Major comments

1. The manuscript needs more explanation on how BiXiao differs from other AI models, some of which can also accept non-continuous data.

Response:

Thank you for this suggestion. We agree that the original manuscript did not sufficiently explain how BiXiao differs from other AI models, especially in distinguishing between using sparse observations and how sparse observations are organized and incorporated into a forecasting system. Therefore, we rewrote the relevant paragraphs in the introduction of the revised manuscript and added a discussion of recent AI weather prediction and AI data-assimilation models that can use sparse or heterogeneous observations, including FengWu-4DVar, DiffDA, Aardvark Weather, OMG-HD, and the observation-space forecasting approach proposed by McNally et al.

In the revised manuscript, we further clarify that the key design difference of BiXiao is not simply that it can accept non-continuous data, but that it adopts a heterogeneous framework combining continuous meteorological grids with non-continuous environmental grids. Unlike methods such as FengWu-4DVar and DiffDA, which use data assimilation to generate gridded analysis fields, BiXiao does not explicitly assimilate pollutant observations. It also differs from models such as Aardvark Weather and OMG-HD, which encode multi-source observations into gridded initial states. BiXiao does not reconstruct a continuous pollutant initial field or a continuous pollutant forecast field.

We also discussed the method proposed by McNally et al. in the revised manuscript. Their approach is more observation-space oriented, directly predicting future observations from historical observations. This idea is close to BiXiao in the sense that it preserves a non-continuous observational representation. However, BiXiao does not rely solely on discrete observations for prediction. Instead, it transfers continuous three-dimensional meteorological background information generated by the meteorological module to the non-continuous environmental grids, allowing the environmental module to use both monitoring-network information and future meteorological evolution. Therefore, BiXiao shares with McNally et al. a similar motivation to avoid conventional continuous-field reconstruction, but its core difference lies in the coupling between continuous meteorological fields and non-continuous environmental grids.

We believe that these revisions more clearly explain the relationship and differences between BiXiao and existing AI forecasting models.

2. **Lines 192–193 state, “specific station information can be found in the Appendix.” However, no appendix or supplementary material has been included in the preprint. I am not able to provide a judgment on the use of**

this dataset since it is not explained. Please ensure the appendix is included in the final version.

3.

Response:

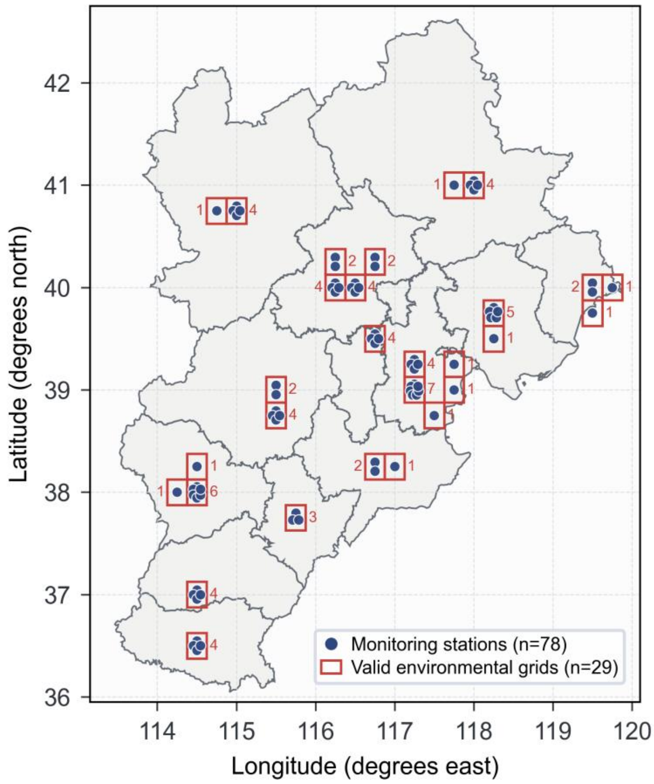

Figure S1. Spatial distribution of the 78 air-quality monitoring stations and the 29 valid ERA5-aligned environmental grids in the Beijing--Tianjin--Hebei region.

Thank you for pointing this out. We agree that the original manuscript referred to an Appendix, but the corresponding appendix or supplementary material was not included in the preprint. This made it difficult for readers to assess the environmental observation dataset and the construction of the environmental grids. Following this comment, we added a Supplement to the revised manuscript and included Fig. S1 to show the spatial relationship between the air-quality monitoring stations and the valid environmental grids.

