Response to Reviewer

August 8, 2024

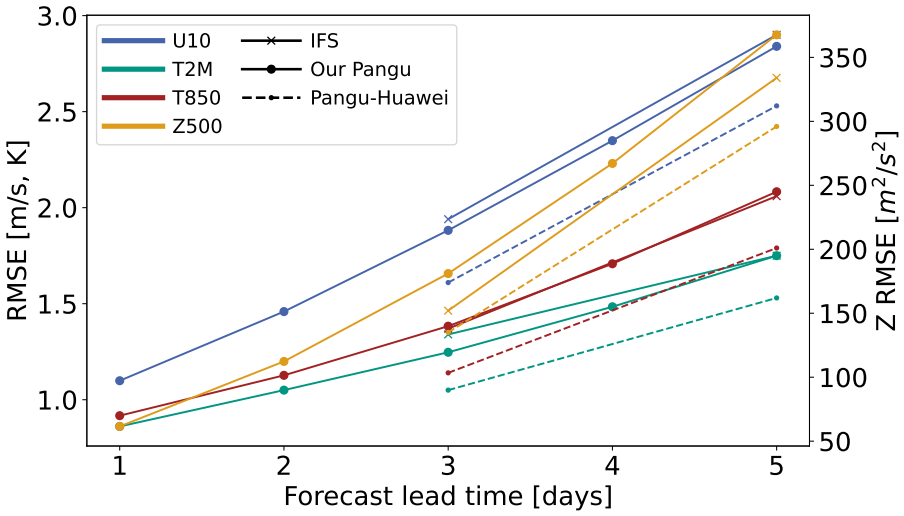
Responses are marked in blue.

General comments

- Figure 3 is an important chart supporting the validity of this research. However, it does not include specific humidity, Z500, or V10, as mentioned earlier in this paper. Including these variables could strengthen the argument for the effectiveness of the 2D-Transformer. Due to the 6-hour subsample, readers ultimately do not know if the 2D-Transformer has improved the forecast accuracy of the original Pangu-Weather model. Including such comparisons could significantly increase the citation rate of this paper. There are still certain changes and clarifications that the authors should address prior to publication. For these reasons, I believe that the manuscript can be accepted for publication. Below, I have some specific comments to the authors.

  Thank you for your insight. The RMSE values from the official Pangu-Weather are obtained from their summary performance information in their Github repository (Bi et al., 2023), where detailed performance metrics of Z500, U10, T2M, and T850 are reported. V10 is missing from this table. We have included the evaluation of Z500 in Figure 3, indicated by the yellow line. Our version of Pangu did not outperform IFS for Z500. We have modified the text accordingly:

  "We observe that our PGW model performs better than IFS for three variables (U10, T2M, T850) over the five-day forecast but performs worse for the Z500 variable. There is still a notable difference between the performance of our model and the published PGW model. We attribute that to training the model on only a 6 h-subsample of the total data trained by the original authors, as well as slight differences in training procedure such as batch size."



Specific comments

- Line #2 - #5, the sentence is too long and difficult to read. It can be revised to "The Transformer-based PGW introduced novel architectural components, including the three-dimensional attention mechanism (3D-Transformer) in the Transformer blocks. Additionally, it features an Earth-specific positional bias term that accounts for weather states being related to the absolute position on Earth."

  Thank you for the suggestion. We have revised the sentence "The Transformer-based PGW introduced novel architectural components including the three-dimensional attention mechanism (3D-Transformer) in the Transformer blocks and an Earth-specific positional bias term..." to "The Transformer-based PGW introduced novel architectural components including the three-dimensional attention mechanism (3D-Transformer) in the Transformer blocks. Additionally, it features an Earth-specific positional bias term..."

- Line #24, "the authors also admit" could be replaced with more specific wording, such as "previous studies have shown". The same issue appears in Line #91, where the architecture described "by the authors" could be replaced with "in this study." This sentence reads as if the ablation study is original to this paper and not derived from the model itself. If this is the case, some references could be cited here as evidence to support the experiment design.

  Thank you for your comment. "The authors", in this case, refers to Bi et al., who published the original Pangu-Weather model. Following your suggestion, we have changed line 24 to "The authors of PGW admit that the models have not arrived at full convergence (Bi et al., 2023a)".

  We have also changed Line 96 to specifically reference the authors of Bi et al. as follows:

"...an ablation study was performed based on the PGW-Lite architecture described by Bi et al. (2023) rather than the full PGW model."

- Line #27, the published model cannot be run, what is the reason? Is it also caused by modularized manner issue? Does it conflict with reproduction introduced in section 2.3?

