

EWMoE: An Effective Model for Global Weather Forecasting with Mixture-of-Experts

Lihao Gan¹, Xin Man^{1,2}, Chenghong Zhang³, Jie Shao^{1,2*}

¹University of Electronic Science and Technology of China, Chengdu, China

²Sichuan Artificial Intelligence Research Institute, Yibin, China

³Institute of Plateau Meteorology, China Meteorological Administration, Chengdu, China
{ganlihao,manxin}@std.uestc.edu.cn, zhangchenghong@cma.gov.cn, shaojie@uestc.edu.cn

Abstract

Weather forecasting is a crucial task for meteorologic research, with direct social and economic impacts. Recently, data-driven weather forecasting models based on deep learning have shown great potential, achieving superior performance compared with traditional numerical weather prediction methods. However, these models often require massive training data and computational resources. In this paper, we propose EWMoE, an effective model for accurate global weather forecasting, which requires significantly less training data and computational resources. Our model incorporates three key components to enhance prediction accuracy: 3D absolute position embedding, a core Mixture-of-Experts (MoE) layer, and two specific loss functions. We conduct our evaluation on the ERA5 dataset using only two years of training data. Extensive experiments demonstrate that EWMoE outperforms current models such as FourCastNet and ClimaX at all forecast time, achieving competitive performance compared with the state-of-the-art models Pangu-Weather and GraphCast in evaluation metrics such as Anomaly Correlation Coefficient (ACC) and Root Mean Square Error (RMSE). Additionally, ablation studies indicate that applying the MoE architecture to weather forecasting offers significant advantages in improving accuracy and resource efficiency.

Introduction

Weather forecasting is the analysis of past and present weather observations, as well as the use of modern science and technology, to predict the state of the Earth atmosphere in the future. It is one of the most important applications of scientific computing and plays a crucial role in key sectors such as transportation, logistics, agriculture, and energy production (Bauer, Thorpe, and Brunet 2015). Traditionally, atmospheric scientists have relied on Numerical Weather Prediction (NWP) methods (Bjerknes, Volken, and Brönnimann 2009; Lorenc 1986), which utilize mathematical models of the atmosphere and oceans to forecast the weather states based on current weather conditions. While modern meteorological forecasting systems have achieved satisfactory results using NWP methods, these methods largely rely on parametric numerical models, which can introduce errors in the parameterization (Beljaars et al. 2018) of complex,

unresolved processes. Additionally, NWP methods face challenges in meeting the diverse needs of weather forecasting due to its high computational cost, the difficulty of solving nonlinear physical processes, and model deviations (Robert 1982; Simmons and Hollingsworth 2002).

To address the above issues of NWP models, researchers have turned their attention to data-driven weather forecasting based on deep learning methods. These methods run very quickly and can easily achieve a balance among model complexity, prediction resolution, and prediction accuracy (Dueben and Bauer 2018; Scher 2018; Weyn, Durran, and Caruana 2019). Denby (2020) first employed Convolutional Neural Network (CNN) for the classification of weather satellite images. Xu et al. (2019) utilized a combination of Generative Adversarial Network (GAN) and Long Short-Term Memory (LSTM) for cloud prediction. While these attempts reveal the potential of deep learning methods in weather forecasting, they are limited by low-resolution data and ineffective models, resulting in limited applications.

Recently, FourCastNet (Kurth et al. 2023) increased the resolution to 0.25° , comparable to the ECMWF Integrated Forecast Systems (IFS). ClimaX (Nguyen et al. 2023) showed superior performance on weather benchmarks for weather forecasting and climate projections, even when pretrained at lower resolutions and with limited computing budgets. Pangu-Weather (Bi et al. 2022) was the first state-of-the-art model to outperform IFS. These models are based on Vision Transformer (ViT) (Dosovitskiy et al. 2020), and use standard ViT embedding to process meteorological data. However, meteorological data is different from general computer vision image input. The channels in meteorological data represent atmospheric variables with intricate physical relationships and have different coordinate information in the earth coordinate system. ViT embedding method cannot effectively extract the physical features between these meteorological variables and the geographical features of the variables themselves (Hu et al. 2023). Moreover, these weather forecasting models usually require a very large amount of data, and their adoption was constrained by the high computational demands required (Chen et al. 2023b) for training. For example, FourCastNet utilized 64 Nvidia A100 GPUs for a training period of 16 hours, highlighting the extensive resources needed (Bi et al. 2023; Chen et al. 2023a) for the development of cutting-edge, deep learning based weather forecasting model. These

*Corresponding author: Jie Shao.

Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

issues motivated us to investigate a novel weather-specific embedding to model the meteorological data and an effective architecture to achieve superior weather forecasting using less training data and computational resources.

In this work, we present an effective data-driven model called EWMoE for global weather forecasting. We start with a Vision Transformer (ViT) architecture and, to address the issues of deep learning based models, our EWMoE consists of three novel components: (1) 3D absolute position embedding that fully models the geographical location features of each atmospheric variable. Different from other weather models that use relative position embedding in ViT or Swin Transformer (Liu et al. 2021), our 3D absolute position encoding can fully represent meteorological variables in terms of longitude, latitude and altitude, improving model prediction performance. (2) a crucial Mixture-of-Experts (MoE) structure that increases the model capacity without increasing compute requirements, greatly improving model prediction accuracy with significantly less training data and computational resources. This important improvement breaks the reliance of previous weather models on massive amounts of training data and enables our model to show superior performance even on less data. (3) The elaborately designed auxiliary loss and position-weighted loss perform specific operations during model training, optimizing our MoE layer and 3D absolute position encoding process respectively. We trained our proposed EWMoE on two years of data from the ERA5 dataset (Hersbach et al. 2020). Experiment results show that EWMoE significantly outperforms FourCastNet and Climax, and achieves a comparable level of forecasting accuracy as Pangu-Weather (Bi et al. 2022) and GraphCast (Lam et al. 2023a) for short-range forecasting (1-3 days). As the forecast time increases, EWMoE exhibits more stable and excellent prediction results compared with them. Notably, EWMoE achieves this performance by training on less data and requiring orders-of-magnitude fewer GPU hours. Finally, we conduct extensive ablation studies to analyze the importance of individual components in EWMoE, demonstrating its potential for facilitating future works.

