovsky, and S. Zagoruyko, "Scaling the scattering transform: Deep hybrid networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 5618–5627.

[26] J. O. Ramsay, "When the data are functions," *Psychometrika*, vol. 47, pp. 379–396, 1982.

[27] S. Rasp, P. D. Dueben, S. Scher, J. A. Weyn, S. Mouatadid, and N. Thuerey, "Weatherbench: A benchmark data set for data-driven weather forecasting," *Journal of Advances in Modeling Earth Systems*, vol. 12, no. 11, p. e2020MS002203, 2020.

[28] A. Vaswani, "Attention is all you need," *Advances in Neural Information Processing Systems*, 2017.

[29] Y. Verma, M. Heinonen, and V. Garg, "Climode: Climate and weather forecasting with physics-informed neural odes," *arXiv preprint arXiv:2404.10024*, 2024.

[30] F. Wang and D. Tian, "On deep learning-based bias correction and downscaling of multiple climate models simulations," *Climate dynamics*, vol. 59, no. 11, pp. 3451–3468, 2022.