  Thank you for this question. The reason why the published model code cannot be run is that the original authors Bi et al. (2023) did not publish their model training code, but only an inference code that is based on the already trained model weights. The pseudocode that is published by the authors outlines the general architecture of the model, but is written in a way that requires heavy modification and re-implementation before the code can be run. Furthermore, all other publicly available re-implementations of Pangu-Weather are either less complete (in terms of implementation or documentation) than ours, and would have required significant modification to perform our targeted ablation study.

  We have modified the body of the text as follows: "The pseudocode outlines the architecture, but is not complete Python code that can be run without major modification."

- Figure 1, 4, 5 and 7 could be appropriately enlarged. Some images are difficult to discern even when enlarged. For Figure 4 (a), the authors can separate the lines by adjusting the y-coordinates.

  We apologize for this and thank you for the suggestions. Based on your feedback, we have changed the y-axis to increase the readibility of Figure 4.

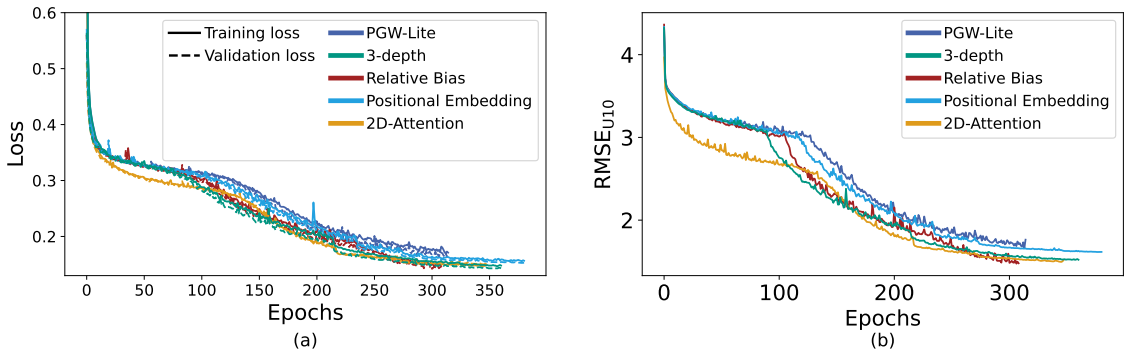


Figure 4: Training curves of ablation study for PGW-Lite, 3-depth model, relative bias, positional embedding, and 2D-Attention models. (a) Training and validation loss values as a function of the epochs. (b): RMSE for the 10m-U-velocity as a function of the epochs, as evaluted for the validation samples.

We have increased the size of Figures 1, 4, 5, 6, and 7.

We have also split up Figure 6 into two subfigures to facilitate parsing the information in the plots more easily:

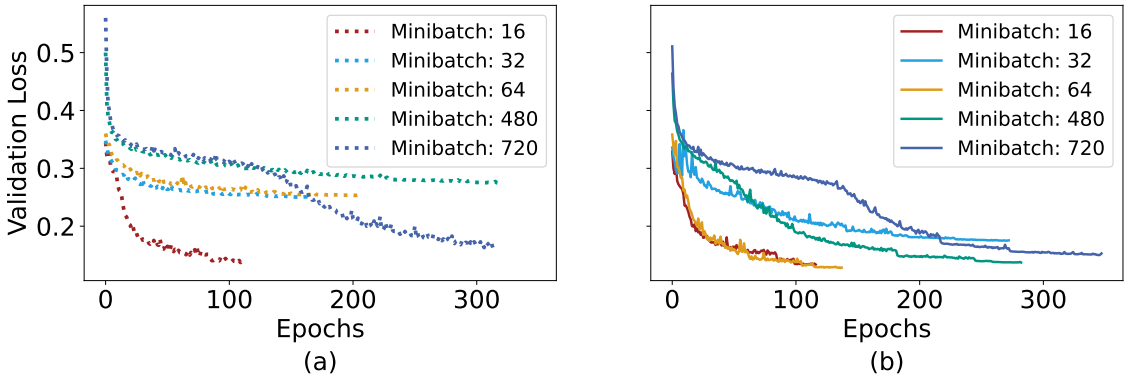


Figure 6. Validation loss of 2D-Attention and PGW-Lite as a function of epochs for global minibatch sizes of 16, 32, 64, 480, and 720.

- Figure 1, the cosine functions for each variable with weights from Bi et al. (2023) could be listed to explain the normalized process at each epoch. In Figure 1(b), it would be helpful to present the equation for the sloped MSP graph to facilitate understanding.

  We thank the reviewer for their observation and apologize for the confusion. The initial weights are distributed according to normal cosine functions. We also design the cosine distributions such that the velocity fields in the surface fields peak at the same time as in the pressure levels. However, at every epoch, we need to normalize the sum of the weights to match those of the original weight sums.

