Response to Reviewer #2

The paper presents the use of an analog ensemble (AnEn) technique to improve air quality forecasts. The AnEn is applied to outputs of a numerical model for air quality, specifically on O3 and PM2.5. It relies on past observations and the corresponding forecasts to draw an ensemble of analog situations. The improvements in the forecast are assessed by means of different scores, and compared to references.

The presented method seems to improve the forecast in different ways and it might be relevant for this kind of applications. However, as I'm not working with air quality, I cannot judge how the method stands against other model output statistics or statistical postprocessing techniques in that context. The paper also gives the impression that the authors do not come from this field, as the provided specific literature on that topic is rather poor. There is no mention of other M.O.S. approaches, whereas there should be some. Moreover, I'm not certain about the novelty of this study compared to previous works of the authors.

Reply: To the best of our knowledge, we are proposing for the first time a novel approach to generate probabilistic predictions for air quality, which is based on a significant shift in paradigm with respect to traditional ensemble methods: i.e., rather than running a numerical model with several different configurations to create the ensemble members, we run the air quality model in real time only once, and then generate the necessary uncertainty quantification by inference from the training data set. Additionally, a strategic selection of the predictors is required for using the analog ensemble method in air quality applications. Specifically, the predictors have to be selected in such a way that they are able to (1) identify air pollution episodes of similar magnitude in the past, and (2) identify the meteorological and chemical conditions leading to similar past air pollution episodes. Following these two criteria, we selected O_3, $PM_{2.5}$, 10-m wind speed and direction, 2-m air temperature, 2-m specific humidity, and cloud cover as the predictor variables in our implementation of analog ensemble for air quality. The rationale for selecting these variables as predictors is now described in the revised manuscript and reproduced in response to your comment on introduction of predictors in the manuscript.

In this study, we compare the performance of analog ensemble against the Persistence ensemble and show that the analog ensemble performs better (section 3.1). We appreciate the reviewer suggestion of comparing our method against other approaches such as MOS but we would prefer to perform a comprehensive comparison of our methods with the others such as MOS using a common dataset for all the methods rather than comparing the results from different studies that focus on different regions with different objectives and model configuration. We have cited many papers on the ensemble modeling in the Introduction Section and also added more references on transport and dispersion modeling following suggestion from Reviewer #1 (e.g., we cite Galmarini et al., 2001; Galmarini et al. 2004; Kioutsioukis and Galmarini et al. 2014; Potempski et al. 2008; Potempski and Galmarini, 2009; Solazzo et al. 2012).

The whole manuscript is not so well written and is often difficult to follow. It should be rewritten in a more fluent way, and it should better describe the methods used. The frequent use of "we" and "our" is inadequate. A substantial work on the language should be done through the whole manuscript.

Reply: We apologize for the imprecise use of language here. The manuscript has been revised thoroughly and proof read by a native English speaker.

The predictors used in the method are introduced very late, in the middle of the results, while they should be introduced earlier. Moreover, there is no justification for the choice of these predictors. Please better explain the choice of the predictors and the method itself.

Reply: We introduced the predictors in Section 2.3 along with the description of CMAQ modeling system. The revised manuscript also provides a justification for the use of predictors. The new text is reproduced here for your ready reference. "The rational for selecting the aforementioned air quality and meteorological variables as predictor variables is as follows. O_3 and $PM_{2.5}$ allow us to identify pollution episodes of similar magnitude in the past. Temperature plays a vital role in several processes relevant to air quality including atmospheric chemical kinetics, biogenic emissions, and mixing. The wind speed and wind direction allow us to ensure that similar transport pathways contributed to the analogous air pollution episodes in the past. Humidity is selected for its key role in the formation and destruction of both O_3 and $PM_{2.5}$. Water vapor (H_2O) in conjunction with O_3 photolysis is the main source of hydroxyl (OH) radical, which in turn initiates photochemical production of O_3 through oxidation of different volatile organic compounds (VOCS). In the case of $PM_{2.5}$, humidity determines the aerosol water content, which is important for secondary aerosol formation. Cloud cover determines the amount of solar radiation available for atmospheric photochemical reactions that produce both O_3 and $PM_{2.5}$. In summary, the predictors are strategically selected in such a way that they are not only able to identify the pollution episodes of similar magnitude in the past but also identify the meteorological and chemical conditions leading to similar air pollution episodes in the past."
Regarding the method description, our previous paper (Delle Monache et al., 2013) already presents a step-by-step description of the basic technique and we have reproduced the necessary details here. Here, we focus on describing the ways in which the application of AnEn to AQ differ from previous applications rather than repeating the information from the published literature.

