Efficient Localized Adaptation of Neural Weather Forecasting: A Case Study in the MENA Region

Muhammad Akhtar Munir[1] , **Fahad Shahbaz Khan**[1,3], **Salman Khan**[1,2]
[1]Mohamed bin Zayed University of AI, [2]Australian National University, [3]Linköping University
akhtar.munir@mbzuai.ac.ae

Abstract

Accurate weather and climate modeling are critical for both scientific advancement and safeguarding communities against environmental risks. Traditional approaches rely heavily on Numerical Weather Prediction (NWP) models, which simulate energy and matter flow across Earth's systems. However, heavy computational requirements and low efficiency restrict the suitability of NWP, leading to a pressing need for enhanced modeling techniques. Neural network-based models have emerged as promising alternatives, leveraging data-driven approaches to forecast atmospheric variables. In this work, we focus on limited-area modeling and train our model specifically for localized region-level downstream tasks. As a case study, we consider the MENA region due to its unique climatic challenges, where accurate localized weather forecasting is crucial for managing water resources, agriculture and mitigating the impacts of extreme weather events. This targeted approach allows us to tailor the model's capabilities to the unique conditions of the region of interest. Our study aims to validate the effectiveness of integrating parameter-efficient fine-tuning (PEFT) methodologies, specifically Low-Rank Adaptation (LoRA) and its variants, to enhance forecast accuracy, as well as training speed, computational resource utilization, and memory efficiency in weather and climate modeling for specific regions. Our codebase and pre-trained models can be accessed at https://github.com/akhtarvision/weather-regional.

1 Introduction

The accurate modeling and prediction of weather and climate patterns hold prominent significance for both scientific research and societal well-being. Primarily, weather and climate modeling can be categorized into two types: numerical (NWP) methods and neural network-based models. The former, usually related to General Circulation Models (GCMs) [1, 2], are designed to simulate the flow of energy and matter across the land, atmosphere, and ocean. However, the ability to perform detailed simulations to forecast weather and atmospheric variables in NWP models is restricted by computational resources and the time required for execution. As a consequence, there is a pressing need to address computational challenges and improve the precision of weather and climate modeling.

Neural network-based models [3,4,5,6,7,8,9,10,11,12,13] are data-centric and can absorb large-scale data to enhance performance. These models undergo comprehensive training on globally accessible datasets, thereby providing a viable alternative for predicting weather and climate patterns. Despite the absence of explicit physics assumptions in these models, the datasets employed for predictive task training inherently incorporate implicit physics assumptions. Within data-driven approaches, transformer-based [3, 7] and graph-based [6] models have been explored. One notable example of a transformer-based model is ClimaX [7], which focuses on developing a foundational model for climate and weather applications. ClimaX introduces a pre-training and fine-tuning paradigm, where during pre-training, it trains on data curated from physics-based models, enhancing its predictive capabilities and capturing relationships among atmospheric variables. This is followed by finetuning for various climate and weather tasks on a comprehensive climate reanalysis dataset. Other prominent

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

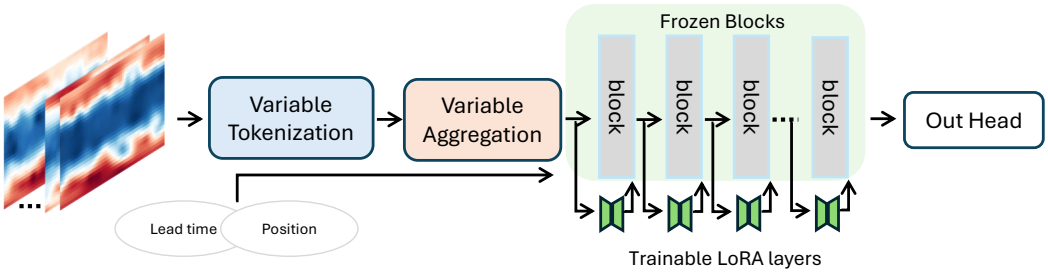

Figure 1: Main architecture: Integration of LoRA involves trainable layers while transformer blocks are frozen. The architecture modifies the main ViT by dealing with each channel separately for tokenization.

methods include Pangu [3] and GraphCast [6], with the latter being a graph neural network based approach for weather prediction. GraphCast operates on a high-resolution latitude-longitude grid, employing an architecture consisting of an encoder, a processor, and a decoder. However, these models focus on global forecasts and do not account for region-specific dynamics.

