Review Comments

This manuscript proposes AtmoNAT, a Neighborhood-Attention-based Transformer for data-driven weather forecasting. Within each Transformer feature-extraction submodule, the authors augment Neighborhood Attention with a Gated Relative Position Encoding (GRPE) and replace the conventional feed-forward network with the proposed Spatial-Aware Feed-Forward Network (SAFN). In general, I find the proposed GRPE and SAFN scientifically interesting, and the experimental results indicate that both components can improve weather-prediction performance.

However, during my reading, I found that the overall narrative and logic of the manuscript are not consistently organized. Different sections sometimes use mismatched wording, make claims at different levels, or present arguments that do not fully connect with one another. Also, I find that the central goal of the paper shifts throughout the manuscript. It is not always clear whether the authors intend to propose AtmoNAT as a complete weather-forecasting model or primarily to demonstrate that GRPE and SAFN can improve a conventional Neighborhood Attention Transformer. The current experimental structure and ablation results tend to support the latter claim.

Overall, I consider the proposed GRPE and SAFN to have valid scientific potential. My main recommendation is that the authors clarify the central contribution, carefully revise the wording and logic throughout the manuscript, and restructure the presentation so that the motivation, methods, experiments, discussion, and conclusion consistently support the same scientific claim.

Below are the detailed line-to-line comment

Line-by-Line Comments

Abstract, Lines 14–16, and Introduction, Lines 40–60

In the abstract, the authors state that existing Transformer-based architectures use positional encodings that “typically embed limited spatial and temporal context, failing to fully account for the time variability, directionality, and location-dependency inherent in atmospheric motions.” Lines 40–60 appear to justify this statement by discussing the positional representations used in several recent data-driven weather models. However, I feel that this limitation is stated too broadly, and the discussion in Lines 40–60 is not sufficiently comprehensive to support such a general claim. Although the limitation is grammatically attributed to the positional encodings rather than to the Transformer architecture itself, the current wording may still blur this distinction. Transformer architectures are not inherently limited in their ability to model spatial or temporal dependencies. There is extensive prior work on explicitly spatiotemporal Transformer architectures, especially in the computer vision field. TimeSformer applies self-attention across the spatial and temporal dimensions of video sequences (Bertasius et al., 2021); ViViT constructs spatiotemporal tokens and investigates factorized spatial and temporal attention (Arnab et al., 2021); and Earthformer develops a dedicated space–time Transformer for Earth-system forecasting (Gao et al., 2022). Lines 40–60 review position-encoding choices in several weather models but don't engage with this broader literature. It would help to clarify whether the stated limitation is about Transformers in general or specifically about the position-encoding schemes surveyed here.

Another point is that the authors move from describing several static or fixed positional directly to the conclusion that these approaches fail to capture time variability, directionality, and location dependency.

As written, it is not fully clear whether “fixed positional bias” is being treated as equivalent to “state-independent attention.” This is not necessarily true: even when the positional bias is fixed, self-attention remains state dependent because the query and key vectors depend on the instantaneous input state:

$$A_{ij}(X_t) = \mathrm{softmax}\left(\frac{Q_i(X_t)K_j(X_t)^\top}{\sqrt{d}} + b_{ij}\right).$$

Therefore, a fixed b_{ij} does not imply that the resulting attention pattern is fixed, time invariant, or non-directional. The specific distinction introduced by GRPE appears to be that the positional bias itself becomes conditional on the atmospheric state:

$$b_{ij} \longrightarrow b_{ij}(X_t).$$

If this is the intended methodological gap, the authors should state it explicitly and narrow the corresponding claims in both the abstract and Lines 40–60.

Lines 60–64:

The statement that Neighborhood Attention “significantly reduces the computational complexity of standard Transformer blocks” should be stated more specifically. First, it is unclear what the authors mean by “standard Transformer blocks.” If this refers to Transformers using global self-attention, the claim is valid. However, this advantage does not necessarily apply when Neighborhood Attention is compared with other localized attention mechanisms. Therefore, I suggest that the authors specify that the claimed complexity reduction is relative to global self-attention and applies specifically to the self-attention operation, rather than referring broadly to “standard Transformer blocks.”