Overall, our contributions can be summarized as follows:

- We propose EWMoE, an effective weather model with MoE for global weather forecasting, which demonstrates superior performance over other state-of-the-art models for short-to-medium-range weather prediction.
- Our EWMoE consists of three main components: (1) a 3D absolute position embedding to fully extract the geographical location features; (2) an MoE layer to increase the model capacity without increasing compute requirements; (3) two loss functions to optimize the training process.
- Unlike other deep learning based models, EWMoE achieves these superior global weather predictions with significantly smaller number of computational resources and training data.

Related Work

Numerical Weather Prediction

Numerical Weather Prediction (NWP) is a method used to forecast atmospheric conditions and weather states by utilizing

systems of partial differential equations (Bauer, Thorpe, and Brunet 2015; Lynch 2008; Kalnay 2002). These equations describe different physical processes and thermodynamics, which can be integrated over time to obtain future prediction results. Although NWP models have good reliability and accuracy in weather forecasting, they face many challenges such as systemic errors (Beljaars et al. 2018; Allen, Kettleborough, and Stainforth 2002) produced by parametrization schemes. NWP methods also involve high computation costs due to the complexity of integrating a large system of partial differential equations (Stensrud 2007), especially when modeling at high spatial and temporal resolutions. Furthermore, more observation data does not improve NWP forecast accuracy since models rely on the expertise of scientists in the meteorological field to refine equations, parameterizations and algorithms (Magnusson and Källén 2013).

In recent years, many efforts have been made to improve the accuracy and efficiency of NWP models. For example, some researchers (Best 2005) have proposed grid refinement techniques to increase the model resolution, while others (Navon 2009) have suggested fine-tuning physical parameterizations to further enhance the accuracy of weather forecasts.

Deep Learning based Weather Forecasting

In order to address the challenges of NWP models, researchers have shown increasing interest in the application of deep learning models to weather forecasting (Ren et al. 2021; Weyn et al. 2021). These models train deep neural networks to predict future weather states using vast amounts of historical meteorological data (Rasp et al. 2020, 2023; Man et al. 2023), such as the ERA5 reanalysis dataset. Compared with the traditional NWP models, deep learning based models have the potential to generate more accurate weather forecasts with less computational cost (Keisler 2022; Bi et al. 2023; Lam et al. 2023b; Gan et al. 2019). Once trained, these models can produce timely forecast in a few seconds, which is considerably faster than NWP models that take hours or even days (Schultz et al. 2021).

Weyn, Durran, and Caruana (2019) proposed an elementary weather prediction model using deep Convolutional Neural Networks (CNNs) trained on past weather data, although their method only achieves a modest resolution of 2.5° and contains no more than three variables per grid. However, rapid progress has been made in recent years. FourCastNet (Kurth et al. 2023), a data-driven weather forecasting model, utilized the Vision Transformer (ViT) architecture and Adaptive Fourier Neural Operators (AFNO) (Guibas et al. 2021), first pushing model resolution to 0.25° as NWP methods can. FourCastNet's predictions are comparable to the IFS model at lead times of up to three days, pointing to the enormous potential of data-driven modeling in complementing and eventually replacing NWP. ClimaX (Nguyen et al. 2023) is the first model that can effectively scale using heterogeneous climate datasets during pretraining and generalize to diverse downstream tasks during fine-tuning, paving the way for a new generation of deep learning models for Earth systems science. Pangu-Weather (Bi et al. 2022), a 3D Earth-specific Transformer model, is the first to outperform operational Integrated Forecasting System (IFS) (Bougeault et al. 2010), producing

Vertical level	Variables
Surface	$U_{10}, V_{10}, T_{2m}, sp, mslp$
10000 hPa	U, V, Z
850 hPa	T, U, V, Z, RH
500 hPa	T, U, V, Z, RH
50 hPa	Z
Integrated	$TCWV$

Table 1: The abbreviations are as follows: U_{10} and V_{10} represent the zonal and meridional wind velocity from the surface, specifically, at a height of 10m; T_{2m} represents the temperature at 2m from the surface; T, U, V, Z and RH represent the temperature, zonal velocity, meridional velocity, geopotential and relative humidity at specified vertical level; $TCWV$ represents the total column water vapor.

even more favorable evaluation results. In GraphCast (Lam et al. 2023a), Graph Neural Network (GNN) layers are employed for modeling weather dynamics and autoregressive finetuning is used for increasing the long-lead prediction. GenCast (Price et al. 2023) is a diffusion-based probabilistic weather model with greater skill and speed than the top operational medium-range weather forecast in the world. Neural-GCM (Kochkov et al. 2024) is a general circulation model with machine learning components and can generate forecasts of deterministic weather, ensemble weather and climate on par with the best machine-learning and physics-based methods.

Preliminaries

Dataset

ERA5 (Hersbach et al. 2020) is a publicly available atmospheric reanalysis dataset produced by the European Centre for Medium-Range Weather Forecasts (ECMWF). It provides comprehensive information about the Earth climate and weather conditions, and is widely used in climate research, climate change analysis, weather forecasting, environmental monitoring, and other fields. The dataset covers the period from 1940 to the present and includes a wide range of meteorological variables such as temperature, humidity, wind speed, precipitation, cloud cover, and more. The spatial resolution of the data is 0.25° latitude and longitude, with hourly intervals, and 37 vertical pressure levels ranging from 1000 hPa to 1 hPa.

In this study, we use the ERA5 reanalysis dataset as the ground-truth for the model training, which has a spatial resolution of 0.25° (721×1440 latitude-longitude grid points). Specifically, we select six-hourly sampled data points (T_0, T_6, T_{12}, T_{18}), with each sample consisting of twenty atmospheric variables across five vertical levels (see Table 1 for details). In addition, to demonstrate the effectiveness of our model in the case of limited data and computing resources, we use two years of data for training (2015 and 2016), one year for validation (2017), and one year for testing (2018).

