Review of "Evaluation and Error Apportionment of an Ensemble of Atmospheric Chemistry Transport Modelling Systems: Multi-variable Temporal and Spatial Breakdown", by Efisio Solazzo et al, ACP, 2016.

My rating for this paper is minor revisions; no additional analysis is necessary. At the same time, I have a number of comments, questions and suggestions for the authors, which would improve the usefulness and "cite-ability" of the paper by the general research community.

Main Points:

(1) Regarding the limitations of KZ filtering (page 5, lines 190 to 199): The authors state that "a clear-cut separation of the components of Equation (8) is not achievable, since the separation is a non-linear function of the parameters m and k … and the leakage among the components mixes together in each component different physical processes". I agree with the authors that the choice of m and k values which have been used to date in their and other analyses quoted, along with the construction of equation (8) from the differences between KZ low-pass filters of relatively close m,k pairs, results in unwanted energy overlap across the spectral components. However, there are other options which could be used to minimize the potential for energy overlap. For example, the frequency analysis of the KZ(103,5), KZ(13,5), KZ(3,3) pairs carried out by Hogrefe et al (2000) (their Figure 1 on page 2086 of that article) shows the nature of the overlap issue – the KZ filter does not have a sharp cut-off in energy as a function of frequency, so that, for example, the low-pass KZ(3,3) passes 100% of the 1/week variation, while the KZ(13,5) passes about 13% of the 1/week variation (with the result that about 13% of the 1/week energy overlaps between the "SY" and "DU" time series, and differences between the two may have interference due to this overlap). The unmodified KZ filter is thus imprecise, though there are strategies which could reduce this imprecision. For example, rather than making use of the KZ filter as a band-pass through differencing, one could choose m,k values which represent the <u>complete</u> elimination of energy for frequencies higher than the given limit. Specifically, the frequency of the KZ filter's 50% energy pass limit is given by the equation below:

$$\omega_0(m,k) = \frac{\sqrt{6}}{\pi}\sqrt{\frac{1-\left(\frac{1}{2}\right)^{\frac{1}{2k}}}{m^2-\left(\frac{1}{2}\right)^{\frac{1}{2k}}}} \qquad (1)$$

From inspection of Hogrefe et al (2000)'s energy diagram, it can be seen that the low frequency cutoff limit (i.e. the frequency above which 99% or more of the energy will be removed by the low-pass filter) is about 2.82 times the 50% frequency from the formula above. One can thus choose values of m,k for which most of the energy is removed (e.g. a KZ(523,3) will remove 99% of the energy corresponding to periods shorter than 30 days, KZ(95,5) will remove 99% of the energy corresponding to periods shorter than 1 week, KZ(17,3) will remove 99% of the energy corresponding to periods shorter than 1 day). Using these KZ(m,k) values (and comparing the analyses for them) will also show the impact of the different time scales just as well as the band-pass approach currently in use by the authors - without the issue of energy overlap due to attempting to use KZ as a band-pass. This as an alternative to attempting a band-pass by

differencing two close low-pass filters. Another option is to use the modified KZ filter known as the KZ Fourier Transform (KZFT), wherein the original moving average is multiplied by a complex exponential function centered on the desired center wavelength. This is a better option for band-pass than the differencing in the references quoted by the authors, though it has the disadvantage of being a very narrow band-pass (see Yang and Zurbenko, WIREs Comp Stat, 2, pp 340-351, 2010).

My point here is not that the authors approach is invalid (it has limitations, and they've stated its limitations accurately) – but there are other ways to make use of the KZ filter which will be less prone to energy overlap (and thus blurring of the impacts of time scale) aside from the strategy used to date. i.e. while a "clear cut separation of the components of equation 8 is not achievable", one doesn't necessarily need to use equation (8) to recover the effects of different time scales with a KZ filter, and there are other strategies which can get around this problem. A few lines of discussion acknowledging these possibilities should be added to the existing discussion.

(2) The discussion on the emissions inventories (lines 211 to 237) was a bit hard to follow. Lines 211 to 220 read like a single inventory was used, while lines 224 to 225 mention two inventories, and which inventories were used for which models is not always clear. Some of this seemed to contradict some of the information about the individual modelling systems appearing later in the manuscript (where modified emissions are mentioned in some model system descriptions), with the result that the reader is not able to determine exactly which emissions inventories were used with which model, and the extent to which emissions were invariant between modelling systems. The authors should clarify this by including the emissions inventory(/ies) employed in each model in their summary table comparing the models, and modify the text accordingly.