Our study focuses on a transformer-based weather and climate forecasting model trained for localized regional forecasting. The model capitalizes on comprehensive fine-tuning of the pre-trained weights of the global foundational model. It is crucial to investigate how these models can adapt to a paradigm of parameter-efficient fine-tuning (PEFT) [14, 15, 16, 17], providing efficient training and maintaining optimal performance. This paper implements various PEFT methods, primarily incorporating Low Rank Adaptation (LoRA) and its variants, over the ClimaX model. LoRA enhances model efficiency by introducing trainable components to each layer, thereby reducing the number of trainable parameters in large-scale models (Fig. 1). This approach addresses challenges related to speed, computational resources, and memory efficiency during the training of large models.

In the pursuit of adapting large-scale models with a parameter-efficient tuning methodology for downstream forecasting tasks, our proposal involves the exploration and investigation of LoRA and its variants across the Middle East and North Africa (MENA). The MENA region has experienced significant climate impacts, with summer temperatures projected to rise at more than twice the global average. This increase in heat, along with extended heat waves, could affect the region's habitat and can affect human health [18]. We aim to corroborate the efficacy of the proposed strategies in enhancing performance metrics, speed, and memory efficiency. Additionally, we integrate the flash attention mechanism into our approach, desirably accelerating the attention calculation process, and thereby facilitating resource-efficient model training for localized regions.

2 Method

2.1 Preliminaries

Notations: To input the neural network model, it takes the input $\mathcal{I}$ of shape $D \times H \times W$, where D is the number of variables spanning the atmospheric or climate ones. Let $\mathcal{F}$ be a neural network operator for gridded prediction tasks which takes input, $\mathcal{F}(\mathcal{I})$ and output $\mathcal{O}$ of shape $\hat{D} \times \hat{H} \times \hat{W}$. Spatial resolution $H \times W$ determines the density of the grid, and here we operate with two levels of resolutions 5.625° (32×64 grid points) and 1.40625° (128×256 grid points).

Architecture: The Vision Transformer (ViT) architecture involves the partitioning of input data into fixed patches, followed by a linear transformation to generate patch embeddings, commonly referred to as tokens [19]. The ClimaX [7] model we use consists of two main features, which will be briefly described in this paragraph. **(i) Variable Tokenization:** The proposed framework aims to mitigate a limitation observed in ViT architecture, which inherently processes a specified number of channels in the input. This approach independently addresses each input atmospheric and climate variable. **(ii) Variable Aggregation:** The variable tokenization approach presents challenges, notably an increase in sequence length corresponding to the rise in atmospheric and climate variables. Consequently, the memory complexity escalates quadratically when containing attention mechanisms. To address these challenges, a cross-attention operation is implemented for each spatial position within the map. This strategy effectively mitigates the computational complexity. For more details, we refer readers to [7].

2.2 Fine tuning using PEFT

Several methods are considered under the PEFT paradigm. We extensively study LoRA and its variants along with GLoRA, to cater to several PEFT mechanisms in one place.

