MS No.: acp-2016-15

Anonymous Referee #2

General comments.

The work presented herein, presents a new approach to model evaluation that attempts to shade light into the processes that influence model errors, rather than traditionally compare modeled ozone concentrations to in-situ measured values. The methodology is scientifically solid and sound and will help the AQ community move toward new ways of error diagnostics and thus, model improvement. The title of the manuscript reflects the contents of the paper and is considered sufficient. The main comments from the review process are related to obscure parts in the discussion of the figures and results. The specific comments that follow are meant to strengthen the communication of the results to readers that may not be as familiar with the history of the AQMEII initiative or the details of the spectral decomposition methodology. I am in favor of publishing this paper with Atmospheric Chemistry and Physics, after addressing the minor specific comments that follow.

Specific comments/suggestions

1. Section 2.1: In the beginning of the error decomposition section, please add references to the original published work (i.e. Wilmott, Murphy and others). The part that was uniquely developed for this work should also be clearly identified in this section.

Response. The content of section 2.1 actually reflects the origin of the methodology we propose, which is derived from many fields and never applied to air quality (or geophysical time series, to our knowledge) before. This is the first work that put together the Theil decomposition and the minimisation of the error for spectrally decomposed time series. We have moved the reference to Murphy (1988) at the beginning of the section, but rather keep the rest unchanged.

2. Page 5: in the minimization of MSE, the authors want to achieve independency of MSE from the model's statistical metrics, since the observed ones are not controllable. Can you please add a brief explanation in the text as to why you chose to differentiate over the mean model value and model standard deviation?

Response. Explanation added to the text

3. Page 6, spectral decomposition: In Rao et al. (1997) and Hogrefe et al. (2000) the ozone time series are log-transformed before the analysis to stabilize the variances. Did the authors use the log-transform in their KZ application? If not, please explain the rationale behind using the original ozone data.

Response. We used the original time series of ozone data. Prior to the analysis, tests have shown that the results of the MSE breakdown were independent from the log transform of the initial data. We have used an approach consistent with Galmarini et al. (2013), where the raw data were also used.

4. Page 6, lines 188-190: what is the meaning of the bias in the discussion of the decomposition components? Equation 10 is applied to modeled and observed values separately.

Response. The bias should intended as presented in section 2.1, as from from Johnson et al (2008): '*the closeness of agreement between the average value obtained from a large series of measurements and the true value',* where the keyword is 'average'. The bias is the off-set of the averaged model results from the averaged measured values. In this sense, the band-pass components ID, DU, SY have zero mean by definition, and are therefore unbiased.

5. Table 2: please denote in the title which table corresponds to which AQMEII phase.

Response. Done

6. Page 8, section 4.1, line 249: the phrase "spatially averaged over the two continental areas" must be rephrased to "spatially averaged over each continental area". I am assuming that the MSE is calculated for each spectral component and each station and then averaged over each continent (there is one value of MSE for each component and each station for the period of May-Sep). Please clarify in the text accordingly.

Response. Done, it has been clarified in the text

7. In Figure 1, the cross components are denoted by subscript cc in the name of the variable. I suggest using the same name convention in the appendix, where the description of the cross components is included. This will avoid confusion to readers that are not familiar with the prior literature.

Response. Done

8. Page 9, lines 281-283: Is this statement based on results from the current or previous published work? Please add a reference to this statement accordingly.

Response. The statement is not derived from previous studies but based on the experience of the current work.

9. Page 9: in the 1st paragraph of section 4.2 the bias is described as influenced by both internal and external model errors (which is true). In the 2nd paragraph (line 292), the authors suggest that the bias of LT shows the externally induced errors. Can you clarify this inconsistency?

Response. The inconsistency is driven by the word 'error' rather than 'bias'. It has been corrected now. All biases (internal and external) are driven by the LT component, thus it is correct to say that the bias of the external drivers is incorporated in the bias of the LT component.

10. The units in Figure 2 are ppb square (ppb2) or ppb? If the former is true, then the MSE breakdown must be bias2, variance and mMSE from equation 9? Please revise the label accordingly.

Response. We have added labels to figure 2 and modified the caption accordingly

11. Section 4.3, figures 3-6: even though I embrace the idea of including a lot of information in one plot, it has been very challenging to read and understand the figures. I don't understand where the under- or over-estimation is indicated. I suggest the inclusion of one example (maybe in the figure caption or in the text) that will describe the results from one specific station (highlighted with a square of circle). That way, it will be easier for the reader to connect the color coded scale with the different components. The plots provide valuable information which must be communicated in the most efficient manner.

Response. Thanks for the valuable suggestion. We have improved the readability of the figure and added, as an example, a scheme on how the figure has to be interpreted (see last panel of figure 3).

12. Page 12, figure 7: the components of figure 7 must be explained in more detail. What are the units? x and y axes? What is shown in the upper plot? The paragraph describing figure 7 and the method behind it needs further improvement to communicate a clear message.

Response. The figures 7 and 8 have been revised and improved. We have also slightly revised the contents of Section 5, which has already a detailed and independent introduction, with examples and review of the results. We acknowledge the topic might be not straightforward to understand and therefore have made extra effort in trying to simplify it.

Anonymous Referee #1

This is an interesting and well written paper that makes a meaningful contribution to model evaluation. A few comments and editorial suggestions are provided below.

1. Wavelet filters can provide better separation of components (i.e., reduced covariances among components).

Response. Eskridge et al. (1997) compared the kz filter against several other methods, including wavelet filters, showing that kz has the same level of accuracy and (often) higher level of confidence. The kz filter has also the advantage of 1. being insensitive to missing values, 2. being supported by extensive literature when applied to ozone, 3. depending on two parameters only, which are quite robust for ozone. It is true, however, that the main shortcoming of method we have developed is the overlapping between the cross components and especially the fact that the error of cross components can be quantified but cannot be apportioned according to the methodology outlined in the current work. Nonetheless, we have preferred to rely on this methodology and possibly exploring wavelet in the future.

2. The spatial support of the model (model grid average) is greater than that of the observations (point scale), and should therefore have a smaller variance, as should all the temporal components. The term σm will typically be less than rσo for this reason.

Response. We have included some comments in the text (see line 306 onwards)

3. The model/observation agreement in the DU component is driven largely by diurnal forcing (similarly, the LT component has a significant amount of annual energy). Model performance metrics for the DU component is misleadingly optimistic because it mostly reflects the 24 hour and annual forcings embedded in both the observations and model values. For periodic processes, metrics derived from the amplitude and phase can be more informative.

Response. We have added the comment to the text (see line 332). We reserve to expand to those metrics in future analysis.

4. The variance of the ID term is very small. Therefore, although the paper shows both the fraction of variance due to each component and the error terms, it should be pointed out that the small errors in the ID component are quite large relative to the total amount of ID variability.

Response. We have included some comments in the text (see line 344)

5. Model/observation correlation as a stand-alone metric can be informative as it shows whether the model can reproduce patterns seen in the observations. For example, the ID component, as noted, has small errors, but for individual monitoring sites (not spatially averaged), correlation between modeled and observed ID is often quite low and insignificant (there often appears to be no relationship between the two). On the other hand, correlation tends to improve as time and space scales increase, often leaving the LT component with the best agreement in terms of correlation.

Response. We have included the values of the correlation coefficient directly into the error breakdown plots, therefore allowing for a compact view of the error magnitude and associativity value.

Editorial comments

There is some confusion in the text when discussing bias. Figure 2 actually shows squared bias, though the discussion seems to be referring to both bias and squared bias.

Response. We have clarified the figure 2 by adding 'bias^2' in the legend and clarified the discussion in the main text

Line 263: should read "has little impact" or "has negligible impact"

Response. Done

Line 283: The statement ending on this line could use a reference.

Response. It is derived based on the analysis of the current study.

Line 305: What is meant by "sparseness of the modeled values"?

Response. The sentence has been removed from the text.

Line 452: should have a period at the end

Line 457: should have a period at the end

Response. Done

Figure 1. Panels do not have 'a)' and 'b)' labels.

Response. Done

Also, if it's not too much trouble, invert the legends so that the colors appear in the same order as they do in the bars.

Response. Done

Figures 3 and 4 are very difficult to look at. When error terms are small it is hard to tell where the intersection is. Zooming in would help with better image resolution.

Response. We have improved the resolution of the figures and added an zoomed example for clarification. Rather than the individual station's error, we wish to convey the message contained in the method.

Figure 8: Caption should read 'from right to left'.

Response. Done

Editor's comments left open from to quick report

Figure 3 – 6: please use larger font to show the title of each panel ("MSE of spectral components …"). It may be sufficient to simply show the model name, AQMEII phase, and continent as title in each panel.
Response. Done

Figure 4, 6: please make sure that the color scale on the right does not overlap the geographical features in the Northeast corner of the map.
Response. Done

Figure 8: Please add a title and units to the y-axis
Response. Done

Figures S4 – S7: Please add a title and units to the y-axis
Response. We have removed the figures for the supplementary material, as they did not add much to the discussion with respect to the ones already presented in the paper

Is Table 2 accidentally split in two sections? (pages 20 and 21)

Response. The table is split in two parts, each one describing the models participating to AQMEII 1 and AQMEII2, respectively. We have specified it in the tables.

ERROR APPORTIONMENT FOR ATMOSPHERIC CHEMISTRY-TRANSPORT MODELS. A NEW APPROACH TO MODEL EVALUATION

E. Solazzo, S. Galmarini

European Commission, Joint Research Centre, Institute for Environment and Sustainability, Air and Climate Unit, Ispra, Italy

Author for correspondence: S. Galmarini, stefano.galmarini@jrc.ec.europa.eu,
Phone: +390332785382

Abstract. In this study, methods are proposed to diagnose the causes of errors in air quality (AQ) modelling systems. We investigate the deviation between modelled and observed time series of surface ozone through a revised formulation for breaking down the mean square error (MSE) into bias, variance, and the minimum achievable MSE (*mMSE*). The bias measures the accuracy and implies the existence of systematic errors and poor representation of data complexity, the variance measures the precision and provides an estimate of the variability of the modelling results in relation to the observed data, and the *mMSE* reflects unsystematic errors and provides a measure of the associativity between the modelled and the observed fields through the correlation coefficient. Each of the error components is analysed independently and apportioned to resolved process based on the corresponding timescale (long scale, synoptic, diurnal, and intra-day) and as a function of model complexity.

The apportionment of the error is applied to the AQMEII (Air Quality Model Evaluation International Initiative) group of models, which embrace the majority of regional AQ modelling systems currently used in Europe and North America.

The proposed technique has proven to be a compact estimator of the operational metrics commonly used for model evaluation (bias, variance, and correlation coefficient), and has the further benefit of apportioning the error to the originating timescale, thus allowing for a clearer diagnosis of the process that caused the error.