How does the method perform for extremes ? I suspect that the peaks, which are the most relevant to forecast, might not be well covered by the ensemble due to the very limited size of the observations that can be used as analogs. Additionally, how would the derived deterministic time series (the mean) work for more extreme values?
Reply: This is an excellent question, and we agree with the reviewer that an analysis of extreme events, which has now been added, significantly enriches the paper. To understand the performance of AnEn for extreme events, we computed the bias, RMSE, and correlation coefficient for both the CMAQ forecasts and AnEn derived deterministic time series of ozone and $PM_{2.5}$ using only the observations above the 95% quantile, computed independently at each lead time and observation location. The estimated bias, RMSE, and correlation coefficient for extreme events are shown in Figures R1, R2, and R3, respectively. A lower RMSE and higher correlation coefficient of AnEn derived deterministic time series for both ozone and $PM_{2.5}$ at all the lead times shows that AnEn outperforms CMAQ for the extreme events. However, the bias in AnEn is higher than CMAQ raw forecasts for extreme events of

$PM_{2.5}$ mainly because of substantial reduction in the number of available quality analogs when we consider only extreme events. A lower RMSE even at the lead times where AnEn bias is higher indicates that AnEn compensate the latter by reducing the random errors, i.e., Centered Root Mean Square Error (CRMSE) in the forecasts. Our future work will focus on a bias correction technique to reduce the AnEn bias for the extreme events. This information has been included in the revised manuscript.

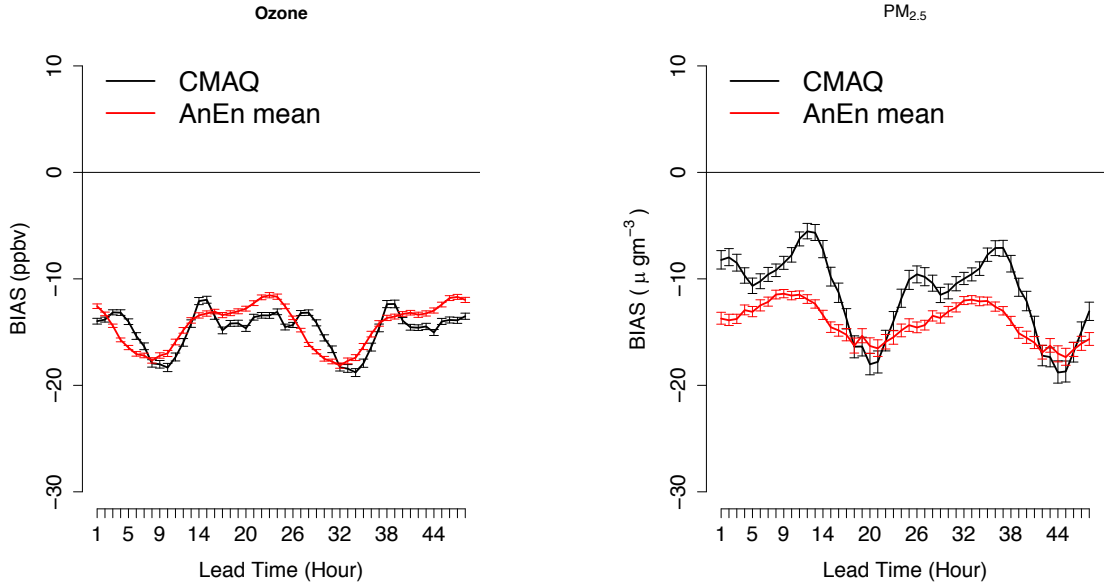

Figure R1: Estimated bias in CMAQ forecasts and AnEn derived deterministic forecasts of ozone (left panel) and $PM_{2.5}$ (right panel) for the extreme events that are identified as observations above 95% quantile of the distribution.