  The reason that the peaks are offset is due to this weight normalization. Taking the U/V in plot (b) for instance: according to the regular cosine scheduling, the peaks should be at epochs 0, 100, and 200. However, since the weight of MSP at these points has a higher proportion of the overall weight, the relative weight of U/V is decreased. At epoch 85, the weights of MSP and T2M are small, meaning that the U/V scales to a greater value.

  We have included a more explicit formulation of the weight distribution in the manuscript as follows:

  "Given an individual variable i at epoch e, for a period of n_{epochs} (in this case, $n_{\text{epoch}} = 100$). For the pressure variables Z, Q, T, U, V, $i = 1, 3, 2, 0, 0$, respectively. For the surface variables MSLP, U10, V10, T2M, $i = 1, 0, 0, 2$, respectively.

$$w(i, e) = \cos\left(\frac{2\pi}{n_{\text{epoch}}}\left(e - \frac{i}{2}\right)\right) + 1.1 \tag{1}$$

  At each epoch, the pressure variables are normalized to sum up to the original weights of $3.00 + 0.60 + 1.50 + 0.77 + 0.54$.

$$W(i, e) = w(i, e) \cdot \frac{3.00 + 0.60 + 1.50 + 0.77 + 0.54}{\sum_i w(i, e)} \tag{2}$$

  A similar procedure is applied to the surface variables, where the numerator is replaced with

$1.5 + 0.77 + 0.66 + 3.0$. The ordering of the weights was designed such that the weights of the velocity and temperature variables (U, U10; V, V10; T, T2M) would peak at similar epochs."

- Figure 6: Could the author explain the reason for the failure to converge based on PGW-Lite structure with minibatch sizes of 32, 64, and 480? PGW-Lite with minibatch size 720 eventually converged. Could the authors explain this unexpected result? Part of the reason is explained in Line #229. It is not necessary to strictly separate the results and discussion sections. Explaining part of the findings in the result section can enhance the content.

  We apologize for the lack of detail in our original submission regarding these results and their interpretation. We gladly shed some more light on that aspect: Given the irregularity in the final loss values for the PGW-Lite model, we hypothesize that the use of different random seeds can have a significant effect on the attainable loss. Due to computational constraints, we did not perform multiple repetitions to study the effect of different seeds on the final loss values of these models. To address this, we have added the following explanation to Section 3.3:

  "In contrast, the PGW-Lite models converge at sub-optimal loss values for minibatch sizes of 16, 32, and 480. In these cases, the models were allowed to train until the learning rate dropped to $3 \cdot 10^{-5}$. The irregularity of this behavior can be attributed to the model's sensitivity to the initial random seed."

- Line #255, the sentence could be updated to "since wind vectors, acting as pressure gradients, can drive certain atmospheric processes, such as advection terms in atmospheric variables."

  We thank the reviewer for their suggestion in improving our phrasing. We have implemented the suggestion in the body of the text:

  "This causes the wind velocity to be a major driver for other atmospheric variables since wind vectors, acting as pressure gradients, can drive certain atmospheric processes, such as advection terms in atmospheric variables"

Other suggestions

Following are suggestions and do not affect the validity of the argument in this paper.

- Line #98: the models could be compared in more details in Table 1, like the hidden dimension, etc.

  We thank the reviewer for their suggestion. To enhance the clarity, we have followed the suggestion and added the hidden dimension into Table 1.

Table 1: Overview of models included in ablation study. All models are based on PGW-Lite.

Model	Number of parameters (millions)	Hidden dimension
Pangu-Weather-Lite (absolute bias)	44.6	192
Relative bias	24.3	192
Positional embedding	24.3	192
2D-Attention	57.2	288
Three-depth	108.9	192

Technical corrections

- Figure 5 y axis could be updated into "normalized RMSE"

  Thank you for your detailed observation. We have modified the figure accordingly.

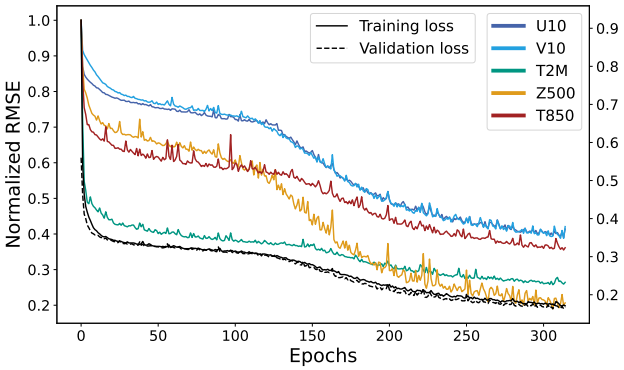


Figure 5. RMSE as a function of epochs for different prognostic variables for the PGW-Lite, as evaluated on the validation dataset. All values are normalized by their maximum RMSE value.