Major comments

1. More suitable baselines are needed.

The current evaluation mainly compares BiXiao with CAMS and WRF-Chem. These are useful references, but they do not show whether BiXiao outperforms simpler station-based forecasting methods. Since the environmental module uses pollutant observations at T+0 as input, short-term skill may partly come from persistence or temporal autocorrelation.

The authors should add simple baselines such as persistence, autoregressive regression, or simple LSTM/GRU/TCN models. These types of models are commonly used in air-quality time-series or station-level forecasting studies (Zheng et al., 2015; Li et al., 2017; Bai et al., 2018). A BiXiao variant without meteorological inputs would be especially useful.

Response:

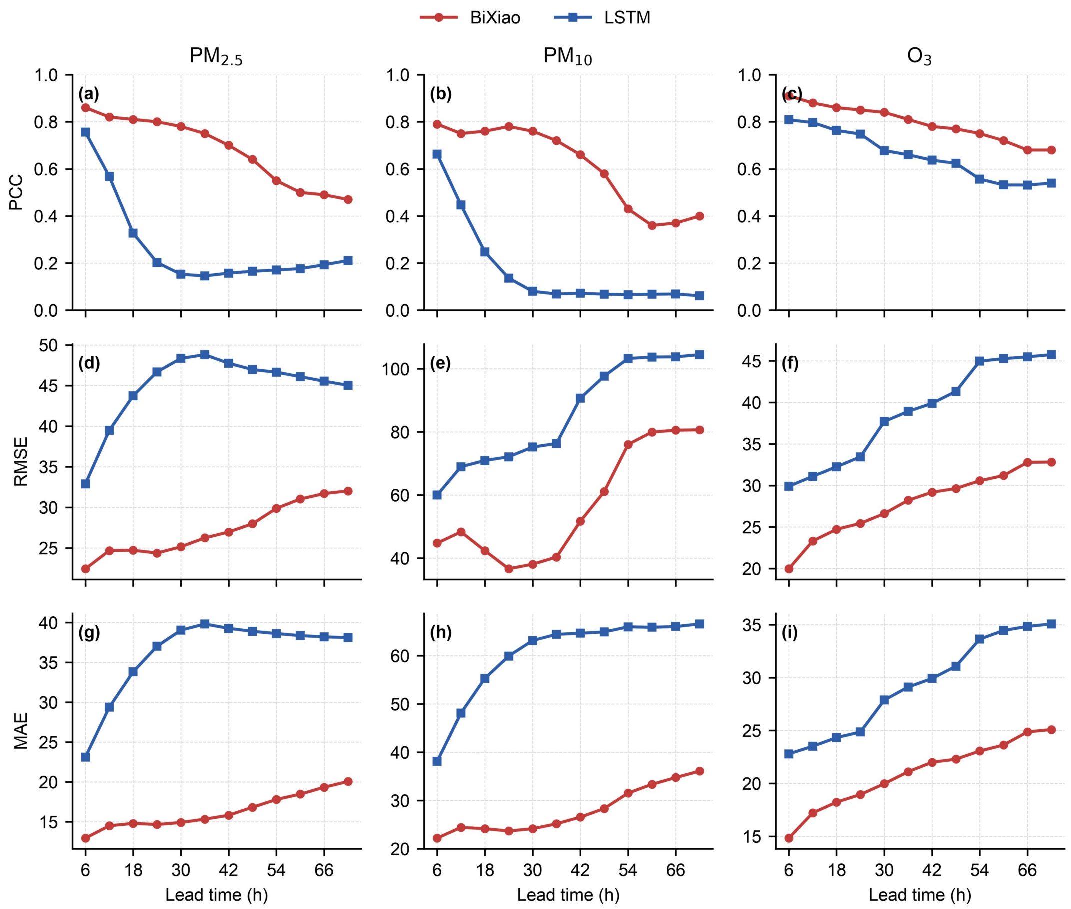

Figure 1. Comparison of 72 h forecast verification results between BiXiao and the LSTM model for $PM_{2.5}$,

PM_{10}, and O_3.

Thank you for this important suggestion. Because the environmental module of BiXiao uses pollutant observations at the current time as input, short-range forecast skill may indeed partly arise from the persistence and temporal autocorrelation of pollutant concentrations. To address this issue, we added an LSTM baseline model based only on historical pollutant observations and compared it with BiXiao. Figure 1 shows the comparison between BiXiao and the LSTM baseline for the main pollutants $PM_{2.5}$, PM_{10}, and O_3 over the 72 h forecast range, including PCC, RMSE, and MAE.