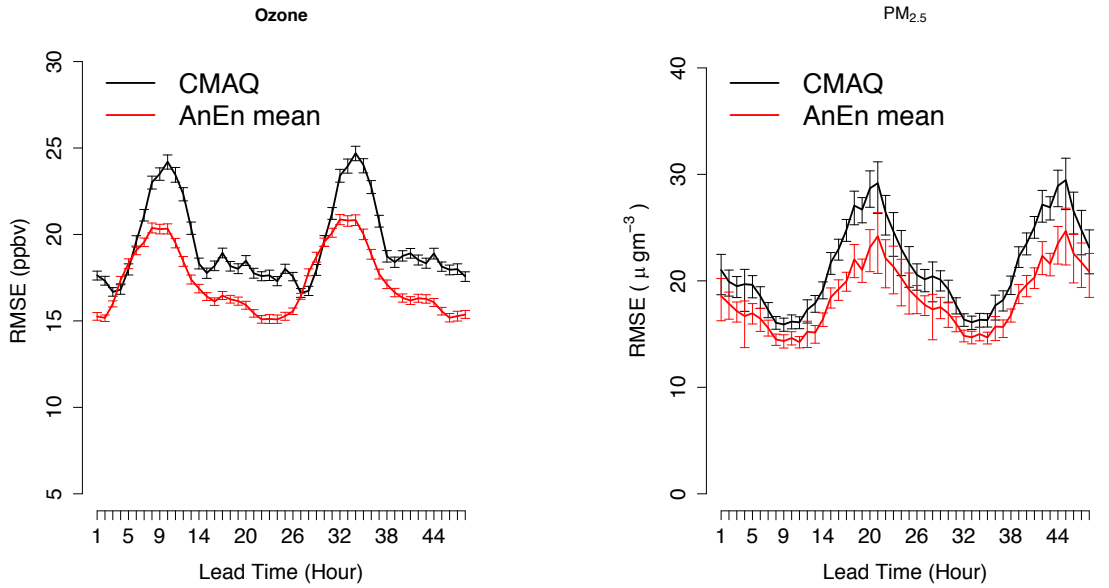

Figure R2: Same as R1 but for the RMSE.

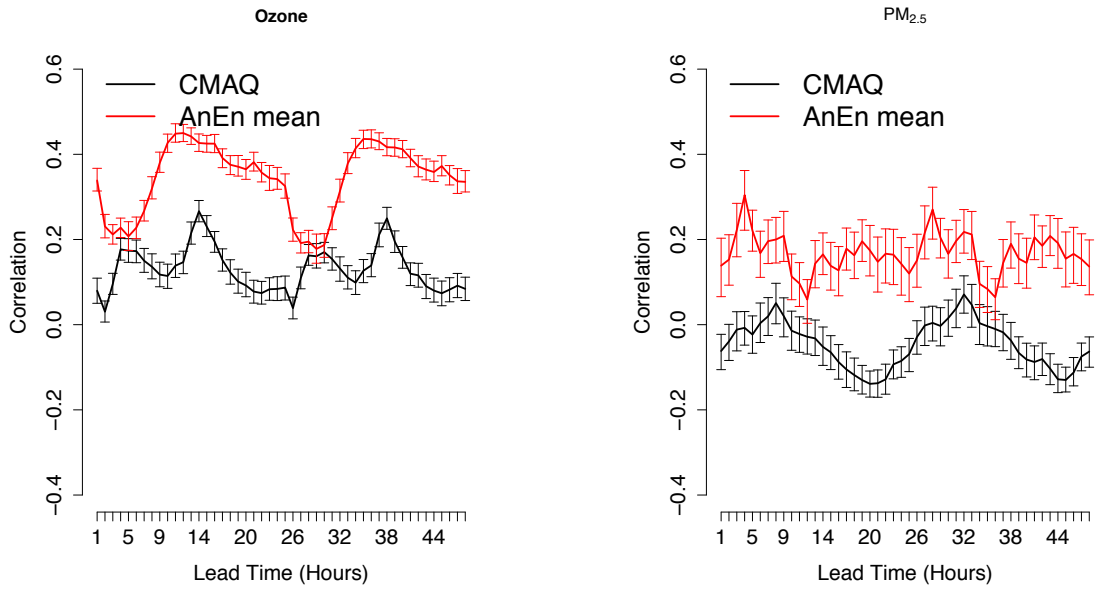

Figure R3: Same as R1 but for the correlation coefficient.

