Dear referee #2:

Thank you very much for your comments on this work. The comments and suggestions are very useful to improve our manuscript. Here we provide a point-by-point response to your comments. The texts in blue are the comments, those in black are our response. The referenced line numbers correspond to the revised manuscript with changes marked.

Specific Comments:

1. While the index I_{dis}^{k} (Eq. 1) effectively quantities EM performance, the threshold for classifying "good" and "bad" EMs is not explicitly defined. In addition, how does the index consider the temporal and spatial performance variations? For instance, some EMs might perform well during specific periods or under particular conditions but poorly in others.

Reply: Thank you very much. In this study, the index I_{dis}^{k} represents the "distance" in model performance of an EM compared to the best performance. The smaller index value indicates better performance of the EM. therefore, we classified EMs with index ranking in the top 10% as "good" EMs, while those in the bottom 10% were classified as "bad" EMs. In this case, out of the 30 EMs, 3 EMs will be designated as good EMs and 3 as bad EMs. We believe this index can consider both temporal and spatial performance for the following reasons:

(1) For this ozone episode, the variation pattern of the ozone time series throughout the episode, along with the extremely high concentration observed in the afternoon of July 31, highlighted the temporal feature of this high ozone episode. These two factors are critical for evaluating whether the simulated ozone performed well. As a result, we implemented R to validate the ozone variation pattern and MNB to validate the high concentration in the afternoon of July 31. Since the index is calculated based on the contributions of both R and MNB, it incorporated the temporal features of the ozone episode.

(2) The ozone concentrations used to calculate R and MNB are derived from the average observed ozone across all stations within the GBA, as well as the average simulated ozone at corresponding grids. The average time series of ozone incorporated the spatial information of surface ozone in the GBA. Therefore, the index also reflects the spatial features of this episode to some extent.

2. More explanation and discussion on the construction of the I_{dis}^{k} index is needed. Why were the correlation coefficient (R) and mean normalized bias (MNB) chosen? Although R and MNB were normalized for comparability, is it reasonable to assign equal weight to both R and MNB since the model can better capture the variation trend of ozone in most EMs?

Reply: Thank you very much. We believe that capturing the variation pattern of the ozone time series is one of the most important factors in demonstrating that the performance of an ozone simulation is "satisfactory" in a specific region. Both the correlation coefficient (R) and the index of agreement (IOA) can evaluate this feature, and since their effects are quite similar, choosing either is acceptable. In this study, we applied R.

For this ozone episode, the extremely high concentration of ozone that occurred in the afternoon of July 31 is another significant feature which the simulated ozone should well reproduce; however, the base simulation failed to capture this aspect. Many statistical metrics can evaluate this feather, for example mean bias (MB), root mean square error (RMSE) and mean normalized bias (MNB). We chose MNB for two reasons:

(1) MNB is a dimensionless quantity and has been normalized, making it similar to R.

(2) The EPA (2005, 2007) provides benchmarks of MNB (±15%), offering a clear reference for evaluating the performance of ozone simulation.

Since both the variation pattern and the extremely high concentration are key features of this high ozone episode. It is hard to determine which is more important. In this context, we assigned equal weight to both R and MNB to calculate the index I_{dis}^{k}. And according to the classification results derived from this index, comparing to the bad EMs and the base simulation, the good EMs demonstrated better model performance in this ozone episode, both in terms of the variation pattern and the high concentrations. This finding suggests that the index and the classification method are effective.

3. Here, the ensemble simulations were conducted with the perturbated meteorological fields. Hence, the physical and chemical discrepancy between the "good" and "bad" EMs is mainly attributed to the change in the meteorological fields. Emissions are another important factor affecting ozone formation. Although biogenic emissions were indirectly perturbed due to the change in meteorology, most emissions remain consistent in this study. This point should be pointed out and further discussed.

Reply: Thanks for this comment. We completely agree with your opinion. Among all the EMs, the only differences in input data and parameterization were the perturbed meteorological variables in the initial and boundary condition files. Therefore, the discrepancy between good and bad EMs can primarily be attributed to the changes in meteorological variables. Based on the model validations and the findings in our study, it is also suggested that more accurate simulations of meteorological variables lead to improved accuracy in the distribution and variations of ozone concentrations. In addition, it also should be noted that emissions are another important factor affecting ozone precursors, which can further affect the ozone through photochemistry. While emission indexes have improved substantially in recent years, there are still considerable uncertainties in them which may lead to uncertainty in ozone simulations. Quantifying the relationship between emissions and ozone concentration through ensemble simulation is also an interesting topic that we plan to study in our future work. Thank you once again for highlighting this point. We have added relevant discussions to the revised manuscript. Please check the details at lines 121-123 in the revised manuscript.

4. The number of ensemble members in this study is 30. How does the ensemble number was decided?

Reply: Thank you very much. Based on the work of Zhang et al. (2009), an ensemble size of 30 has been found to be affordable and reasonable for simulating typical cyclones. Therefore, we decided to conduct our ensemble simulation with 30 members in this study.

5. The index of the base simulation can be added into Figure 3. It could facilitate comparison between base and ensemble simulation.

Reply: Thank you very much. The index of the base simulation has been added into the new figure. Please check Figure 3 in the revised manuscript.

6. The caption of some figures needs to be modified for better understanding. For instance, "ΔCHEM" in Figure 5c, does it represent the differences between the CHEM of "good" and "bad" EMs? or just the contribution of the chemical process in the "good" EMs? Similar problem should be checked for other figures.

Reply: Thank you very much. We have double-checked all the figures relating to the issue you mentioned. To enhance clarity, we have added “Good-Bad” to the figures that illustrate the differences between good and bad EMs or their captions. Please check the

updated figures and captions in the revised manuscript.

7. Some of the subfigures are small and need to be adjusted, as in Figure 5c. It should be mentioned at least once that the time zone on the x-axis is local time.

Reply: Thank you for your comment. We have double-checked all the figures related to this issue. For the relevant figures, we have added "LT" (abbreviation for local time) to indicate that the time is in local time. Please check the updated figures in the revised manuscript.

References

EPA, U.S.: Guidance on the Use of Models and Other Analyses in Attainment Demonstrations for the 8-hour Ozone NAAQS, EPA-454/R-05-002, 2005.

EPA, U.S.: Guidance on the Use of Models and Other Analyses for Demonstrating Attainment of Air Quality Goals for Ozone, PM2.5, and Regional Haze, EPA-454/B-07-002, 2007.

Zhang, F., Weng, Y., Sippel, J.A., Meng, Z., and Bishop, C.: Cloud-resolving hurricane initialization and prediction through assimilation of Doppler radar observations with an ensemble Kalman filter, Monthly Weather Review, 137: 2105-2125, 10.1175/2009MWR2645.1 2009.