Paraformer: Parameterization of Sub-grid Scale Processes Using Transformers

Shuochen Wang
Department of Civil and Environmental Engineering
Northeastern University
Boston, MA 02115, USA
wang.shuoc@northeastern.edu

Nishant Yadav
Microsoft Corporation
Redmond, WA 98052, USA
nishantyadav@microsoft.com

Auroop Ganguly
Department of Civil and Environmental Engineering
Northeastern University
Boston, MA 02115, USA
a.ganguly@northeastern.edu

Abstract

One of the major sources of uncertainty in the current generation of Global Climate Models (GCMs) is the simulation of sub-grid scale physical processes. Over the years, with significantly improved computational performance, a series of Deep Learning (DL) parameterization schemes have been developed and incorporated into GCMs. However, these schemes use classic architectures whereas the new attention mechanism is not widely investigated. We proposed a "memory-aware" transformer-based model on ClimSim, the largest-ever dataset for climate parameterization. Our results show that the attention mechanism successfully captures the complex non-linear dependencies of sub-grid scale variables and reduces the prediction error.

1 Introduction

Understanding and modeling complex physical processes in the earth system is of great importance for making prompt and precise decisions regarding climate change issues. The Global Climate Model (GCM) is a tool to simulate the changes in the Earth's climate due to both natural and anthropogenic activities. Since the end of the last century, GCMs have been developed rapidly with joint efforts from Earth system scientists and computer scientists. With the help of supercomputers, there was a large improvement in terms of the model complexity and prediction accuracy in the current generation of GCMs compared to their early counterparts. However, the current generation of GCMs still has a coarse spatial resolution of 100-200 km and a large spread in simulations due to the presence of sub-grid scale physical processes. For example, studies have shown that one of the major sources of uncertainty in the current generation of GCMs is the simulation of small-scale clouds [1, 2]. These processes cannot be explicitly resolved by the model and thus must be parameterized.

Over the years, a series of parameterization schemes were developed using the traditional data assimilation method, integrating observational data with a numerical model to estimate the state of the Earth's system. However, in recent years, there has been a rising trend in developing parameterization schemes using data-driven approaches like Deep Learning (DL) [3, 4]. One critical challenge of this method is the acquisition of sub-grid information since it requires high-resolution model runs. Although some studies took an alternative approach by testing these schemes on a highly idealized model such as the Lorenz 96 system, which consists of two sets of differential equations representing

Tackling Climate Change with Machine Learning: workshop at NeurIPS 2024.

the change in sub-grid and grid-scale variables [5, 6, 7], it is preferable to test parameterization schemes directly on real-geography GCMs. The newly published ClimSim dataset addressed this challenge, providing the largest and most physically comprehensive testbed for DL parameterization schemes on climate simulations [8].

In this work, we proposed a transformer-based "memory-aware" parameterization scheme on the ClimSim dataset. Transformer is a new generation of DL model first introduced in the field of Natural Language Processing [9] and applied to machine translation [10] and text generation [11]. The attention mechanism in transformers performs well at handling long-range dependencies in sequential data and thus has been used for time series prediction recently [12, 13, 14]. Current DL parameterization schemes on weather and climate primarily use classic architectures such as Convolutional Neural Networks [15, 16, 17], Multi-Layer Perceptron [4, 18], and Generative Adversarial Networks [6, 19, 20], whereas the attention mechanism is not widely investigated. The attention mechanism allows transformer models to focus on specific parts of the input data, assigning importance scores to previous data points. In this way, a transformer model can generate a dynamic memory bank while predicting the next item. For parametrization schemes, it translates to the model capturing inter-state and intra-state temporal dependencies while approximating sub-grid scale processes. Our results suggest that the attention mechanism effectively captures the complex non-linear dependencies of sub-grid processes and achieves lower prediction error compared to classic architectures.

2 Data and Methods

2.1 The ClimSim Dataset

ClimSim is a dataset for developing machine-learning emulators on climate parameterization [8]. The input and output variables are generated by the Energy Exascale Earth System Model (E3SM) with a Multiscale Modeling Framework (MMF), representing the grid-scale and sub-grid scale physical processes in the Earth's system, respectively. Details of these variables are provided in Table A1. We use the low-resolution, real-geography dataset from the ClimSim suite, which contains 10 years of data at 384 unstructured spatial grids with a temporal resolution of 20 minutes. We further subsample the dataset to have an effective temporal resolution of 140 minutes while keeping the spatial structure intact. The training set covers 7 years (0001-02 to 0008-01) and the validation set includes 1 year (0008-02 to 0009-01). The test set has the same temporal coverage as the validation set but with a temporal resolution of 120 minutes to approximate a different climate model run. We use this subsampled version for two reasons: there is a memory limitation to input all the data without subsampling; the variables used are consistent with the original ClimSim paper [8], which is ideal for comparing the performance to the baseline models provided.