The LSTM baseline does not use any meteorological inputs or the future three-dimensional meteorological forecasts provided by the BiXiao meteorological module. Its input consists of observations of six pollutants at the 29 valid environmental grids over five 6 h time steps during the previous 24 h, i.e., [T-24, T-18, T-12, T-6, T0], and its output is the pollutant concentration at the next lead time, T+6. To generate 72 h forecasts, the LSTM is applied autoregressively during testing from T+6 to T+72. The model consists of two LSTM layers with a hidden dimension of 256 and a dropout rate of 0.1. Both the input and output correspond to six pollutants at the 29 valid environmental grids.

To avoid temporal leakage, we adopted a strict data-splitting strategy. For each official test initialization time, all environmental observation times from T-24 to T+72 were defined as a protected test window. Any training or validation sample overlapping with this protected window was removed. The training and validation sets were separated by complete monthly blocks, and they did not share any environmental observation times. The normalization statistics were calculated only from the training samples. The LSTM baseline was trained and evaluated using the same test initialization times and evaluation range as BiXiao.

The comparison figure between BiXiao and the LSTM baseline shows the changes in PCC, RMSE, and MAE for $PM_{2.5}$, PM_{10}, and O_3 from T+6 to T+72. The results show that the LSTM has some short-range forecast skill, confirming that pollutant persistence and temporal autocorrelation contribute to short-term prediction. For example, at T+6, the PCC values of the LSTM for $PM_{2.5}$, PM_{10}, and O_3 are 0.756, 0.663, and 0.808, respectively. However, as the forecast lead time increases, the performance of the LSTM decreases rapidly, especially for $PM_{2.5}$ and PM_{10}. Averaged over the 6--72 h forecast range, the PCC values of BiXiao for $PM_{2.5}$, PM_{10}, and O_3 are 0.681, 0.613, and 0.794, respectively, clearly higher than those of the LSTM baseline (0.269, 0.170, and 0.656). Compared with the LSTM baseline, BiXiao reduces the mean RMSE by 39.2%, 33.8%, and 28.2%, and reduces the mean MAE by 54.9%, 53.7%, and 28.5% for $PM_{2.5}$, PM_{10}, and O_3, respectively.

These results indicate that the forecast skill of BiXiao does not simply come from persistence or temporal autocorrelation in pollutant observations. Although the LSTM based only on historical pollutant observations provides useful short-range skill, it cannot maintain stable performance over the full 72 h forecast range. In contrast, by incorporating continuous three-dimensional meteorological background fields and transferring their information to the non-continuous environmental grids, BiXiao achieves more stable performance at medium-range lead times and for particulate matter forecasts.

In the revised manuscript, we also reconstructed Section 4, Routine Environmental Forecasting, and added the comparison between BiXiao and the LSTM baseline.

2. The station-level claim needs clarification.

The manuscript states that BiXiao directly uses discrete station observations and predicts station-level/site-level concentrations. However, the actual processing maps 79 stations to 29 ERA5-aligned valid grids, with multiple stations in one grid averaged.

This weakens the claim of direct station-level prediction. The authors should either provide true station-level experiments or revise the wording to describe the output as station-derived, ERA5-aligned non-continuous grid predictions.

Response:

