Response to Referee #2: We would like to thank the editor for the careful review throughout the paper that helps to improve our paper.

Our Reply follows (*the editor's comments are in italics and blue*)

General Comments

This manuscript develops an EnKF-fusion method combining RFSML prediction and chemical transport model outputs to improve air quality forecasts, which has some good applications. However, the manuscript needs some major revision for publication. First, English languages need to be improved overall. Some of the sentences and words are causing confusion. Also, the organization of the manuscript can be improved to be more concise and clear. Please see the specific comments below.

Reply: Thanks for the comments. We have polished the English languages thoroughly and reorganize the manuscript accordingly.

Specific comments

1. The title is kind of confusing, it doesnot show the purpose of the work, yes, it is a method, but what it is used for?

Reply: The title was changed to comply with the journal policies according to the editor's suggestion. We now changed it from "***EnKF-based fusion of site-available machine learning air quality predictions from RFSML v1.0 and gridded chemical transport model forecasts from GEOS-Chem v13.1.0***" to "***A gridded forecast through fusing site-available machine learning air quality predictions from RFSML v1.0 and chemical transport model forecasts from GEOS-Chem v13.1.0 using EnKF*** " to make the purpose of manuscript more clearly.

2. The abstract needs to be revised. It is not clear what the method used is and what the improvement is like from this method. Also, there is no need to discuss a lot of the technical details in the abstract. It should also focus on the significance of this work.

Reply: To highlight the significance of the manuscript, **ABSTRACT** is now revised to "***Statistical methods, particularly machine learning models, have gained significant popularity in air quality predictions. These prediction models are commonly trained using the historical measurement datasets independently collected at the environmental monitoring stations, and their operational forecasts onward by the inputs of the real-time ambient pollutant observations. Therefore, these high-quality machine learning models only provide site-available predictions and cannot be used as the operational forecast solely. In contrast, deterministic chemical transport models (CTM), which simulate the full***

life cycles of air pollutants, provide forecasts that are continuous in the 3D field. Despite their benefits, CTM forecasts are typically biased particularly on a fine scale owing to the complex error sources due to the emission, transport, and removal of pollutants. In this study, we proposed a fusion of site-available machine learning prediction, which is from our RFSML v1.0, and a CTM forecast. Compared to the normal pure machine learning model, the fusion system provides a gridded prediction with relatively high accuracy. The prediction fusion was conducted using the Bayesian theory-based ensemble Kalman filter (EnKF). Background error covariance was an essential part in the assimilation process. Ensemble CTM predictions driven by the perturbed emission inventories were initially used for representing their spatial covariance statistics, which could resolve the main part of the CTM error. In addition, a covariance inflation algorithm was designed to amplify the ensemble perturbations to account for other model errors next to the uncertainty in emission inputs. Model evaluation tests were conducted based on independent measurements. Our EnKF-based prediction fusion presented superior performance compared to the pure CTM. Moreover, covariance inflation further enhanced the fused prediction particularly in the cases of severe underestimation."

3. P2, Line 1-2: the atmospheric environment has been improved, you mean air quality has been improved?

Reply: Accepted. "***atmospheric environment***" is now changed to "***air quality***".

4. P2, Line 3: how do you know if it is primary PM5 or secondary PM2.5?

Reply: We meant the primary pollutant $PM_{2.5}$. To make it clear, "*primary*" is now removed.

5. P2, Line 13: what do you mean by typical cities?

Reply: Here typical cities mean megacities in China, including Shanghai and Chengdu. "***Cheng et al. (2021c) successfully predicted ground daily maximum 8-h ozone concentrations in typical cities of China by utilizing wavelet decomposition and two machine learning models.***" is now replaced with "***Cheng et al. (2021c) successfully predicted ground-level daily maximum 8-hour ozone concentrations in several megacities in China, Shanghai and Chengdu, by utilizing wavelet decomposition and two machine learning models.***" in page 2, line 16-18.

6.P2, Line 15-25: These few sentences are quite confusing, I am not sure what the purpose is. Are you trying to say the RFSML is good or not?