2.2 Data Preprocessing and DL Architecture

The DL model used in the paper is an attention-based, encoder-only transformer. The model is defined by its context window size (i.e. the number of sequential inputs it processes in a single pass). The context window size controls how much memory the model holds at a time. In ClimSim, we generate the context of a set size (hyperparameter) by creating sequential windows over the time dimension. We tested two approaches: the first involves creating a series of sliding windows that move one step at a time from the beginning to the end of the time series. This approach resembles a text generation problem, where a word or letter is generated based on the preceding sequence of text. We tested different window sizes ranging from 5 to 40. However, one issue with this method is data duplication, which significantly increases the dataset size depending on the selected window size, thereby increasing training time. Therefore, we tested a second method, where the time dimension is split into independent, non-overlapping windows. We found a window size of 5 produced the best results and creating a sliding window does not outperform simply splitting the time series. A window size of 5 is equal to roughly 12 hours of climate memory in the model at the given temporal resolution.

3 Results

We performed a hyperparameter search process summarized in B. The best hyperparameter configuration is: an embedding dimension of 256, 6 transformer encoder layers, 4 attention heads, a batch size of 512, and AdamW optimizer. The model's loss function is taken as mean square error (MSE) and the learning rate is defined using a ReduceLROnPlateau learning rate scheduler with a patience of 10 epochs, a reduction factor of 0.5 for a total of 200 epochs, and an initial learning rate of 1×10^{-4}.

Table 1: MAE and R^2 for target variables using the six baseline models provided in [8] and the transformer model. Metrics for MLP are from the reimplemented version while values for other baselines can be found in [8]. Large negative global R^2 are not shown for dq/dt and PRECSC due to the variabilities in the upper atmosphere and tropics, respectively. Units of non-energy flux variables are converted to a common energy unit, W/m^2 [8]. The best model for each variable is bolded.

Variables	**MAE [W/m^2]**							**R^2**						
	CNN	ED	HSR	MLP	RPN	cVAE	Transformer	CNN	ED	HSR	MLP	RPN	cVAE	Transformer
dT/dt	2.585	2.684	2.845	2.673	2.685	2.732	**2.332**	0.627	0.542	0.568	0.594	0.617	0.59	**0.681**
dq/dt	4.401	4.673	4.784	4.519	4.592	4.68	**4.049**	-	-	-	-	-	-	-
NETSW	18.85	14.968	19.82	13.753	18.88	19.73	**9.739**	0.944	0.98	0.959	0.982	0.968	0.957	**0.990**
FLWDS	8.598	6.894	6.267	5.410	6.018	6.588	**4.471**	0.828	0.802	0.904	0.917	0.912	0.883	**0.940**
PRECSC	3.364	3.046	3.511	2.687	3.328	3.322	**2.285**	-	-	-	-	-	-	-
PRECC	37.83	37.25	42.38	33.838	37.46	38.81	**22.199**	**0.077**	-17.909	-68.35	-34.545	-67.94	-0.926	-1.764
SOLS	10.83	8.554	11.31	8.163	10.36	10.94	**6.529**	0.927	0.96	0.929	0.959	0.943	0.929	**0.972**
SOLL	13.15	10.924	13.6	10.562	12.96	13.46	**8.950**	0.916	0.945	0.916	0.945	0.928	0.915	**0.958**
SOLSD	5.817	5.075	6.331	4.603	5.846	6.159	**3.701**	0.927	0.951	0.923	0.955	0.94	0.921	**0.969**
SOLLD	5.679	5.136	6.215	4.841	5.702	6.066	**4.318**	0.813	0.857	0.797	0.863	0.837	0.796	**0.888**