(3) The text descriptions of the models were uneven in the level of detail – some described all of the individual model parameterizations with references, some were much shorter, some overlapped the information in the table, some did not, some described processes not described in others. This makes it difficult for the reader to understand the differences between the different modelling systems, hence draw inferences for the differences in model results. Rather than repeat the table, could the authors use the text in this section to describe only those components of the models which are unique from the others, particularly for the case of multiple implementations of the same model (e.g. have one WRF-CHEM main description followed by a paragraph describing the variations used in the study, ditto for WRF-CMAQ, etc.)? Part of what readers of the article will want to do is determine which key differences between the models are responsible for some of the differences in model results – this is difficult to do with the current formatting.

(4) Data analysis methodology, lines 441 – 443 and 449 – 451: the means of hole-filling for data gaps in the temporal records for the accepted stations should be described (e.g. local interpolation for smaller gaps? Average over all values for all gaps?). Lines 449 to 451 are a bit unclear: why was spatial averaging carried out and what were the domains? I think this may need a line or two at this point in the text to the effect of "hierarchical clustering was used to determine sub-regions with similar characteristics – spatial averaging within these sub-regions

was carried out due to the similarity of the observation data within these regions implying they will experience common chemistry"… or words to that effect.

(5) For the analysis itself (sections 3 and 4): the analysis tended to focus on how the models performed, as opposed to why differences in performance took place. The former is a valuable service in describing the state of the science, which has now appeared in all three phases of AQMEII – but the latter is of interest for those wishing to use the comparisons to further improve model performance. I'm hoping that the authors could take the time (I'm thinking a few days of discussion followed by an additional page of text in the manuscript) to delve a little bit deeper in their evaluation to suggest/speculate why certain models had poor performance for some predicted variables while others had better performance, in order to provide guidance to the community on how to move the science behind these simulations forward. Some examples:

   a. Lines 518-522: This subset of models had the worst performance for wind speed – what makes them different from the other models in this regard? A particular variation of the met driver? Different surface characteristics?
   b. Lines 548-550: This is an important result – a common problem across many models. For those models which seemed to be the least affected by this problem – what makes them different from the other models?
   c. Dry deposition discussion (section 3.2): WRF-DEHM was different from the other models – why? What is different about that model's deposition setup which might give rise to this result?
   d. Lines 573 – 576: There is a factor of 7 difference between the different model's PM2.5 deposition for the EU – what are the main differences in model PM2.5 processes between the models which could contribute to these differences?
   e. Section 3.3.1 – most of the error seems to reside in the LT component as bias – but not all models are the same; can the authors suggest to what components of the models the differences might be attributed?
   f. Lines 720-724: The common model EU negative bias of the mean NO2 is an important result – noting that the winter bias is usually positive, this implies that the summer bias may be quite negative. What possible causes might contribute to this bias, based on the different models' performance? Common positive bias of the PBL height (except in winter) perhaps? Photolysis rates too high? Shading effects missing, forest canopy or urban canopy? Emissions estimates for residential combustion low? – Line 751 suggests emissions as the key feature – but there is variation across the models which might give some insights into other factors.
   g. Lines 869-878: Most SO2 emissions are due to large stack sources. How are SO2 emissions distributed in the vertical in the different models? Are they all using the same plume rise algorithm? Is there any correlation between model vertical resolution and SO2 performance (LT bias)? The ECMWF-L-EUROS, WRF-WRF/Chem2, and ECMWF-chimere models had a large negative bias – are there any commonalities between these models that might account for this common negative bias? For that matter, what are the main differences between WRF-WRF/Chem1 and WRF-WRF/Chem2 which might

account for the substantial difference in SO2 bias between these two relatively similar models? Meanwhile WRF-CMAQ3 has a large positive bias – what makes it different from the other implementations?

h. Section 3.3.6: the SY correlation for PM2.5 is poor for three specific models (WRF-CAMx, WRF-Chem1, and WRF-Chem2) – why? What do these models have in common and/or are different from the other models?

i. Section 4 – the models' performance for this covariance analysis seemed to show the most variation across northern Germany and the Benelux countries; compare WRF-CAMx and ECMWF-L-EUROS to WRF-CMAQ3, CCLM-CMAQ-N. The ECMWF based models seemed to get positive numbers there, WRF based models negative. The implication is a meteorological driver bias leading to a difference in O3 memory. What met factors might be having this effect? Is there a corresponding regional temperature bias, for example? WRF-Chem1 and WRF-Chem2 had different performance – what's different between these implementations which might lead to these differences.

