Authors' Response (in red) to Referee #2's Comments and Changes in Manuscript (in blue)

The paper presents some interesting findings about inherent uncertainties in chemistry transport model simulations of ozone concentrations. It is very well written, but sometimes hard to follow. In my opinion, some terms should be explained in more detail, before the paper can be published.

We thank the reviewer for the positive feedback on our paper.

General comments:

The authors should explain why they think the data set they constructed by combining measured base line ozone with meteorology related short term variations from a 21 years CMAQ run could be seen as the output of a "perfect model". They claim that there is some inherent variability in the meteorological data that cannot be captured by any model system. However, reanalysis data may represent meteorologically related variations on time scales of few days very well, i.e. part of the variation included in the SY component of the time series may be modelled quite well.

We have revised the discussion in Section 3.2 to clarify that this analysis combining the observed baseline component with 21 CMAQ synoptic components is meant to quantify the amount of model error present in the current simulations that could conceivably be reduced through improving the representation of synoptic-scale processes and/or increased horizontal resolution. As part of this revision, we no longer refer to this combination of the measured baseline and modeled synoptic component as "perfect model". Moreover, we have added text at the end of the introduction to clarify that when we refer to the errors that can be expected even from a "perfect" model with "perfect" inputs throughout the manuscript, we consider these errors to be those arising from atmospheric stochasticity which we estimate in Section 3.1 using historic observations. We also should have noted in our original manuscript that four-dimensional data assimilation (FDDA) was utilized following the methodology suggested by Gilliam et al. (see *Atmos. Environ*., Vol.53, pp 186-201, 2012) and modified for fully-coupled meteorology-chemistry model applications as described in Hogrefe et al. (see *Atmos. Environ.*, Vol. 115, pp 683-694, 2015). As the reviewer pointed out, the SY component in the model output contains some meteorologically related variations on time scales of few days. The following revisions and additions have been made to address the reviewer's comment:

Revised and expanded the end of Section 1 (page 2, lines 29 – 32 in the original manuscript) as follows: "Also, no assessments have been made to date on the errors that are to be expected even from "perfect" regional-scale air quality modeling systems. To estimate such irreducible model errors due to atmospheric stochasticity (which we consider to be the errors that are expected even from a "perfect" model with "perfect" inputs), we analyzed the observed daily maximum 8-hr (DM8HR) ozone time series data at monitoring locations across the contiguous United States (CONUS) during the 1981-2014 time period and present the results of this analysis in Section 3.1. In Section 3.2, we illustrate how this information could be used in guiding model development specifically aimed at addressing reducible errors in the synoptic component by contrasting the results from Section 3.1 with analysis using the synoptic component from a 21-year simulation performed with the fully coupled WRF-CMAQ simulations covering the 1990-2010 period."

Expanded the model description in Section 2 (added after page 3 line 7 in the original manuscript) as follows: "To ensure better characterization of the prevailing meteorology (i.e., synoptic forcing) in the retrospective 21-year WRF-CMAQ simulations, four-dimensional data assimilation (FDDA) was utilized following the methodology suggested by Gilliam et al. (2012) and modified for fully-coupled meteorology-chemistry model applications as described in Hogrefe et al. (2015). The model set-up and performance evaluation of these historical multiyear WRF-CMAQ simulations have been published by Xing et al. (2015), Gan et al. (2015), and Astitha et al. (2017)."

Modified the discussion in Section 3.2 (page 5, line 23 – page 6, line 9 in the original manuscript) as follows: "In this section, we analyze long-term records of model simulations in an attempt to quantify the error associated with the modeled SY component that results both from not explicitly representing stochastic variations in atmospheric dynamics and emission variability in the current generation regional air quality models and from other reducible sources of model error. … To isolate the impact of model imperfections on only the SY time scale on errors across the ozone distribution, we assume that the model perfectly reproduces the 'true' BL depicted by the observed 2010 BL. We then use this 'perfect' modeled BL and reconstruct 'pseudo-simulated' ozone time series, similar to what was done in Fig. 3, except for using the SY component embedded in the 21 years of coupled WRF-CMAQ simulations. The rationale for this analysis is to quantify the amount of model error present in the current simulations that could conceivably be reduced through improving the representation of synoptic and mesoscale processes and/or increased horizontal resolution with appropriate data assimilation techniques."