Table 1 shows the mean absolute error (MAE) and coefficient of determination (R^2) for the target sub-grid scale variables using our transformer-based model and the other six baseline models provided in [8]. The transformer-based model outperforms other structures on all except one variable in terms of the global-mean MAE and R^2, suggesting that the attention mechanism effectively captures the complex temporal dependencies of the sub-grid scale physical processes. To visualize the prediction improvement of variables with vertical structures, we reimplement the identical Multi-Layer Perceptron (MLP) model in [8]. The results are shown in Figure 1. In the upper levels of the atmosphere (level index 0 to 20), both models have similar performances for heating and moistening tendencies, whereas curves diverge with the decrease in height. The transformer model mainly reduces MAE and root mean square error (RMSE) in the lower levels of the atmosphere (level index around 40), with a general improvement of R^2 across all levels from 20 to 60 for heating tendency and 30 to 60 for moistening tendency. In Figure 2, we further examine the performances over different latitudes. Both models have the highest prediction accuracy over the low-latitude, 400-hPa region, whereas the values are lower around 50°S and 50°N. In line with Figure 1, the transformer model generally increases the overall accuracy but with limited improvement over the high-level atmosphere.

4 Discussion

Our analysis in this paper is based on a subset of variables from the ClimSim dataset. State-of-the-art GCMs model more complex grid-scale and sub-grid scale physical processes, and building a DL framework on the full set of variables may improve the result [8]. A recent study that used a UNet structure on ClimSim observed improvements using the full input and output variables [15]. Future research could also explore the high-resolution version with consideration of high computational cost. In addition, the ClimSim dataset can serve as a testbed for advanced transformer architectures designed for spatial and temporal data [21], which may further enhance prediction accuracy. At the same time, integrating the physical constraints of sub-grid scale processes into the model is also a promising approach. Recently, various Physics-Informed Neural Networks (PINNs) have been applied to weather and climate data analysis [22]. Incorporating these constraints could lead to a new model that improves global predictions and provides a better understanding of the complex physical processes in upper-level sub-grid variables.

5 Conclusion

In this paper, we proposed a "memory-aware" transformer-based model that outperforms other classic DL architectures on the ClimSim dataset. By creating sequences along the time dimension of the

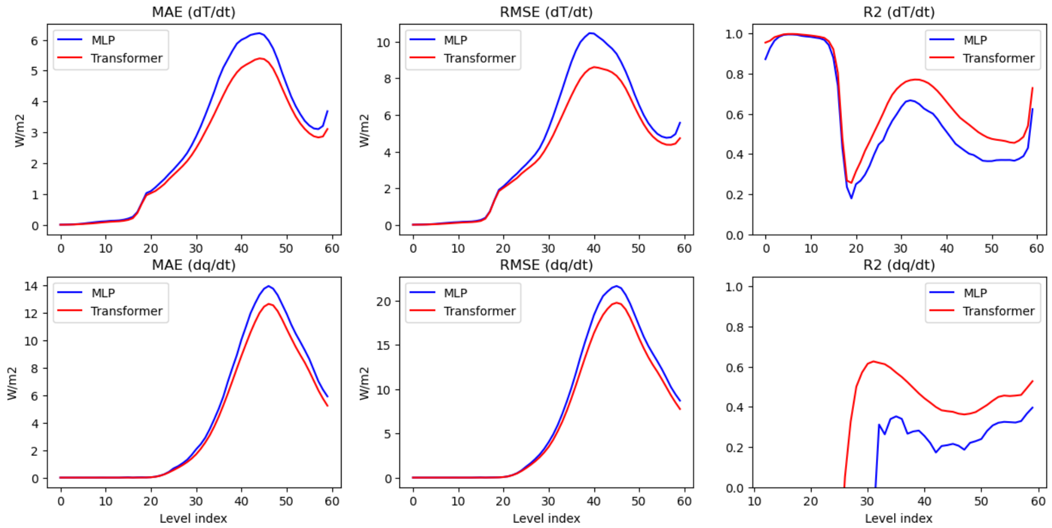

Figure 1: MAE, RMSE and R^2 of heating tendency dT/dt and moistening tendency dq/dt using Multi-Layer Perceptron (MLP) and our transformer-based model. The details of the configuration of the MLP are provided in [8]. Each index on the x-axis represents a vertical level in the atmosphere starting from the top (i.e. level index 0 represents the top of the atmosphere). Units of non-energy flux variables are converted to a common energy unit, W/m^2 [8]. Negative R^2 values are not shown.