Specific comments: -

The calibration / verification periods should be clearly explained in the beginning of the manuscript, and the independence of the verification period specifically detailed. It is not clear which results are provided for the calibration or verification period. Is Fig. 2 in the verification period ?

Reply: The verification and training periods are now defined right before section 3.1. The following text has been added to the manuscript. "The verification periods are selected as 1 June to 30 September 2015 for O_3 because O_3 is a major air quality problem during summertime; and 1 December 2015 to 29 February 2016 for $PM_{2.5}$ because $PM_{2.5}$ pollution is higher during wintertime. Consequently, the training periods for O_3 and $PM_{2.5}$ are selected as 1 July 2014 to 31 May 2015, and 1 July 2014 to 30 November 2015, respectively." Yes, Fig. 2 represents the verification period.

- The number of references are unbalanced. There are too many for some points (e.g. P2L14-17), while some assertions have no reference.

Reply: We agree that we have a large number of references at this line but we think it is really important to acknowledge the previous research conducted on this subject.

- P3L4-5: not clear
Reply: This sentence has now been changed to:

"A *well resolved* ensemble is one that provides a probability close to 100% on occasions when an event (e.g., ozone above 100 ppb) occurs and forecast close to 0% when the event

does not occur. I.e., it is specific from case to case about what conditions will and will not occur.

- P3L16-17: issue with the ()
Reply: Thanks for pointing this out. Correct ( ) are placed now.

- P5L5: geopotential height at 500hPa is generally is generally a predictor in analog methods, not a predictand (likely the same for 2-m dew point)
We agree with the Reviewer that Geopotential height at 500 hPa has been used as a predictor in past applications of AnEn, and it is an important variable to consider when generating weather forecasts. In fact, through and ridges in its field are a proxy for indicating area of instability and/or underling low/high pressure systems. It can be relevant for air quality as well, but likely not as much as the predictors we have selected. We now recognize that the list of predictors we selected may not be exhaustive (at the beginning of the second paragraph of section 2.3).

- P6L19: t=1: what is the unit ? days ?
Reply: We apologize for the confusion and this has now been corrected; 1 represents hours.

- P7L2: specify the section where the number of analogs is optimized

Reply: The section number is mentioned now.

- P10L8: Figure 3

Reply: Yes, we meant Figure 3. Corrected.

- P11L2: sensitivities ?

Reply: Changed to sensitivity.

- Figure 2: On the verification period? Is it the best reproduced days, or are they representative of the skill of the method?
The verification period has been chosen independently of the model's performance. We chose a winter period for $PM_{2.5}$ and a summer period for O_3 because these are the season when these pollutants have higher concentrations.

- Figures 3, 4, . . .: a) b) c) d) not present on the figures
Reply: Corrected.

- P14L3: May and Nov or May to Nov ?

Reply: This are May 2015 and Nov 2015 and not May to Nov 2015.

- P17L11-14: + they might not sample the observation archive uniformly

We agree with the reviewer. As already mentioned, these testing periods are those most significant because they usually include high pollution episodes for the two pollutants considered.

- P20L2-5: Not clear how you process it

Reply: The sentence has been broken down into two parts now to improve readability. The CPRS is equivalent to the mean absolute error of deterministic predictions relative to the observations.

- P20L12: They are = or very close in some cases!

Reply: Yes, the CPRS for AnEn and PeEn are similar at lead time of around 72 h and 135 h. That is why we wrote that AnEn has a better (i.e., lower) CRPS than PeEn for "most of the lead times" of O_3 predictions and not for "all the lead times".

- P21L18-19: not clear
We added some text to better clarify the point.

- P22L2: a slightly better resolution, but not much. . .
Reply: The improvement of the AnEn over PeEn in resolution are about 10% and 15% for ozone and PM2.5 respectively. We specified this in the paper.

- P24L16-18: Not clear which spread you are taking about
Reply: We specified that the spread is defined as the standard deviation of the members about the ensemble mean.

- The summary should not contain the details of the periods, but the results should be more discussed.
Reply: We believe that it is helpful to remind the readers of the period for which the study is conducted in the Summary section of the manuscript.