Keywords: Model evaluation; Time series analysis; Bias-variance decomposition; AQMEII

1. INTRODUCTION

Due to their use for regulatory applications and to support legislation, air quality (AQ) models must model correctly and be correctly applied, justifying the need for a thorough evaluation. A framework for the operational and scientific evaluation of geophysical models was already envisaged in the early ‘80s (Fox, 1981; Wilmott et al., 1985), the former being ‘*a comparison with data exclusively within a particular application context*’, and the latter defined as ‘*some understanding of cause-and-effect relationship that relies on testing model components and extensively detailed data collection*’ (Fox, 1981). Thirty years later, as AQ models became more and more complex and their range of applicability widened, Dennis et al. (2010) further elaborated the concept of model evaluation by proposing a four-level evaluation, according to which different complementary aspects of the models should be tested, namely:

Efisio Solazzo 26/4/2016 14:38
Deleted: Denis

a. Operational: the level of agreement of model results with observations;

b. Dynamic: ability of the modelling system to respond to changes (in emissions, or in meteorological events);

c. Diagnostic: identify and attribute the source of the error to the relevant process;

d. Probabilistic: confidence and uncertainty levels of the modelled results.

In the framework originally designed by Dennis et al. (2010), the diagnostic component plays a central role. It *i)* answers the fundamental issue left open by the operational screening, in other words whether the model provides the right answer for the right reason, *ii)* provides feedback to developers to help make model improvements, and *iii)* sets the basis for the probabilistic evaluation (Figure 1 of Dennis et al., 2010).

Over the years, and despite the increasing relevance of modelling systems for AQ applications, model evaluation continues to rely almost exclusively on operational evaluation, which basically involves gauging the model's performance using distance, variability, and associativity metrics. This common practice has little or no impact on model improvement, as it does not target the source of the modelling error and does not discriminate between the reasons for appropriate or inappropriate performance.

Such a requirement is even more pressing these days, with current state-of-the-science AQ modelling systems accounting for an increasing number of coupled physical processes and being described using hundreds of modules, which are the result of decades of targeted and, generally, independent investigations. Furthermore, AQ modelling systems typically depend on external sources for the inputs of meteorology and emissions data, as well as for boundary conditions. These fields are generally produced by other models (which, in turn, depend on external sources for initial and/or boundary conditions) and, after substantial processing, are used by the AQ modelling systems with no guarantee of being unbiased and/or accurate. The bias introduced by these inputs, along with the uncertainty associated with model error, the linearisation of non-linear processes, and omitted and unresolved variables and processes, all contribute to the model error. The extensive use of AQ models for AQ assessment and planning is equally important, and requires a good knowledge of the model capabilities and deficiencies that would allow for a more educated use of the modelling systems and their results.

Recently, the AQMEII (Air Quality Model Evaluation International Initiative) activity (Rao et al., 2011) applied the approach proposed by Dennis et al. (2010), by organising model evaluation activities (AQMEII 1, 2 and 3) using operational (Solazzo et al., 2012a,b; Solazzo et al., 2013a; Im et al., 2015a,b), probabilistic (Solazzo et al., 2013b; Kioutsioukis et al., 2014), and diagnostic (Hogrefe et al., 2014; Makar et al., 2015) evaluation frameworks.

The study we present here follows and complements the previous investigations based on the AQMEII models collected in the first and second phases of the activity (AQMEII1 in 2006 and AQMEII2 in 2010). The main aim is to introduce a novel method that combines operational and diagnostic evaluations. This method helps apportion the model error to its components, thereby identifying the space/timescale at which it is most relevant and, when possible, to infer which process/es could have generated it. This work is designed to support the analysis of the currently ongoing third phase of the AQMEII activity (Galmarini et al., 2015).

2. Mean square error as a comprehensive metric

For the model evaluation strategy proposed, we start by breaking down the Mean Square Error (MSE) (used here as unique metric to evaluate model performance) into the sum of the variance (and covariance) and the squared bias. The error and its components are then calculated on the spectrally decomposed time series of modelled and observed hourly ozone mixing ratios. The advantage of this evaluation strategy is twofold:

- With respect to a conventional operational evaluation, the new method allows for a more detailed assessment of the distance between model results and observations given the breakdown of the error into bias, variance and covariance and their associated interpretations.
- Decomposing the MSE into spectral signals allows for the precise identification of where each portion of the model error predominantly occurs. Given that specific processes are associated with specific scales, the apportionment of the error components to their relevant scales helps to more precisely identify which processes described in the model could be responsible for the error. Information about the nature of the error and the class of process can significantly help modellers and developers to improve model performance.

The data used are produced by the modelling communities participating in AQMEII1 and AQMEII2 over the European (EU) and North American (NA) continental scale domains for the years 2006 (AQMEII1) and 2010 (AQMEII2).

2. 1. Error decomposition

The MSE is the squared difference of the modelled (*mod*) and observed (*obs*) values:

$$MSE = E(mod - obs)^2 = \frac{\sum_{i=1}^{nt}(mod_i - obs_i)^2}{n_t} \qquad \textbf{EQ 1}$$

where E(·) denotes expectation and n_t is the length of the time series. The bias is:

$$bias = E(mod - obs) \tag{EQ 2}$$

i.e. $bias = \overline{mod} - \overline{obs}$. Thus, the following relationship holds:

Efisio Solazzo 26/4/2016 14:38
Deleted: *s*

$$MSE = var(mod - obs) + bias^2 \tag{EQ 3}$$

which is a well-known property of the MSE, (var(·) is the variance operator). By using the property of the variance for correlated fields:

$$var(mod - obs) = var(mod) + var(obs) - 2cov(mod, obs) \tag{EQ 4}$$

the final formulation for the MSE components reads:

$$MSE = bias^2 + var(mod) + var(obs) - 2cov(mod, obs), \tag{EQ 5}$$

where the covariance term (last term on the right-hand side of Eq 5) accounts for the degree of correlation between the modelled and observed time series. When the covariance term is zero, *var(obs)* is referred to as the *incompressible part of the error* and represents the lowest limit that the MSE of the model can achieve. When dealing with model evaluation, the modelled and observed time series are typically highly correlated and therefore, within the limits of the perfect match (correlation coefficient of unity), *cov(mod, obs)* = *cov(obs,obs)* = *cov(mod,mod)* = *var(mod)* = *var(obs)* and the MSE can be reduced to only the bias term. That implies that the development of a high-quality model needs to ensure:

a. the highest possible precision in order to maximise the *cov(mod, obs)* term, and

b. the highest possible accuracy, in order to minimise the bias.

Elaborating on Eq 5, Theil (1961) derived the following:

$$MSE = (\overline{mod} - \overline{obs})^2 + (\sigma_{mod} - \sigma_{obs})^2 + 2(1 - r)\Box_{mod}\Box_{obs} \tag{EQ 6}$$

In Eq 6, the variance term is expressed as the difference between the standard deviation of the model and that of the observations, and the covariance term (last term on the right) includes *r*, the coefficient of correlation between the observed and modelled time series. The ratios of the three terms on the right-hand side of Eq 6 to the overall MSE are known as *Theil's coefficients* (Pindick and Rubinfeld, 1998). Murphy (1988) provided examples of the scores that can be developed using the components of the MSE.

Efisio Solazzo 26/4/2016 14:38
Moved (insertion) [1]

The bias measures the departure of the modelled from the observed results, and is a measure of systematic error, since it measures the extent to which the average modelled values deviate from the observed ones. The bias is commonly used to express the degree of 'trueness', i.e. "the closeness of agreement between the average value obtained from a large series of measurements and the true value" (Johnson, 2008). The variance shows whether the modelled variability is compatible with that observed. Finally, the covariance term represents the unexplained proportion of the MSE due to the remaining unsystematic errors, i.e. it represents the remaining error after deviations from the mean values have been accounted for. This latter term is a measure of the lack of correlation of the model with comparable observations, and is considered the least 'worrisome' portion of the error (Pindick and Rubinfeld, 1998).

Aiming at minimising the MSE, the only controlled variables in Eq 6 are $\overline{mod}$ and σ_{mod}, and differentiating with respect to them yields the conditions that minimise the MSE::

Efisio Solazzo 26/4/2016 14:38
Deleted: Elaborating on

Efisio Solazzo 26/4/2016 14:38
Deleted: ,

Efisio Solazzo 26/4/2016 14:38
Deleted: are:

$$\begin{cases} \dfrac{\partial MSE}{\partial \overline{mod}} = 2\left(\overline{mod} - \overline{obs}\right) = 0 \\ \dfrac{\partial MSE}{\partial \sigma_{mod}} = 2(\sigma_m - \sigma_{obs}) + 2(1-r)\sigma_{obs} = 0 \end{cases}$$

i.e. the best agreement between modelled and observed values is achieved by:

$$\begin{cases} \overline{mod} = \overline{obs} \\ \sigma_m = r\sigma_{obs} \end{cases} \qquad \text{EQ 7}$$

which analytically corresponds to the aforementioned items *a* and *b*. By inserting Eq 7 into Eq 6, the minimum achievable MSE (*mMSE*) is

$$mMSE = \sigma_{obs}^2(1 - r^2) \qquad \text{EQ 8}$$

which is the unexplained portion of the error, as it reflects the share of observed variance that is not explained by the model (r^2 is the coefficient of determination). The presence of an unexplained part of the error suggests a modification of the MSE decomposition in Eq 6 in such a way as to explicitly include *mMSE*:

$$MSE = \left(\overline{mod} - \overline{obs}\right)^2 + (\sigma_{mod} - r\sigma_{obs})^2 + mMSE \qquad \text{EQ 9}$$

The decompositions in Eq 5, Eq 6, and Eq 9 contain all the relevant operational metrics usually applied to score modelling systems (bias, variance, correlation coefficient), and therefore prove to be a compact estimator of accuracy (bias), precision (variance) and associativity (unexplained portion through the correlation coefficient). Eq 9 has been explicitly derived in this study to help evaluate AQ models.

Efisio Solazzo 26/4/2016 14:38
Deleted:

Efisio Solazzo 26/4/2016 14:38
Moved up [1]: Murphy (1988) provided examples of the scores that can be developed using the components of the MSE.

Ideally, the entire error should be attributable to unsystematic fluctuations. From a model development perspective, the variance and covariance are possibly more revealing of model deficiencies than is the bias term, as they are produced by the AQ model itself, while the bias is also due to external sources (e.g. emissions, boundary conditions). From the application viewpoint, however, it is the overall error that counts, which is mostly made up of the bias.

2.2. Spectral decomposition of modelled and observed time series

Hourly time series of (modelled and observed) ozone concentrations have been decomposed using an iterative moving average approach known as the Kolmogorov-Zurbenko (kz) low-pass filter (Zurbenko, 1986), whose applications to ozone are vastly documented in the literature (Rao et al., 1997; Wise and Comrie, 2005; Hogrefe et al., 2000 and 2014; Galmarini et al., 2013; Kang et al., 2013; Solazzo and Galmarini, 2015). The kz filter depends on two parameters: the length of the moving average window *m* and the number of iterations *k* ($kz_{m,k}$). Since the kz is a low-pass filter, the filtered time series consists of the low-frequency fluctuating component, while the difference between two filtered time series provides a band-pass filter. This latter property is used to decompose the ozone concentration time series as:

$$O_3 = LT(O_3) + SY(O_3) + DU(O_3) + ID(O_3) \qquad \textbf{EQ 10}$$

where LT is the long-term component (periods longer than 21 days); SY is the synoptic component (weather processes that last between 2.5 and 21 days); DU is the diurnal component (day/night alternation period between 0.5 and 2.5 days); and ID is the intra-day component accounting for fast-acting processes (less than 12 hours). The decomposition presented in Eq 10 is such that the original time series is perfectly returned by the summation of the components (see Appendix for details). Dealing with one year of data, any filter longer than the LT component would not be meaningful. The periods of the components correspond to well-defined peaks in the power spectrum of ozone, e.g. as detailed in Rao et al. (1997) and Hogrefe et al. (2000).