We also corrected and clarified the description of the dataset in the revised manuscript. The original manuscript stated that 79 monitoring stations were used. However, in the final dataset used for model training and evaluation, 78 monitoring stations were actually included in the construction of the valid environmental grids. We therefore corrected the number of stations in the main text from 79 to 78. The revised data section now explains that these stations are mapped to 29 valid ERA5-aligned non-continuous environmental grids. For grids containing multiple monitoring stations, the pollutant concentration of the grid is calculated as the arithmetic mean of the observations from the stations assigned to that grid.

The newly added Fig. S1 shows the spatial distribution of the 78 monitoring stations and the 29 valid environmental grids in the Beijing--Tianjin--Hebei region. This figure allows readers to visually inspect the monitoring-network coverage and the correspondence between stations and environmental grids. We believe that this addition improves the transparency of the data-processing procedure and addresses the concern regarding the missing Appendix/Supplement.

4. The authors explain that this is a regional study, focused on the BTH region, and they justify this in the manuscript, for instance, by explaining the computational demands of the model. I would suggest that the regionality of this study be more clearly stated in the abstract and early in the introduction. I also suggest that the authors elaborate on the scalability of the model for future applications on a larger-scale.

Response:

Thank you for this helpful suggestion. We revised both the abstract and the early part of the introduction to strengthen the description of the regional scope of this study.

First, in the abstract, we now explicitly state that BiXiao is developed and evaluated as a regional AI-based atmospheric environmental forecasting model for the Beijing--Tianjin--Hebei region. We also revised several broad expressions in the abstract. For example, operational urban-scale forecasts and across all key cities were replaced with more accurate regional descriptions, such as regional forecasts and valid environmental grids. In addition, we avoided using the term assimilates to describe the use of monitoring-station observations by the environmental module, in order to avoid confusion with the concept of data assimilation in traditional forecasting systems.

Second, we expanded the introduction to explain why the Beijing--Tianjin--Hebei region was selected as the regional demonstration area. The revised manuscript states that this region is densely populated and industrialized, with complex air-pollution episodes driven by local emissions, regional transport, and meteorological variability. It also has a relatively dense air-quality monitoring network, making it a suitable test bed for developing and evaluating a regional atmospheric environmental forecasting model. We also emphasize when introducing BiXiao that the model is developed and evaluated within the BTH region in the present study.

Finally, we added a more cautious discussion of the scalability of the model. The revised manuscript states that the heterogeneous design of BiXiao, which combines continuous meteorological grids with non-continuous environmental grids, provides a basis for future extension to larger domains. However, larger-scale applications will require an expanded meteorological input domain, broader monitoring-network coverage, additional training data, and systematic validation under different emission structures, terrain conditions, and meteorological regimes. Through these revisions, the manuscript now more clearly defines the regional scope of the current study while discussing future scalability in a more cautious manner.

5. Figure 10: Please indicate times on each of the panels.

Response:

Thank you for pointing this out. We have redrawn both Fig. 7 and Fig. 10 and added the corresponding time labels to each panel, so that the timing of the meteorological fields can be clearly identified.

Minor comments

- **Throughout the manuscript, there are many erroneous quotation marks, including around the model name BiXiao, and they are used inconsistently. I noticed this, for example, on lines 68, 69, 126, 167, 206, 212, 249, 254, 259, 264, 410, 434, etc. These quotation marks are not needed, and I suggest the authors go through the manuscript and remove the unnecessary quotation marks for better readability and a more professional looking article.**

Response:

Thank you for this careful comment. We removed the unnecessary quotation marks around the model name BiXiao and other expressions, and we also checked and standardized the use of quotation marks throughout the manuscript.

- **Line 11: CAMS would be more accurately described as a gridded reanalysis dataset, not a numerical model. I suggest the authors adjust this terminology to be more accurate.**

Response:

We agree with this comment and have revised the terminology accordingly.

- **Line 16: I'm not seeing a connection between the statement made by the authors and the Meng et al. (2023) reference. I suggest that the authors find a more relevant reference.**

Response:

Thank you for pointing this out. Following this suggestion, we replaced the original citation with Wang et al. (2008, 2024) and revised the sentence to more accurately describe the mechanistic basis of traditional physics-based models in atmospheric environmental forecasting and their computational-cost challenges.