Reply: Sorry for the confusion. Our intention was to demonstrate both the advantages and disadvantages of the RFSML. Our prediction fusion is specifically designed to address the unavoidable disadvantages of the RFSML. Now we have reorganized these few sentences by replacing "~~***Recently, we successfully***~~

~~*performed air quality prediction with high accuracy using the regional feature selection-based machine learning model (RFSML). Our forecast system is capable of providing national-scale short-term prediction over 1262 sites in China. Moreover, ensemble-SAGE feature selection algorithm was developed that can exclude those redundant inputs and efficiently improve the forecasting abilityy (Fang et al., 2022). These models were trained via the historical measurement datasets collected at the air quality monitoring stations independently, and their operational forecasts forwards with the inputs of the real-time air quality observations. Unlike the gridded forecast, these predictions are only available over the location of the air quality monitoring sites. Meanwhile, the spatial distribution of existing environmental monitoring stations is rather uneven. For example, the monitoring network in China is dense in east and very sparse in west as shown in Fig. 1*~~" with "***We have successfully used the regional feature selection-based machine learning model (RFSML) to predict air quality with high accuracy. Our forecast system can provide short-term predictions for over 1262 sites across China at a national scale. We developed the ensemble-SAGE feature selection algorithm to exclude redundant inputs, which efficiently improves our forecasting ability (Fang et al., 2022). We trained these models using historical measurement datasets collected at independent air quality monitoring stations, and they operate using real-time air quality observations as inputs. However, unlike gridded forecasts, our predictions are only available for the location of the air quality monitoring sites. Meanwhile, the spatial distribution of existing environmental monitoring stations is rather uneven in China, with a dense monitoring network in the east and a sparse network in the west, as shown in Fig. 1. Therefore, our RFSML predictions limited to these few monitoring stations cannot accurately represent the $PM_{2.5}$ concentrations on a national scale.**"

7.Paragraph 3 and 4 can be combined to discuss why model forecasts and data assimilation are needed. For example, line 30-35 on P2 mentioned a lot of previous modeling work but did not mention how well the models did, which is the point.

Reply: Thanks for the comment. The Paragraph 3 and 4 are of course closely related. The reason for separating them are as following:

(a) We would like to emphasis that air quality forecasts from CTMs are not perfect and list the main error sources in paragraph 3. Conclusive descriptions for the model forecast are added in page 3, line 4-9 by saying "***While those air quality forecast models can capture the spatiotemporal variations in ambient pollutants to some extent, they are susceptible to systematic bias, particularly at a fine scale, due to multi-source uncertainties in emission inventories (Keenan et al., 2009; Fan et al., 2018), initial and boundary conditions, and parameterization of physical and chemical processes such as transport and removal (Croft et al., 2012; Solazzo et al., 2017). This makes the CTM forecast less reliable for localized air quality predictions (Bi et al., 2022).***"

(b) In paragraph 4, we aimed to illustrate the both the CTM models and the observations have weakness when they are used in the reanalysis product. Data assimilation could combine them together and resulted in a gridded reanalysis dataset with high accuracy.

8.P3, line 5: what are the same challenges here?

Reply: More descriptions are now added in page 3 line 10-12 by saying "***Both the machine learning and CTM models have weakness when they are solely used in operational air quality forecasting. The same challenges exist when observations and simulation models are used to describe the atmospheric dynamics in reanalysis products.***"

9.P3, line 17-26: this paragraph is not clear. I cannot tell what the innovation is and what exactly the method is.

Reply: Thanks for the comment. We now rewrite this paragraph in page 3 line 24-35 as "***This study introduces the Bayesian theory-based assimilation method to fuse the regional-feature-selection machine learning forecast (RFSML v1.0) (Fang et al., 2022) and the deterministic chemical transport model (CTM) air quality prediction. The prediction fusion aims to achieve a gridded prediction with less bias and higher accuracy than the pure CTM prediction. It is continuous in 3D field unlike the machine learning forecast that is only site-available. To the best of our knowledge, this is the first time that the assimilation method has been applied in this way, as it is typically used for nudging model simulations with observations. The specific assimilation algorithm used is the ensemble Kalman filter (EnKF). The background error covariance of the CTM prior forecast is the fundamental term in the assimilation-based fusion. Ensemble CTM forecasts, which are driven by perturbed emission inventories, are forwarded in parallel to represent the potential distribution of ambient pollutant levels and the spatial covariance statistics. To avoid model divergence, an additional covariance inflation algorithm is developed that accounts for model errors other than uncertainties in emission inputs. The uncertainty of the other prior, the machine learning forecast, is also an essential part of the assimilation-based fusion. To accurately quantify the errors, dynamic covariance is designed.***"