LoRA: It presents several advantages, including the capability to integrate multiple LoRA modules for specific tasks without modifying the underlying base foundation model. By freezing the weights of the base model and optimizing low-rank learnable matrices, significant reductions in storage requirements can be accomplished. Additionally, there is no added inference latency compared to using a fully fine-tuned model. Because of these advantages, we also explore a variant of LoRA that incorporates residuals to retain information from previous blocks. However, empirical observations indicate that the simple LoRA method continues to outperform these variants. Neural network layers typically possess full rank, while LoRA attempts to project weights into a smaller subspace. Let $\mathcal{W}_p$ represent the pretrained weight matrix, and its modification $\Delta\mathcal{W}$ is replaced by low-rank decomposed matrices, denoted as BA, resulting in the equation:

$$\mathcal{W}_p + \Delta\mathcal{W} = \mathcal{W}_p + BA \tag{1}$$

Here, $B \in \mathbb{R}^{d\times r}$ and $A \in \mathbb{R}^{r\times k}$ decompose the matrix $\mathcal{W}_p \in \mathbb{R}^{d\times k}$, where r is the rank determined as $r \leq \min(d, k)$. To explain further, it is notable that gradients are not updated for $\mathcal{W}_p$. For more details, interested readers are encouraged to consult [14].

GLoRA: This methodology presents a unified framework that integrates fine-tuning approaches within a singular formulation. The architecture features a supernet, which is efficiently optimized through evolutionary search techniques. These conventional methods often rely on resource-intensive hyperparameter searches, dependent upon data availability. Employing an implicit search mechanism prevents the requirement for manual hyperparameter tuning, relaying the simultaneous increase in training time. We refer to App. **??** for more details on GLoRA.

Flash-attention: In advancing large models like language models, the flash attention algorithm, introduced by [20], optimizes attention computation by reducing memory usage. Its successor, flash attention-2 [21], further improves efficiency by minimizing non-matrix multiplication operations and parallelizing over sequence length.

3 Experiments and Results

The ERA5 reanalysis, developed by the European Center for Medium-Range Weather Forecasting (ECMWF), is utilized as a fundamental data source for training and evaluating weather forecasting systems. ERA5 integrates state-of-the-art Integrated Forecasting System [22] model outputs with observational data to generate comprehensive records of atmospheric, and land surface conditions. More details regarding data and implementation can be found in the App. **??** and App. **??** respectively.

3.1 Experiments with Global and Regional Modeling

To evaluate our forecasting models, we employ the ERA5 dataset, which provides global atmospheric reanalysis data at different resolutions. We compare ClimaX with the settings of global forecasting as well as in regional forecasting. We assess performance at both 5.625° and 1.40625° resolutions to

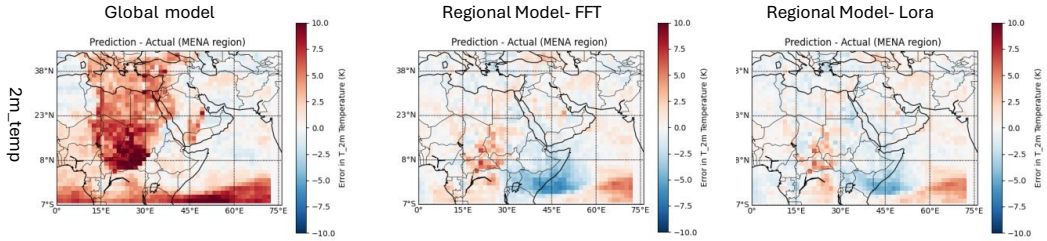

Figure 2: Qualitative: Error/Bias in Predictions and Actual measurements for temperature_2m (K). Dated, 11^{th} April 2017, lead time 3 days.

Metric	Model	geop@500	2m_temp	r_hum@850	s_hum@850	temp@850	10m_u_wind	10m_v_wind
ACC (↑)	Global	0.292	0.230	0.255	0.282	0.246	0.287	0.238
	Regional	0.585	0.804	0.502	0.623	0.620	0.570	0.517
RMSE (↓)	Global	674.295	3.349	23.308	0.003	3.561	3.733	4.162
	Regional	411.125	1.518	18.945	0.002	2.366	2.931	3.219

Table 1: ***Global vs Regional***: Forecasting on MENA region for 72 hrs prediction. Resolution is 1.40652°. The global model performs worse whereas regional model, specific to the needs of the local region performs better.