Weather Forecasting Task

Given a dataset of historical weather data, the task of global weather forecasting is to forecast the future global atmosphere states based on the current atmosphere conditions

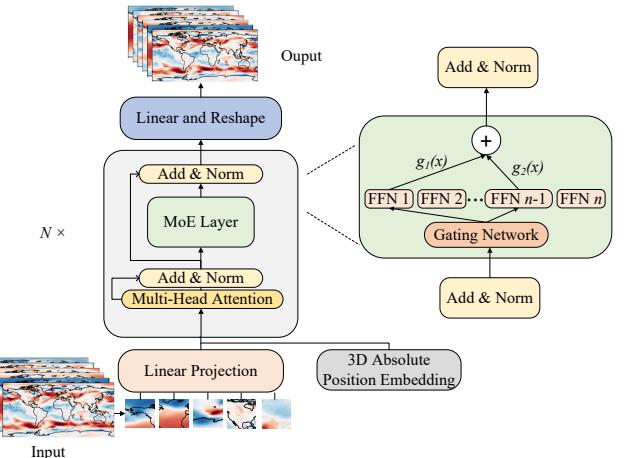


Figure 1: Overall architecture of the proposed EWMoE. Based on the standard encode-decoder design (Vaswani et al. 2017), EWMoE first uses a linear projection layer to extract the feature embeddings of input weather images and add the 3D absolute position embedding. Then, an MoE layer routes the tokens to top- k experts and integrates the outputs by the gating network. Finally, the feature representation is used to reshape the model output.

(Abbe 1901). Specifically, we denote the initial weather state as $X_i \in \mathbb{R}^{C \times H \times W}$, where C represents the number of atmosphere variables or channels, H and W are the height and width, respectively. Our model aims to generate 8-day forecasts $\{\hat{X}_{i+1}, \hat{X}_{i+2}, \dots, \hat{X}_{i+32}\}$ with a time interval of six hours. However, it is challenging to train the model to directly forecast the future weather state $\hat{X}_T = f_\theta(X_i)$ for each target lead time T . Since the weather system is chaotic, forecasting the future weather directly for large T is difficult (Rasp and Thuerey 2021; Clare, Jamil, and Morcrette 2021; Weyn, Durran, and Caruana 2020). Moreover, it requires training one network for each lead time, which can be computationally expensive when the dataset is very large. To avoid this issue, we train our model to produce forecasts in an autoregressive manner. For longer forecasts, we unfold the model by iteratively feeding its predictions back to the model as input, e.g., $\hat{X}_{i+1} = f_\theta(X_i), \hat{X}_{i+2} = f_\theta(\hat{X}_{i+1}), \dots, \hat{X}_{i+32} = f_\theta(\hat{X}_{i+31})$.

Methods

In this section, we introduce EWMoE, an effective weather model with Mixture-of-Experts (MoE) for global weather forecasting, as illustrated in Figure 1. Our EWMoE consists of three main components, which include: 1) 3D absolute position embedding; 2) the structure of the MoE layer; 3) the auxiliary loss and the position-weighted loss for model training optimization.

Pre-processing

The information contained in weather data is very different from the natural images used in computer vision tasks. The channels in weather data represent different meteorological

variables, and there are complex physical relationships among these variables. For example, there is a close relationship between temperature and relative humidity, while temperature and pressure obey the ideal gas law and are positively correlated. Therefore, effectively extracting the internal relations between these meteorology variables is the key to accurate weather forecasting. We denote the input image as a high dimension tensor $X \in \mathbb{R}^{C \times H \times W}$, and the module divides the input image into a sequence of patches, where the size is $p \times p$. Each input patch of size p^2 is linearly embedded to a vector of dimension D , where D is the embedding size. This results in $C \times (H/p) \times (W/p)$ patches in total. Then, a learnable query vector is used to perform cross-attention operation at each position to conduct the interactions between the meteorological variables of each channel, which is proposed by Climax (Nguyen et al. 2023) and is applied in Stormer (Nguyen et al. 2024). The cross-attention layer outputs a sequence of shape $(H/p) \times (W/p)$, significantly reducing the sequence length and lowering the computational cost.

3D Absolute Position Embedding

For global weather forecasting, each input token corresponds to an absolute position on the Earth's coordinate system. More importantly, some meteorology variables are closely related to their absolute position. For example, geopotential height is closely related to the latitude, while the wind speed and temperature are closely related to height. In this situation, using relative position embedding or 1D/2D position embedding does not capture this intrinsic feature well. Therefore, we use a 3D absolute position embedding for meteorology-specific position embedding, taking the 3D position information (longitude, latitude and altitude) of the patch into account. Specifically, for each input D dimensional vector, we train three sets of learnable position embedding vectors with dimension of $D/3$. Each set corresponds to the absolute position of a patch on the Earth's coordinate system, which are altitude, longitude and latitude respectively. After concatenating these three sets of vectors, we obtain the final 3D absolute position embedding vector with a dimension of D .

Structure of The MoE Layer

Following the position embedding, we leverage sparsely activated Mixture-of-Experts (MoE) in our EWMoE model, which allows increasing the model capacity (total number of available parameters) without increasing computing requirements (number of active parameters) and is widely used in Natural Language Processing (NLP) tasks (Shazeer et al. 2017; Lepikhin et al. 2021). There are N encoder blocks in EWMoE and we replace the dense Feed-Forward Network (FFN) layer present in encoder with a sparse MoE layer, as shown in Figure 1. Each MoE layer consists of a collection of independent feed forward networks as the "experts". A gating network then uses a softmax function to route the input tokens to the best-determined top- k experts. This means that for each given input token, only a small number of experts are activated, giving our model more flexibility and capacity to complete complex weather forecasting tasks with strong performance.

Given N experts and input token x , the output y of the MoE layer can be written as follows:

$$y = \sum_{i=1}^N g_i(x) E_i(x), \quad (1)$$

where $g_i(x)$ is the output of the i -th element of gating network and $E_i(x)$ is the output of the i -th expert network. According to the formulation above, we can save computation based on the sparsity of the output $g(x)$. When $g(x)$ is a sparse vector, only a few experts would be activated and updated by back-propagation during training. Wherever $g_i(x) = 0$, we need not compute the corresponding $E_i(x)$.

