Response to Reviewer

August 8, 2024

Responses are marked in blue.

Comments

- The model names used in the tables and in the text are not consistent. Table 1 and Table 2 might be merged with unique model names.

  We apologize for our oversight. We have now consolidated the names of the models within Tables 1 and 2.

Table 1: Overview of models included in ablation study. All models were 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

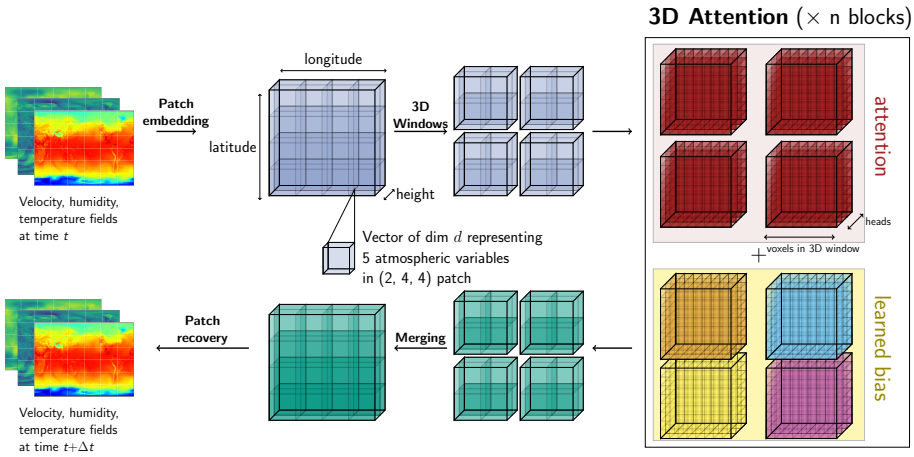
Table 2: Best validation loss and total epochs for the different models

Model	Best Validation loss	Epoch #
Pangu-Weather-Lite (absolute bias)	0.143	360
Relative Bias	0.143	300
Positional Embedding	0.152	358
2D-Attention	0.148	263
Three-depth	0.143	346

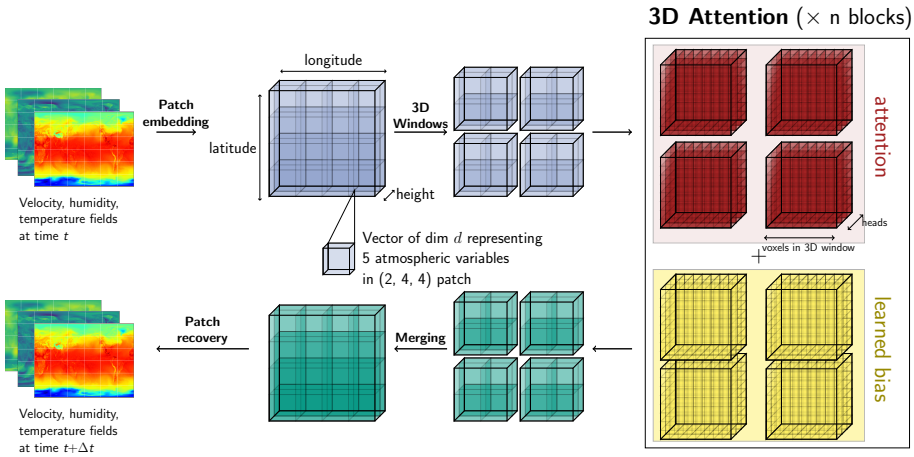
- Visualisation of the modification in model architecture: I find the modularised code in the repository very helpful. Could the code be included as pseudo code in the paper giving a comparative overview of the architectural details of the different models? This would be helpful in connection with the graphics in the original Pangu Publication (Fig. 2).

  Thank you for your comment. We have provided a visualization of the different architectures

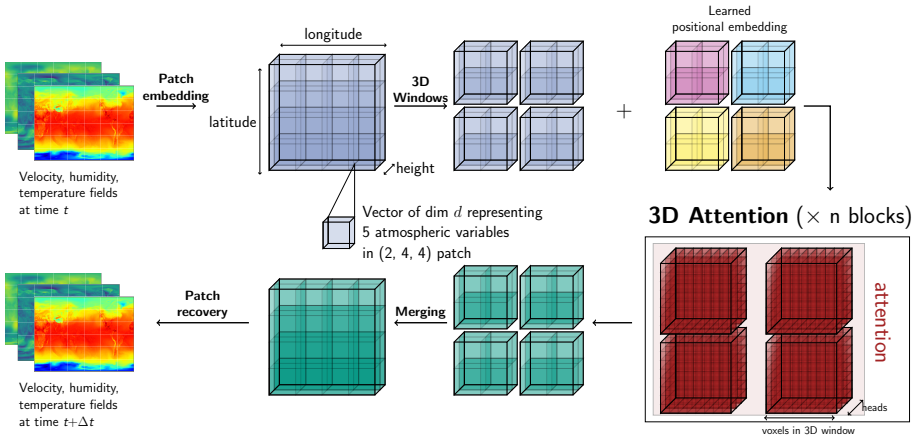
in the following figure. For simplification and clarity's sake, these diagrams neither include the up and downsampling layers, nor the Sliding-Window Transformer (SWIN) architecture. This is also why the Three-depth model is not depicted here, as its architecture is largely the same as Pangu-Weather Lite, albeit with an additional up/downsampling layer.