Why do you use 30+ years of ozone measurements for analyzing the observations while only 21 years can be used for comparisons to the model results? Wouldn't it be enough to look at the data set 1990 to 2010 for the observations, too? And why do you need to construct "pseudo ozone observations" and cannot use the observational data set as such? Please explain this in the text.

As noted in our paper, any observation at a given percentile represents an event or a single realization out of a population. The object of our paper is to quantify the inherent variability in the observations due to the stochastic nature of the atmosphere. To this end, we thought that the use of 30+ years of historical data rather than 21 years would help in making more robust estimates of the expected errors even from "perfect" models driven with "perfect" input. As demonstrated in our previous research (see Porter, et al., 2017 *Atm. Poll. Res.*; Astitha, et al., 2017 and Luo, et al., 2019 in *Atm. Env*.), the baseline forcing can be viewed as the deterministic part in observations while the SY forcing is the near-stochastic part. We superimposed 30+ adjusted SY forcing on the baseline embedded in 2010 ozone observations to generate 30+ representations of ozone concentrations, which are equally likely to occur at any given probability point stemming from the stochastic nature of the atmosphere. Details on this approach can be found in Luo, et al. (2019 AE). The discussion in Section 3.1 of the revised manuscript was modified in response to the reviewer's comments.

Updated the second paragraph of Section 3.1 (page 4, line 25 – page 5, line 11 in the original manuscript) as follows: "Once the scale separation is achieved with the KZ5,5, we superimposed the SY forcing imbedded in 30+ years of historical DM8HR ozone time series measured at a given location on the baseline component of the ozone time series at that location to generate 30+ reconstructed or pseudo ozone distributions. Illustrative results using eq. (3) at a suburban location in Altoona, PA are presented for 2010 base year in Fig. 3a … note, it is equally likely for any of these 30+ CDFs to occur because of the stochastic nature of the atmosphere even though the individual event in 2010 yielded the CDF shown in red. As mentioned before, ozone mixing ratio at any given probability point on the red line in Fig. 3a reflects an individual event while ozone values at the same probability in different CDFs (gray lines) reflect the population stemming from the stochastic nature of the atmosphere. In other words, there are 30+ dynamically consistent ozone time series attributable to the 2010 baseline (given 2010 emissions) for examining the inherent variability due to atmospheric stochasticity. … Using these 30+ pseudo-observation ozone mixing ratios and the actual observed ozone values at each percentile, statistical metrics such as Bias, RMSE, coefficient of variation (CV=standard deviation/mean), normalized mean error (NME) and normalized mean bias (NMB) are presented in Fig. 3b and c (see Emery et al. (2016) for the description of the statistical metrics considered here). … The extreme values are better described in statistical terms rather than in deterministic sense (Hogrefe and Rao, 2001)."

Specific comments:

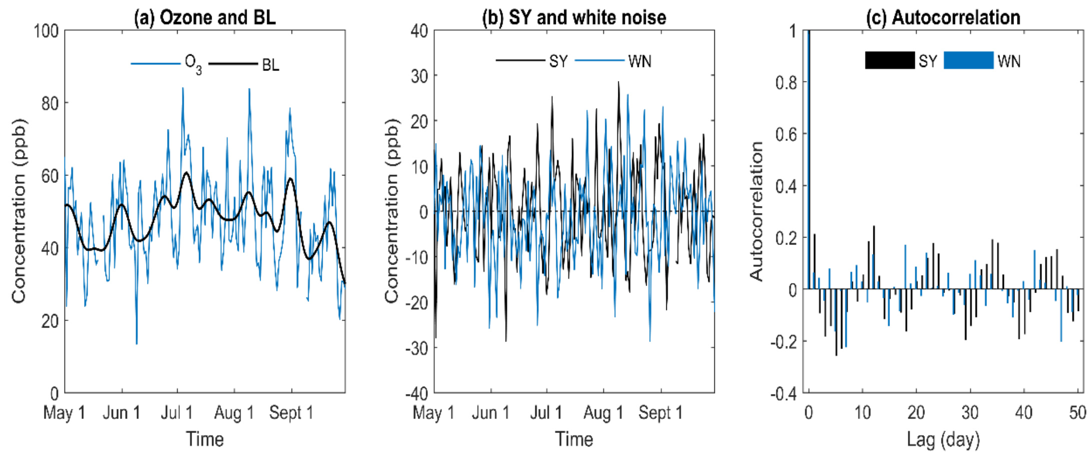
Page 3, line 26: In equation (1) it should be made clear that the filter KZ(5,5) is applied to the ozone time series O3(t).