Top- k routing. We use top- k routing to select the top ranked experts, keeping only the top- k gate values while setting the rest to $-\infty$ before taking the softmax function. Then, the following $g(x)$ can be formulated as:

$$g(x) = \text{Softmax}(\text{top-}k(x \cdot W + \epsilon, k)), \quad (2)$$

$$\text{Top-}k(m, k)_i = \begin{cases} m_i & \text{if } m_i \text{ is in top-}k \text{ elements,} \\ -\infty & \text{otherwise,} \end{cases} \quad (3)$$

where W is a trainable weight matrix and $\epsilon \sim \mathcal{N}(0, \frac{1}{e^2})$ is a Gaussian noise for exploration of expert routing (e is the mathematical constant). When $k \ll N$, most elements of $g(x)$ would be zero so that our model can achieve greater capacity while using less computation. In the MoE layer, we train our model with $k = 2, N = 20$.

Loss Function for Model Training Optimization

Auxiliary loss for load balancing. In the MoE layer, we dispatch each token to k experts. There is a phenomenon that most tokens may be dispatched to a small portion of experts, as the favored experts are trained more rapidly and thus are selected even more by the gating network. Such an unbalanced distribution would decrease the throughput of our model, and as most experts would not be fully trained, the flexibility and performance of the model would be reduced. To resolve this issue, we use a differentiable load balancing auxiliary loss instead of separate load-balancing and importance-weighting losses for a balanced loading in routing. Given E experts and a batch B with L tokens, the following auxiliary loss is added to the total model loss during training:

$$l_{aux} = E \cdot \sum_{i=1}^E h_i \cdot P_i, \quad (4)$$

where h_i is the fraction of tokens dispatched to expert i :

$$h_i = \frac{1}{L} \sum_{x \in B} \mathbb{1}\{\text{argmax } g(x) = i\}, \quad (5)$$

and P_i is the fraction of the router probability distributed for expert i :

$$P_i = \frac{1}{L} \sum_{x \in B} g_i(x). \quad (6)$$

The goal of the auxiliary loss is to achieve a balanced distribution. When we minimize l_{aux} , we can see both h_i and P_i would close to a uniform routing.

Position-weighted loss. In the weather forecasting tasks, it is crucial to correctly predict the atmospheric variables at different locations, which has a very large social impact on human activities. We use a position-weighted function to represent the weights of variables at different locations and employ the latitude-weighted mean squared error as our objective function. Given the prediction $\hat{X}_{i+\Delta t}$ and the ground-truth $X_{i+\Delta t}$, the loss is written as:

$$\mathcal{L} = \frac{1}{CHW} \sum_{c=1}^C \sum_{i=1}^H \sum_{j=1}^W f(v) L(i) \left(\hat{X}_{i+\Delta t}^{cij} - X_{i+\Delta t}^{cij} \right)^2, \quad (7)$$

where $f(v)$ is a learnable parameter related to the absolute position of variable v , and $L(i)$ is the latitude-weighting factor at the coordinate i :

$$L(i) = \frac{\cos(\text{lat}(i))}{\frac{1}{H} \sum_{i'=1}^H \cos(\text{lat}(i'))}, \quad (8)$$

where $\text{lat}(i)$ denotes the latitude value.

Experiments

We first introduce the training details of EWMoE and then compare it with other state-of-the-art weather forecasting models, and show the results on predicting multiple meteorological variables. We also provide visualization examples to demonstrate the superiority of EWMoE in global weather forecasting. Additionally, we conduct extensive ablation studies to analyze the importance of each component in our model.

Implementation Details of Model Training

For each input data sample from the ERA5 dataset, it can be represented as an image with 20 channels. We set the patch size as 8×8 , and the EWMoE model consists of encoders with depth=6, dim=768 and decoders with depth=6, dim=512. Each encoder has a MoE layer, and each MoE layer consists of 20 independent experts. Specifically, in the gating network of each MoE layer, we use top-2 routing to select the top-2 ranked experts for forward propagation of training data. We employ the AdamW optimizer with two momentum parameters $\beta_1=0.9$ and $\beta_2=0.95$, and set the weight decay to 0.05. Our implementation code is available at <https://github.com/technomii/EWMoE>.

Evaluation Metrics

Following the previous deep learning based methods, the accuracy of deterministic forecast is computed by two quantitative metrics, namely, the latitude-weighted Root Mean Square Error (RMSE) and latitude-weighted Anomaly Correlation Coefficient (ACC).

The latitude weighted ACC for a forecast variable v at forecast time-step l is defined as follows:

$$\text{ACC}(v, l) = \frac{\sum_m L(m) \hat{X}_{pred} \hat{X}_{true}}{\sqrt{\sum_m L(m) (\hat{X}_{pred})^2 \sum_m L(m) (\hat{X}_{true})^2}}, \quad (9)$$

where $\hat{X}_{pred/true}$ represents the long-term-mean-subtracted value of predicted or true variable v at the location denoted

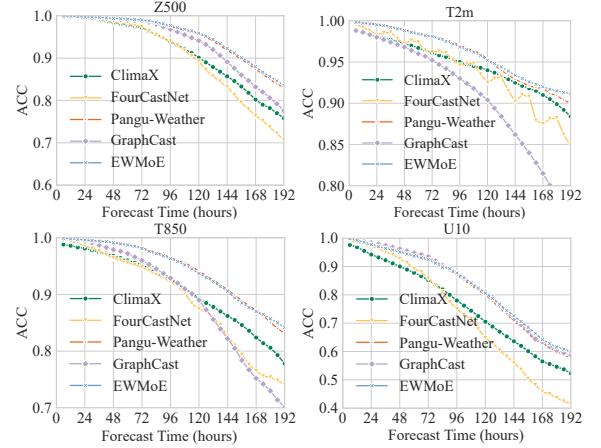


Figure 2: Latitude-weighted ACC results of EWMoE and the baselines predicting four key variables $Z500$, $T2m$, $T850$ and $U10$ in 2018 (higher ACC is better).

by the grid co-ordinates at the forecast time-step l . The long-term mean of a variable is just the mean value of it over a large number of historical samples in the training dataset. The long-term mean-subtracted variables $\hat{X}_{pred/true}$ represent the anomalies of those variables that are not captured by the long term values. $L(m)$ is the latitude weighting factor at the co-ordinate m which is defined in Eq. (8). The latitude-weighted RMSE for a forecast variable v at forecast time-step l is defined by the following equation:

$$\text{RMSE}(v, l) = \sqrt{\frac{1}{NM} \sum_m^M \sum_n^N L(m) (X_{pred} - X_{true})^2}, \quad (10)$$

where $X_{pred/true}$ represents the value of predicted or true variable v at the location denoted by the grid co-ordinates at the forecast time-step l .