The LT component is the baseline and incorporates the bias of the original (undecomposed) time series. The other components (SY, DU, and ID) are zero-mean fluctuations around the LT time series and are therefore unbiased. The band-pass nature of the SY, DU, and ID components is such that they only account for the processes occurring in the time window the filter allows the signal to ‘pass’. For instance, the DU component is insensitive to processes outside the range of 0.5 to 2.5 days.

Further properties of the spectrally decomposed ozone time series of AQMEII derived by Galmarini et al. (2013), Hogrefe et al. (2014), and Solazzo and Galmarini (2015) are as follows:

- The DU component accounts for more than half of the total variance, followed by the LT and SY components;
- The ID component has the smallest influence due to the small amplitude of its fluctuations;

- The variance of the spectral component is neither strongly nor systematically associated with the area-type of the monitoring stations (i.e. rural, urban, suburban);
- Due to the bias, most of the error is accounted for by the LT component, followed by the DU component. The ID contributes very little to the overall MSE.

Further important technicalities of the spectral decomposition, including a method to estimate the contribution of the spectral cross-components (the overlapping regions of the power spectrum) to the total error, are reported in the Appendix.

The signal decomposition of Eq 10 is applied to the full-year time series. However, to evaluate the model performance with regard to ozone, the analysis is restricted to the months of May to September, i.e. when the production of ozone due to photochemistry is most relevant.

3. Data and Models used

The observational dataset derived from the surface AQ monitoring networks operating in the EU and NA constitutes the same dataset used in the first and second phases of AQMEII to support model evaluation. Only stations with over 75% valid records for the whole periods and located at altitudes below 1000 m have been used for this analysis. Details of the modelled regions and number of receptor stations are reported in Table 1.

Since the main scope of this study is to introduce the error apportionment methodology (rather than to strictly evaluate the models), the analysis is presented for continental areas for convenience and easier display of the results. However, given the size of the domains and the heterogeneity of climatic and emission conditions, dedicated analyses for three sub-regions in both continents are proposed in the Supplementary material (Figure S1 to Figure S3).

Efisio Solazzo 26/4/2016 14:38
Deleted: extension

There are profound differences between the modelling systems that participated in AQMEII1 and AQMEII2. The two sets of models have been applied to different years (2006 for phase 1 and 2010 for phase 2) and are therefore dissimilar with respect to the input data of emissions and boundary conditions for chemistry. The AQ models of the second phase are coupled (online chemistry feedbacks on meteorology), while those of the first phase are not. The effect of using online models for simulating ozone accounts for the impact of aerosols on radiation and therefore on temperature and photolysis rates (Baklanov et al., 2014).

The model settings and input data for phase I are described in Solazzo et al. (2012a, b; 2013a), Schere et al. (2012), and Pouliot et al. (2012); for phase II, similar information is presented in Im et al. (2015a, b), Brunner et al. (2015), and Pouliot et al. (2015).

Table 2 summarises the features of the modelling systems analysed in this study with regard to ozone concentrations in the EU or NA. The modelling contribution to the two phases of AQMEII consists of 12 and 9 models and of 8 and 3 models for EU and NA, respectively.

Efisio Solazzo 26/4/2016 14:38
Deleted: ... [1]

Efisio Solazzo 26/4/2016 14:38
Deleted: results

Detailed analysis of the main differences in emissions, boundary conditions, and meteorology between the modelled years of 2006 (AQMEII1) and 2010 (AQMEII2) is presented in Stoeckenius et al. (2015). A summary of the performance of the two suites of model runs is provided in Makar et al. (2015), showing that the AQMEII1 models generally performed better than the AQMEII2 models, based on standard operational metrics. However, the use of standard evaluation methods does not allow for the assessment of whether the feedback processes have an effect on the deterioration of model performance, or rather the different sets of emissions and boundary conditions. We try to assess the problem using the error apportionment methods outlined above.

4. Results for the Spatially Averaged Time series

4.1 MSE of Spectral Components

Figure 1 reports the MSE share of the spectral components and cross components for each model, for both phases of AQMEII, derived from the ozone time series spatially averaged over each continental area.

Efisio Solazzo 26/4/2016 14:38
Deleted: the two

Efisio Solazzo 26/4/2016 14:38
Deleted: areas

The LT share of the total MSE is the largest in absolute value for both continents and both simulated years. The LT share ranges between 9.9% (GEM-AQ, AQMEII1, NA) and 86.7% (WRF/Chem, AQMEII1, NA), and averages at ~34% and ~46.5% for the EU and ~50.6% and ~47% for NA (AQMEII1 and AQMEII2, respectively).

The second largest share of the total MSE is of the DU component, accounting for ~20% (all cases), followed by the SY component. Depending on the model, the MSE share of the remaining spectral components and cross-components varies significantly. Being the intermediate time scales, the overlap of the DU and SY components is likely to be more significant than the overlap of the LT and ID scales. The contribution of DU_{cc} and SY_{cc} to the total error can be as high as 17% (DU_{cc} for GEM-AQ, AQMEII1, NA) and 16% (SY_{cc} for MM5-CAMx, AQMEII1, EU). Overall, the DU_{cc} terms (interaction of DU with the neighbouring SY and ID scales) are significant in both continents (~10%), while the share of the SY component and cross-components is more significant in the EU.

The ID component has a little impact on the total MSE (negligible in some instances), exceeding the 3% share only for the two EU instances of the L.-Euros model

Efisio Solazzo 26/4/2016 14:38
Deleted: or negligible

Efisio Solazzo 26/4/2016 14:38
Deleted: .

The results of Figure 1 help identify the time-scales and associated processes for which the largest improvement in model accuracy can be achieved. The LT component has the largest share of the error due to the bias (error breakdown is discussed in the next section), but 'internal' chemical processes, transport, and deposition also occur at this timescale. Diurnal processes are the second largest source of error, including, among others, chemistry, boundary layer dynamics, radiation forcing, and their interactions. The processes in the SY band bridge meteorological and chemical processes, and discern between the fast-acting diurnal processes and the baseline. As such, although the SY signal is not as strong as that of the DU components (variance of SY is comparable to the variance of ID, see Hogrefe et al., 2014), it accounts for a significant portion of the total error, as discussed next.

4.2 The quality of the error: Error apportionment

The error breakdown (Eq 9) of each spectral component complements the analysis presented in the previous section, and is reported in **Figure 2** (please note that results in Figure 2 are reported in ppb^2 for reason of clarity). The bias (only included in the LT component) is the average amount by which the modelled time series is displaced with respect to the observed time series, and is the main source of error. The bias can be either due to 'internal' model errors, or inherited from external drivers (emissions, meteorology, boundary conditions). Based on the experience matured within AQMEII, while the internal model errors are of interest for model development because they are generated by systematic modelling errors, the bias introduced by external drivers is responsible for the largest share of modelling errors.

Efisio Solazzo 26/4/2016 14:38
Deleted: .

Efisio Solazzo 26/4/2016 14:38
Deleted: While the former

From the continental average error breakdown of **Figure 2** we can conclude that the majority of EU models (in both AQMEII phases) have small bias (continental-wide average), with the important exceptions of CCLM-CMAQ and Muscat models in AQMEII1, and CMAQ in AQMEII2, which introduced large positive biases. The bias for the NA continent is more uniformly distributed across the models (model over-prediction in both AQMEII phases), possibly indicating a common source of (external) bias in the NA models. The bias introduced by external fields is reflected by the bias of the baseline component (LT). For the period between May and September, the error in modelled ozone due to the boundary condition is typically small (Solazzo et al., 2012; Im et al., 2015; Giordano et al., 2015; Hogrefe et al., 2014), while the emissions of ozone precursors and VOCs are problematic, especially in the EU (Makar et al., 2015; Brunner et al., 2015). We further notice that the absence of bias in some models may be caused by the presence of compensating bias, i.e. spatially distributed biases of opposite signs. The spatial distribution of the MSE is discussed in the next section. In all cases, the MSE_{best} model is, by definition, the model with lowest MSE and thus the one with the smallest LT bias.

Efisio Solazzo 26/4/2016 14:38
Deleted: are poorly biased

Efisio Solazzo 26/4/2016 14:38
Deleted: error

Efisio Solazzo 26/4/2016 14:38
Deleted: should not

Efisio Solazzo 26/4/2016 14:38
Deleted: confused with

The variance share of LT error is generally small (~1 - 2.5 ppb). This is not entirely unexpected, as the LT component has a high signal-to-noise ratio with a well-structured seasonal cycle, peaking in summer. While such a cycle is typically well reproduced by the models, its phase and/or the amplitude are not always well captured (Solazzo et al., 2012; Im et al., 2015), leading to the variance error. The variance error also originates from the different spatial support (incommensurability) of point measurements vs. gridded model outputs. The latter have typically larger spatial support, while receptors are more likely to detect local scale effects that enhance the observed variance.

Efisio Solazzo 26/4/2016 14:38
Deleted: In detail, the

The *mMSE* error of the LT component outweighs the variance error in most cases (in both the EU and NA), and is due to the unexplained portion of observed variance. The processes

Efisio Solazzo 26/4/2016 14:38
Deleted: , thus to the sparseness of the modelled values.

responsible for the *mMSE* error of the LT component (such as deposition, transport, stratospheric mixing and photochemistry) act at timescales of more than 21 days.

The DU error (on average 3-4 ppb for AQMEII1 and 2-3 ppb for AQMEII2) makes up the second highest contribution to the total error. The portioning between variance and the *mMSE* error varies greatly from model to model. However, a comparison of the two AQMEII phases shows that the *mMSE* is predominant for AQMEII2, while the variance error (typically due to model under-prediction of the observed variability) is most relevant in several cases of AQMEII1. Therefore, at the DU scale, the 'quality' of the error of the AQMEII2 phase is higher than that of its AQMEII1 counterpart. One possible explanation is the fact that coupled models were used in AQMEII2, while AQMEII1 exclusively used non-coupled models. As already mentioned (end of section 3), Makar et al. (2015) found that AQMEII1 models performed better overall with respect to AQMEII2. An analysis of the LT component showed that the bias in the AQMEII2 models is higher, possibly due to the 2010 emission inventory, while an analysis of the DU error found that the variance error in the AQMEII2 models is significantly reduced with respect to the AQMEII1 models, and is almost null. We postulate that the inclusion of feedback effects may have been beneficial, and that the reduced performance of AQMEII2 models is likely due to external bias. The residual *mMSE* error of the DU component (~1-2 ppb on average for both continents) is mostly likely generated by a number of processes, including chemistry, cloudiness, boundary layer transition and vertical mixing. From Figure 2, the values of the correlation coefficient for the DU component are very high (exceeding 0.8 in the majority of the cases). Such a high performance can be misleadingly optimistic though, because it mostly reflects the 24-hour and annual forcing embedded in both the observations and model values. Further analysis on the amplitude and phase of the error can reveal more informative.