10.P4, line 1-5: this paragraph is repetitive from the last paragraph of the introduction. Instead, you can briefly mention the study domain and period here to give an overview.

Reply: This paragraph is a duplicate and we have removed it accordingly. The study domain and period are introduced in ***Section 2.1 Study domain and observations***.

11.P4, line 10: six categories, you mean 6 regions?

Reply: The six categories refer to six regions, and we now use '***six regions***' to avoid any potential misunderstanding.

12.P4, line 15: the winter of 2019, can you give the specific study period instead?

Reply: It should be explicitly mentioned in this work. To explain this, remarks are now added in page 4, line 18-20 "***The winter of 2019 (from 15 October to 30 December 2019) was selected as the test period following the choice in our recently work (Li et al., 2022) as winter suffers the most severe haze pollution than other seasons in China.***"

13.This work has tested the predication skill of the method for 6, 12, 18, 24h, would it work for longer time periods such 48h or more?

Reply: We aimed to combine the accurate short-term machine learning forecast and the CTM forecast. The fused prediction then relied on the machine learning prediction results. As we had previously tested a prediction horizon of up to 24 hours with success, we also tested the same prediction horizon in this work. Based on our experience, it was observed that the RFSML predictive performance tends to deteriorate rapidly when predicting beyond a 24-hour horizon. Hence, we do not recommend using RFSML for predicting longer time periods. The notation is in page 17, line 23-25: "***Note that our fused prediction skill generally declines with an increase in prediction length following the RFSML. Therefore, prediction fusion with a longer forecasting horizon (>24 h) was not carried out.***"

14.I am not sure if Figure 2 can really explain much of the fusion method. For section 2.3, it is not clear how the RFSML prediction works, for example, what are the inputs used? Also, how do you first acquire the RFSML prediction and then do the fusion?

Reply: Thanks for the comments. We agree that the inputs and their relationships would make the schematic plot more comprehensible. We have modified the Figure 2 as shown below.

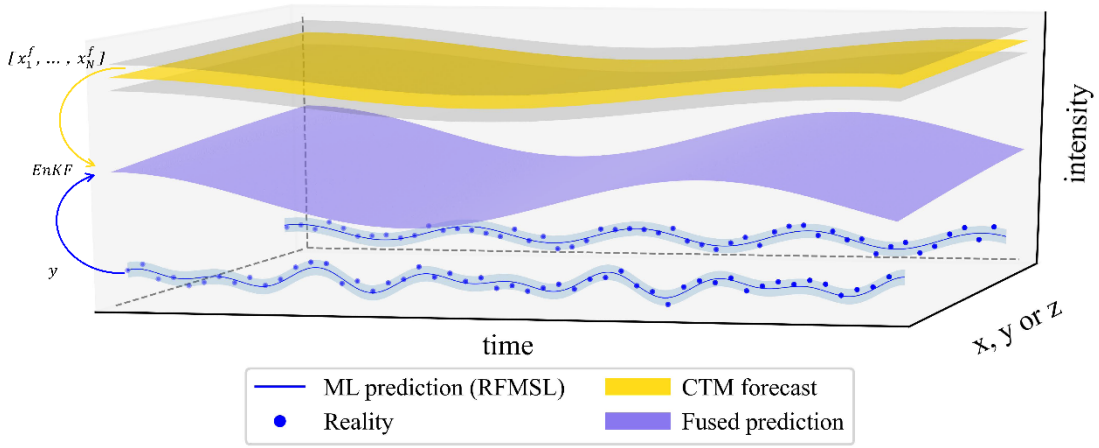

Figure 2. Framework of EnKF-based prediction fusion. The blue lines and their shadows imply for the RFSML prediction and its uncertainty at air quality monitoring stations, which are very close to the $PM_{2.5}$ reality. The gold surface and its surrounding grey surfaces stand for the CTM forecast and its uncertainty. The medium slate blue surface represents the fusion prediction of RFSML and CTM forecast.