Yes, the reviewer is correct that the SY component is estimated by applying the KZ filter with a window length of 5 days and 5 iterations to the ozone time series O3(t) as described in Porter et al. (2015), Rao et al. (2011), and Luo et al. (2019). We updated the notation in equations (1) and (2) and also changed occurrences of KZ(5,5) in the text to $KZ_{5,5}$ to better reflect the operation of the filter.

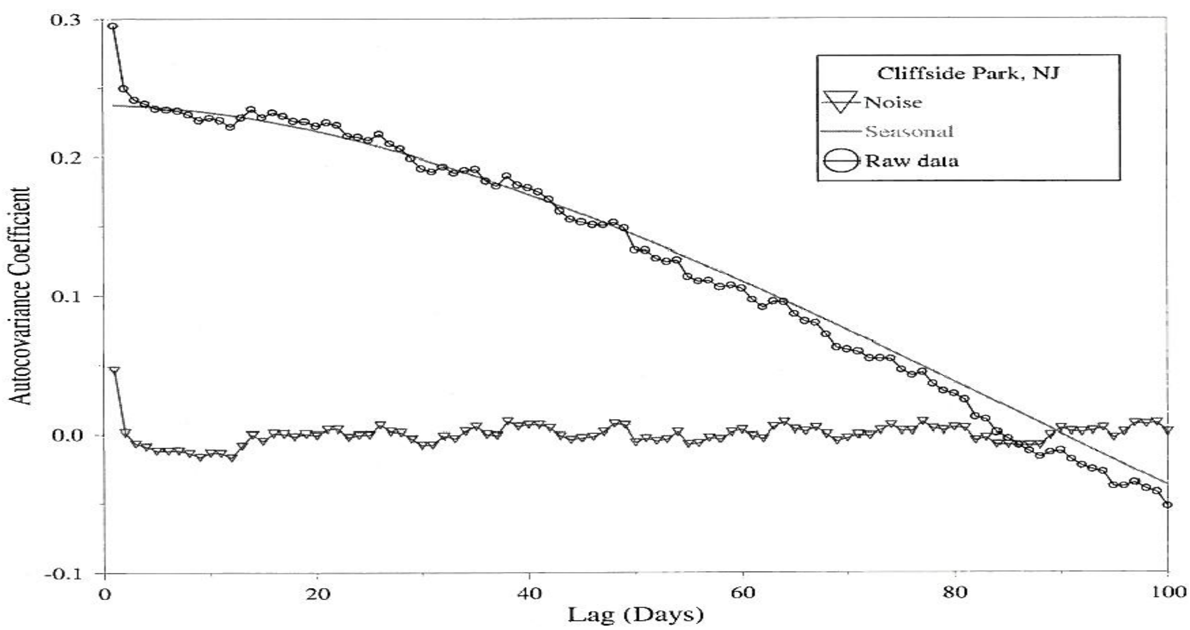
Equations 1 and 2: changed "KZ(5,5)" to "$KZ_{5,5}(O_3(t))$". Also changed all occurrences of KZ(5,5) in the text to $KZ_{5,5}$

Page 4, line 17: Can you show by statistical evaluation that the SY component represents white noise.

Because we haven't used the concept of white noise process to statistically model the SY forcing in our paper, we've removed the reference to white noise in the revised manuscript. However, to respond to the referee's question, we display below the autocorrelation function for time series in SY forcing and white noise. It is evident that the correlation drops off after 1-day lag.