- **Line 22: In the context of BiXiao, I would suggest the term "large-scale modelling" rather than "Earth system modelling".**

Response:

We agree with this suggestion and have revised the wording accordingly.

- **Line 39: Inness et al. (2019) is a good reference for the CAMS reanalysis datasets, but this sentence should also have a reference for Aurora.**

Response:
Thank you for this suggestion. We added the citation to the Aurora paper by Bodnar et al. (2025) in the revised manuscript.

- **The paragraph from lines 58-61 is not needed.**

Response:
We agree with this comment and have removed this paragraph from the revised manuscript.

- **Lines 90 – 91: Additional references are needed to support this background info.**

Response:
Thank you for this suggestion. We added relevant references in the revised manuscript to support the discussion of the original Swin Transformer, the Video Swin Transformer, and related three-dimensional Transformer-based weather forecasting studies. These additions provide stronger literature support for the description of three-dimensional window-based attention and its relevance to atmospheric three-dimensional field modelling.

- **Please expand on section 2.2.3, as the current version is too brief.**

Response:
Thank you for this suggestion. We expanded Section 2.2.3 to describe in more detail how the meteorological module separately processes upper-air and surface meteorological fields, including the input-variable dimensions, the parallel patch-embedding branches, the use of three-dimensional and two-dimensional convolutions, and the horizontal alignment between the two feature representations.
In addition, we described the dimensional reconciliation, spatial alignment, and preliminary fusion of multi-source meteorological data in the BiXiao meteorological module. We further clarified how surface meteorological features are expanded by adding a latent vertical dimension and then concatenated with upper-air meteorological features to form a unified three-dimensional feature tensor. The revised manuscript specifically emphasizes that this additional surface-feature layer is a representation in the latent feature space and does not correspond to an actual physical pressure level.

- **Please also expand on section 2.2.4 and provide more references in this section. For** instance, please explain what SwinTRansformerBlock3D module is.

Response:

Thank you for this suggestion. We expanded Section 2.2.4 by adding a description of the three-dimensional feature extraction process and further explaining the structure, function, and role of the SwinTransformerBlock3D module in meteorological-field modelling.

We first explain that SwinTransformerBlock3D is a module that extends the window-based attention mechanism of the Swin Transformer to three-dimensional feature spaces. Each block consists of three-dimensional window-based multi-head self-attention, a feed-forward network, Layer Normalization, and residual connections. We added references to the original Swin Transformer, the Video Swin Transformer, the Transformer self-attention mechanism, and related three-dimensional Transformer-based weather forecasting studies to better support the origin and applicability of this structure.

We also expanded the explanation from a meteorological modelling perspective. Three-dimensional window attention can capture local and regional covariability among variables such as wind, temperature, humidity, and geopotential height across both the latent vertical dimension and the horizontal dimensions. The shifted-window mechanism further facilitates information exchange among neighbouring regions and different latent vertical layers. This design does not explicitly solve atmospheric physical equations; instead, it learns statistical relationships associated with horizontal transport, vertical stratification, and the evolution of weather systems from the spatiotemporal relationships in the training data.

- **Section 2.2.6 needs more clarification. Is this the part that outputs a 3D grid similar to WRF-Chem? Please expand on this section to be more clear about the methodology.**

Response:

Thank you for this suggestion. The environmental module used in BiXiao does not output a three-dimensional continuous pollutant field like WRF-Chem. To avoid this possible misunderstanding, we expanded and clarified Section 2.2.6 in the revised manuscript.

The revised description now explicitly states that although the environmental module uses three-dimensional meteorological fields containing vertical information as inputs, it does not generate a complete three-dimensional pollutant concentration field and does not cover all continuous grids in the study region. Its outputs are restricted to the surface concentrations of six pollutants at the 29 valid ERA5-aligned non-continuous environmental grids. We also added the input and output tensor dimensions for the single-step prediction process, explaining that the model uses the meteorological background fields at Ti and Ti+1, together with the environmental state at Ti, to predict surface pollutant concentrations at Ti+1.

