Reply to Reviewer #01

We thank the reviewer for the suggestions and comments to improve our manuscript. We have accordingly revised our manuscript to address these concerns. Please find detailed point-by-point responses (in blue) to the reviewer's issues below.

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 comments.

Reply: Thank you for your insight on the central goal of our paper. We agree that "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." We reconstruct the manuscript to support the latter claim that GRPE and SAFN can improve the Neighborhood Attention Transformer. The main revisions are in the abstract, introduction, discussion and conclusion sections.

In the abstract (L14-L30 of the submitted manuscript), we introduce background, GRPE, and SAFN's structure. We stress on how they perform on ablation study and the potential of them on next generation weather forecasting. We weaken the importance of AtmoNAT, omitting its structural details. Performance of AtmoNAT only serves as validation of those two structures.

In the introduction, we claim that:

1. “In this work, we introduce Gated Relative Position Encoding (GRPE) to improve the Neighborhood Attention Transformer in weather forecasting.” (L77 of the submitted manuscript).
2. “We introduce a Spatial-Aware Feed-Forward Network (SAFN) to enhance NAT in weather forecasting.” (L98 of the submitted manuscript).
3. “In this study, we choose NAT as the backbone Transformer block of our data-driven weather model to test GRPE and SAFN performance.” (L104 of the submitted manuscript).

We stress that GRPE and SAFN are what we propose to enhance NAT for weather forecasting. AtmoNAT is not what we want to propose. It is just a platform to validate the performance of GRPE and SAFN.

We move the computational efficiency section to discussion. In the discussion, we claim that “From results shown previously, GRPE and SAFN achieve substantial error reductions with omittable model size and FLOPs increase, which contributes a lot to AtmoNAT's high efficiency” (L413 of the submitted manuscript). We attribute the high efficiency of AtmoNAT to GRPE and SAFN.

In the conclusion, we state that:

1. “In this study, we propose GRPE and SAFN to enhance NAT-based data-driven weather forecasting.” (L448 of the submitted manuscript).
2. “The performance of the GRPE- and SAFN-enhanced model is first evaluated in deterministic forecasting experiments.” (L455 of the submitted manuscript).
3. “GRPE and SAFN consistently improve predictive accuracy while introducing only negligible increases in parameter count and FLOPs, which contributes a lot to AtmoNAT's high efficiency. Comparisons with alternative positional encoding schemes and feed-forward network designs show that the NAT architecture equipped with GRPE and SAFN produces the lowest overall forecasting errors, indicating that explicitly modeling both flow-dependent spatial relationships and direct feature interactions is beneficial for data-driven weather prediction.” (L464 of the submitted manuscript).

We again stress we propose GRPE and SAFN to enhance NAT-based data-driven weather forecasting. The performance of AtmoNAT is attributed to GRPE and SAFN.

We think now the central contribution is that GRPE and SAFN can improve the Neighborhood Attention Transformer for weather forecasting.

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:

$$A_{i,j}(X_t) = softmax\left(\frac{Q_i(X_t)K_j(X_t)^T}{\sqrt{d}} + b_{i,j}\right).$$

Therefore, a fixed $b_{i,j}$ 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:

$$b_{i,j} \rightarrow b_{i,j}(X_t).$$

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.

Reply: Thank you for your advice. As revealed by the comments, we make a claim "typically embed limited spatial and temporal context, failing to fully account for the time variability, directionality, and location-dependency inherent in atmospheric motions", which is too strong. And the current contents in L40-60 cannot fully support our claim. And we do not stress whether it is because of the Transformer block or the positional encoding that limit the spatial and temporal modeling. We should also explicitly claim that the positional bias should become conditional on the atmospheric state to avoid misunderstanding.

In the abstract, we delete the original strong expression, and claim that "Transformer-based weather forecasting models often use static positional encodings, limiting their representation of flow-dependency and geographical heterogeneity. We introduce Gated Relative Position Encoding (GRPE) to address these limitations", which is a much more conservative expression (L14 of the submitted manuscript).