Comparison with State-of-the-art Models

We compare the forecast performance of EWMoE with FourCastNet, ClimaX, Pangu-Weather and GraphCast, four leading deep learning methods for global weather forecasting. Some of the methods used for comparison (such as Pangu-Weather) did not provide source code for training, so it is infeasible to compare them with our model under the same conditions. Actually, the open-source FourCastNet model has already been trained on the same two-year dataset and the performance was reported in W-MAE (Man et al. 2023). We can see that the inference results of EWMoE are greatly superior to that of FourCastNet with two-year dataset reported in W-MAE. To highlight the computation savings of EWMoE, we finally compare EWMoE trained on two-year dataset with other leading models trained on more than 30 years. Figures 2 and Figure 3 evaluate different methods on forecasting four key weather variables at lead time from 1 to 8 days in terms of ACC and RMSE, respectively. The results show that EWMoE has both higher ACC and lower RMSE than FourCastNet (Kurth et al. 2023) and ClimaX (Nguyen et al. 2023) for all the variables analyzed. For short-range forecasting (1-3

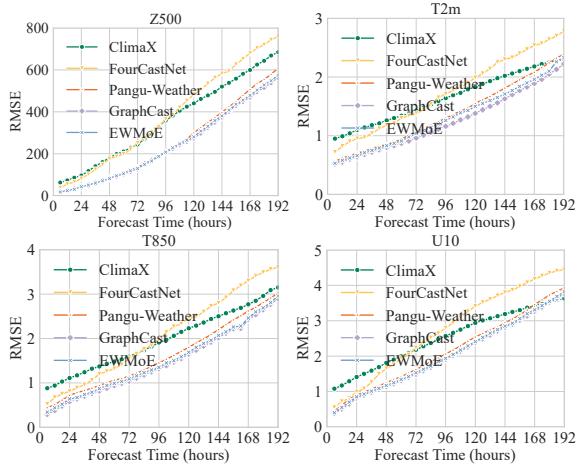


Figure 3: Latitude-weighted RMSE results of EWMoE and the baselines predicting four key variables $Z500$, $T2m$, $T850$ and $U10$ in 2018 (lower RMSE is better).

days), EWMoE demonstrates a comparable level of forecasting accuracy as Pangu-Weather (Bi et al. 2022). In addition, as the forecast time increases, significant improvement with our EWMoE is observed and EWMoE outperforms Pangu-Weather from day 3, demonstrating EWMoE’s remarkable ability and stability for short-to-medium-range weather forecasting. Compared with GraphCast, each model has its own advantages. In terms of the ACC metric, EWMoE performs better than GraphCast, while in terms of the RMSE metric, GraphCast is slightly better than EWMoE. This result may be attributed to the fact that GraphCast uses a 12-step autoregressive finetuning strategy to reduce the error accumulation in long lead predictions but increases the consumption of training resources at the same time.

Moreover, we note that EWMoE achieves this strong performance with much less training data and computing resources compared with the baselines. We train our EWMoE on 2 years of training data, which is approximately $18\times$ less data than FourCastNet and ClimaX’s 37 years of training data, and $120\times$ less than that used for Pangu-Weather and GraphCast, which use 39 years of training data with 13 pressure levels. The training of EWMoE was completed under 9 days on 2 Nvidia 3090 GPUs. In contrast, FourCastNet took 16 hours to train on 64 A100 GPUs, ClimaX took 7 days on 80 V100 GPUs, Pangu-Weather took 64 days on 192 V100 GPUs and GraphCast took 4 weeks on 32 Google Cloud TPU v4 devices. Our novel weather forecasting MoE model generates accurate forecasts with much less training data and computational cost, which will facilitate future works that build upon our proposed framework.

Visualization

We visualize the predicted results of EWMoE at lead days 1, 3, 7 for two variables, $Z500$ (geopotential at the pressure level of 500 hPa) and $U10$ (the 10m zonal wind velocity), and compare the results with the ERA5 ground-truth. The initial time is 00:00 UTC, January 15th, 2018. In Figure 4 and

Figure 5, the first column shows the ERA5 ground-truth at that lead day, the second column shows the prediction result, and the third column shows the bias, which is the difference between the prediction result and the ground-truth. These visualizations validate our model’s ability to predict future weather states close to the ground-truth.

Ablation Studies

We analyze the importance of individual elements in EWMoE by removing one component at a time and observing the performance difference.

Effect of 3D absolute position embedding. We conduct experiments to compare the performance of EWMoE with and without 3D absolute position embedding to evaluate its effectiveness in extracting the 3D geographical location features. Figure 6a shows the superior performance of 3D absolute position embedding compared with the standard ViT position embedding at all forecast time, indicating that it is a crucial component in modeling geographical characteristics of different meteorological variables.

Effect of the MoE layer. We evaluate the effectiveness of the MoE layer in EWMoE. As shown in Figure 6b, EWMoE with the MoE layer significantly outperforms model with a standard feed-forward network, and the performance gap becomes larger as the forecast time increases. We attribute this result to the ability of the MoE layer, which allows increasing the model capacity and flexibility. The total number of model parameters with only one FFN layer is 43 million. After using the MoE layer with 20 experts, the total number of model parameters has reached to 580 million. The total number of model parameters has increased by 13 times, allowing EWMoE to better extract and model meteorological data features. We also note that EWMoE achieves this improvement without increasing computing requirements, as only a small portion of parameters are activated during training. This suggests that applying MoE structure to weather forecasting is promising. Moreover, we also conduct extensive experiments to evaluate the importance of auxiliary loss used in the MoE layer routing and the number of selective top- k experts, as shown in Figure 7.