The SY error (almost entirely due to *mMSE* in AQMEII2) is comparable across all models applied to the same continental domain (except for GEM-AQ and WRF/Chem, NA), indicating that a possible common source of error may be due to missing processes in the models related to the interaction between chemistry and transport.

Efisio Solazzo 26/4/2016 14:38
Deleted: GEMAQ in

Finally, the error of the ID component is less than 1 ppb (on average ~0.2 ppb for AQMEII2) and is generated by both variance (most commonly model over-prediction) and *mMSE*. The fast-acting photochemical processes are, therefore, modelled with satisfactory precision, , although the small errors in the ID component can be quite large relative to the total amount of ID variability.

Efisio Solazzo 26/4/2016 14:38
Deleted: .

4.3. Spatial distribution of the spectral error components

Maps of MSE by spectral components are reported in Figure 3 to Figure 6. As anticipated by the error analysis, the LT is the most problematic source of error for both continents, although the variety in the models' behaviour does not allow for generalisation.

Some of the cases presented in Figure 2, where the bias was null (MM5-CAMx, MM5-DEHM for AQMEII1 and CosmoArt for AQMEII2, both in EU), show bias compensation, typically due to model underestimation in the central part of the EU (Germany, eastern France) and model overestimation in the rest of the continent. The case of the CosmoArt model (Figure 5c) clearly shows the effect of the spatial averaging in masking the error that is only cancelled when a continental average is calculated. The model is in fact affected by severe bias and component errors.

Efisio Solazzo 26/4/2016 14:38
Deleted: most

The Po valley in Italy and the southern part of the EU are the most problematic areas, affected by severe LT errors (Figure 3 and Figure 5). The central and northern parts of the EU are less problematic, especially for AQMEII2. The other components of the error are significantly smaller than the LT error, with some exceptions (especially for the DU component). The length of the segment is in fact normalised to the largest error for each model, to facilitate the interpretation and the relative weight of each error component.

Concerning NA (Figure 4 and Figure 6), the DU error has more weight and competes with the LT error in the central and south-eastern parts of the continent. For AQMEII2, the SY error is as significant as the LT error on the East Coast (Wrf/Chem, Figure 6c). The greatest LT error is observed in the coastal areas (east and west) and across the north-eastern border between the US and Canada (due primarily to model underestimation in the east and north, and model overestimation in the west).

The analysis presented provides a detailed breakdown of the error in terms of error components, spectral decomposition and spatial distribution, thereby avoiding the pitfalls of extreme averaging and providing a comprehensive analysis of where the error occurs and the associated timescales and processes, and whether the error is internally generated or stems from the model's input data.

Efisio Solazzo 26/4/2016 14:38
Deleted: analyses

5. MSE DECOMPOSITION AND COMPLEXITY

In regression analysis and statistical learning theories, the problem of under- and over-fitting complex systems is at the root of the MSE decomposition into bias and variance. The trade-off between bias and variance is strictly dependent on the complexity of the model. Over-fitting occurs when too many parameters and modules are added to the model: each new module added to describe a process is a new source of variance due to internal parameterisation and linearisation. In other words, over-fitting is associated with the stochasticity inherent to the data/model, and contributes to the increase in variance and consequent decrease in bias. Under-fitting occurs due to an oversimplification of the modelled processes, and is an important source of bias as it is associated with the deterministic property of the modelling activity (Hastie et al., 2009).

The problem of the bias-variance trade-off becomes markedly more complicated when dealing with complex models with many degrees of freedom, such as AQ modelling systems. Adding new modules to cope with unexplained physical processes can lead to a reduction in the bias due to that specific process, but also feeds new variance and possibly new bias into the model due to the non-linear interaction of the new module with existing ones, since reducing the bias while preserving the variance is non-trivial.

Rao (2005), in the context of dispersion modelling, provided the theoretical variations of the total model uncertainty by exploiting the components of the difference between the modelled and observed variance (Figure 1 of Rao et al., 2005). Rao (2005) used the number of meteorological parameters in the model as a measure of model complexity, and concluded that the optimal model complexity could not be defined a priori, but is a trial-and-error combination of the model, the measurement error and the stochastic uncertainty.

In this study we attempt to derive the curves of the MSE components (bias, variance and covariance) as a function of model complexity, providing a first-time attempt to analysis the error of a regional AQ model as function of its complexity. The aim is to find the time scale dominated by the error (and hat type of error) and, if exists, the time window where the error decreases. The information obtained is of immediate usefulness for model development, as provides a clear temporal cut-off that discriminates the dynamics of the error.

Efisio Solazzo 26/4/2016 14:38
Deleted: .

Figure 7 shows an example of the approach used to break down model complexity, which basically relies on the resolved timescale of the model. The complexity of the model is assumed to increase when the resolved timescale is shortened: the shorter the timescale, the more complex the model. The timescale of the resolved processes is thus used as a measure of the complexity, and is obtained by recursively applying the *kz* filter to the ozone time series. The minimum complexity is assumed to be represented by a model that cannot resolve any temporal scale below ~1 month (far right of Figure 7), while the maximum complexity corresponds to the hourly time series, i.e. the standard model's output (far left of Figure 7).

Efisio Solazzo 26/4/2016 14:38
Deleted: (

Efisio Solazzo 26/4/2016 14:38
Deleted:).

In Figure 8, we report the spatially averaged curves of bias, variance, and covariance according to Eq 6 as a function of model complexity. According to the regression analysis theories outlined above, we would expect the variance to increase according to the complexity ($\frac{d\sigma_m^2}{dcomplexity} > 0$), and the distance between the modelled and observed variance to decrease $\left(\frac{d(\sigma_{m-}\sigma_o)^2}{dcomplexity} < 0\right)$, and the opposite for the bias. The curves of variance in Figure 8 indeed turn downwards as predicted by the theory, while the curves of bias have a mixed behaviour but are, basically, constant $\left(\frac{d(\overline{mod}-\overline{obs})^2}{dcomplexity} \approx 0\right)$.

More specifically:

- The $(\sigma_{m-}\sigma_o)^2$ term decreases steadily but slowly to a timescale of ~1 day, after which it drastically drops to significantly lower values. This indicates that *i)* the complexity of the AQ systems increases exponentially at the DU timescales (not entirely surprising, given the day/night behavioural properties of ozone); *ii)* the efforts made to improve the model capabilities on the short-term processes governing the ozone dynamics improve the model precision; *iii)* there is a possible lack of parameterisation and modelling of the processes of transport and chemical transformation over periods longer than 1-2 days.

- The fact that the bias varies only by small amounts indicates that a fully evolved model, capable of reproducing processes at the shortest timescales (turbulent dispersion, fast chemical reactions, even day/night variability, etc.) is no more accurate than a basic model that only accounts for long-term processes. This might indicate that *i)* the bias at the shorter timescales is introduced entirely by the larger timescales, and/or *ii)* the bias is continuously fed into the model by an external source acting at all scales, as for example the emissions data or boundary conditions.

Summarising, in most cases (both continents, both AQMEII phases), the $(\sigma_{m-}\sigma_o)^2$ term decreases sharply after a timescale of resolved processes of ~1 day; the bias term is surprisingly independent on complexity; the covariance is complementary to the variance. Thus, the bias seems the error term more urgently needing attention and current studies are carried out to diagnose more precisely its origin within AQ modelling systems.

Efisio Solazzo 26/4/2016 14:38
Deleted: In Figure S4 to Figure S7 we propose the same analysis as that in Figure 8 but replicated for all receptors individually (with no spatial average). In

Efisio Solazzo 26/4/2016 14:38
Deleted: confirms the independency to

Efisio Solazzo 26/4/2016 14:38
Deleted: at all receptors

5. Conclusions

This study presents a novel approach to model evaluation, and aims to combine standard operational statistics with the time allocation of the component error. The methodology we propose tackles the issue of diagnostic evaluation from the angle of the spectral decomposition and error breakdown of model/data signals, introducing a compact operator for the quantification of bias, variance, and the correlation coefficient.

When the analytical decomposition of the error into bias, variance and *mMSE* is applied to the decomposition of the signals into long-term, synoptic, inter-diurnal and diurnal components, information can be gathered that helps reduce the spectrum of possible sources of errors and pinpoint the processes that are most active at a particular scale which need to be improved. The procedure is denoted here as *error apportionment* and provides an improved and more powerful capacity to identify the nature of the error and associate it with a specific part of the spectrum of the model/measurement signal. The AQMEII set of models and measurements have been used in the evaluation procedure.

After analysing the ozone concentrations gathered in the two phases of AQMEII, which cover a number of modelling systems in two different years and geographical areas, we conclude that:

- The bias component of the error is by far the most important source of error, and is mainly associated with long-term processes and/or input fields (likely emissions data or boundary conditions). With regard to the model application, any effort to improve the current capabilities of AQ modelling systems are likely to have little practical impact if this primary issue is not addressed and solved;

- Most relevant to model development, the variance error (the discrepancy between modelled and observed variance) is mainly associated with the DU component. At timescale of ~1-2 days, the complexity of modelling systems increases substantially and many processes are involved; the fact that the variance error of the DU component for the AQMEII2 runs is reduced with respect to the AQMEII1 runs might indicate the benefits of including feedback in the models. Such a conclusion could not be drawn with simpler operational evaluation strategies;
- The limited magnitude of the variability of the SY and LT signals produces little variance errors for these two components, and only becomes comparable to the LT or DU error when the bias is negligible or the total MSE is small;
- The *mMSE* error is predominant in some instances of the analysed models, and is due to the random distribution of modelled values. There are many causes of *mMSE* error, including all 'internal' processes that produce non-systematic errors such as noise, representativeness, the linearisation of non-linear process, and turbulence closure;
- The analysis of the spatial distribution of the error highlights the diversity in the behaviour of each modelling system. The common spatial structures of the LT error (for example in the central and southern EU) may reveal common sources of error (e.g. emissions data), while the error of the other components (especially DU and SY) are peculiar to each model and need to be assessed individually.

Analyses of the modelling results for the third phase of AQMEII are currently building on the methodology outlined in this study, with specific attention being given to the diagnostic of the error of the LT component in relation to external forcing (emissions and boundary conditions) and of the DU component with respect to the variance error.

APPENDIX

As in Hogrefe et al. (2000) and Galmarini et al. (2013), the time windows (*m*) and the smoothing parameter (*k*) have been selected as follows:

$$\begin{gathered} \mathrm{ID}(t) = \mathbf{x}(t) - \mathrm{kz}_{3,3}(\mathbf{x}(t)) \\ \mathrm{DU}(t) = \mathrm{kz}_{3,3}(\mathbf{x}(t)) - \mathrm{kz}_{13,5}(\mathbf{x}(t)) \\ \mathrm{SY}(t) = \mathrm{kz}_{13,5}(\mathbf{x}(t)) - \mathrm{kz}_{103,5}(\mathbf{x}(t)) \\ \mathrm{LT}(t) = \mathrm{kz}_{103,5}(\mathbf{x}(t)) \\ \mathbf{x}(t)=\mathrm{ID}(t)+\mathrm{DU}(t)+\mathrm{SY}(t)+\mathrm{LT}(t) \end{gathered} \qquad \text{EQ. S.1}$$

where **x**(t) is the time series vector.

