Response to Referee #1: We would like to thank the referee for the careful review throughout the paper that help to improve our paper.

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

General Comments

In this article, the author proposes a high-precision grid prediction by fusing regional-feature selection machine learning prediction (RFSML v1.0) and CTM prediction. The set CTM prediction driven by the disturbance emission inventory is used to represent its spatial covariance statistics to solve the error, and the covariance expansion algorithm is designed to amplify the integrated disturbance and reduce the error, and the model evaluation is carried out on the basis of independent measurement. The prediction fusion based on EnKF is significantly improved than pure CTM. The manuscript has a good innovation, the content of the manuscript is detailed, the discussion is explicit, and the conclusion is clear. It is suggested to add some content for minor revision and publish it in the journal Geoscientific Model Development.

Comments

What are the input variables of machine learning model (RFSML)?

Reply: Thanks for the comments and this was indeed not clearly explained in our previous version. In our prior work, we identified the top three significant features for each region, as outlined in Table 6 of our publication. Additionally, we utilized the preceding nine hours for the forecast, resulting in a 27-dimensional input matrix. To explain this, remarks are now added in page 4, line 14-16 "***We identified the top three significant features for each region, as outlined in Table 6 of our publication (Fang et al., 2022). Additionally, we utilized the preceding nine hours for the forecast, resulting in a 27-dimensional input matrix for models.***"

It is suggested to add time-based verification results, that is, to use the data of two months for modeling and the data of the third month for verification.

Reply: Time-based verification is crucial for machine learning. In RFSML, we utilized 15,690 hours (from January 1, 2018, to October 15, 2019) for model training and cross-validation. The remaining 1824 hours of data from October 15 to December 30, 2019, were reserved for actual testing.

To further verify the predictive capability of RFSML in a rolling fashion, we tested the forecast fusion for a less polluted month, April 2020. The details could be found in the response to Question "*What is the effect if the model is applied to other seasons?*" in the below.

It is suggested to use a schematic diagram to show the variable input and output relationship between the machine learning model and the chemical transfer model, so as to facilitate readers' understanding.

Reply: We agree with the referee that including the input relationships would make the picture more comprehensible. We added the two inputs of EnKF in Figure 2 as shown below.

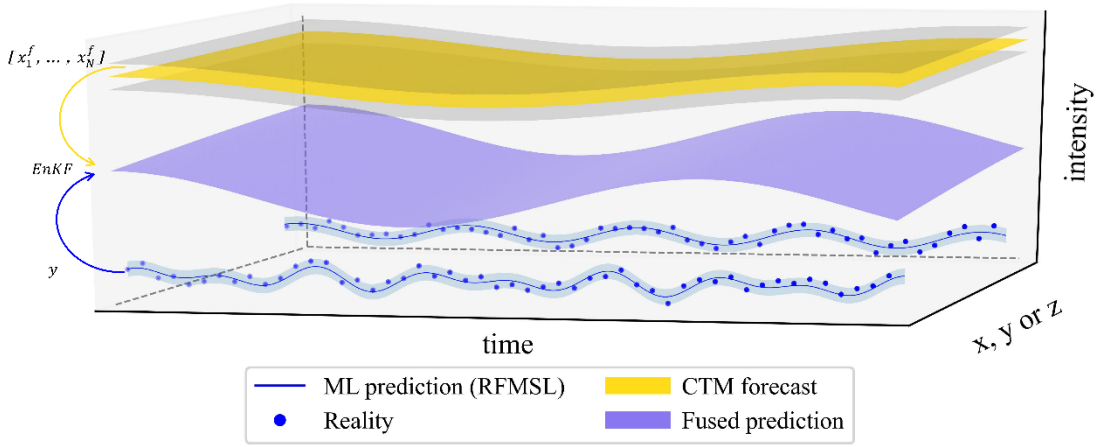

Figure 2. Framework of EnKF-based prediction fusion. The blue lines and their shadows imply for the RFSML prediction and its uncertainty at air quality monitoring stations, which are very close to the $PM_{2.5}$ reality. The gold surface and its surrounding grey surfaces stand for the CTM forecast and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

In what aspects is the feature selection of machine learning region reflected?

Reply: In our previous study, we demonstrated the significant role of feature selection in machine learning and conducted it for different regions. As the impact of feature selection varies across regions, the prediction performance of machine learning also varies across regions. Therefore, the use of prediction results from previous work affected the prediction performance in different regions in our study. Furthermore, the prediction performance of the CTM forecast varied across regions and also had an impact on our method.