Conclusion

In this paper, we introduce EWMoE, an advanced and effective deep learning model for weather forecasting. By integrating three novel components, 3D absolute position embedding, an MoE layer and two specific loss functions, it excels at a resolution of 0.25° and forecast time of up to 8 days, outperforming the leading models such as FourCastNet and ClimaX, and competing well with Pangu-Weather and GraphCast in short-range forecasting. It is worth mentioning that EWMoE achieves this superior performance with significantly less training data and computing resources, addressing the challenges of computational efficiency and prediction accuracy. Our study also provides insights for modeling the interactions among atmospheric variables, demonstrating the feasibility and potential of implementing the MoE paradigm in weather forecasting tasks. We hope that our work will inspire future work on applying effective MoE architecture to a wider range of climate researches.

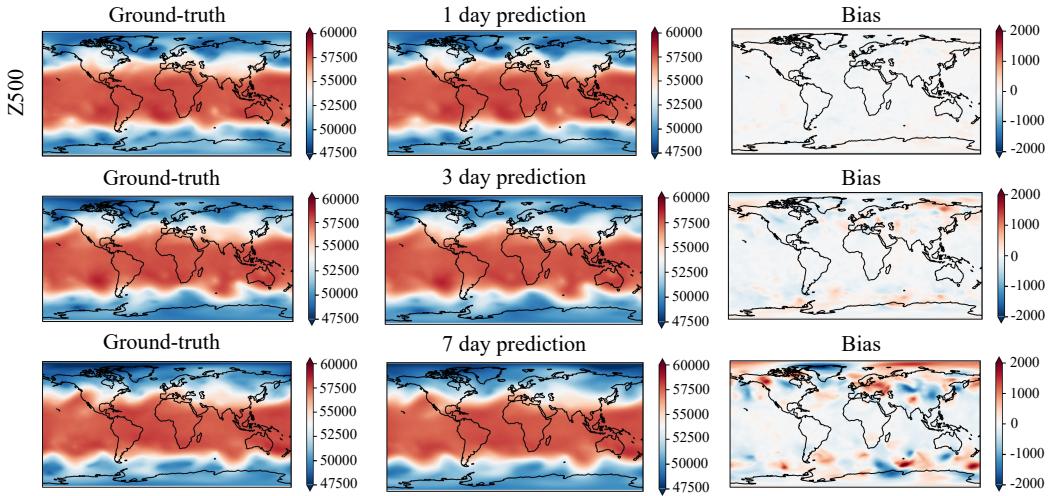


Figure 4: Visualization examples of future state prediction for Z500 compared with ground-truth.

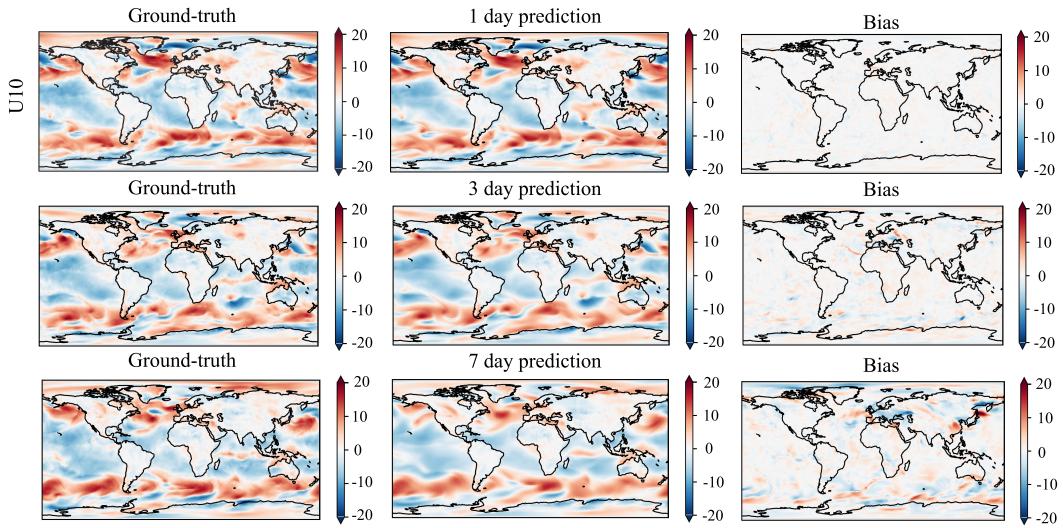


Figure 5: Visualization examples of future state prediction for U10 compared with ground-truth.

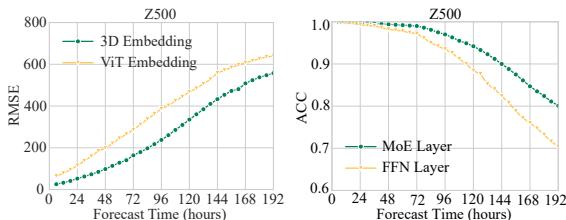


Figure 6: Ablation studies showing the importance of each component in EWMoE. Similar trends are observed across different variables.

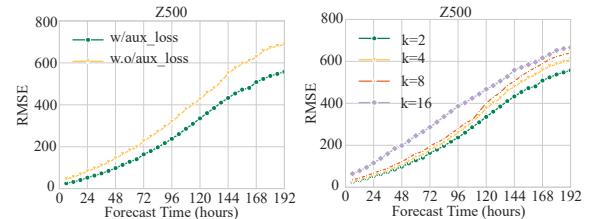


Figure 7: EWMoE improves consistently with auxiliary loss (left) and smaller k (right).

Acknowledgements

This work was partially supported by the National Key R&D Program of China (No. 2022YFC3003904), Yibin Science and Technology Program (No. 2024SF011), Heavy Rain and Drought-Flood Disasters in Plateau and Basin Key Laboratory of Sichuan Province (No. SCQXKJYJXMS202314) and Southwest Sichuan (Ya'an) Rainstorm Laboratory (No. CXNBYSYSZD202404).