L34-70 of the original manuscript is now changed into (L46-L84 of the submitted manuscript):

“Over the past decade, the Transformer (Vaswani et al., 2017) has emerged as a powerful architecture for modeling long-range spatial and temporal dependencies. TimeSformer applies self-attention separately along the spatial and temporal dimensions of video sequences (Bertasius et al., 2021), while ViViT constructs spatiotemporal tokens and explores factorized spatial and temporal attention (Arnab et al., 2021). For Earth-system applications, Earthformer further develops a dedicated space – time Transformer for forecasting geophysical fields (Gao et al., 2022). Benefiting from their strong capacity for spatiotemporal representation, Transformers have also become a common backbone of data-driven weather models and have achieved state-of-the-art forecast skill (Bi et al., 2023; Bodnar et al., 2025; Chen et al., 2023; Chen et al., 2025; Couairon et al., 2024; Du et al., 2025; Nguyen et al., 2024; Niu et al., 2025).

Because self-attention itself does not preserve token order or spatial arrangement, positional information must be explicitly introduced. FuXi (Chen et al., 2023) adopts the continuous relative position bias of Swin Transformer (Liu et al., 2021), where attention scores are adjusted according to the relative displacement between tokens. FengWu uses rotary position embedding (RoPE), which applies position-dependent rotations to query and key vectors before computing their dot-product attention (Chen et al., 2025; Su et al., 2024). Both approaches impose geometry-based positional priors that are independent of the instantaneous atmospheric state: identical spatial configurations in attention windows across locations and forecast times receive the same relative bias or rotation rule. Such static representations may be restrictive for Earth-system forecasting, where interactions among atmospheric variables are spatially heterogeneous and dynamically modulated by geographical setting, circulation regime, season, and evolving weather conditions.

Several recent models have introduced additional spatial or temporal information to address this issue. Aurora (Bodnar et al., 2025), for example, uses additive embeddings derived from latitude, longitude, patch area, and absolute time, while Baguan (Niu et al., 2025) incorporates forecast lead time into its token representations. These approaches provide additional context but still represent only a limited subset of the factors governing the spatial and temporal variability of atmospheric dynamics. Pangu-Weather (Bi et al., 2023) instead introduces an Earth-Specific Positional Bias (ESPB) into its 3D Transformer, assigning attention biases according to absolute spatial coordinates to account for the spherical geometry and vertical structure of the atmosphere. However, To et al. (2024) showed that ESPB can be replaced by relative positional bias without degrading validation loss while reducing model size by nearly 40%, suggesting that its large coordinate-dependent parameterization is not an efficient representation of location-dependent atmospheric interactions. These findings motivate the development of positional encodings that can represent atmospheric dynamics more efficiently while adapting to both location and flow state.

In this work, we introduce Gated Relative Position Encoding (GRPE) to improve the Neighborhood Attention Transformer in weather forecasting (NAT; Hassani et al., 2023). Rather than prescribing a fixed positional relationship, GRPE dynamically adjusts relative positional biases according to the evolving atmospheric state. Specifically, GRPE represents the positional bias as a gated linear combination of learnable prototypes,

providing a compact yet flexible representation of flow-dependent spatial relationships. It further retains coordinate-dependent biases to capture persistent geographical heterogeneity, analogous to the role of ESPB, but with only a negligible increase in model size. When integrated into NAT, GRPE produces substantial improvements on WeatherBench 2 (Rasp et al., 2024) compared with otherwise identical models without GRPE, demonstrating the benefit of combining geographical dependence with dynamically varying flow state."