A clear-cut separation of the components of EQ. S.1 cannot be achieved, as the separation is a non-linear function of the parameters *m* and *k* (Rao et al., 1997). It follows that the components of EQ. S.1 are not completely orthogonal and that some level of overlapping energy exists (Kang et al., 2013). Galmarini et al. (2013) found that the explained variance by the spectral components account for 75 to 80% of the total variance, the remaining portion being explained by the interactions between the components.

Assuming a spectral decomposition which is valid for the modelling and the observational time series, the MSE formulation outlined in Galmarini et al. (2013) holds:

$$MSE(O_3) = MSE(LT + SY + DU + ID) = \sum MSE(spec\ comp) + \sum MSE\ (cc)) \qquad \text{EQ. S.2}$$

Efisio Solazzo 26/4/2016 14:38
Deleted: $(O3) =$

Efisio Solazzo 26/4/2016 14:38
Deleted: $\sum MSE(cross\ comp)$

Where *spec comp* are the diagonal terms, and *LT, SY, DU, ID* and cc identifies the cross components, i.e. the off-diagonal terms deriving from the squared nature of the MSE: LT_oSY_m, SY_oLT_m, SY_oDU_m, DU_oSY_m, DU_oID_m, ID_oDU_m, LT_mSY_m, LT_oSY_o, DU_mSY_m, DU_mID_m, DU_oSY_o, DU_oID_o (*o* and *m* represent observed and modelled fields, respectively). For simplicity, the cross-components are assumed to be symmetric, so the *o* and *m* subscripts are dropped. This simplification has little impact on the MSE breakdown since, as shown by Galmarini et al. (2013), the diagonal terms alone account for over 80% of the total variance.

Efisio Solazzo 26/4/2016 14:38
Deleted: comp are

To isolate the contribution to MSE of a single spectral component, we proceed as follows. We subtract a component (e.g. LT) from the whole time series:

$$MSE(O_3\text{-}LT(O_3)) = $$
$$MSE(SY)+MSE(DU)+MSE(ID)+2MSE(IDDU)+2MSE(IDSY)+2MSE(DUSY) \qquad \text{EQ. S.3}$$

By removing EQ. S.3 from EQ. S.2, the contribution of LT and its cross-component is isolated:

$$\text{EQ. S.2- EQ. S.3} = MSE(LT) + MSE(LTID) + MSE(LTSY) + MSE(LTDU) \qquad \text{EQ. S.4}$$

We can further elaborate on EQ. S.4 to isolate the contribution of each cross-component. For instance, the case of SYLT:

$$MSE(SY\text{-}ID\text{-}DU) - MSE(SY) - MSE(LT) = [MSE(SY) + MSE(LT) + 2MSE(SYLT)] - MSE(SY) - MSE(LT) = 2MSE(SYLT) \quad \textbf{EQ. S.5}$$

The procedure in EQ. S.5 has been applied to derive the contribution of all cross-components.

Acknowledgements

We would like to thank the community of modellers and data providers of the first and second phases of AQMEII.

References

Baklanov, A., and et al., 2014. Online coupled regional meteorology chemistry models in Europe: current status and prospects. Atmospheric Chemistry and Physics 14, 317-398.

Brunner, D., Jorba, O., Savage, N., Eder, B., Makar, P., Giordano, L., Badia, A., Balzarini, A., Baro, R., Bianconi, R., Chemel, C., Forkel, R., Jimenez-Guerrero, P., Hirtl, M., Hodzic, A., Honzak, L., Im, U., Knote, C., Kuenen, J.J.P., Makar, P.A., Manders-Groot, A., Neal, L., Perez, J.L., Pirovano, G., San Jose, R., Savage, N., Schroder, W., Sokhi, R.S., Syrakov, D., Torian, A., Werhahn, K., Wolke, R., van Meijgaard, E., Yahya, K., Zabkar, R., Zhang, Y., Zhang, J., Hogrefe, C., Galmarini, S., 2015. Evaluation of the meteorological performance of coupled chemistrymeteorology models in phase 2 of the air quality model evaluation international initiative. Atmos. Environ

Dennis, R., Fox, T., Fuentes, M., Gilliland, A., Hanna, S., Hogrefe, C., Irwin, J., Rao, S.T., Scheffe, R., Schere, K., Steyn, D., Venkatram, A., 2010. A framework for evaluating regional-scale numerical photochemical modeling systems. Environ. Fluid Mech. (Dordr.) 10, 471-489. http://dx.doi.org/10.1007/s10652-009-9163e2.

Fox, D.G., 1981. Judging air quality model performance. Bulletin of the American Meteorological Society 62, No.5, 599-609.

Galmarini, S. Solazzo, E., Im, U., Kioutsioukis, I., 2015. AQMEII 1, 2 and 3: Direct and Indirect Benefits of Community Model Evaluation Exercises. 34th International Technical Meeting on Air Pollution Modelling and its Application, Montpellier (France) 4-8 May 2015.

Galmarini, S., Kioutsioukis, I., Solazzo, E., 2013. E pluribus unum: ensemble air quality predictions. Atmos. Chem. Phys. 13, 7153-7182.

Giordano, L., Brunner, D., Flemming, J., Hogrefe, C., Im, U., Bianconi, R., and et al., 2015. Assessment of the MACC reanalysis and its influence as chemical boundary conditions for regional air quality modelling in AQMEII-2. Atmospheric Environment 115, 371-388.

Hastie, T., Tibshirani, R., Friedman, J., 2009. The elements of statistical learning (2nd edition). Springer-Verlag. 763 pages.

Hogrefe, C., Rao, S.T., Zurbenko, I.G., Porter, P.S., 2000. Interpreting the information in ozone observations and model predictions relevant to regulatory policies in the Eastern United States. Bull. Am. Meteorol. Soc. 81, 2083e2106. http:// dx.doi.org/10.1175/1520-0477(2000)0812.3.CO;2.

Hogrefe, C., Roselle, S., Mathur, R., Rao, S.T., Galmarini, S., 2014. Space-time analysis of the Air Quality Model Evaluation International Initiative (AQMEII) phase 1 air quality simulation. J. Air Waste Manag. Assoc. 64, 388-405.

Im, U., Bianconi, R., Solazzo, E., Kioutsioukis, I., Badia, A., Balzarini, A., Baro, R., Bellasio, R., Brunner, D., Chemel, C., Curci, G., Denier van der Gon, H., Flemming, J., Forkel, R., Giordano, L., Jimenez-Guerrero, P., Hirtl, M., Hodzic, A., Honzak, L., Jorba, O., Knote, C., et al., 2015a Evaluation of operational onlinecoupled regional air quality models over Europe and North America in the context of AQMEII phase 2. Part II: particulate matter. Atmos. Environ. 115, 421-441

Im, U., Bianconi, R., Solazzo, E., Kioutsioukis, I., Badia, A., Balzarini, A., Baro, R., Bellasio, R., Brunner, D., Chemel, C., Curci, G., Flemming, J., Forkel, R., Giordano, L., Jimenez-Guerrero, P., Hirtl, M., Hodzic, A., Honzak, L., Jorba, O., Knote, C., Kuenen, J. J.P., et al., 2015b. Evaluation of operational on-line-coupled regional air quality models over Europe and North America in the context of AQMEII phase 2. Part I: ozone. Atmos. Environ. 115, 404-420

Johnson, R. 2008 Assessment of Bias with Emphasis on Method Comparison. Clin Biochem Rev Vol 29 Suppl (i) S37–S42.

Kang, D., Hogrefe, C., Foley, K.L., Napelenok, S.L., Mathur, R., Rao, S.T., 2013. Application of the Kolmogorov-Zurbenko filter and the decoupled direct 3D method for the dynamic evaluation of a regional air quality model. Atmos. Environ. 80, 58-69.

Kioutsioukis, I., Galmarini, S., 2014. De praeceptis ferendis: good practice in multi-model ensembles. Atmospheric Chemistry and Physics 14, 11791–11815.

Makar, P.A., Gong, W., Hogrefe, C., and et al., 2015. Feedbacks between air pollution and weather, part 2: effects on chemistry. Atmospheric Environment 115, 499-526

Murphy, A.H., 1988. Skill scores based on the mean square error and their relationship to the correlation coefficient. Monthly Weather Review 116, 2417-2424

Pindyck, R.S., Rubinfeld, D.L., 1998. Econometric Models and Economic Forecast, Irwin/McGraw-Hill, Singapore, 388 pg

Pouliot, G., Denier van der Gon, H., Kuenen, J., Makar, P., Zhang, J., Moran, M., 2015. Analysis of the emission inventories and model-ready emission datasets of Europe and North America for phase 2 of the AQMEII project. Atmos. Environ. 115, 345-360.

Pouliot, G., Pierce, T., Denier van der Gon, H., Schaap, M., Moran, M., and Nopmongcol, U., 2012. Comparing Emissions Inventories and Model-Ready Emissions Datasets between Europe and North America for the AQMEII Project. Atmos. Environ. 53, 4–14.

Rao, K.S., 2005. Uncertainty analysis in atmospheric dispersion modelling. Pure and Applied Geophysics 162, 1893-1917.

Rao, S.T., Galmarini, S., Puckett, K., 2011. Air quality model evaluation international initiative (AQMEII). Bull. Am. Meteorol. Soc. 92, 23-30. http://dx.doi.org/ 10.1175/2010BAMS3069.1.

Rao, S.T., Zurbenko, I.G., Neagu, R., Porter, P.S., Ku, J.Y., Henry, R.F., 1997. Space and time scales in ambient ozone data. Bull. Am. Meteorol. Soc. 78, 2153e2166. http://dx.doi.org/10.1175/1520-0477(1997)0782.0.CO;2.

Schere, K., Flemming, J., Vautard, R., Chemel, C., Colette, A., Hogrefe, C., Bessagnet, B., Meleux, F., Mathur, R., Roselle, S., Hu, R.-M., Sokhi, R. S., Rao, S. T., and Galmarini, S.: Trace gas/aerosol concentrations and their impacts on continental-scale AQMEII modelling sub-regions, Atmos. Environ., 53, 38–50, 2012.

Solazzo, E., Bianconi, R., Vautard, R., Appel, K.W., Moran, M.D., Hogrefe, C., Bessagnet, B., Brandt, J., Christensen, J.H., Chemel, C., Coll, I., van der Gon, H.D., Ferreira, J., Forkel, R., Francis, X.V., Grell, G., Grossi, P., Hansen, A.B., Jericevic, A., Kraljevic, L., Miranda, A.I., Nopmongcol, U., Pirovano, G., Prank, M., Riccio, A., Sartelet, K.N., Schaap, M., Silver, J.D., Sokhi, R.S., Vira, J., Werhahn, J., Wolke, R., Yarwood, G., Zhang, J., Rao, S.T., Galmarini, S., 2012a. Model evaluation and ensemble modelling and for surface-level ozone in Europe and North America. Atmos. Environ. 53, 60-74.