Lines 72–80:

The statement that “gating mechanisms have not yet been explicitly applied in data-driven weather models” is too broad and should be revised. Gating has already been used in data-driven atmospheric modeling: FengWu-Adas employs gated convolution and gated cross-attention Chen et al., 2023. AtmosMJ introduces a Gated Residual Fusion mechanism Cheon, 2025. I'd suggest a more comprehensive literature review before keeping this claim.

Lines 80–85:

Based on the full manuscript, my understanding is that the authors are trying to introduce AtmoNAT as a new weather-forecasting architecture, rather than extending a previously established AtmoNAT model. However, the current wording, “We introduce a Spatial-Aware Feed-Forward Network (SAFN) to AtmoNAT”, can be read as though AtmoNAT already exists and the primary novelty of this work is simply the addition of SAFN. I suggest restructuring the introduction section, especially this part, so that the hierarchy of contributions is stated explicitly.

Lines 103–122

The statement that the latitude weight “decreases as latitude increases” should also be clarified, for example, what kind of weighting decrease strategies. The pressure-level weighting should likewise be specified. In addition, please clarify whether all preprocessing and normalization statistics are derived only from the training period,

Figure 1

The relationship between panels (a)–(c) of Figure 1 is a bit confusing. Panels (b) and (c) represent the detailed implementations of the GRPE-NA and SAFN submodules contained within each Transformer block in panel (a), but at first glance they can be interpreted as subsequent stages of the overall architecture. Adding an explicit zoom-in/inset connection from one Transformer block in panel (a) to panels (b) and (c), or visually indicating that GRPE-NA and SAFN are the two internal components of each Transformer submodule should really helps the understanding.

Section 2.2.3 (SAFN)

The novelty boundry of the proposed SASFN is not very clear. Based on reading, the authors seems to try presenting the entire SAFN formulation as a new module. However, severa;of its main architectural components resemble some existed gated convolutional feed-forward networks. For example, Restormer introduces a Gated-Dconv Feed-Forward Network with channel expansion, parallel branches, depthwise convolution, and multiplicative gating Zamir et al., 2022. And also NAFNet, which uses channel splitting and multiplicative feature gating through its SimpleGate mechanism Chen et al., 2022.

I therefore suggest that the authors clarify precisely which aspect of SAFN they intend to claim as novel. I do see potential novelty in applying the global positional bias β to the gating branch, which, to my knowledge, has not appeared in prior work. If this is intended to be the primary point of novelty, I suggest stating it explicitly and properly citing the prior work underlying the remaining components of the module.

Figure 2 line 230-245

Using ECMWF HRES as a normalization reference can be reasonable if the authors clarify and justify why this particular baseline was chosen. Additionally, a more convincing evaluation would include validation against independent observations. If the authors believe that using ECMWF HRES as a normalization reference is sufficient to substitute for such independent observational validation, they should clarify this and explain their reasoning.

Figure 3 line 245 - 264

The claim that AtmoNAT's ACC for wind velocity is "significantly higher" is not supported by any statistical test. If "significantly" is only used to describe the visual difference, the wording should be weakened. Similarly, the claims that AtmoNAT ranks second for Z500 and outperforms Stormer on T2m are difficult to verify because several curves overlap, especially between 24 and 48 h.

Figure 4 line 265 - 270:

I understand that the abstract emphasizes that AtmoNAT performs especially well over land, and that the later discussion connects this behavior to SAFN and the global positional bias β. Therefore, including a land-only experiment is useful for supporting that specific claim. However, AtmoNAT is presented throughout the paper as a global weather forecasting model, not a land-only model. To support that broader claim, the evaluation should also include performance over ocean areas, or at least the full global domain. If the model does not show the same advantage over ocean regions, this limitation should also be stated explicitly rather than leaving the reader with only the land-focused comparison.