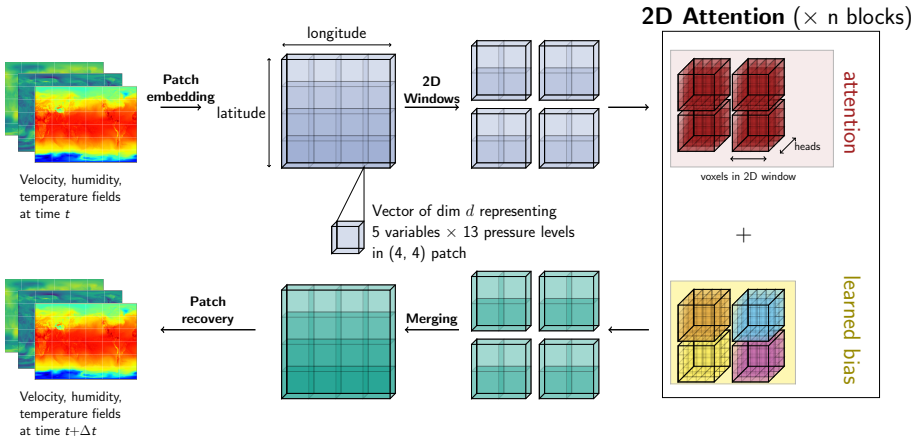
(a) Pangu-Weather



(b) Relative bias



(c) Learned positional embedding



(d) 2D-Attention

Figure: Architecture of Pangu-Weather model and variations considered in the ablation study. The sliding window mechanism (SWIN) is not depicted in the figure for simplicity. The up- and downsampling layers are also excluded from the figure.

- Parameter numbers in Table 1: The numbers of parameters for the 2d attention model s larger than for the 3d attention (PanguLite) in Table 1. This is counter intuitive. Is it due to the fact that the hidden dimension C was enlarged? What was the reasoning behind that choice? Could it be chosen such that the overall parameter size would match that of PanguLite again? Is this dimension C the same for PanguLite and Pangu? Could the authors extrapolate the parameter numbers in Table 1 for the original Pangu model with the original batch size?

  Thank you for your insight. This is explained in the end of Section 2.5 with the following text:

  "To compensate for the reduction in parameters associated with a 2D-Transformer model, the hidden dimension was increased from C = 192 to C = 288, increasing the dimension of the hidden layer to 48 per attention head as opposed to the original 32. Note that the latent space in the 2D-Transformer now encodes 5 variables × 13 pressure levels instead of just 5 variables × 1 pressure level as in the 3D case."

  To enhance the clarity, we have followed your suggestion and added the hidden dimension into Table 1.

Table 1: Overview of models included in ablation study. All models were 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

- Parameter numbers in Table 3 (relating to Remark 3): In paragraph 2.5 the authors state that reducing the model size allows for larger local batches. Hence, PanguLite should have larger local batches than the 2d version. In Table 3 it is the other way round. Could the authors please clarify this?

  We apologize for the confusion. While it is true that increasing the model size will allow for more data samples to fit on a single GPU, the main memory cost in training Transformer models comes from the attention block. When the models are performing the forward and backward passes, many intermediate states need to be retained in memory on the GPU—this explains why even though the model weights have a size of about 500 MB, and input data has a size of about 250 MB, 20 GB of memory is required on the GPU during one forward and backward pass. As the 2D-Attention mechanism essentially halves the memory requirement of attention (since attention is computed over $1 \times 6 \times 12$ windows instead of $2 \times 6 \times 12$), 2D-Attention requires less memory on the GPU.

  This is explained in section 4.1 as follows:

  "by using 2D-attention, the individual attention blocks require much less memory. Specifically, the attention mechanism requires an $\mathcal{O}(n^2)$ memory requirement with respect to the size of the sequence (Vaswani et al., 2017). Reducing the attention blocks to perform attention over

patches from (2, 6, 12) to (1, 6, 12) in the 2D-Transformer has allowed us to double the local minibatch size per GPU, reducing the total computational requirements."