Solazzo, E., Bianconi, R., Pirovano, G., Matthias, V., Vautard, R., Moran, M.D., Appel, K.W., Bessagnet, B., Brandt, J., Christensen, J.H., Chemel, C., Coll, I., Ferreira, J., Forkel, R., Francis, X.V., Grell, G., Grossi, P., Hansen, A.B., Hogrefe, C., Miranda, A.I., Nopmongco, U., Prank, M., Sartelet, K.N., Schaap, M., Silver, J.D., Sokhi, R.S., Vira, J., Werhahn, J., Wolke, R., Yarwood, G., Zhang, J., Rao, S.T., Galmarini, S., 2012b. Operational model evaluation for particulate matter in Europe and North America. Atmos. Environ. 53, 75-92.

Solazzo, E., Bianconi, R., Pirovano, G., Moran, M., Vautard, R., Hogrefe, C., Appel, K.W., Matthias, V., Grossi, P., Bessagnet, B., Brandt, J., Chemel, C., Christensen, J.H., Forkel, R., Francis, X.V., Hansen, A., McKeen, S., Nopmongcol, U., Prank, M., Sartelet, K.N., Segers, A., Silver, J.D., Yarwood, G., Werhahn, J., Zhang, J., Rao, S.T., Galmarini, S., 2013a. Evaluating the capabilities of regional scale air quality models to capture the vertical distribution of pollutants. Geophys. Model Dev. 6, 791-818.

Solazzo, E., Riccio, A., Kioutsioukis, I., Galmarini, S., 2013b. *Pauci ex tanto numero*: reduce redundancy in multi-model ensemble. Atmos. Chem. Phys. 13, 8315-8333.

Solazzo, E., Galmarini, S., 2015. Comparing apples with apples: Using spatially distributed time series of monitoring data for model evaluation. Atmos. Environ. 112, 234-245

Stoeckenius, T.E., Hogrefe, C., Zagunis, J., Sturtz, T.M., Wells, B., Sakulyanontvittaya, T., 2015. A comparison between 2010 and 2006 air quality and meteorological conditions, and emissions and boundary conditions used in simulations of the AQMEII2 North American domain. Atmospheric Environment, 115, 389-403.

Theil, H., 1961. Economic forecast and policy. North-Holland, Amsterdam

Willmott, C.J., and et al., 1985. Statistics for the evaluation and comparison of models. Journal of Geophysical research 90, No. C5, 8995-9005.

Wise, E.K., Comrie, A.C., 2005. Extending the KZ filter: application to ozone, particulate matter, and meteorological trends. J. Air Waste Manag. Assoc. 55 (8), 1208e1216.

Zurbenko, I.G., 1986. The Spectral Analysis of Time Series. North-Holland, Amsterdam, 236 pp.

FIGURES

Figure 1. Share (in %) of the total MSE in the main spectral components and the cross components (see Appendix for detail) for *a)* AQMEII1 and *b)* AQMEII2. Top panel: EU; lower panel: NA.

Figure 2. MSE (ppb^2) breakdown in bias squared, variance and *mMSE* of the spectral components ID, DU, SY, LT, based on Eq 9. The bias is entirely accounted for by the LT component. The sign within the share of bias and variance indicates model overestimation (+) or underestimation (-) of mean concentration (bias) and variance. The colour of the *mMSE* share of the error is coded based on the values of r, the correlation coefficient, according to the colour scale at the bottom of each plot. *a)* AQMEII1 and *b)* AQMEII2. Top panel: EU; lower panel: NA.

Figure 3. Spatial distribution of the MSE in the spectral components for the EU models of AQMEII1. The segments are centred at the rural receptors' position (clockwise from north: MSE of ID, DU, SY, and LT). Their length is proportional to the MSE magnitude, coded according to the colour scale. For each model, the colour scale extends from zero up to the 75th percentile, and the last value of the scale is the maximum MSE. The colour of the MSE values above the 75th percentile represents the maximum value. The tick-dashed LT segment indicates model underestimation (low model bias), while thin continuous segment indicates model overestimation (high model bias). The example in the last panel indicates how the maps reports the error of the spectral components at each receptor (the colours are arbitrary). The example on the left represents the error at a receptor where the LT component is biased high, while the example on the right refers to a case where the bias is negative. The other components do not change.

Efisio Solazzo 26/4/2016 14:38
Deleted: The thick dashed LT segment indicates model underestimation (low model bias).

Figure 4 As in **Figure 3,** but for the NA models of AQMEII1.

Figure 5. As in **Figure 3,** but for the EU models of AQMEII2.

Figure 6 As in **Figure 3,** but for the NA models of AQMEII2.

Figure 7 Example of the model complexity as time-resolved scale of the transport and dispersion processes: the minimum complexity (far right) is a poor time-resolving time series obtained as *kz(250,5)* (> 1 month). The complexity increases towards the left, with the scale of resolved processes becoming finer up to the maximum complexity (far left), which represents the full time series. The upper panel shows an example of how the curves of the error for covariance, variance and bias vary according to complexity.

Figure 8 Evolution of error components (red: bias; Blue: variance; Black: covariance) as a function of model complexity. Complexity increases from right (min) to left (MAX) and is calculated as the temporal scale of the resolved process using the kz filter on the modelled signal: kz(i,5), i=2,…,250.

Efisio Solazzo 26/4/2016 14:38
Deleted: Red

Efisio Solazzo 26/4/2016 14:38
Deleted: left (min.) to

Efisio Solazzo 26/4/2016 14:38
Deleted: max.)

FIGURE S1. Sub-regions of the two continental domains a) EU, and b) NA. Overlaid are the ozone monitoring stations for the year 2010 classified based on the network**.**

FIGURE S2. MSE (ppb^2) breakdown in bias, variance and mMSE of the spectral components ID, DU, SY, LT (based on Eq 9) for the models of AQMEII1 and the three sub-regions of Figure S1. The sign within the share of bias and variance indicates model overestimation (+) or underestimation (-) of mean concentration (bias) and variance. Top three panels: EU; lower three panels: NA.

FIGURE S3. As in Figure S2 for the AQMEII2 models

TABLES

Efisio Solazzo 26/4/2016 14:38
Deleted: **FIGURE S4.** Variance (left column), Bias (central), and Covariance (right) for each modelling system and at all receptors as function of model complexity (decreasing from left to right on the horizontal axis). Vertical dotted lines represent the separation scale between SY and LT (103) and between ID and DU (13). Vertical axis is in pp ... [2]

Table 1. Features of the modelled domains

	Europe		**North America**	
	phase 1	**phase 2**	**phase 1**	**phase 2**
Simulated year	2006	2010	2006	2010
Extension	(-10,39)W; (30,65)N		(-125,-55)W; (26,51)N	
Number of receptors (min validity=75%; max altitude = 1000 m)	1339	1360	672	652

Table 2. Modelling systems participating in the first (Table a) and second (Table b) phases of AQMEII for Europe and North America

a)

Model			Grid(km)	Emissions	Chemical BC
Code	**Met**	**AQ**			
EUROPE – AQMEII 1					
DK1	MM5	DEHM	50	Global emission databases, EMEP	Satellite measurements
FR3	MM5	Polyphemus	24	Standard§	Standard
HR1	PARLAM-PS	EMEP	50	EMEP model	From ECMWF and forecasts
UK2	WRF	CMAQ	18	Standard§	Standard
US4	WRF	WRF/Chem	22.5	Standard§	Standard
FI1	ECMWF	SILAM	24	Standard anthropogenic; In-house biogenic	Standard
FR4	MM5	Chimere	25	MEGAN, Standard	Standard

PL1	GEM	GEM-AQ	25	Standard over AQMEII region; Global EDGAR/GEIA over the rest of the global domain	Global variable grid setup (no boundary conditions)
NL1	ECMWF	Lotos-EUROS	25	Standard[§]	Standard
DE1	COSMO	Muscat	24	Standard[§]	Standard
US3	MM5	CAMx	15	MEGAN, Standard	Standard
DE3	COSMO-CLM	CMAQ	24	Standard[§]	Standard
NORTH AMERICA- AQMEII 1					
CA1	GEM	AURAMS	45	Standard*	Climatology
PL1	GEM	GEM-AQ	25	Standard over AQMEII region; Global EDGAR/GEIA over the rest of the global domain	Global variable grid setup (no boundary conditions)
PT1	MM5	CAMx	24	Standard	LMDZ-INCA
US1	WRF	CAMQ	12	Standard	Standard
US3	WRF	CAMx	12	Standard	Standard
FR4b	WRF	CHIMERE			
DK1	MM5	DEHM	50	Global emission databases, EMEP	Satellite measurements
DE3	COSMO-CLM	CMAQ	24	Standard[§]	Standard
ES3	WRF	WRF/Chem	23	Standard	Standard

[§] Standard anthropogenic emissions and biogenic emissions derived from meteorology (temperature and solar radiation) and land use distribution implemented in the meteorological driver.

*Standard anthropogenic inventory but independent emission processing, exclusion of wildfires, and different versions of BEIS(v3.09) used.

Refer to Solazzo et al. (2012a-b) and references therein for details.

b)

Model			Grid	Emissions	Chemical BC
Code	Met	AQ			
EUROPE – AQMEII 2					
AT1	WRF	WRF/Chem	23 km	Standard	Standard
CH1	COSMO	Cosmo-ART	0.22°	Standard	Standard
ES2a	NMMB	BSCCTM	0.20°	Standard	Standard
ES3	WRF	WRF/Chem	23 km	Standard	Standard
NL2	RACMO	LOTOS-EUROS	0.5° x 0.25°	Standard	Standard
UK5	WRF	CMAQ	18 km	Standard	Standard
UK4	MetUM	UKCA RAQ	0.22°	Standard	Standard
DE3	COSMO	Muscat	0.25°	Standard	Standard
NORTH AMERICA – AQMEII 2					
ES1	WRF	WRF/CHem	36 km	Standard	Standard
US6	WRF	CMAQ	12km	Standard	Standard