Figure 5 line 275-285

Please clarify why the spatial forecast/error comparison is restricted to AtmoNAT, HRES, and Pangu-Weather, while the broader evaluation in Section 3.1 includes several other competing models such as NeuralGCM, Stormer, ArchesWeather, FuXi, and GraphCast.

Section 3.2 / Table 1

The ablation results shows that Neighborhood Attention itself provides little to no improvement over the Swin Attention baseline. And the more substantial gains come from the Multi-Scale Head, GRPE, and SAFN. This makes it unclear whether Neighborhood Attention is a critical component for the proposed AtmoNAT, or whether similar gains could be achieved by applying GRPE and SAFN into the original Swin-based backbone instead. I'd suggest the authors either clarify this distinction in the text, or provide extra experiment about applying GRPE and/or SAFN within a Swin-based backbone, to determine whether the reported gains are from the Neighborhood Attention backbone itself or can be transferable to other local-attention backbones. This would also help readers understand how much of AtmoNAT's overall competitive performance in Section 3.1 should be attributed to GRPE/SAFN specifically versus the backbone choice.

Section 4 / Figure 10, Lines 390–405:

Figure 10 demonstrates that GRPE learns asymmetric spatial patterns whose contributions vary with geographic location and time. However, the statement that the prototypes "explicitly encode the underlying directionality of atmospheric motions" implies a physical correspondence between the learned GRPE orientation and actual atmospheric flow direction. The current analysis only supports the claim that the prototypes exhibit geometric structure and spatial/temporal weighting. Unless an additional physical-matching experiment is provided, the authors should adjust the wording to avoid this implication.

Figure 11 / Section 4 line 405 – line 415:

Please use a unified color-bar range across the four panels. From the current visualization, the bias appears relatively smooth over most land and ocean regions, with stronger variations mostly over some high-altitude/mountainous areas. Therefore, the statement that the bias “strongly correlates with global terrain” seems too strong based on the figure alone.

minor format comment

Abstract, Lines 22: The format of the words” 1.5° spatial resolution, AtmoNAT’s deterministic” are different and it didn’t make sense to use different format for these words as well

References

Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M., & Schmid, C. (2021). ViViT: A Video Vision Transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (pp. 6836–6846). https://arxiv.org/abs/2103.15691

Bertasius, G., Wang, H., & Torresani, L. (2021). Is Space-Time Attention All You Need for Video Understanding? In *Proceedings of the 38th International Conference on Machine Learning*, *Proceedings of Machine Learning Research, 139*, 813–824. https://proceedings.mlr.press/v139/bertasius21a.html

Chen, K., Bai, L., Ling, F., Ye, P., Chen, T., Luo, J.-J., Chen, H., Xiao, Y., Chen, K., Han, T., & Ouyang, W. (2023). Towards an End-to-End Artificial Intelligence Driven Global Weather Forecasting System. *arXiv preprint arXiv:2312.12462*. https://arxiv.org/abs/2312.12462

Chen, L., Chu, X., Zhang, X., & Sun, J. (2022). Simple Baselines for Image Restoration. In *Computer Vision – ECCV 2022*. https://arxiv.org/abs/2204.04676

Cheon, M. (2025). AtmosMJ: Revisiting Gating Mechanism for AI Weather Forecasting Beyond the Year Scale. *arXiv preprint arXiv:2506.09733*. https://arxiv.org/abs/2506.09733

Gao, Z., Shi, X., Wang, H., Zhu, Y., Wang, Y., Li, M., & Yeung, D.-Y. (2022). Earthformer: Exploring Space-Time Transformers for Earth System Forecasting. In *Advances in Neural Information Processing Systems, 35*. https://proceedings.neurips.cc/paper_files/paper/2022/hash/a2affd71d15e8fedffe18d0219f4837a-Abstract-Conference.html

Zamir, S. W., Arora, A., Khan, S., Hayat, M., Khan, F. S., & Yang, M.-H. (2022). Restormer: Efficient Transformer for High-Resolution Image Restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. https://openaccess.thecvf.com/content/CVPR2022/html/Zamir_Restormer_Efficient_Transformer_for_High-Resolution_Image_Restoration_CVPR_2022_paper.html