Metric	ACC (↑)			RMSE (↓)		
Model	geop@500	2m_temp	temp@850	geop@500	2m_temp	temp@850
Global	0.276	0.579	0.316	615.428	2.111	3.343
Regional$_{fft}$	0.554	0.816	0.597	453.256	1.448	2.536
Regional$_{Lora}$	0.582	0.823	0.614	438.213	1.414	2.476
Regional$_{resLora}$	0.580	0.823	0.613	438.669	1.411	2.476
Regional$_{GLora}$	0.557	0.806	0.591	445.437	1.482	2.535

Table 2: ***Lora, variants and fft***: Forecasting on MENA region for 72 hrs prediction. Resolution is 5.652°. Lora gives a competitive performance with a full fine-tune (fft) version and also results are reported with our proposed variant of Lora (resLora) and generalized Lora (GLora).

understand the impact of spatial granularity on forecasting accuracy. The performance of each model is evaluated using latitude-weighted root mean squared error (RMSE) and latitude-weighted anomaly correlation coefficient (ACC), standard metrics in weather prediction literature, reflecting forecast accuracy and consistency with observed data. In addition to global forecasting, we extensively extend our analysis to regional forecasting focusing on Middle East and North Africa (MENA). By selecting a regional dataset (ERA5-MENA) with the same set of variables, we assess ClimaX's ability to forecast weather conditions specifically within this region. We compare regional ClimaX with PEFT paradigms and a global version of ClimaX. Furthermore, we explore the impact of training ClimaX on data at different resolutions and evaluate its performance in regional forecasting tasks.

Regional vs Global: A comparative analysis between global and regional models unveils their respective behaviors and notably, the regional model consistently demonstrates superior performance when tasked with regional prediction objectives (Table 1). The regional model is specialized to predict more accurate forecasts due to the localized features learned during training. **Comparison with regional model variants:** In Table 2, our findings demonstrate that ClimaX with LoRA shows superior performance as compared to the global model and competitive performance compared to the full fine-tune method in predicting key atmospheric variables indicating its efficacy in weather forecasting applications. The number of trainable parameters is reduced from 108M (fft) to 16.2M (LoRA). Our study also presents findings obtained using GLora and a proposed modified version of Lora, which integrates residuals from similar blocks while aggregating information into subsequent transformer blocks. However, our observations indicate that while it does not exhibit superior performance, it does deliver reasonable results. We show qualitative results in Fig. 2 for one of the atmospheric variables. Bias is referred to as the difference between prediction and ground truth. We show further experiments in App.**??**, which include predictions over ranges and ablations, primarily focusing on the impact of rank r, LoRA's significance on attention and FC layers, memory and time comparison, and more qualitative results.

4 Conclusion

Neural network based models offer a promising alternative as these data-driven approaches, such as transformer-based models, leverage comprehensive training on large-scale datasets to forecast atmospheric variables with reasonable accuracy. Our study focuses on enhancing the transformer-based forecasting model, particularly in the Middle East and North Africa (MENA) region, through parameter-efficient fine-tuning methods such as Low-Rank Adaptation (LoRA). By investigating the efficacy of LoRA and its variants, alongside the integration of flash attention mechanisms, we improve model prediction accuracy and efficiency in terms of speed and memory. This effort represents an initial step towards efficiently adapting foundational weather models for localized regional dynamics.

References

[1] P. Lynch, "The origins of computer weather prediction and climate modeling," *Journal of computational physics*, vol. 227, no. 7, pp. 3431–3444, 2008.

[2] P. Bauer, A. Thorpe, and G. Brunet, "The quiet revolution of numerical weather prediction," *Nature*, vol. 525, no. 7567, pp. 47–55, 2015.

[3] 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.

[4] P. D. Dueben and P. Bauer, "Challenges and design choices for global weather and climate models based on machine learning," *Geoscientific Model Development*, vol. 11, no. 10, pp. 3999–4009, 2018.

[5] A. Grover, A. Kapoor, and E. Horvitz, "A deep hybrid model for weather forecasting," in *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, 2015, pp. 379–386.