The RFSML prediction is acquired from our last work "***Development of a regional feature selection-based machine learning system (RFSML v1.0) for air pollution forecasting over China***" (Fang et al., 2022). We will clearly explain it in ***Section 2.3 RFSML prediction & uncertainty*** in page 8, line 1-2 by saying "***Our RFSML is capable of providing the operational air quality prediction with a maximum horizon of 24 h. The RFSML prediction results used in this study are directly acquired from our last work (RFSML v1.0).***"

15.Figure 4, time series of surface PM5, right? This should be added in the caption. Also, where are the ground observations on the plot?

Reply: The mistake in the title is now corrected as shown below. Additionally, we have identified similar mistakes in the titles of Figure 4, Figure 6, and Figure S4, which are also corrected accordingly. Figure 4 was to demonstrate that the posterior forecast error sometimes remained high, particularly when the prior chemical transport model (CTM) significantly underestimated pollution levels. Mathmatically, this is because the CTM spread (shown as a silver shadow) was much lower than the uncertainty estimated by the random forest statistical machine learning (RFSML) model (shown as a blue shadow). Therefore, we did not include the ground observation here.

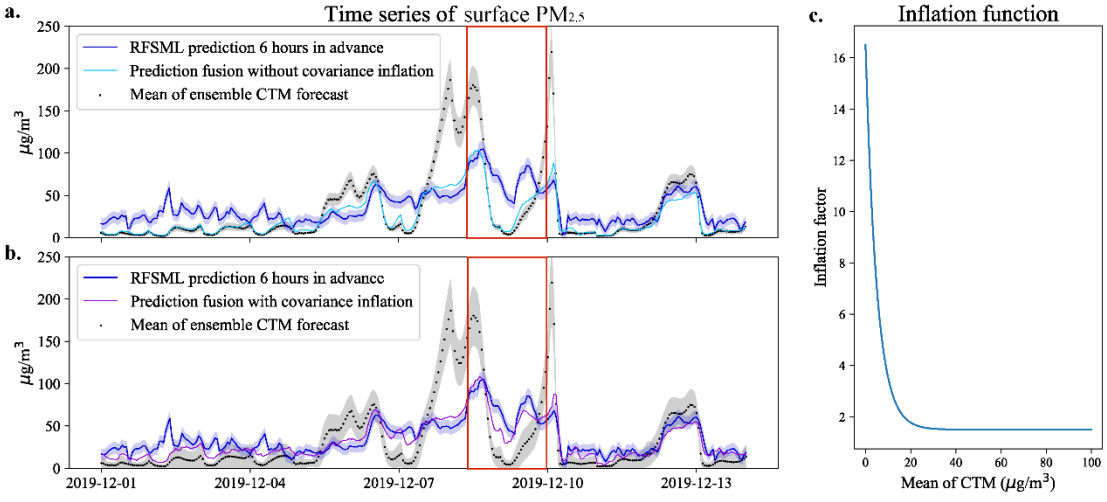

Figure 4. Panels a and b are the time series of an environmental monitoring station (Latitude:40.98°N, Longitude:117.95°E) in Chengde, Hebei Province. RFSML prediction is available at this station and will be assimilated. Markers of medium blue solid line, deep sky-blue solid line, dark violet solid line, and black dot represent the RFSML prediction, prediction fusion without covariance inflation, prediction fusion with covariance inflation, and mean of ensemble CTM predictions respectively. The

silver and blue shadows represent the uncertainty of the CTM and RFSML respectively. Panel c is the inflation function.

16.For Figure 6, can you also use scatter plots or Taylor diagrams to show the overall performance compared with ground observations?