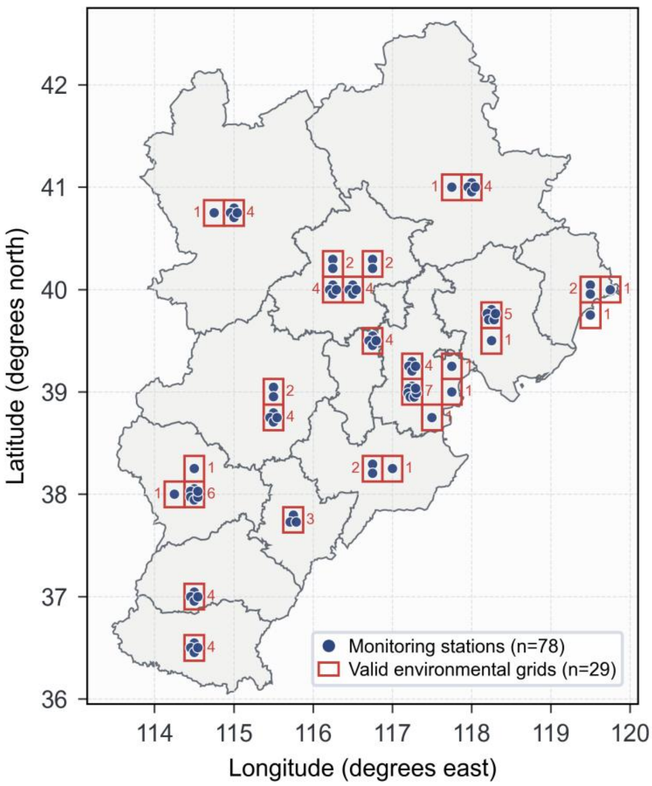

Figure S1. Spatial distribution of the 78 air-quality monitoring stations and the 29 valid ERA5-aligned environmental grids in the Beijing--Tianjin--Hebei region.

Thank you for this important comment. We clarify that BiXiao does not directly predict concentrations at individual monitoring stations. In fact, the environmental module does not output station-by-station concentrations. Instead, station observations are mapped to valid ERA5-aligned environmental grids, and predictions are made on these non-continuous environmental grids. To avoid misleading wording, we systematically revised the relevant expressions throughout the manuscript.

Specifically, we first revised the description of the BiXiao environmental module in the abstract, introduction, and overall model framework. We no longer describe the model as directly producing station-level predictions. Instead, we clarify that the environmental module uses station-derived environmental information or station-derived environmental grid data constructed from the monitoring network, and predicts pollutant concentrations at valid ERA5-aligned non-continuous environmental grids.

Second, we further clarified the model output in the methodology section describing the environmental module. The revised manuscript now states explicitly that the environmental module does not generate a continuous three-dimensional pollutant concentration field similar to WRF-Chem, nor does it cover all continuous grids in the study region. Its outputs are restricted to the 29 valid ERA5-aligned environmental grids constrained by observations. We also clarified that the output layer maps the fused features to these 29 valid environmental grids, and that the results should be interpreted as station-derived non-continuous grid predictions rather than true station-level predictions.

Third, we expanded the description of the data-processing procedure used to construct the environmental grids. The revised manuscript states that this study uses observations from 78 air-quality monitoring stations, which are mapped to 29 valid ERA5-aligned environmental grids. If multiple stations fall within the same grid, the pollutant concentration for that grid is constructed as the arithmetic mean of the observations from those stations. We also clarify that this value represents the mean concentration of the monitoring stations assigned to the grid, rather than a true area-averaged concentration. Therefore, BiXiao is evaluated at these station-derived valid environmental grids, not at individual monitoring stations. To improve transparency, we added Fig. S1 in the Supplement to show the spatial distribution of the 78 monitoring stations and the 29 valid environmental grids.

Finally, we revised potentially misleading wording in the results analysis, the CAMS/WRF-Chem comparisons, and the conclusion. Expressions that could be interpreted as station-level prediction were replaced with terms such as valid environmental grids, station-derived environmental grid data, and station-derived environmental grid observations. Through these revisions, the manuscript now more accurately describes the environmental predictions of BiXiao as station-derived, ERA5-aligned non-continuous grid predictions, rather than strict individual station-level predictions.

3. The generality of the claims should be toned down.

The experiments are limited to the Beijing – Tianjin – Hebei region and 29 valid environmental grids. This is a useful regional demonstration, but it is not enough to support broad claims such as a “new paradigm” for fine-scale urban environmental forecasting or future nationwide applications.

The authors should soften these claims or add broader tests, such as spatial holdout experiments across stations, grids, or cities.

Response:

Thank you for this suggestion. Because of current computational limitations, the present experiments are mainly focused on the Beijing--Tianjin--Hebei region, and the environmental module predicts only 29 valid ERA5-aligned non-continuous environmental grids. To avoid overgeneralization, we toned down and clarified the relevant statements in the revised manuscript.

First, we further clarified the regional nature of this study. In both the abstract and the introduction, the revised manuscript now states that BiXiao is developed and evaluated as a regional AI-based atmospheric environmental forecasting model for the Beijing--Tianjin--Hebei region. In the study-region description, we also present the BTH region as a regional test bed for evaluating the BiXiao framework. We explicitly state that the current BTH experiment alone does not demonstrate the transferability of BiXiao to other regions, and that future multi-region validation is still required.