We review the Transformer's application in spatiotemporal modeling according to your comments and clarify that Transformer has strong ability in spatiotemporal modeling. We thoroughly review how positional biases are used in FuXi and FengWu. We find that FengWu actually use RoPE for their positional bias. We revise the original manuscript. We claim that both approaches in FuXi and FengWu impose geometry-based positional priors that are independent of the instantaneous atmospheric state, which may be restrictive for Earth-system forecasting. We think the wording now is logically coherent. Statement is not board. Discussion can support our claim. There is no confusion that "stated limitation is about Transformers in general or specifically about the position-encoding schemes". We also stated explicitly that "Rather than prescribing a fixed positional relationship, GRPE dynamically adjusts relative positional biases according to the evolving atmospheric state" to confirm the positional bias itself is conditional on the atmospheric state.

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".

Reply: Thank you for your advice. We now state that "Neighborhood attention bridges the gap between the powerful modeling capabilities of Transformers and the efficient, localized characteristics of convolutional neural networks. It significantly reduces the computational complexity of global self-attention operation and has already been successfully applied to data-driven weather forecasting." (L106 of the submitted manuscript).

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.

Reply: Thanks for the comments. Gating mechanism has used in data-driven model. This line should be changed.

The changed contents are in L92-L96 of the submitted manuscript:

"Interactions among atmospheric variables are central to weather evolution, motivating the use of gating to represent these dependencies in feature space. For example, FengWu-Adas incorporates gated convolution and gated cross-attention into its assimilation system (Chen et al., 2023), while AtmosMJ uses gated residual fusion to mitigate long-term error accumulation (Cheon, 2025). However, the specific contribution of gating to data-driven weather forecasting remains insufficiently understood."

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.

Reply: In the reply of your first comment, we reconstruct the introduction that we focus on introducing SAFN to enhance NAT in weather forecasting. The changed contents are in L98-L102 of the submitted manuscript:

"We introduce a Spatial-Aware Feed-Forward Network (SAFN) to enhance NAT in weather forecasting. Inspired by the Gated-Dconv Feed-Forward Network in Restormer (Zamir et al., 2022), SAFN combines channel expansion, parallel branches, and depth wise convolution with multiplicative gating to model feature interactions. We further introduce a global positional bias to capture spatial context associated with terrain. Extensive ablation experiments demonstrate that SAFN improves AtmoNAT's forecasting performance."

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.

Reply: We thank the reviewer for this comment and have clarified both points.

Regarding the weighting strategies, following GraphCast (Lam et al., 2023), Pangu-Weather (Bi et al., 2023), and ArchesWeather (Couairon et al., 2024), the latitude weight is proportional to the grid-cell area (i.e., cosine-latitude weighting), and the pressure-level weight is approximately proportional to pressure, decreasing from the lowest to the highest level. Variable-specific weights are also applied to balance the contributions of different variables.

Regarding the normalization statistics, all preprocessing and normalization statistics are derived from the training and validation periods only, and no data from the test period (2020) are used. Specifically, they are computed from the most recent 20 years (2000–2019) rather than the full record, because long-term climate trends (e.g., global warming) would bias full-record statistics toward earlier climate states. The recent statistics therefore better represent the atmospheric states at forecast time.

We now explicitly state that AtmoNAT is trained with a weighted Charbonnier loss (Charbonnier et al., 1994); this loss is adopted to improve training stability, and its full form is given in the revised Eq. (1). The changed contents are in L128-L144 of the submitted manuscript:

"The dataset is chronologically partitioned into training (1979–2018; 29,200 samples), validation (2019; 730 samples), and testing (2020; 730 samples) sets. All variables are standardized using statistics computed from the most recent 20 years (2000–2019) to better reflect the recent climate state, and no test data are used.