References

- Abbe, C. 1901. The physical basis of long-range weather forecasts. *Monthly Weather Review*, 29(12): 551–561.
- Allen, M.; Kettleborough, J.; and Stainforth, D. 2002. Model error in weather and climate forecasting. In *ECMWF Predictability of Weather and Climate Seminar*, 275–294.
- Bauer, P.; Thorpe, A.; and Brunet, G. 2015. The quiet revolution of numerical weather prediction. *Nature*, 525: 47–55.
- Beljaars, A.; Balsamo, G.; Bechtold, P.; Bozzo, A.; Forbes, R.; Hogan, R. J.; Köhler, M.; Morcrette, J.-J.; Tompkins, A. M.; Viterbo, P.; and Wedi, N. 2018. The numerics of physical parametrization in the ECMWF model. *Frontiers in Earth Science*, 6: 137.
- Best, M. J. 2005. Representing urban areas within operational numerical weather prediction models. *Boundary-Layer Meteorology*, 114: 91–109.
- Bi, K.; Xie, L.; Zhang, H.; Chen, X.; Gu, X.; and Tian, Q. 2022. Pangu-weather: A 3d high-resolution model for fast and accurate global weather forecast. *arXiv preprint arXiv:2211.02556*.
- Bi, K.; Xie, L.; Zhang, H.; Chen, X.; Gu, X.; and Tian, Q. 2023. Accurate medium-range global weather forecasting with 3D neural networks. *Nature*, 619: 533–538.
- Bjerknes, V.; Volken, E.; and Bronnimann, S. 2009. The problem of weather prediction, considered from the viewpoints of mechanics and physics. *Meteorologische Zeitschrift*, 18(6): 663–667.
- Bougeault, P.; Toth, Z.; Bishop, C.; Brown, B.; Burridge, D.; Chen, D. H.; Ebert, B.; Fuentes, M.; Hamill, T. M.; Mylne, K.; Nicolau, J.; Paccagnella, T.; Park, Y.-Y.; Parsons, D.; Raoult, B.; Schuster, D. C.; Dias, P. S.; Swinbank, R.; Takeuchi, Y.; Tennant, W.; Wilson, L.; and Worley, S. 2010. The THOR-PEX interactive grand global ensemble. *Bulletin of the American Meteorological Society*, 91(8): 1059–1072.
- Chen, K.; Han, T.; Gong, J.; Bai, L.; Ling, F.; Luo, J.-J.; Chen, X.; Ma, L.; Zhang, T.; Su, R.; Ci, Y.; Li, B.; Yang, X.; and Ouyang, W. 2023a. Fengwu: Pushing the skillful global medium-range weather forecast beyond 10 days lead. *arXiv preprint arXiv:2304.02948*.
- Chen, L.; Zhong, X.; Zhang, F.; Cheng, Y.; Xu, Y.; Qi, Y.; and Li, H. 2023b. FuXi: a cascade machine learning forecasting system for 15-day global weather forecast. *npj Climate and Atmospheric Science*, 6: 190.
- Clare, M. C. A.; Jamil, O.; and Morcrette, C. J. 2021. Combining distribution-based neural networks to predict weather forecast probabilities. *Quarterly Journal of the Royal Meteorological Society*, 147(741): 4337–4357.
- Denby, L. 2020. Discovering the importance of mesoscale cloud organization through unsupervised classification. *Geophysical Research Letters*, 47(1): e2019GL085190.
- Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.
- Dueben, P. D.; and Bauer, P. 2018. Challenges and design choices for global weather and climate models based on machine learning. *Geoscientific Model Development*, 11: 3999–4009.
- Gan, L.; Man, X.; Li, C.; She, L.; and Shao, J. 2019. W-MRI: A Multi-output Residual Integration Model for Global Weather Forecasting. In *Web and Big Data - 7th International Joint Conference, APWeb-WAIM 2023*, 209–222.
- Guibas, J.; Mardani, M.; Li, Z.; Tao, A.; Anandkumar, A.; and Catanzaro, B. 2021. Adaptive fourier neural operators: Efficient token mixers for transformers. *arXiv preprint arXiv:2111.13587*.
- Hersbach, H.; Bell, B.; Berrisford, P.; Hirahara, S.; Horányi, A.; Muñoz-Sabater, J.; Nicolas, J.; Peubey, C.; Radu, R.; Schepers, D.; Simmons, A.; Soci, C.; Abdalla, S.; Abellan, X.; Balsamo, G.; Bechtold, P.; Biavati, G.; Bidlot, J.; Bonavita, M.; and Thepaut, J.-N. 2020. The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730): 1999–2049.
- Hu, Y.; Chen, L.; Wang, Z.; and Li, H. 2023. SwinVRNN: A Data-Driven Ensemble Forecasting Model via Learned Distribution Perturbation. *Journal of Advances in Modeling Earth Systems*, 15(2): e2022MS003211.
- Kalnay, E. 2002. *Atmospheric modeling, data assimilation and predictability*. Cambridge: Cambridge University Press.
- Keisler, R. 2022. Forecasting global weather with graph neural networks. *arXiv preprint arXiv:2202.07575*.
- Kochkov, D.; Yuval, J.; Langmore, I.; Norgaard, P.; Smith, J.; Mooers, G.; Klöwer, M.; Lottes, J.; Rasp, S.; Düben, P.; Hatfield, S.; Battaglia, P.; Sanchez-Gonzalez, A.; Willson, M.; Brenner, M. P.; and Hoyer, S. 2024. Neural general circulation models for weather and climate. *Nature*, 632: 1060–1066.
- Kurth, T.; Subramanian, S.; Harrington, P.; Pathak, J.; Mardani, M.; Hall, D.; Miele, A.; Kashinath, K.; and Anandkumar, A. 2023. Fourcastnet: Accelerating global high-resolution weather forecasting using adaptive fourier neural operators. In *Proceedings of the Platform for Advanced Scientific Computing Conference, PASC 2023*, 13:1–13:11.
- Lam, R.; Sanchez-Gonzalez, A.; Willson, M.; Wirnsberger, P.; Fortunato, M.; Alet, F.; Ravuri, S.; Ewalds, T.; Eaton-Rosen, Z.; Hu, W.; Merose, A.; Hoyer, S.; Holland, G.; Vinyals, O.; Stott, J.; Pritzel, A.; Mohamed, S.; and Battaglia, P. 2023a. Learning skillful medium-range global weather forecasting. *Science*, 382: 1416–1421.
- Lam, R.; Sanchez-Gonzalez, A.; Willson, M.; Wirnsberger, P.; Fortunato, M.; Alet, F.; Ravuri, S.; Ewalds, T.; Eaton-Rosen, Z.; Hu, W.; Merose, A.; Hoyer, S.; Holland, G.; Vinyals, O.;