Reply: Thanks for the suggestion. Scatter plots are now added to highlight the superior performance more clearly. Remarks about the plot will be added in page 13, line 26-30 by saying: "***To further highlight the superior forecast skill, we present the distributions of both prediction fusion methods and ground observations for a 6-hour prediction horizon in supplementary Figure S5. The results clearly indicate that the prediction fusions align closely with the ground observations, and that the prediction fusion with covariance inflation effectively addresses the underestimation issue present in the prediction fusion without covariance inflation.***"

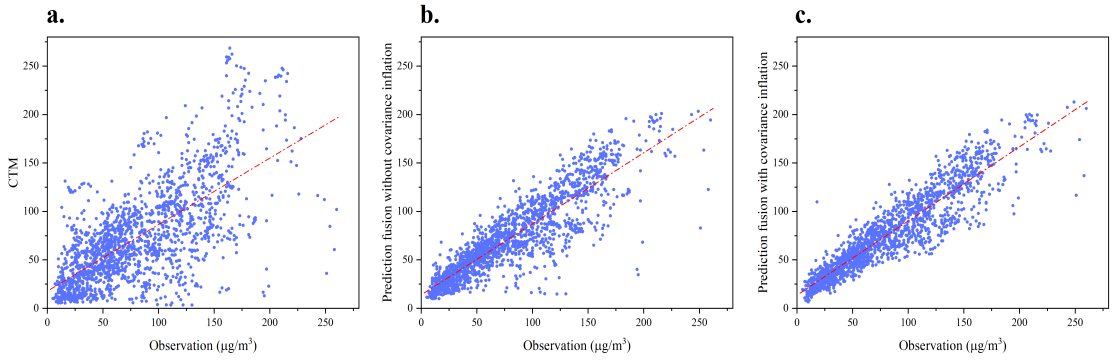

Figure S5. The distributions of model predictions and observations for a 6-hour predicting horizon. Panels (a), (b), and (c) illustrate the distributions between CTM, prediction fusion without covariance inflation, and prediction fusion with covariance inflation, and their corresponding ground observations, respectively.

17.Section 2.5: the radius tested is 3°, 5° and 10°. The distance is quite far, even using 3°, the distance between a ground station and grid would be about 300 km, the correlation between these two places would be low, if I understand the method correctly. If you reduce the radius, I assume you can see some impacts in the city area where ground observations are clustered. There is still value using the interpolation method. Also, what would be the computational cost for the interpolation method vs the fusion method? It is not quite convincing here just to make this comparison and I think it depends on the application.

Reply: Reducing the radius of spatial interpolation of course can improve its performance, especially in areas with dense ground observations. However, our work was primarily focused on addressing the challenge of producing gridded predictions across China. Cressman interpolation with radius less than

3° could meet the requirement. Therefore, we believe it is suitable for this specific task. So, we did not discuss the computational cost of the interpolation method vs the fusion method and just passed the interpolation method. We agree with the referee that it is worth noting that the interpolation method is computationally less expensive. In contrast, the fusion method requires running ensemble CTMs, which is much more computationally intensive. We now made these points more clearly by saying: "***It is worth noting that the interpolation method is computationally less expensive than the fusion method, and it can be a powerful tool for gridded prediction when there are plenty of ground observations available.***" in page 12, line 13-15.

18.What is the computational cost of this fusion method? What is the application of this method? Some discussions are needed for the implication of this work.

Reply: A new section for describing the computational complexity in ***Section 3.4*** by saying: "***In this study, all computations related to the prediction fusion were carried out on nodes equipped with 4 x 16-core 2.1 GHz Intel Xeon E5-2620 v4 CPUs and with memory 64 GB. The RFSML was demonstrated to be relatively efficient in computation as illustrated in Fang et al., (2022). While the ensemble CTM forecasts takes the most computation power, which however could be implemented in parallel. Each CTM model takes approximately 30 minutes to run a 24-hour simulation on average, with only 16 cores. The computational cost for EnKF fusion is also low, with an average time of three minutes for a prediction fusion. Overall, the proposed prediction fusion is time affordable.***"