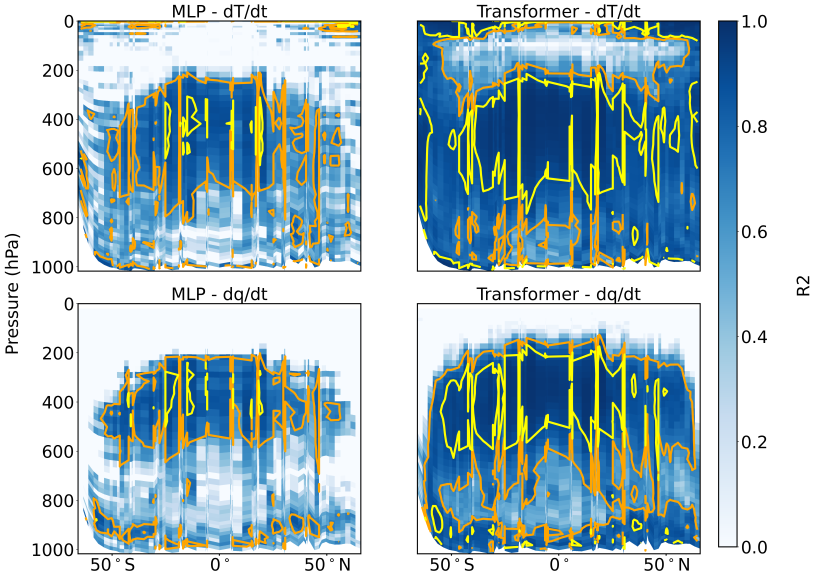

Figure 2: R^2 of daily-mean, zonal-mean heating tendency dT/dt and moistening tendency dq/dt for MLP and transformer at different pressure levels. Yellow contours cover regions of $> 0.9R^2$, orange contours cover regions of $> 0.7R^2$.

data, the attention mechanism in the transformer successfully captures the change in the sub-grid scale physical processes. Our model fills the gap of utilizing the attention mechanism in the climate parameterization problem and provides insights for developing future DL parameterization schemes.

References

[1] Tapio Schneider, João Teixeira, Christopher S Bretherton, Florent Brient, Kyle G Pressel, Christoph Schär, and A Pier Siebesma. Climate goals and computing the future of clouds. *Nature Climate Change*, 7(1):3–5, 2017.

[2] Mark D Zelinka, Timothy A Myers, Daniel T McCoy, Stephen Po-Chedley, Peter M Caldwell, Paulo Ceppi, Stephen A Klein, and Karl E Taylor. Causes of higher climate sensitivity in CMIP6 models. *Geophysical Research Letters*, 47(1):e2019GL085782, 2020.

[3] Stephan Rasp, Michael S Pritchard, and Pierre Gentine. Deep learning to represent subgrid processes in climate models. *Proceedings of the national academy of sciences*, 115(39):9684–9689, 2018.

[4] Pierre Gentine, Mike Pritchard, Stephan Rasp, Gael Reinaudi, and Galen Yacalis. Could machine learning break the convection parameterization deadlock? *Geophysical Research Letters*, 45(11):5742–5751, 2018.

[5] Daniel S Wilks. Effects of stochastic parametrizations in the Lorenz'96 system. *Quarterly Journal of the Royal Meteorological Society: A journal of the atmospheric sciences, applied meteorology and physical oceanography*, 131(606):389–407, 2005.

[6] David John Gagne, Hannah M Christensen, Aneesh C Subramanian, and Adam H Monahan. Machine learning for stochastic parameterization: Generative adversarial networks in the Lorenz'96 model. *Journal of Advances in Modeling Earth Systems*, 12(3):e2019MS001896, 2020.

[7] HM Arnold, IM Moroz, and TN Palmer. Stochastic parametrizations and model uncertainty in the Lorenz'96 system. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 371(1991):20110479, 2013.

[8] Sungduk Yu, Walter Hannah, Liran Peng, Jerry Lin, Mohamed Aziz Bhouri, Ritwik Gupta, Björn Lütjens, Justus C Will, Gunnar Behrens, Julius Busecke, et al. ClimSim: A large multi-scale dataset for hybrid physics-ML climate emulation. *Advances in Neural Information Processing Systems*, 36, 2024.

[9] Ashish Vaswani. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017.

[10] Qiang Wang, Bei Li, Tong Xiao, Jingbo Zhu, Changliang Li, Derek F Wong, and Lidia S Chao. Learning deep transformer models for machine translation. *arXiv preprint arXiv:1906.01787*, 2019.

[11] Junyi Li, Tianyi Tang, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. Pre-trained language models for text generation: A survey. *ACM Computing Surveys*, 56(9):1–39, 2024.