In addition, we display below the autocovariance function as a function of lag (days), extracted from the article titled "Dealing with the ozone non-attainment problem in the Eastern United States" by Rao et al. on Page 20 in the January 1996 issue of the EM Magazine, a publication of the Air & Waste Management Association. These results reveal that the short-term variation (SY component) in observed ozone time series data is statistically indistinguishable from "white noise" with an autocorrelation coefficient that drops to zero at a lag of 1 day.



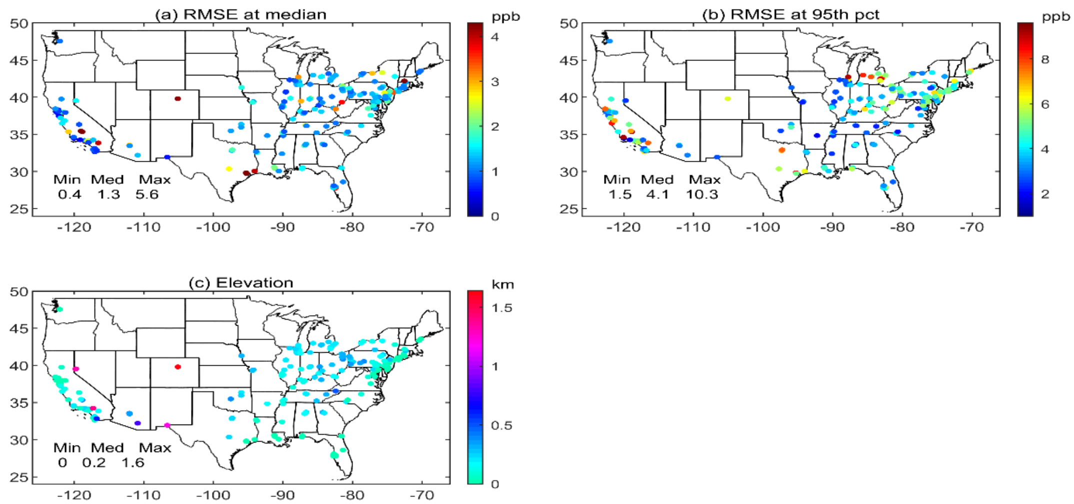
Page 4, line 18: please explain AR(1)

An AR(1) autoregressive process is the first-order process, meaning that the current value is based on the immediately preceding value. However, since we removed the references to "white noise" and AR(1) from our manuscript, this explanation has not been added to the revised manuscript.

Page 5, line 18-20: It isn't obvious for the reader which stations are at elevated sites.

We agree that adding this information would be useful for the reader and have added a panel to Figure 5 that shows the elevation of each monitoring site over CONUS.

The following new figure (Figure 5c) has been added to the revised manuscript:



Page 5, line 25: Define the "strength" of the SY component (being the standard deviation of the time series) here.

The definition has been added to the revised manuscript as follows:

"it should be noted that the linear relationship between the strength of SY (defined as the standard deviation of the data in the synoptic component) and the magnitude of the BL (defined as the mean of the data in the baseline component) has been taken into account in generating 30+ years of adjusted SY forcing as illustrated in Luo et al. (2019)"

Page 6, line 8: Is there any reason why you selected this site?

It has complete data for 30+ years. We could have picked another site; they all exhibit same features.

Page 6, line 10/11: Is there an explanation why the model doesn't perform well for low concentrations?

The CMAQ team at the U.S. Environmental Protection Agency is investigating the reasons for model's poor performance at the low end of the concentration distribution. Plausible causes include 36-km grid spacing not resolving the effects of NO titration in urban areas, errors in atmospheric deposition, representation of variability in background concentrations of O_3, precursor and reservoir species, etc. Also, it should be noted that the stochastic variability affects mostly the lower and upper tails of the pollutant concentration distribution.

Page 6, line 31/32: Couldn't this also be caused by emissions missing the correct temporal variation?

Yes, it is possible. To address the reviewer's comment, we have expanded an earlier sentence that also discusses the limitations of the current model setup as follows:

Revised sentence (page 6, lines 1-2 in the original manuscript): "The 36-km grid may be better representing the large-scale synoptic forcing associated with the translation of weather systems than the meso-scale weather and urban influences (both dynamics and emission driven) that are embedded in the observed SY component."