These above are a few examples I noticed from the work – which shows in detail the extent to which the models differed, and at different time scales, but doesn't discuss why they might be different to any great extent. I recommend the authors include a paragraph or three in the conclusions suggesting possible causes for these differences, and recommendations for their investigation.

(6) Several times in the discussion, the authors attribute common poor diurnal (DU timescale) performance on poor meteorological performance, since the latter has a significant diurnal variation. I agree that this may be one possible cause of the problem – another might be poor quality of the diurnal portion of the temporal variation in the driving emissions (c.f. Makar, P.A., Nissen, R., Teakles, A., Zhang, J., Zheng, Q., Moran, M.D., Yau, H., diCenzo, C., Turbulent transport, emissions and the role of compensating errors in chemical transport models, Geosci. Model Dev., 7, 1001-1024, 2014), where we showed some examples of the impact of poor temporal splitting of specific source types on model performance). How well does the temporal variation in the input CO emissions in the EU (see lines 607-616) correspond to observed near-source variations? Also, DU and smaller time-scale performance may correspond to errors in the wind direction taking the modelled plumes from sources in a different direction from reality. In that respect, a wind direction comparison in addition to wind speed would be very useful (is this do-able with the submitted data)?

Minor issues:

Line 397: HZG has not been defined.

Line 441: the means of hole filling for data gaps should be outlined – were averages of the entire period used for all gaps, or were smaller gaps filled by local interpolation, for example?

The inset map figures are I think supposed to show the station locations for the vertical profiles – these locations are very difficult to make out. I don't see why the inset maps need to show any sort

of concentration field (impossible to read that for their size anyway) – please replace with a white background with a large symbol showing the station location.

Lines 560 to 565: Not really clear to the reader how the deposition figures were generated; please clarify. A total accumulation in deposition would be a single number for each model, while these are distributions. The different models had different horizontal resolutions – were the deposition outputs from the models accumulated to a common grid prior to calculating the distributions shown? Otherwise this may be an apples to oranges comparison; a model with a higher resolution would tend to have a greater variability than a lower resolution model due to less spatial averaging of surface characteristics.

Line 711-712: this lack of dependence on the NO2/NOx emissions ratio should not be a great surprise given the fast chemistry between NO2 and NO.

Lines 781-784, lines 830-834: the SY component low precision is interesting – is there a seasonality that might be linked to downslope winds in mountainous areas? EU3 being surrounded by mountains – this made me wonder about tropopause fold events. These can sometimes have a big impact on ozone downwind, if a mechanism (such as convection or foehn wind circulation) exists to transfer the ozone further towards the surface from the middle troposphere – cf Makar, P.A., Gong, W., Mooney, C., Zhang, J., Davignon, D., Samaali, M., Moran, M.D., He, H., Tarasick, D.W., Sills, D., and Chen, J., Dynamic adjustment of climatological ozone boundary conditions for Air-Quality Forecasts, *Atmos. Chem. Phys*. 10 (6), 8997-9015. Do the different met models have a mechanism to parameterize troposphere/stratosphere exchange events? What was the upper boundary condition employed by the models for ozone (and other species)? Those with a higher top and a more detailed meteorology might capture fold events better than those with a lower top and/or less detailed meteorology.

Lines 805 – 808: my own work suggests that the bias error may be due to the absence of forest shading in most air-quality models (EGU presentation and ITM conference proceedings so far, paper under review) – this would also be consistent with the NO2 underprediction showing up in the EU results.

Text on Figure 21 is too small to read.

Section 3.3.4: This makes sense in terms of the chemistry, but the driving causes for those chemical changes are less clear. Temperature gradient or PBL height might be worth checking – is the bias due to too stable / low PBL in winter (too high in summer)?

Line 1081: probably should be “conclusions” rather than “considerations” in this sentence.