Implemented in PyTorch (Paszke et al., 2019), AtmoNAT is trained with a weighted Charbonnier loss (Charbonnier et al., 1994) to improve training stability, as defined in Eq. (1):

$$L = \frac{1}{C_A HWL}\sum_{c=1}^{C_A}\sum_{i=1}^{H}\sum_{j=1}^{W}\sum_{k=1}^{L} w_c a_i b_k \rho(\hat{X}_{c,i,j,k}^{t+1} - X_{c,i,j,k}^{t+1}) + \frac{\lambda}{C_S HW}\sum_{c=C_A+1}^{C_A+C_S}\sum_{i=1}^{H}\sum_{j=1}^{W} w_c a_i \rho\left(\hat{X}_{c,i,j}^{t+1} - X_{c,i,j}^{t+1}\right), \quad (1)$$

where $\rho(x) = \sqrt{x^2 + \epsilon}$ is the Charbonnier penalty (Charbonnier et al., 1994) with a small constant ϵ, and $\lambda = 0.1$ balances the contributions of the surface and upper-air terms. C_A, C_S, H, W, and L denote the number of upper-air variables, surface variables, and grid points in the latitudinal, longitudinal, and vertical directions, respectively, and c, i, j, and k are the corresponding indices. $\hat{X}_{c,i,j,k}^{t+1}$ and $\hat{X}_{c,i,j}^{t+1}$ denote the upper-air and surface variables predicted by AtmoNAT at a 12-hour lead time $(t+1)$ from the current state (t), and $X_{c,i,j,k}^{t+1}$ and $X_{c,i,j}^{t+1}$ are the corresponding ERA5 ground truth values. Following GraphCast (Lam et al., 2023), Pangu-Weather (Bi et al., 2023), and ArchesWeather (Couairon et al., 2024), the latitude weight a_i is proportional to the grid-cell area (i.e., cosine-latitude weighting), the pressure-level weight b_k is approximately proportional to pressure and thus decreases from the lowest to the highest level, and the variable-specific weight w_c accounts for the varying physical importance and training difficulty of different variables."

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 help the understanding.

Reply: We thank the reviewer for this constructive suggestion. We have redrawn Figure 1 so that the hierarchical relationship is explicit, following both options you proposed:

1. Zoom-in connection. Panels (b) and (c) are now enclosed in a single dashed frame labeled "Transformer Block", which is linked to the corresponding Transformer Block in panel (a) by dashed zoom-in lines. The two panels are therefore interpreted as an enlargement of the interior of one block rather than as subsequent stages.

2. Visual indication of the internal components. The Transformer Block in panel (a) is drawn as a container holding its two submodules, and the submodules are colored consistently across the figure: GRPE-NA in blue and SAFN in yellow, matching the background of panels (b) and (c) respectively.

3. Caption. The caption has been rewritten to state explicitly that each Transformer Block contains two submodules that are enlarged on the right, and to distinguish arrows (data flow) from dashed lines (zoom-in relations).

Section 2.2.3 (SAFN)

The novelty boundary of the proposed SAFN is not very clear. Based on reading, the authors seem to try presenting the entire SAFN formulation as a new module. However, several 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, depth wise 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 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.

Reply: Thank you for your advice. Novelty of SAFN module needs to be clear. We now admit that our SAFN is inspired by Gated-Dconv Feed-Forward Network of Zamir et al. (2022) and introduce a global positional bias β when computing the gating coefficients. Now the changed contents are in L243-L249 of the submitted manuscript:

"SAFN builds on the Gated-Dconv Feed-Forward Network in Restormer (Zamir et al., 2022). To account for terrain's persistent, location-dependent influence on atmospheric motion, we introduce a global positional bias β when computing the gating coefficients. Multiplicative interactions between the input features I and gating features G provide a way to represent interaction among atmospheric variables. These interactions also extend across space: for example, local transport is influenced by the surrounding wind field. Parallel 3×3 depth wise separable convolutions capture this spatial context by incorporating information from neighboring locations.

The ablation studies in Section 3.2 assess SAFN's contribution to AtmoNAT's forecasting performance."

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.

Reply: Thank you for your comments. The reason why we use ECMWF HRES as normalization reference is that almost all the data-driven weather models choose this as the benchmark for their model performance, including Pangu-Weather (Bi et al., 2023), NeuralGCM (Kochkov et al., 2024), GraphCast (Lam et al., 2023), FuXi (Chen et al., 2023), Aurora (Bodnar et al., 2025), FengWu (Chen et al., 2025), Baguan (Niu et al., 2025), ArchesWeather (Couairon et al., 2024), and Stormer (Nguyen et al., 2024). Choosing ECMWF HRES as reference makes it much easier to compare our model's performance with other models.