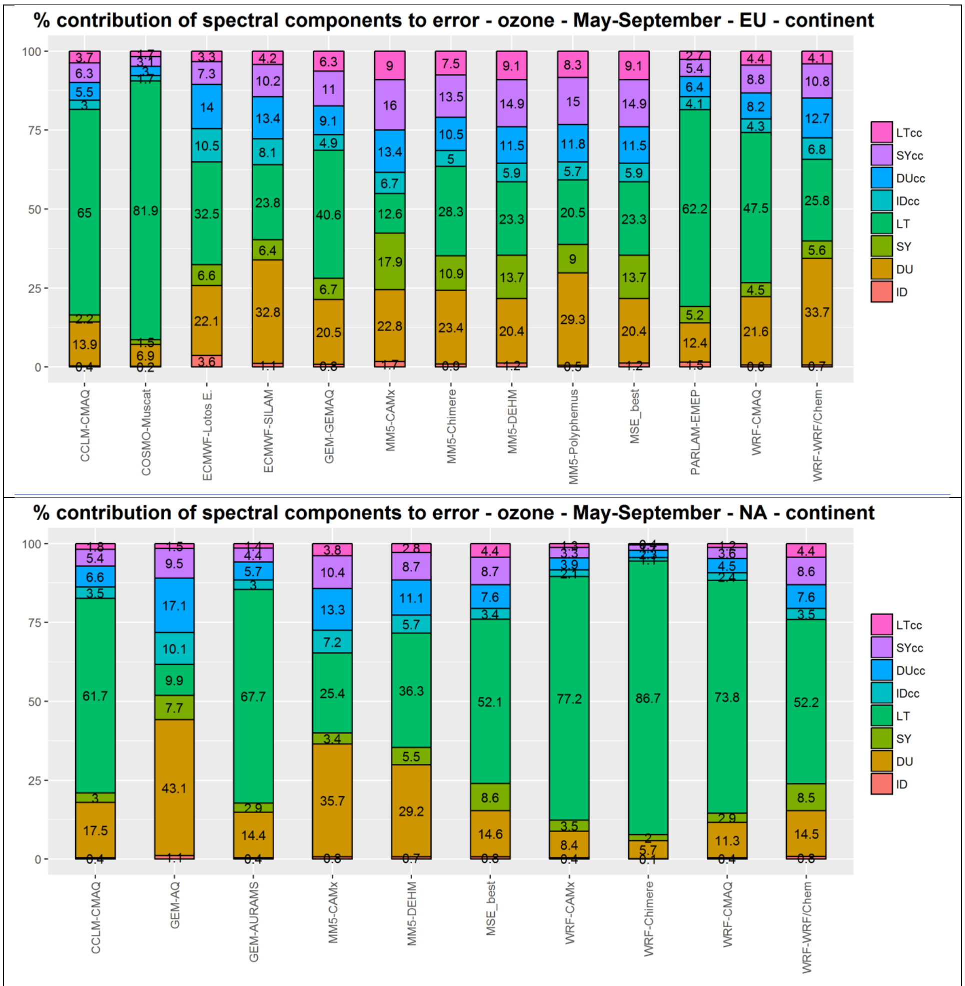
CA2f	GEM	MACH	15 km	Standard	Standard

Standard Boundary conditions: 3-D daily chemical boundary conditions were provided by the ECMWF IFS-MOZART model run in the context of the MACC-II project (Monitoring Atmospheric Composition and Climate - Interim Implementation) at 3-hourly and 1.125 spatial resolution. Refer to Im et al. (2015a-b) for details.

Standard Emissions: based on the TNO-MACC-II (Netherlands Organization for Applied Scientific Research, Monitoring Atmospheric Composition and Climate - Interim Implementation) framework for Europe and by the US EPA (Environmental Protection Agency) and Environment Canada for North America. The 2008 National Emissions Inventory (http://www.epa.gov/ttn/chief/net/2008inventory.html) and the 2008 Emissions Modeling Platform (http://www.epa.gov/ttn/chief/ emch/index.html#2008) with year-specific updates for 2006 and 2010 were used for the US portion of the modelling domain. Canadian emissions were derived from the Canadian National Pollutant Release Inventory (http://www.ec.gc.ca/inrp-npri/) and Air Pollutant Emissions Inventory (http://www.ec.gc.ca/inrp-npri/ donnees-data/ap/index.cfm?lang¼En) values for the year 2006. Refer to Im et al. (2015a-b) for details.

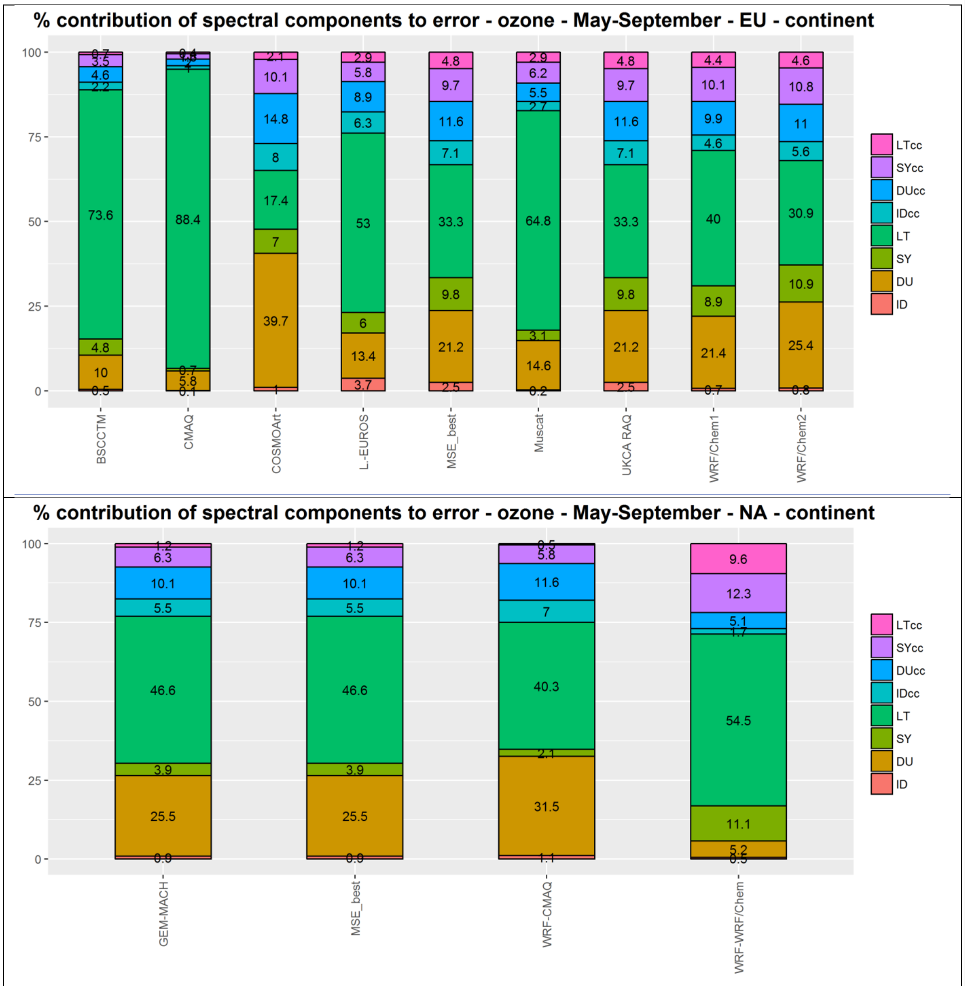
FIGURES

AQMEII1


% contribution of spectral components to error - ozone - May-September - EU - continent
100
75
50
25
0
CCLM-CMAQ: 3.7, 6.3, 5.5, 3, 65, 2.2, 13.9, 0.4
COSMO-Muscat: 1.7, 3.1, 3, 1.7, 81.9, 1.5, 6.9, 0.2
ECMWF-Lotos E.: 3.3, 7.3, 14, 10.5, 32.5, 6.6, 22.1, 3.6
ECMWF-SILAM: 4.2, 10.2, 13.4, 8.1, 23.8, 6.4, 32.8, 1.1
GEM-GEMAQ: 6.3, 11, 9.1, 4.9, 40.6, 6.7, 20.5, 0.8
MM5-CAMx: 9, 16, 13.4, 6.7, 12.6, 17.9, 22.8, 1.7
MM5-Chimere: 7.5, 13.5, 10.5, 5, 28.3, 10.9, 23.4, 0.9
MM5-DEHM: 9.1, 14.9, 11.5, 5.9, 23.3, 13.7, 20.4, 1.2
MM5-Polyphemus: 8.3, 15, 11.8, 5.7, 20.5, 9, 29.3, 0.5
MSE_best: 9.1, 14.9, 11.5, 5.9, 23.3, 13.7, 20.4, 1.2
PARLAM-EMEP: 2.7, 5.4, 6.4, 4.1, 62.2, 5.2, 12.4, 1.5
WRF-CMAQ: 4.4, 8.8, 8.2, 4.3, 47.5, 4.5, 21.6, 0.8
WRF-WRF/Chem: 4.1, 10.8, 12.7, 6.8, 25.8, 5.6, 33.7, 0.7
LTcc
SYcc
DUcc
IDcc
LT
SY
DU
ID
% contribution of spectral components to error - ozone - May-September - NA - continent
100
75
50
25
0
CCLM-CMAQ: 1.8, 5.4, 6.6, 3.5, 61.7, 3, 17.5, 0.4
GEM-AQ: 1.5, 9.5, 17.1, 10.1, 9.9, 7.7, 43.1, 1.1
GEM-AURAMS: 1.4, 4.4, 5.7, 3, 67.7, 2.9, 14.4, 0.4
MM5-CAMx: 3.8, 10.4, 13.3, 7.2, 25.4, 3.4, 35.7, 0.8
MM5-DEHM: 2.8, 8.7, 11.1, 5.7, 36.3, 5.5, 29.2, 0.7
MSE_best: 4.4, 8.7, 7.6, 3.4, 52.1, 8.6, 14.6, 0.8
WRF-CAMx: 1.3, 3.3, 3.9, 2.1, 77.2, 3.5, 8.4, 0.4
WRF-Chimere: 0.4, 2.7, 2.3, 1.1, 86.7, 2, 5.7, 0.1
WRF-CMAQ: 1.2, 3.6, 4.5, 2.4, 73.8, 2.9, 11.3, 0.4
WRF-WRF/Chem: 4.4, 8.6, 7.6, 3.5, 52.2, 8.5, 14.5, 0.8
LTcc
SYcc
DUcc
IDcc
LT
SY
DU
ID

a)

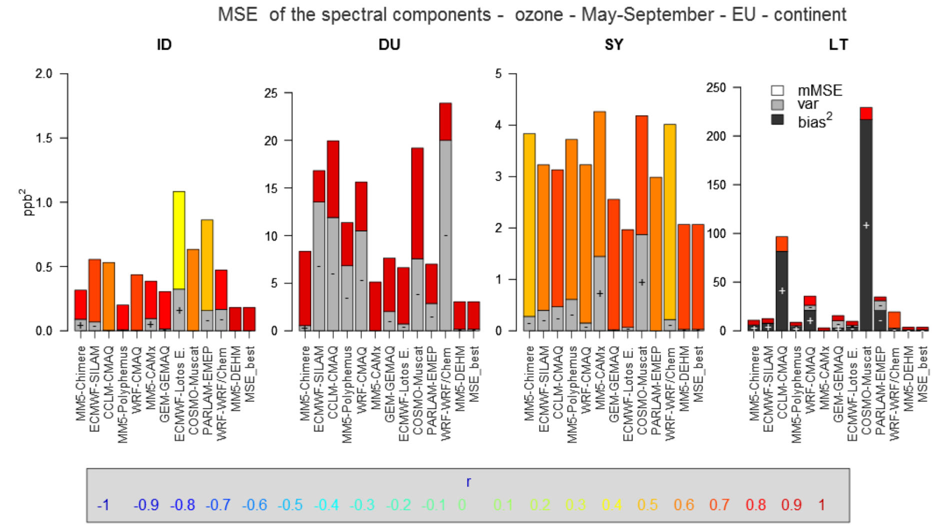
AQMEII2

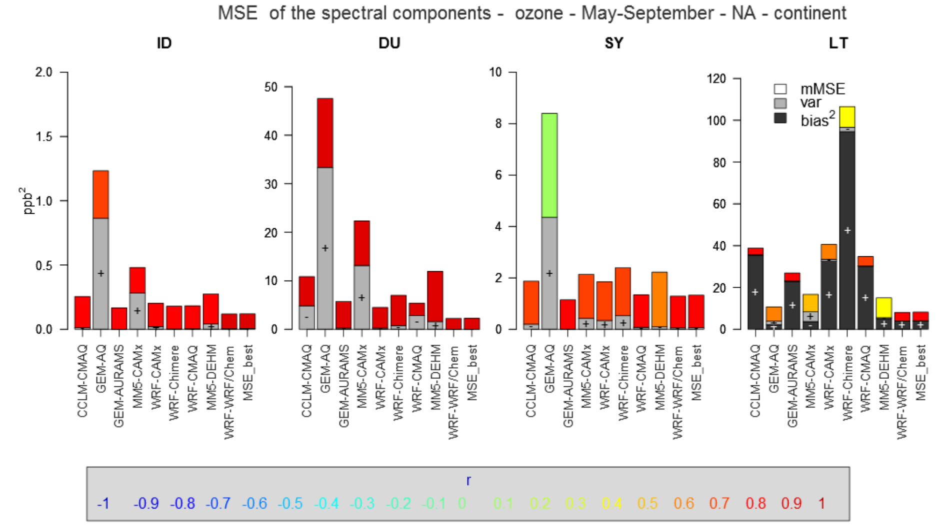


b)

Figure 9. Share (in %) of the total MSE in the main spectral components and the cross components (see Appendix for detail) for a) AQMEII1 and b) AQMEII2. Top panel: EU; lower panel: NA.