- **Section 2.2.7 is also too brief. Please provide some description/examples of the dynamic features captured by the model.**

Response:

Thank you for this suggestion. We expanded the relevant part of Section 2.2.7 to further explain the meteorological feature extraction process and clarify what dynamic features the model can learn from the meteorological inputs.

During model inference, the environmental module uses meteorological background fields at two consecutive times and jointly provides them to the network, rather than explicitly calculating simple differences. This design allows the model to learn dynamic features related to pollutant evolution from consecutive meteorological states, including changes in wind fields, thermodynamic and moisture conditions, and the lower-tropospheric vertical structure. These features are closely related to pollutant transport, dispersion, chemical transformation, and vertical mixing.

- **Lines 187–189: Can the authors please clarify, was the BiXiao model trained only on the lowest four ERA5 model levels, and are only the four lowest levels used in forecasting?**

Response:

Thank you for this question. BiXiao does not use only the four lowest pressure levels across all modules. The meteorological module uses six pressure levels, namely 500, 600, 700, 850, 925, and 1000 hPa, during both training and inference. Because the prediction target of the environmental module is surface pollutant concentration, the environmental module uses only the four lower pressure levels, namely 700, 850, 925, and 1000 hPa, during both training and inference. During training, the environmental module uses ERA5 meteorological fields at these four levels. During coupled inference, the same four levels are selected from the six-level meteorological forecasts generated by the meteorological module. We have clarified the six-level configuration of the meteorological module in Section 3.1.2 and the four-level configuration of the environmental module in Section 2.2.7 of the revised manuscript.

- **Section 3.1.3: Are these datasets from CNEMC? From the description, I think that is the case, but without the appendix, I can’t confirm. If these datasets are from CNEMC, then they should be properly cited. For example, Song et al. (2017) and Tao et al. (2016) are suitable references, and there are likely newer references also available.**

Response:

Thank you for this suggestion. The air-quality observation data used in this study are from the national air-quality monitoring network operated by the China National Environmental Monitoring Centre (CNEMC). We revised Section 3.1.3 to explicitly state the source of the original data and added relevant references describing the CNEMC monitoring network. In addition, we added Fig. S1 to the Supplement to show the spatial distribution of the monitoring stations.

- **Lines 198–201: How do the authors justify averaging all stations in a grid, even though the stations may not be evenly spaced in the grid and therefore may not**

be equally representative of the grid? Why did the authors not choose to use a weighted average or an ML-based method to determine the representative value for each grid box?

Response:
Thank you for this comment. We agree that a simple arithmetic mean cannot fully account for differences in spatial representativeness caused by uneven station distributions within the same grid cell. The purpose of this study is not to estimate a strict area-averaged grid-box concentration, but to construct a station-derived target aligned with the ERA5 grid. Because the stations come from the same monitoring network, we used equal weighting to keep the data-processing procedure transparent and reproducible, and to avoid introducing additional spatial assumptions or interpolation models into the training labels. We have clarified the definition and limitations of this target in the revised manuscript. In future work, we will further consider distance-based, land-use-based, or data-driven weighting methods.

- **Line 234: The sentence starting with "These three metrics" is not necessary.**

Response:
We agree with this comment and have removed this unnecessary sentence from the revised manuscript.

- **Lines 273–275: Is there a formatting error in the PDF for these lines? They appear to be out of place.**

Response:
We agree with this comment and have corrected the formatting issue in the revised manuscript.

- **Figure 3: The results in this figure appear to be presented as one result for each district. However, the methodology of this is not clearly explained. I suggest the authors explain how the model output was regridded by district for presentation in this figure.**

Response:
Thank you for this suggestion. In this study, the model outputs were not regridded to the administrative-district scale. We added a clarification in the revised manuscript to explain the statistical and visualization procedure used for Fig. 3 and to avoid misunderstanding. Specifically, the original outputs of BiXiao remain predictions at the 29 valid non-continuous environmental grids, and they were not spatially regridded into continuous city-scale fields. For the city-level visualization in Fig. 3, the PCC was first calculated independently for each valid environmental grid over the evaluation period. Each valid grid was then assigned to the prefecture-level city containing its grid centre. For cities containing multiple valid grids, the city-level PCC was calculated as the

arithmetic mean of the corresponding grid-level PCC values. These aggregated city-level values were used only to colour the administrative polygons and do not indicate that the model generated continuous administrative-scale forecast fields.

