Reply to the 1st reviewer

General Comments:

It is well known that air quality model outputs contain errors because of errors in the model input data, inadequate representation of the physical and chemical processes in the model, and numerical schemes chosen to solve the relevant equations. Also, if the initial state of the atmosphere is not known, its future state cannot be predicted. Therefore, it is necessary to develop and apply methods for correcting errors in model outputs. Several studies using methods such as data fusion, Kalman filtering, and ensemble modeling have been published in the literature for correcting the errors in the modeled concentrations. This paper presents the application of the KZ filter to examine the performance of individual models in an ensemble and develop bias-corrected pollutant concentrations. Also, a technique is presented for improving air quality forecast with the KZ filter. The results of this study should be of interest to the research community.

We thank the referee for the many helpful suggestions. Our responses are detailed below. In order to better clarify our statements and enhance the readability of the paper, we would like to include to the manuscript some additional plots and a table. They deal with some properties of the kz model, discussed in section 3.3 as well as in the new section 4.4. In particular, the new plots address the fractional decomposition of the mean square error (7d), the decomposition of the mean bias error (7e), the skill of the modeled spectral components (7f) and the independence of the components (table 7, discussed in new section 4.4).

Specific Comments:

• *Note, the KZ filter has been applied before by Kang et al (2008) to produce bias-corrected air quality forecast. Previous studies have demonstrated that regional-scale meteorology and air quality models are not capable of simulating the intraday variations seen in the observations. If the week-to-week variation of the ID component is negligible (i.e., if ID is nearly invariant in time), why not replace the modeled intra-day forcing with that seen in the observations in coming with the best model?*

Compared to the other components, the ID is indeed less successful in capturing the pattern (Figure 7f). But there is also high spread between the members skill. As a result, for the majority of the sub-regions, generally one-two models (variable by region & aggregation) contribute to the ID component of *kz* (Figure 11). Considering the low error fraction the component entails (Figure 7d), this suggestion would not affect appreciably the *kz* skill. However, the use of a harmonized approach across all components, as outlined in six steps in section 2.2, does not make use of any prior knowledge on model performance (besides last week) aiming to facilitate better the interpretation of the results.

• *The higher correlation seen for the diurnal forcing is attributable to day and night differences. How well do these models in the ensemble simulate the amplitude of the diurnal oscillation in the observations?*

Figure 7f shows that the variance of the modelled DU signal is reasonably represented by many models. The spread of the values in the Taylor plot is variable across the sub-regions and in general, it is correlated to the variation of the modelled shortwave radiation (Vautard et al., 2012).

• Because of the leakage of energy across neighboring spectral bands (i.e., intraday and diurnal), would it be better to employ the wavelet technique rather than the KZ filter to extract the diurnal forcing in time series data? Also, why not consider only two spectral components, namely, the short-term (ID+DU+SY) and long-term (seasonal and trend) components, for this analysis because the KZ filter can better separate these two forcings in time series data?

We chose the KZ filter because of its simplicity and optimality. The use of wavelets would yield independent components at the expense of a multi-parametric model. This suggestion would be ideal for an analysis of the spectral bands but its performance in *forecasting* (article objective) is not straightforward. Focusing on variants of KZ forecasting, we have added a new table with links to component independence and error of other spectral combinations generated with the KZ filter (Table 7). The separation between short and long term indeed gives more independent components but results in forecasts with higher error (due to the negative error covariance in the case of dependent components).

• What's the explanation for the large differences seen between the modeled and observed long-term (i.e., baseline) components? If a regional model doesn't properly simulate the longer-term forcing seen in the observations, should it even be treated as a member of the ensemble because it could yield an inappropriate ensemble product (e.g., median model mm)?

Some models tend to allocate dissimilar variances to the spectral components, in comparison with the observed decomposition. As this property is not evident through usual time-series analysis, neither it states that their forecast skill is low, we made use of all available members seeking always for the optimal components.

• There are edge effects associated with the moving average filter, making the information at the tails of the time series unreliable. Hence, how does this affect the results if the KZ filter is used for the forecast period? From the operational perspective, how could this scheme be implemented since it relies on calculating all components right through the present (actually, right through the end of the forecast period in the forecast "F-Step") while we know that the SY and BL components estimated with the KZ filter are heavily influenced by edge effects for the last half length of the KZ?

An investigation of the edge effects for one sub-region demonstrated that if distortions could disappear, we could acquire up to 10% gain in the RMSE of kz. This is attributable to the higher persistence seen in SY and LT components. However, even with those distortions, kz still performs better on average. In addition, one may only consider the first 5 forecast days of the forthcoming week to minimize such distortions.

• The proposed method and analysis is framed in a forecasting context, but its usefulness was demonstrated with hindcast model runs that utilized data assimilation. This distinction matters because the method seems to rely heavily on the idea of "error persistence", e.g. the model with the lowest SY/BL RMSE during the last seven days is assumed to predict a good SY/BL component during the next seven days. I would think that meteorological data assimilation tends to make a given model's ozone error more persistent in time as that error is now more influenced by how well (or poorly) that model represents chemistry rather than how well the met model can "forecast" transport. In other words, I would expect a meteorology model's skill to vary more with changing conditions so that one week's skill is quite different from next week's skill when that meteorological model is run in the forecast mode. Therefore, the finding that the method works well when applied to these hindcast model runs does not necessarily imply that it will work well with forecast simulations.

The skill of the models varies with sub-region, synoptic conditions and chemical conditions. The persistence approach is a simple way to extract members on the basis of the most accurate recent representation of the observed state. More sophisticated approaches (utilizing eg synoptic clustering) could be used in cases of meteorological uncertainty.

• *The paper contains sentences throughout that are either too wordy or vague. For example, I don't understand the sentence "...This new approach to ensemble analysis is motivated by the illusory conception that the statistical treatment would account for the process variability and by the fatal assumption that model results are independent" in the introduction.*

We have rephrased the corresponding sentences.