What is the maximum forecast time for the prediction effect of the model to ensure certain reliability? Can it be applied to short, medium and long term forecast?

Reply: Thanks for the comments. Our method's implementation relies on relatively accurate machine learning prediction results, and therefore, we believe that the prediction time span is primarily constrained by the accuracy and duration of machine learning's future predictions. 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. The notation is in page 17, line 23-25: "***Note that our fused prediction skill also generally declines with an increase in prediction length following the RFSML. Therefore, prediction fusion with***

a longer forecasting horizon was not carried out." Based on our experience, we have observed that the predictive performance of hourly $PM_{2.5}$ prediction tends to deteriorate rapidly when predicting beyond a 24-hour horizon. We have noticed that some studies involve daily predictions of atmospheric pollutants, and we are of the opinion that our methods could be adapted for such applications. To summarize, we believe that our method can be applied to short-, medium-, and long-term forecasts provided that we have the corresponding predictions from machine learning.

What are the main reasons for the different prediction errors in different regions in Figure 9? For example, why does the prediction effect of PRD region have greater error and smaller R than that of other regions?

Reply: Similar to other assimilation systems, the proposed method fundamentally relies on the Bayesian theory to determine the optimal posterior that aligns with the two priors quantified by their respective covariance matrices. In our case, these priors are the RFSML and CTM predictions, along with their corresponding covariance matrices. Figure 9 shows that the CTM prediction of PRD (black star) has greater error and a smaller R value. The similar results found in Figure 8 of our previous work (Fang et al., 2022), where the RFSML prediction of PRD had the smallest R value. We believe that both factors contributed to the greater error and smaller R value observed in PRD. Remarks about this plot will be added in page 17, line 11-12 by saying: "***The smallest R value of PRD can be attributed to both the smallest R value of RFSML and CTM predictions.***"

What is the role of the integrated Kalman filter (EnKF) based on Bayesian theory?

Reply: The Ensemble Kalman Filter is an extension of the standard Kalman filter that uses an ensemble of model simulations to represent the system state and its uncertainty, and the implementation follows the Bayesian theory. To explain this, extra description is added in page 5, line 1-4 "***The specific sequential assimilation system that is used to combine the site-available RFSML prediction and CTM prediction is the EnKF that was originally proposed by Evensen (1994) and further corrected by (Evensen, 2004). Similar to other assimilation algorithms, this assimilation system fundamentally relies on the Bayesian theory for finding the optimal posterior that fits the two priors quantified by their covariance matrices (Evensen et al., 2022).***"

Does Figure 2 show that the prediction effect of fusion is worse than that of RFMSL according to actual observation.

Reply: We agree with the referee that our prediction fusion might not perform as good as the RFSML over the site where RFSML prediction is available. However, the focus of our work is to predict at grids where RFSML prediction is not available, hence the significance of fusion prediction. Because the

RFSML prediction is only site specific as explained in page 3, line 23-28: "***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 true $PM_{2.5}$ concentrations on a national scale.***"

It is suggested to add the advantages, disadvantages and prospects of the article.

Reply: We appreciate your suggestions and agree that adding a section on the advantages, disadvantages, and prospects of the article would enhance the analysis of the topic. We now added a new paragraph in the conclusion that will provide an in-depth analysis of the advantages, disadvantages, and prospects of the topic. To explain this, remarks are now added in page 19, line 26-30 "***In summary, the proposed fused prediction effectively overcomes the weakness of machine learning, which can only predict at specific sites. However, our method has some drawbacks, such as 32 ensemble CTM forecasts which are still computationally expensive. Additionally, the site-based RFSML prediction may have unavoidable errors in representing atmospheric dynamics of the grid mean, which we will address in our future work. This method can be extended to predict the concentrations of other airborne pollutants.***"

What is the effect if the model is applied to other seasons?