Page 7, line 9: How could these "slow changing processes" be improved in the models? What is the role of stratosphere/troposphere exchange which – to my knowledge – isn't well represented in the CMAQ model runs.

As already indicated in the manuscript discussion, one needs to pay more attention to properly specifying the lateral boundary conditions, duration/strength of stratosphere-troposphere exchanges, Madden-Julian Oscillation (MJO), ENSO, climate change, control policies, spatio-temporal variability in emissions loading, etc. The hemispheric CMAQ simulations used to drive the regional CMAQ runs used here employed potential vorticity-based scaling to represent O_3 in the model's upper troposphere-lower stratosphere (UTLS). The method was subsequently enhanced to represent seasonal and latitudinal dependencies in the relationship between potential vorticity and ozone and improved the 3-dimensional O_3 distribution represented by the model as well as the seasonal impacts of STE on lower tropospheric and surface-level O_3 as detailed in analyses presented in Xing et al. (2016) and Mathur et al. (2017). Interested readers are pointed to these studies and references are also included in the discussion.

Modified the second paragraph of Section 2 (Page 3, Lines 11-19 in the original manuscript) and added two references as follows: "It has been shown that time series of the daily maximum 8-hr ozone concentrations contain fluctuations operating on different time scales (e.g., intra-day forcing induced by the fast-changing emissions and atmospheric boundary layer evolution; diurnal forcing induced by the day and night differences; synoptic forcing induced by the passage of weather systems across the country, sub-seasonal forcing due to Madden-Julian Oscillation (MJO), and long-term forcing induced by emissions, El-Nino-Southern Oscillation (ENSO), climate change, and other slow-varying processes such as seasonal and sub-seasonal variations in the atmospheric deposition and stratosphere-troposphere exchange processes) as noted by Rao et al. (1997), Vukovich, (1997), Hogrefe et al. (2000), Porter et al. (2015), Astitha et al. (2017), Xing et al. (2016), and Mathur et al. (2017))."

Mathur, R.; Xing, J.; Gilliam, R.; Sarwar, G.; Hogrefe, C.; Pleim, J.; Pouliot, G.; Roselle, S.; Spero, T.L.; Wong, D.C.; Young, J. Extending the Community Multiscale Air Quality (CMAQ) modeling system to hemispheric scales: overview of process considerations and initial applications. *Atmos. Chem. Phys*. 2017, 17, 12449-12474, https://doi.org/10.5194/acp-17-12449-2017.

Xing, J., R. Mathur, J. Pleim, C. Hogrefe, J. Wang, C.-M. Gan, G. Sarwar, D. Wong, and S. McKeen, Representing the effects of stratosphere-troposphere exchange on 3D O_3 distributions in chemistry transport models using a potential vorticity based parameterization, Atmos. Chem. Phys., 16, 10865-10877, doi:10.5194/acp-16-10865-2016, 2016

Page 7, line 13: See my comment above: I couldn't fully understand how you constructed the "perfect model data". As far as I understood it, they would in any case only be perfect for this model setup and model grid. Could you comment on this?

We have revised the discussion in Section 3.2 to clarify that this analysis combining the observed baseline component with 21 CMAQ synoptic components is meant to quantify the amount of model error present in the current simulations that could conceivably be reduced through improving the representation of synoptic-scale processes and/or increased horizontal resolution. As part of this revision, we no longer refer to this combination of the measured baseline and modeled synoptic component as "perfect model". Moreover, we have added text at the end of the introduction to clarify that when we refer to the errors that can be expected even from a "perfect" model with "perfect" inputs throughout the manuscript, we consider these errors to be those arising from atmospheric stochasticity which we estimate in Section 3.1 using historic observations. We also should have noted in our original manuscript that four-dimensional data assimilation (FDDA) was utilized following the methodology suggested by Gilliam et al. (see *Atmos. Environ*., Vol.53, pp 186-201, 2012) and modified for fully-coupled meteorology-chemistry model applications as described in Hogrefe et al. (see *Atmos. Environ.*, Vol. 115, pp 683-694, 2015). As the reviewer pointed out, the SY component in the model output contains some meteorologically related variations on time scales of few days. The following revisions and additions have been made to address the reviewer's comment:

Revised and expanded the end of Section 1 (page 2, lines 29 – 32 in the original manuscript) as follows: "Also, no assessments have been made to date on the errors that are to be expected even from "perfect" regional-scale air quality modeling systems. To estimate such irreducible model errors due to atmospheric stochasticity (which we consider to be the errors that are expected even from a "perfect" model with "perfect" inputs), we analyzed the observed daily maximum 8-hr (DM8HR) ozone time series data at monitoring locations across the contiguous United States (CONUS) during the 1981-2014 time period and present the results of this analysis in Section 3.1. In Section 3.2, we illustrate how this information could be used in guiding model development specifically aimed at addressing reducible errors in the synoptic component by contrasting the results from Section 3.1 with analysis using the synoptic component from a 21-year simulation performed with the fully coupled WRF-CMAQ simulations covering the 1990-2010 period."

Expanded the model description in Section 2 (added after page 3 line 7 in the original manuscript) as follows: "To ensure better characterization of the prevailing meteorology (i.e., synoptic forcing) in the retrospective 21-year WRF-CMAQ simulations, four-dimensional data assimilation (FDDA) was utilized following the methodology suggested by Gilliam et al. (2012) and modified for fully-coupled meteorology-chemistry model applications as described in Hogrefe et al. (2015). The model set-up and performance evaluation of these historical multiyear WRF-CMAQ simulations have been published by Xing et al. (2015), Gan et al. (2015), and Astitha et al. (2017)."

Modified the discussion in Section 3.2 (page 5, line 23 – page 6, line 9 in the original manuscript) as follows: "In this section, we analyze long-term records of model simulations in an attempt to quantify the error associated with the modeled SY component that results both from not explicitly representing stochastic variations in atmospheric dynamics and emission variability in the current generation regional air quality models and from other reducible sources of model error. … To isolate the impact of model imperfections on only the SY time scale on errors across the ozone distribution, we assume that the model perfectly reproduces the 'true' BL depicted by the observed 2010 BL. We then use this 'perfect' modeled BL and reconstruct 'pseudo-simulated' ozone time series, similar to what was done in Fig. 3, except for using the SY component embedded in the 21 years of coupled WRF-CMAQ simulations. The rationale for this analysis is to quantify the amount of model error present in the current simulations that could conceivably be reduced through improving the representation of synoptic and mesoscale processes and/or increased horizontal resolution with appropriate data assimilation techniques."

Caption of Figure 2: Explain the meaning of the number 420130801. Explain that the "observed" line in Fig 2a is for 2010.

The number represents the site # in EPA's AQS database; it has no specific meaning other than being an identifier for the location of the monitoring site.

The figure caption (Figure 2 in the original manuscript, Figure 3 in the revised manuscript) has been updated as follows: "Figure 3a: Comparison between the observed cumulative distribution function (CDF) for 2010 shown in red with 30+ pseudo-observations CDFs generated from historical DM8HR ozone time series shown in gray at a suburban site at Altoona in PA (AQS station identifier 420130801). The blue line represents the average of the 30+ gray lines; Figure 3b: Display of various statistical metrics (standard deviation (std), root mean square error (RMSE), bias) derived by comparing the actual observed and pseudo ozone values in Fig. 3a; Figure 3c: Normalized statistical metrics of normalized mean error (NME), normalized mean bias (NMB), coefficient of variation (CV). Notice the large variability occurring at the lower and upper percentiles"

Caption of Figure 3: Give the equations you refer to somewhere in this paper (e.g. in an appendix).

The paper by Emery et al. (2016) and many other papers included in the references list recommended the statistical metrics such as RMSE, NME, Bias, NMB, CV, etc. whose definitions are well known.

Updated the second paragraph of Section 3.1 (page 5 line 9 of the original manuscript) in which Figure 3 (Figure 2 in the original manuscript) is discussed as follows: "see Emery et al. (2016) for the description of the statistical metrics considered here"

Figure 5: Use same colors for observations and model in both graphs.

Done.

Caption of Figure 6: "5)" should be "c)". "Light blue" in Fig 6d) appears to be grey. Figure 4 and Figure 7: Give units (ppb).

Done.