[12] Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 11106–11115, 2021.

[13] George Zerveas, Srideepika Jayaraman, Dhaval Patel, Anuradha Bhamidipaty, and Carsten Eickhoff. A transformer-based framework for multivariate time series representation learning. In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*, pages 2114–2124, 2021.

[14] Qingsong Wen, Tian Zhou, Chaoli Zhang, Weiqi Chen, Ziqing Ma, Junchi Yan, and Liang Sun. Transformers in time series: A survey. *arXiv preprint arXiv:2202.07125*, 2022.

[15] Zeyuan Hu, Akshay Subramaniam, Zhiming Kuang, Jerry Lin, Sungduk Yu, Walter M Hannah, Noah D Brenowitz, Josh Romero, and Michael S Pritchard. Stable machine-learning parameterization of subgrid processes with real geography and full-physics emulation. *arXiv preprint arXiv:2407.00124*, 2024.

[16] Thomas Bolton and Laure Zanna. Applications of deep learning to ocean data inference and subgrid parameterization. *Journal of Advances in Modeling Earth Systems*, 11(1):376–399, 2019.

[17] Pablo Rozas Larraondo, Luigi J Renzullo, Inaki Inza, and Jose A Lozano. A data-driven approach to precipitation parameterizations using convolutional encoder-decoder neural networks. *arXiv preprint arXiv:1903.10274*, 2019.

[18] Vladimir M Krasnopolsky, Michael S Fox-Rabinovitz, and Alexei A Belochitski. Using ensemble of neural networks to learn stochastic convection parameterizations for climate and numerical weather prediction models from data simulated by a cloud resolving model. *Advances in Artificial Neural Systems*, 2013(1):485913, 2013.

[19] Balasubramanya T Nadiga, Xiaoming Sun, and Cody Nash. Stochastic parameterization of column physics using generative adversarial networks. *Environmental Data Science*, 1:e22, 2022.

[20] Pavel Perezhogin, Laure Zanna, and Carlos Fernandez-Granda. Generative data-driven approaches for stochastic subgrid parameterizations in an idealized ocean model. *Journal of Advances in Modeling Earth Systems*, 15(10):e2023MS003681, 2023.

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

[22] Karthik Kashinath, M Mustafa, Adrian Albert, JL Wu, C Jiang, Soheil Esmaeilzadeh, Kamyar Azizzadenesheli, R Wang, Ashesh Chattopadhyay, A Singh, et al. Physics-informed machine learning: case studies for weather and climate modelling. *Philosophical Transactions of the Royal Society A*, 379(2194):20200093, 2021.

A Summary of Variables

Table A1: A subset of the input and target variables in the low-resolution, real-geography ClimSim dataset. Some variables have a size of 60, corresponding to the number of vertical levels in the atmosphere, others are scalar variables.

Input (grid-scale variables)	Size	**Output (sub-grid variables)**	Size
Temperature [K]	60	Heating tendency, dT/dt [K/s]	60
Specific humidity [kg/kg]	60	Moistening tendency, dq/dt [kg/kg/s]	60
Surface pressure [Pa]	1	Net surface shortwave flux, NETSW [W/m^2]	1
Insolation [W/m^2]	1	Downward surface longwave flux, FLWDS [W/m^2]	1
Surface latent heat flux [W/m^2]	1	Snow rate, PRECSC [m/s]	1
Surface sensible heat flux [W/m^2]	1	Rain rate, PRECC [m/s]	1
		Visible direct solar flux, SOLS [W/m^2]	1
		Near-IR direct solar flux, SOLL [W/m^2]	1
		Visible diffused solar flux, SOLSD [W/m^2]	1
		Near-IR diffused solar flux, SOLLD [W/m^2]	1

B Hyperparameter Search

In this work, we tested different combinations of hyperparameters in a grid search process summarized as follows: the number of encoder layers: [2, 4, 6, 8, 10, 12], the embedding dimension: [64, 128, 256, 512], the number of attention heads: [4, 8], the optimizer: [SGD, Adam, AdamW], and the learning rate scheduler [CosineAnnealingLR, ReduceLROnPlateau]. We use a traditional sine-cosine positional encoding method but disable the dropout due to a negative impact on the accuracy. Details of the transformer's encoder layer can be found in [9]. We performed the hyperparameter search using multiple nodes on a high-performance computing cluster, with one NVIDIA A100 GPU per node.