Second, we softened the wording related to spatial refinement. Because the output of the environmental module consists of 29 ERA5-aligned non-continuous environmental grids constructed from 78 monitoring stations, rather than high-resolution continuous urban grids, we removed wording such as a new paradigm for fine-scale urban atmospheric environment forecasting, which could imply overgeneralization. The revised manuscript now more accurately describes BiXiao as a regional AI-based atmospheric environmental forecasting framework for station-derived non-continuous environmental grids.

Finally, we revised the statements related to future applications. The original wording regarding nationwide applications and broader deployment has been removed or rewritten. The revised manuscript only retains a cautious discussion of future extensions, stating that future work will evaluate the transferability of BiXiao to additional regions and larger domains. Such extensions will require broader monitoring-network coverage, additional training data, and systematic validation under different emission structures, terrain conditions, and meteorological regimes. Through these revisions, the conclusions of the manuscript are now restricted to what can be supported by the current BTH regional demonstration and the 29 valid environmental grids.

4. Ablation experiments would improve interpretation.

The paper argues that meteorological background fields are important, but it does not show which meteorological inputs matter most. As an optional but valuable improvement, the authors could add grouped ablations, such as removing wind, temperature, humidity, surface variables, or T+1 meteorology.

These tests would help explain the physical relevance of the model and why performance differs among O_3, $PM_{2.5}$, PM_{10}, and other pollutants.

Response:

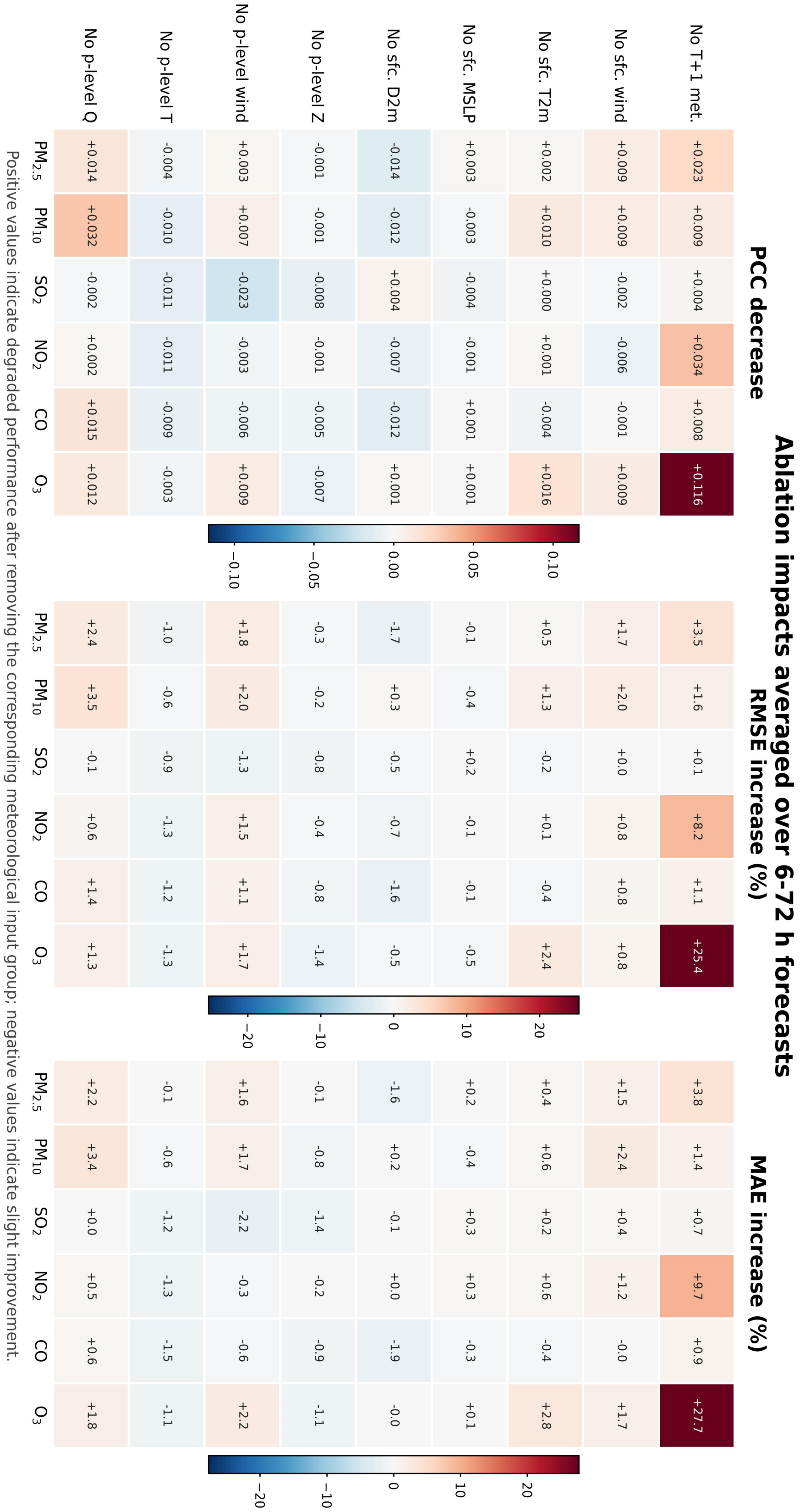

Figure 2 Ablation impacts averaged over 6-72 h forecasts

Thank you for this helpful suggestion. We agree that reporting only the overall forecast performance is not sufficient to explain the contribution of different meteorological inputs to the model results. Following this suggestion, we conducted grouped ablation experiments and provide the corresponding result figure in this response. Specifically, using the full BiXiao model as the baseline, we removed the T+1 future meteorological fields, surface wind speed, 2 m temperature, mean sea-level pressure, 2 m dew-point temperature, and several pressure-level meteorological variable groups, including geopotential height, wind fields, temperature, and specific humidity. All ablation experiments used the same test samples, forecast lead times, and 29 valid environmental grids, and were evaluated using PCC, RMSE, and MAE.

As shown in Fig. 2, removing the T+1 future meteorological fields causes the most pronounced performance degradation, especially for O_3 and NO_2. For example, the mean PCC of O_3 decreases noticeably, while its RMSE and MAE increase by approximately 25.4% and 27.7%, respectively. The errors for NO_2 also increase clearly. This indicates that the forecast skill of BiXiao does not simply come from the persistence or temporal autocorrelation of pollutant concentrations at T+0, but also benefits from information on future meteorological evolution. In comparison, removing pressure-level specific humidity also degrades the performance for some pollutants, especially PM_{10} and $PM_{2.5}$, suggesting that humidity-related processes contribute to particulate matter forecasting. The effect of removing surface temperature is more evident for O_3, which is consistent with the sensitivity of ozone formation to thermal conditions and photochemical processes.

We also note that removing some variables has only a small effect, and even leads to slight improvements for individual pollutants. This may be related to correlations and redundancy among meteorological variables, as well as model uncertainty under the limited training samples. Therefore, we avoid over-interpreting the importance of individual variables and focus instead on the most robust and consistent results, namely the contributions of future meteorological fields and humidity-related variables to pollutant forecasting. Because this comment was framed as an optional additional analysis and the main manuscript is already relatively long, we currently present the full ablation results in the response material. If the editor or reviewer considers it necessary, we would be willing to include this figure and the associated discussion in the Supplement or in the main manuscript.

Minor comments

- Line 1: There is a typo in the title: “AI-dirven” should be “AI-driven”.

Response:

Corrected.

- Lines 249–253 / Fig. 3 caption: The text says Fig. 3 analyzes PCC at 6 h, 48 h, and 72 h, but the Fig. 3 caption labels the columns as 24 h, 48 h, and 72 h. Please make the lead times consistent.

Response:
Thank you for pointing this out. The statement in the main text was a typo, and the correct lead times are 24 h, 48 h, and 72 h, as shown in the figure. We have corrected the text accordingly. It should also be noted that this paragraph has been rewritten in response to another reviewer's comment.

- Lines 342–344: Please check the spelling of the aerosol scheme "MADE/SOGARM"; it is commonly written as "MADE/SORGAM".

Response:
Corrected.

- Fig. 8 : the legend in the top panel should be "O_3 Obs".

Response:
Corrected.

- Line 400: The sentence beginning with "Figure 12. Average pollutant concentrations..." appears to be a figure caption accidentally placed in the main text. Please move it to the Fig. 12 caption or rewrite it as normal prose.

Response:
Corrected.

- Lines 407–412: The manuscript uses "non-continuous grids" elsewhere, but the conclusion switches to "non-uniform grid design". These terms may imply different grid structures. Please use consistent terminology throughout the paper.

Response:
Corrected. The term non-uniform is no longer used in the manuscript, and we now consistently use non-continuous grids or non-continuous environmental grids throughout the text.