- **Lines 330–333: The WRF-Chem model should be properly cited in this section.**

Response:
This has been revised. The WRF-Chem model is now properly cited when it is first introduced in the WRF-Chem experimental setup section.

- **Lines 363–366: Underestimation of pollutants, especially during haze periods, is a well-known issue in WRF-Chem, as well as other large-scale models, such as CMAQ, GEOS-Chem, etc (Gao et al., 2022; Sokhi et al., 2022; Saide et al., 2020; Li et al., 2023). I would suggest that the authors discuss this in further detail because the improvement compared to WRF-Chem that BiXiao offers is an interesting result. Providing further details on what BiXiao improves would be of interest to the large-scale modelling community.**

Response:
Thank you for this valuable suggestion. Further discussion of the differences between BiXiao and WRF-Chem during pollution episodes helps clarify the significance of our results for regional air-quality modelling. Therefore, we added a new Section 5.2.3 to systematically discuss possible reasons for the underestimation of $PM_{2.5}$ pollution peaks by WRF-Chem, including uncertainties in emission inventories, secondary aerosol formation, boundary-layer processes, spatial resolution, and aerosol parameterizations. We also added relevant references, including Gao et al. (2022), Sokhi et al. (2022), Saide et al. (2020), and Li et al. (2023).
In addition, we cited a previous study involving one of the co-authors of the present manuscript (Qu et al., 2023). That study improved the representation of aerosol optical feedbacks and heterogeneous reactions in WRF-Chem and showed that these coupled processes can substantially affect O_3 simulations and enhance sulfate and nitrate formation in winter. This further indicates that insufficient representation of aerosol-photolysis feedbacks and heterogeneous chemistry may affect the simulated intensity and evolution of O_3 and $PM_{2.5}$ pollution episodes.
We also further explain the possible reasons for the improvement achieved by BiXiao. BiXiao uses pollutant observations at the initial time and future meteorological fields, allowing the environmental module to learn nonlinear relationships between evolving meteorological conditions and pollution states shaped by the combined influence of different emission sources. However, we also explicitly state that this result does not imply that BiXiao provides a better physical or chemical representation of atmospheric processes than WRF-Chem. Its advantage may partly arise from observation-based initialization and the temporal autocorrelation of pollutant concentrations. The current conclusions are limited to the study region, the selected pollution episodes, and the WRF-Chem configuration used in this study.

- **Line 400: Is this line supposed to be a figure caption, or is this a typo?**

Response:
This was a typo, and it has been corrected in the revised manuscript.

- **Lines 415–416: Most traditional, physics-based models are written to run on CPUs, not GPUs. Are the authors saying that the AI model, run on a GPU, is more efficient because of the fact that it is run on a GPU instead of a CPU? Please clarify this statement.**

Response:
Thank you for pointing this out. Our intention was not to make a hardware-normalized comparison between BiXiao and a specific WRF-Chem experiment. Rather, we intended to emphasize that BiXiao runs on a single GPU and can rapidly generate 72 h forecasts. Compared with traditional numerical forecasting workflows that typically rely on CPU-based computing systems, BiXiao can provide faster operational forecasts with a more compact computational setup. We have revised the relevant statement to make this point clearer.

- **Figure 12: BiXiao seems to show better performance in northern areas and not quite as good of performance towards the east and south of the BTH region. Interestingly, WRF-Chem seems to show approximately the same pattern. Could the authors speculate on why this could be?**

Response:
Thank you for this suggestion. The similar spatial pattern shown by BiXiao and WRF-Chem may reflect differences in the predictability of pollution processes across different parts of the Beijing--Tianjin--Hebei region. The northern part of the region is dominated by high-elevation mountainous terrain, with relatively fewer urban and industrial emission sources. As a result, the pollution background and its temporal variation are generally more stable, and the mountainous terrain may also partly limit the northward transport of pollutants from the southern plains. In contrast, the southern plains and eastern coastal areas contain denser emission sources and are affected by regional transport, secondary pollutant formation, and local circulations. These factors lead to stronger spatial heterogeneity and temporal variability in pollutant concentrations, thereby increasing forecasting difficulty. We have added this explanation to the discussion of Fig. 12 in Section 5.2.2. We also explicitly state that, because no emission-source or transport sensitivity experiments were conducted in this study, this explanation should be regarded as a plausible interpretation rather than a quantified attribution.