Reply: Thanks for the comment. We hold the belief that our fusion prediction is capable of handling other seasons. As demonstrated above, the prediction fusion is determined by the RFSML and CTM predictions, along with their corresponding covariance matrices. In this study, we only showcased highly polluted seasons. To further demonstrate the robustness of the proposed method, we will also conduct testing for a less polluted month (April 2020). The overall performance is illustrated in the Taylor diagram, as presented in Supplementary Figure S6. Remarks about this plot will be added in page 17, line 27-30 by saying: "***To further showcase the robustness of the proposed prediction fusion approach, we conducted testing for a less polluted month (April 2020) using prediction horizons of 6 hours and 18 hours. The overall performance of each region is illustrated in Supplementary Figure S6 using a Taylor Diagram. Our results demonstrate that the prediction fusion method outperforms CTM in all regions, and the incorporation of covariance inflation further enhances this advantage.***"

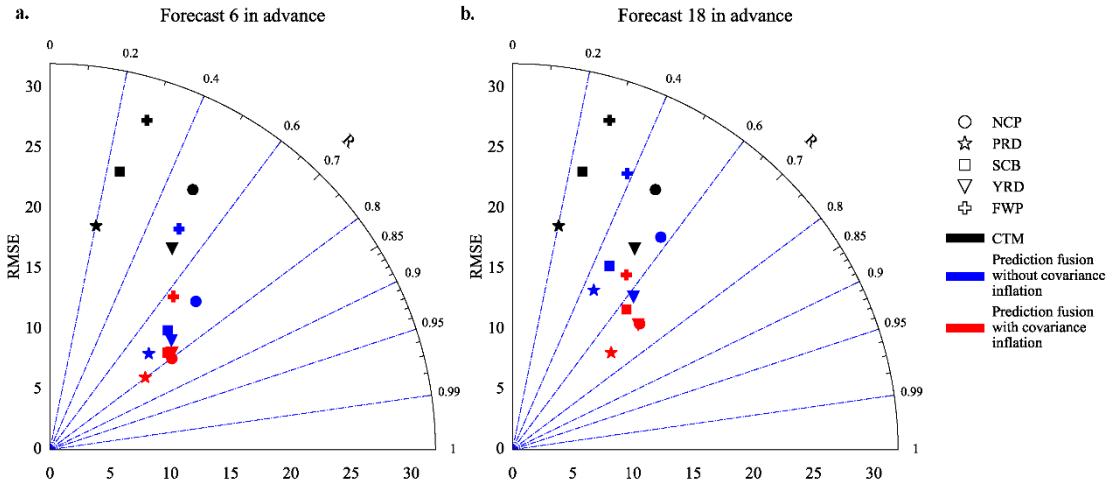

Figure S6. A modified Taylor Diagram that illustrates RMSE and R together. The regions of interest, including NCP, PRD, SCB, YRD, and FWP, are differentiated by unique markers of circle, star, square, triangle down, plus, and diamond, respectively. Additionally, the results from different approaches are visualized by distinct colors, with black, blue, and red indicating the results from CTM, prediction fusion without covariance inflation, and prediction fusion with covariance inflation, respectively. The diagram consists of two panels, representing forecasts for 6-hour and 18-hour time horizons, arranged in a left-to-right sequence.

You'd better analyze the prediction effect in different regions. Which regions are better predicted?

Reply: We have discussed the performance of our method in different regions in page 17, line 18-22, by saying: "***For example, in panel a, the CTM has the worst prediction in terms of R (<0.1) in the PRD region, but it increases to 0.62 when the RFSML forecast is assimilated and further increases to 0.85 when covariance inflation is implemented. In terms of the RMSE, the most remarkable improvement is obtained in SCB region. Our EnKF-based fusion reduced the RMSE from 43 μg·m^{-3} to 12.73 μg·m^{-3} and 10.97 μg·m^{-3} (with covariance inflation).***". Remarks about the prediction performance for each region will be added in page 17, line 10-16 by saying: "***In terms of the Pearson correlation coefficient, SCB shows the best predictability among the regions, while PRD has the poorest performance. However, no significant differences were observed in the predictive performance of these regions. Regarding the Root Mean Squared Error (RMSE), NCP and FWP exhibited the largest values. This outcome can be attributed to the fact that NCP and FWP are located in the northern region of China, where the frequency of pollution days is higher due to adverse meteorological conditions and high emissions during winter. It should be noted that the metric RMSE is directly influenced by the atmospheric pollution levels, wherein higher $PM_{2.5}$ concentrations tend to yield larger RMSE.***"

In the introduction and other section, it is suggested to add some references on the application of machine learning and prediction models (such as: https://doi.org/10.1016/j.scitotenv.2022.160928).

Reply: The suggested reference is indeed relevant and benefit our article. We now included the reference in the appropriate section of Introduction and ensure that it is properly cited in the text and included in the reference list. Remarks are now added in page 2, line 11-12 "***For instance, Chen et al. (2023) fully estimates hourly near-surface ozone concentration in China using a new geostationary satellite with the help of machine learning.***"