We think validation against independent observations is not appropriate. Our model is trained on 1.5° resolution. Independent observations cannot align with this coarse resolution. We think validation against ERA5 data is sufficient to test the performance of our model.

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.

Reply: We agree that the statements are too strong, but AtmoNAT outperforms Stormer on T2m is easy to see. The changed contents are in L288-L293 of the submitted manuscript:

"Figure 3 presents the ACC of the coarse-resolution models for key variables up to a 72-hour lead time. These ACC trends are highly consistent with the RMSE results in Figure 2. AtmoNAT achieves the highest ACC for MSLP, T850, U850, V850, U10m, and V10m. Its ACC for wind velocities (U850, V850, U10m, V10m) is visually higher than that of the competing models. It outperforms Stormer on T2m, while ACC values for Q700 and Z500 are nearly identical across all models. Extended ACC results up to a 120-hour lead time (Figure S10) reveal a performance degradation at day 5, again attributable to the 48-hour fine-tuning limit imposed by computational constraints."

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. 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.

Reply: Thank you for your comments. WeatherBench 2 only delivers the data-driven weather model's performance on each geographical region and land area of those regions. You can refer to the "Region" menu at the website of WeatherBench 2: https://sites.research.google/gr/weatherbench/deterministic-scores/. AtmoNAT's performance on ocean areas cannot be compared with other models since no data is provided. So, we do not draw figures about its performance on ocean areas. Instead, we include additional figures in supplementary material to show AtmoNAT's performance on each geographical region and land area of those regions, and compare it with all other models including those trained on 0.25° resolution (Figure S3, full global domain; Figure S11-S22, each geographical region and land area of those regions). To stress that our model cannot perform equally well in ocean area, we change the relevant contents in L301-L304 of the submitted manuscript:

"Figure S3 shows the relative RMSE evaluated over global land areas, including ocean areas. AtmoNAT still cannot beat GraphCast and FuXi on all variables. Its superior performances on Q700, U850, V850, U10m, and V10m indicates that it cannot perform equally well in ocean areas. RMSE analyses on different geographical regions including those only on land area are presented in S11-S22."

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.

Reply: Thank you for your comments. We include the new models' results including those from NeuralGCM, FuXi and GraphCast in Figure 5. Figure S23 and Figure S24 are also changed. Stormer and ArchesWeather's results are not included because their outputs are not open source.

Section 3.2 / Table 1

The ablation results show 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.

Reply: Thank you for your advice. As revealed by your comments, the NAT module's contribution to AtmoNAT's performance is not important. It further confirms the contribution of GRPE and SAFN, which is the central contribution of this paper. We don't see there is any problem here. We add a sentence to clarify: "It means that the performance of AtmoNAT is not due to the backbone choice" (L337 of the submitted manuscript).

As implementing GRPE into Swin Transformer, GRPE need to be modified dramatically to accommodate its structure to Swin Transformer. Swin Transformer's attention is not defined at every location in the spatial domain. And the attention calculation at the boundary area is intractable and needed to be specially attended. We think even we do this experiment, the comparison is not fair, because the two GRPE are totally different things. As for SAFN, it is not that difficult. But the lack of GRPE experiment makes SAFN experiments alone unnecessary.

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.

Reply: Thank you for your insight. We admit that alignment of direction of the GRPE to actual atmospheric motion is not very clear. We check the whole manuscript and delete the expression "directionality of GRPE", retain "the location and flow dependency of GRPE". You can see the revision in L426-L431 of the submitted manuscript.

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.

Reply: Thank you for your advice. Now the changed sentence is "These biases correlate with global terrain, exhibiting absolute values that are generally proportional to altitude" (L439 of the submitted manuscript). The figure's color bar is unified.

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.

Reply: Abstract has been changed.

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