- Stott, J.; Pritzel, A.; Mohamed, S.; and Battaglia, P. W. 2023b. Learning skillful medium-range global weather forecasting. *Science*, 382(6677): 1416–1421.
- Lepikhin, D.; Lee, H.; Xu, Y.; Chen, D.; Firat, O.; Huang, Y.; Krikun, M.; Shazeer, N.; and Chen, Z. 2021. Gshard: Scaling giant models with conditional computation and automatic sharding. In *9th International Conference on Learning Representations, ICLR 2021*.
- Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, 10012–10022.
- Lorenc, A. C. 1986. Analysis methods for numerical weather prediction. *Quarterly Journal of the Royal Meteorological Society*, 112(474): 1177–1194.
- Lynch, P. 2008. The origins of computer weather prediction and climate modeling. *Journal of computational physics*, 227(7): 3431–3444.
- Magnusson, L.; and Källén, E. 2013. Factors influencing skill improvements in the ECMWF forecasting system. *Monthly Weather Review*, 141(9): 3142–3153.
- Man, X.; Zhang, C.; Feng, J.; Li, C.; and Shao, J. 2023. W-mae: Pre-trained weather model with masked autoencoder for multi-variable weather forecasting. *arXiv preprint arXiv:2304.08754*.
- Navon, I. M. 2009. *Data assimilation for numerical weather prediction: a review*, 21–65. Berlin, Heidelberg: Springer.
- Nguyen, T.; Brandstetter, J.; Kapoor, A.; Gupta, J. K.; and Grover, A. 2023. Climax: A foundation model for weather and climate. In *International Conference on Machine Learning, ICML 2023*, 25904–25938.
- Nguyen, T.; Shah, R.; Bansal, H.; Arcomano, T.; Madireddy, S.; Maulik, R.; Kotamarthi, V.; Foster, I.; and Grover, A. 2024. Scaling transformer neural networks for skillful and reliable medium-range weather forecasting. In *ICLR 2024 Workshop: Tackling Climate Change with Machine Learning*.
- Price, I.; Sanchez-Gonzalez, A.; Alet, F.; Andersson, T. R.; El-Kadi, A.; Masters, D.; Ewalds, T.; Stott, J.; Mohamed, S.; Battaglia, P.; Lam, R.; and Willson, M. 2023. Gencast: Diffusion-based ensemble forecasting for medium-range weather. *arXiv preprint arXiv:2312.15796*.
- Rasp, S.; Dueben, P. D.; Scher, S.; Weyn, J. A.; Mouatadid, S.; and Thuerey, N. 2020. WeatherBench: a benchmark data set for data-driven weather forecasting. *Journal of Advances in Modeling Earth Systems*, 12(11): e2020MS002203.
- Rasp, S.; Hoyer, S.; Merose, A.; Langmore, I.; Battaglia, P.; Russel, T.; Sanchez-Gonzalez, A.; Yang, V.; Carver, R.; Agrawal, S.; Chantry, M.; Bouallegue, Z. B.; Dueben, P.; Bromberg, C.; Sisk, J.; Barrington, L.; Bell, A.; and Sha, F. 2023. Weatherbench 2: A benchmark for the next generation of data-driven global weather models. *arXiv preprint arXiv:2308.15560*.
- Rasp, S.; and Thuerey, N. 2021. Data-driven medium-range weather prediction with a resnet pretrained on climate simulations: A new model for weatherbench. *Journal of Advances in Modeling Earth Systems*, 13(2): e2020MS002405.
- Ren, X.; Li, X.; Ren, K.; Song, J.; Xu, Z.; Deng, K.; and Wang, X. 2021. Deep learning-based weather prediction: a survey. *Big Data Research*, 23: 100178.
- Robert, A. 1982. A semi-Lagrangian and semi-implicit numerical integration scheme for the primitive meteorological equations. *Journal of the Meteorological Society of Japan. Ser. II*, 60(1): 319–325.
- Scher, S. 2018. Toward data-driven weather and climate forecasting: Approximating a simple general circulation model with deep learning. *Geophysical Research Letters*, 45(22): 12,616–12,622.
- Schultz, M. G.; Betancourt, C.; Gong, B.; Kleinert, F.; Langguth, M.; Leufen, L. H.; Mozaffari, A.; and Stadtler, S. 2021. Can deep learning beat numerical weather prediction? *Philosophical Transactions of the Royal Society A*, 379(2194): 20200097.
- Shazeer, N.; Mirhoseini, A.; Maziarz, K.; Davis, A.; Le, Q.; Hinton, G.; and Dean, J. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer.
- Simmons, A. J.; and Hollingsworth, A. 2002. Some aspects of the improvement in skill of numerical weather prediction. *Quarterly Journal of the Royal Meteorological Society*, 128(580): 647–677.
- Stensrud, D. J. 2007. *Parameterization schemes: keys to understanding numerical weather prediction models*. Cambridge: Cambridge University Press.
- Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017*, 5998–6008.
- Weyn, J. A.; Durran, D. R.; and Caruana, R. 2019. Can machines learn to predict weather? Using deep learning to predict gridded 500-hPa geopotential height from historical weather data. *Journal of Advances in Modeling Earth Systems*, 11(8): 2680–2693.
- Weyn, J. A.; Durran, D. R.; and Caruana, R. 2020. Improving data-driven global weather prediction using deep convolutional neural networks on a cubed sphere. *Journal of Advances in Modeling Earth Systems*, 12(9): e2020MS002109.
- Weyn, J. A.; Durran, D. R.; Caruana, R.; and Cresswell-Clay, N. 2021. Sub-seasonal forecasting with a large ensemble of deep-learning weather prediction models. *Journal of Advances in Modeling Earth Systems*, 13(7): e2021MS002502.
- Xu, Z.; Du, J.; Wang, J.; Jiang, C.; and Ren, Y. 2019. Satellite image prediction relying on GAN and LSTM neural networks. In *2019 IEEE International Conference on Communications, ICC 2019*, 1–6.