Responses to the comments of Referee #1:

We would like to thank the reviewer for his/her time and effort for reviewing this manuscript. Please find below our responses.

Referee #1:

The paper provides two case studies critiquing CNN models trained for AQF applications. The first ML model is directly an estimator, the second is used as a corrector for a CMAQ model. The authors use wavelet modal decomposition and a shape-invariant distance metric as analysis tools to find discrepancies in the model predictions and trace them back to environmental factors. This analysis is valuable and interesting in itself. Both positive and negative results are provided.

I encourage the authors to rethink the vision of this paper. What is the central thesis of the paper? Does CNNs work better as post-processing tools rather than raw predictors? Are model biases inevitable in these applications no matter the configuration?

Response:

To respond to your suggestion and comments, the following statements are offered:

- Despite the enormous success of the convolutional neural network (CNN) algorithm in numerous applications, certain issues related to its applications in air quality forecasting (AQF) require further analysis and discussion. Our main goal in this paper was to discuss some of these issues is a few practical applications. In order to discuss these issues analytically, we used wavelet transform and dynamic time warping (DTW), as powerful mathematical tools for time-series analysis and models. Based on the findings that were presented in the paper, these tools are extremely helpful not only in understanding the issues with machine learning models but also in fine-tuning them to improve their performances with a scientific point of view. Awareness of the limitations in CNN models will enable scientists to develop more accurate regional or local air quality forecasting systems by identifying the affecting factors in high concentration episodes.
- We discuss the general issues of the CNN model in two common applications: (i) a real-time AQF model, and (ii) a post-processing tool in a dynamical AQF model, the Community Multi-scale Air Quality Model (CMAQ). As the referee correctly stated, these examples are fundamentally different in terms of execution, one being raw predictor (statistical approach) while other being a post-processor (hybrid approach). Since both models are commonly being used as a real-time air quality prediction systems, we discussed their issues individually to broaden researchers' view on certain issues that one may encounter in executing either of them. Thus, it will prove both machine learning researchers and atmospheric scientists with multiple candidate models and analytical tools to develop any specific model of their choice.

The authors diagnose important limitations of CNN models trained on their data but very few thoughts are offered for interested researches as to how to fix these issues. (except maybe in the conclusions). For example if there is a significant difference in the accuracy of the model in nighttime vs daytime, how does a single model compare to two models trained separately on subsets of data (day/night). If your analysis shows hidden correlations between the error and RH% how can you incorporate that into the input data?

When the model is not performing well, insufficient training (as suspected by authors) is only one possible cause. Another possibility may be under parametrization, such that the model is not complex enough to capture the details of special cases. I think providing error measures on the training data and comparing them with test data can illuminate the source of underperformance.

Response:

To respond to your suggestion and comments, following explanations are stated:

- Based on our findings in the base studies presenting the aforementioned CNN models, in both cases, the CNN model shows promising accuracy for ozone prediction, 24 hours in advance, in both the United States and South Korea. However, similar to other data-driven prediction tools, in a CNN model, the out-of-sample prediction error is almost always greater than the in-sample prediction error. Thus, since both CNN models were designed as a real-time air quality prediction models, the prediction error is inevitable, even though (i) both models were configured for optimum performance (based on the input or training samples), and (ii) in development of both models, careful cross-validation processes were followed to mitigate any systematic biases. In addition, a comprehensive explanation can be found in our previous works, including but not limited to, the potential reasons for underperformance of the CNN model, modeling configuration and fine-tuning processes, training and validation process, arrangements of input variables, scenarios to improve the modeling accuracy, etc. Please refer to Eslami et al. (2019a, 2019b, 2019c), Choi et al. (2019), Sayeed et al. (2020), and Lops et al. (2019), and the discussion within. The authors will be delighted to provide additional explanations if nessaccry to accommodate the referee's comments and suggestions.
- For one case (raw prediction model), we use the wavelet transform to determine the reasons behind the poor performance of CNN during the nighttime, cold months, and high ozone episodes. We find that when fine wavelet modes (hourly and daily) are relatively weak or when coarse wavelet modes (weekly) are strong, the CNN model produces less accurate forecasts. For the other case (post-processing model), we use the DTW distance analysis to compare post-processed results with their CMAQ counterparts (as a base model). For CMAQ results that show a consistent DTW distance from the observation, the post-processing approach properly addresses the modeling bias with predicted IOAs exceeding 0.85. When the DTW distance of CMAQ-vs-observation is irregular, the post-processing approach is unlikely to perform satisfactorily. We are currently working on an individual study to use the findings of this study to fine-tune both CNN models. In our work-in-progress study, we will provide a

78 practical approach in infusing scientific angel of air quality time-series into CNN prediction
79 models using advanced analytical tools.

80
81 *The authors state on line 46 : "Inevitably, a consequence of such enthusiasm in the field is the risk*
82 *of exaggerated expectations, fueled by results focusing on the general performance of ML models*
83 *compared to that of conventional statistical models" and give their previous works as examples.*
84 *At the very least this assertion needs a more detailed explanation.*
85
86 • To address the referee's comment on line 46, we changed the sentence as highlighted in the
87 following, and we applied the changes in the manuscript:
88 ○ However, the focus of these studies was the general performance of the model ML models
89 compared to that of conventional statistical models rather than identifying the shortcoming
90 of such models in explaining the uncertainties of prediction models. Such examples can be
91 found in studies by Eslami et al. (2019a, 2019b, 2019c), Choi et al. (2019), Sayeed et al.
92 (2020), and Lops et al. (2019). To achieve more reasonable outcomes, we must first explore
93 the current challenges we face when forecasting ambient air quality and then assess how or
94 even whether ML models can address the challenges to produce more accurate forecasting.