- Fig. 1: As the curves for U and V are indistinguishable, one colour for both curves would render the figure clearer.

  Thank you for your attention to detail. We have combined the color of U and V (in subfigure (a)) and the color of U10 and V10 (in subfigure (b)).

- Furthermore, plot a) and b) should display temperature and wind in the same colour.

  We have chosen to use different colors for T and T2M; and U and U10; because these colors are consistent with earlier figures from the rest of the manuscript, i.e., Figure 1 and 3.

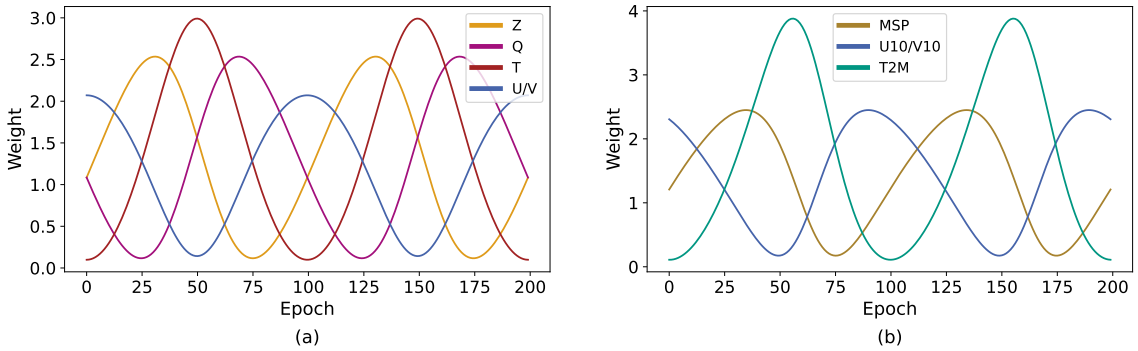


Figure 1. The weights for variable-specific cosine loss scheduling. The period is 100 epochs. The U- and V-velocity fields for both the pressure-level variables and surface variables are superimposed on one another, since they always receive the same values. The weights are normalized to sum up to the sum of the original weights presented in Bi et al. (2023a). (a) Pressure-level variables. (b) Surface variables.

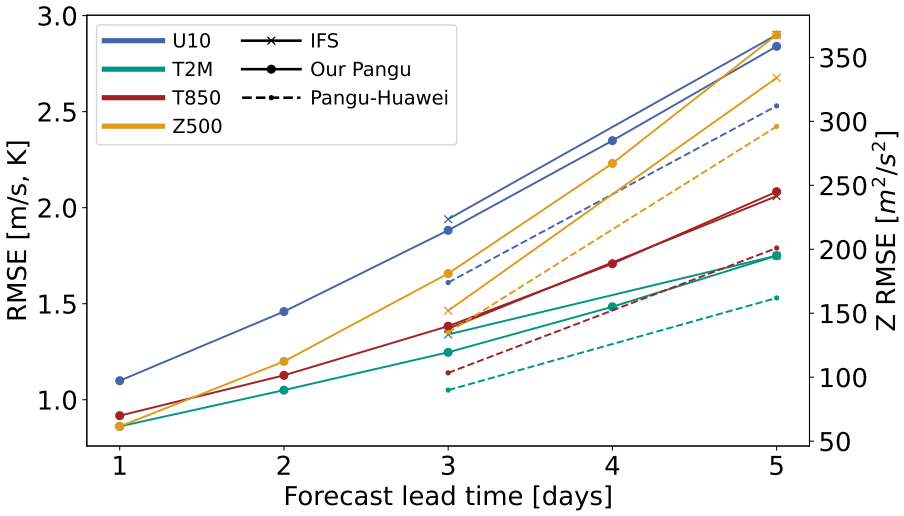


Figure 3. RMSE as a function of forecast lead time of key prognostic variables (T2M, U10, T850) in our model compared to IFS and Pangu-Weather on testing data from 2020–2021.

- Figure 7: What does it mean that the reference model is PanguLite? The text implies that the two curves show both the 2d model with different training losses.

  We apologize for the mistake. We have updated the figure caption to the following:

  "Error metrics for variable-specific weighted cosine loss scheduling applied to the 2D-Attention model. The 2D-Attention model with PyTorch's ReduceLROnPlateau scheduler was maintained as the reference model and compared to the variable-specific weighted cosine loss scheduling model. (a): RMSE values of the 10m-U- and V-velocity, as evaluated on the validation dataset, as a function of the epochs. (b): training and validation losses for the two models."