[6] R. Lam, A. Sanchez-Gonzalez, M. Willson, P. Wirnsberger, M. Fortunato, F. Alet, S. Ravuri, T. Ewalds, Z. Eaton-Rosen, W. Hu *et al.*, "Learning skillful medium-range global weather forecasting," *Science*, vol. 382, no. 6677, pp. 1416–1421, 2023.

[7] T. Nguyen, J. Brandstetter, A. Kapoor, J. K. Gupta, and A. Grover, "Climax: A foundation model for weather and climate," *arXiv preprint arXiv:2301.10343*, 2023.

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

[9] T. Kurth, S. Subramanian, P. Harrington, J. Pathak, M. Mardani, D. Hall, A. Miele, K. Kashinath, and A. Anandkumar, "Fourcastnet: Accelerating global high-resolution weather forecasting using adaptive fourier neural operators," in *Proceedings of the platform for advanced scientific computing conference*, 2023, pp. 1–11.

[10] Z. Ben-Bouallegue, M. C. Clare, L. Magnusson, E. Gascon, M. Maier-Gerber, M. Janousek, M. Rodwell, F. Pinault, J. S. Dramsch, S. T. Lang *et al.*, "The rise of data-driven weather forecasting," *arXiv preprint arXiv:2307.10128*, 2023.

[11] S. Scher and G. Messori, "Weather and climate forecasting with neural networks: using general circulation models (gcms) with different complexity as a study ground," *Geoscientific Model Development*, vol. 12, no. 7, pp. 2797–2809, 2019.

[12] M. Schultz, C. Betancourt, B. Gong, F. Kleinert, M. Langguth, L. Leufen, A. Mozaffari, and S. Stadtler, "Can deep learning beat numerical weather prediction?, philos," in *Roy. Soc. A*, vol. 379, no. 2194, 2021, pp. 10–1098.

[13] T. Weber, A. Corotan, B. Hutchinson, B. Kravitz, and R. Link, "Deep learning for creating surrogate models of precipitation in earth system models," *Atmospheric Chemistry and Physics*, vol. 20, no. 4, pp. 2303–2317, 2020.

[14] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," *arXiv preprint arXiv:2106.09685*, 2021.

[15] B. Lester, R. Al-Rfou, and N. Constant, "The power of scale for parameter-efficient prompt tuning," *arXiv preprint arXiv:2104.08691*, 2021.

[16] M. Jia, L. Tang, B.-C. Chen, C. Cardie, S. Belongie, B. Hariharan, and S.-N. Lim, "Visual prompt tuning," in *European Conference on Computer Vision*. Springer, 2022, pp. 709–727.

[17] A. Chavan, Z. Liu, D. Gupta, E. Xing, and Z. Shen, "One-for-all: Generalized lora for parameter-efficient fine-tuning," *arXiv preprint arXiv:2306.07967*, 2023.

[18] J. Lelieveld, Y. Proestos, P. Hadjinicolaou, M. Tanarhte, E. Tyrlis, and G. Zittis, "Strongly increasing heat extremes in the middle east and north africa (mena) in the 21st century," *Climatic Change*, vol. 137, no. 1, pp. 245–260, 2016.

[19] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2021.

[20] T. Dao, D. Y. Fu, S. Ermon, A. Rudra, and C. Ré, "FlashAttention: Fast and memory-efficient exact attention with IO-awareness," in *Advances in Neural Information Processing Systems*, 2022.

[21] T. Dao, "FlashAttention-2: Faster attention with better parallelism and work partitioning," 2023.

[22] N. Wedi, P. Bauer, W. Denoninck, M. Diamantakis, M. Hamrud, C. Kuhnlein, S. Malardel, K. Mogensen, G. Mozdzynski, and P. Smolarkiewicz, *The modelling infrastructure of the Integrated Forecasting System: Recent advances and future challenges*. European Centre for Medium-Range Weather Forecasts, 2015.