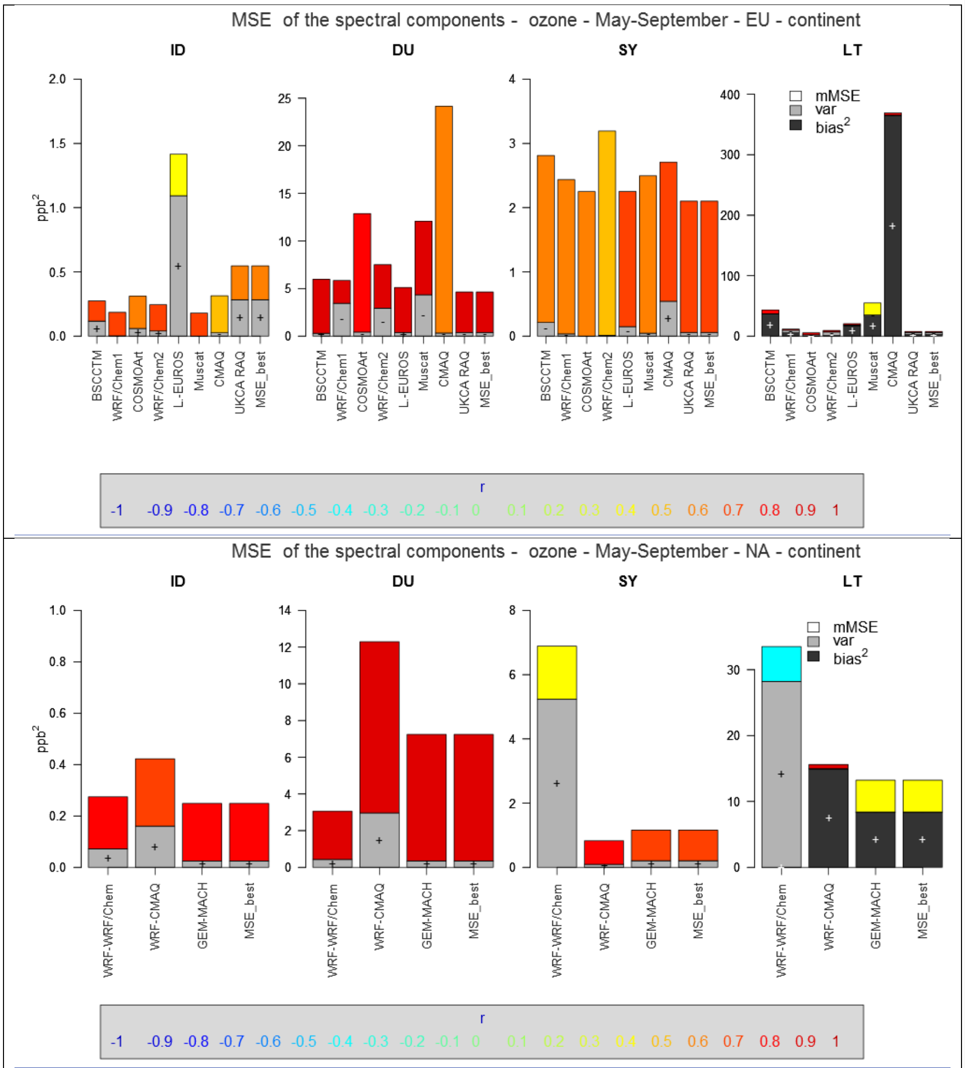
AQMEII1





a)

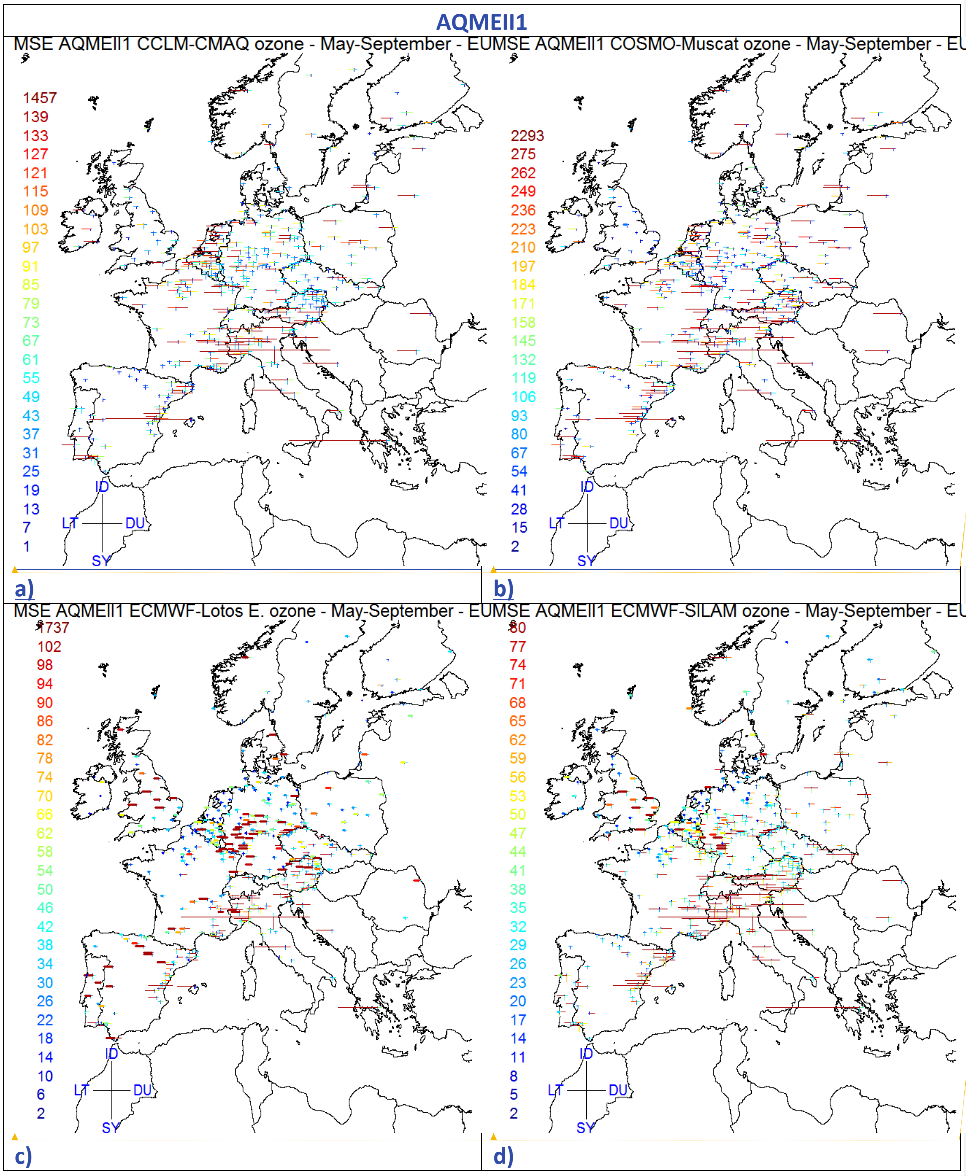
AQMEII2



b)

Figure 10. MSE (ppb^2) breakdown in bias squared, variance and *mMSE* of the spectral components ID, DU, SY, LT, based on Eq 9. The bias is entirely accounted for by the LT component. The sign within the share of bias and variance indicates model overestimation (+) or underestimation (-) of mean concentration (bias) and variance. The colour of the *mMSE* share of the error is coded based on the values of *r*, the correlation coefficient, according to the colour scale at the bottom of each plot.

a) AQMEII1 and b) AQMEII2. Top panel: EU; lower panel: NA.



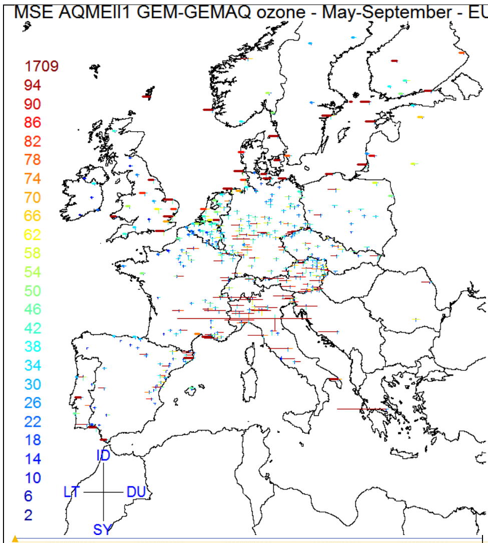
Unknown
Formatted: Font:Bold

Unknown
Formatted: Font:Bold

Unknown
Formatted: Font:Bold

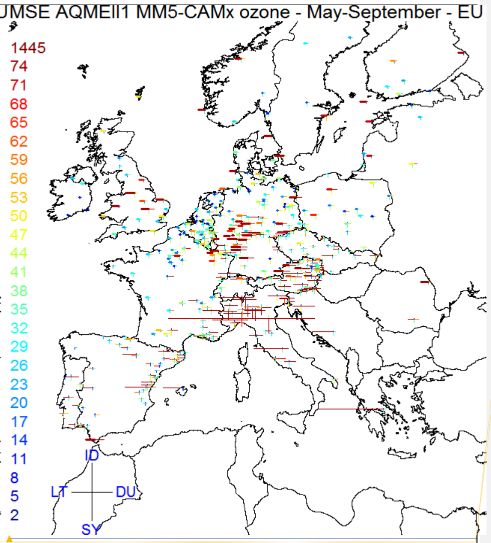
Unknown
Formatted: Font:Bold

MSE AQMEII1 GEM-GEMAQ ozone - May-September - EU



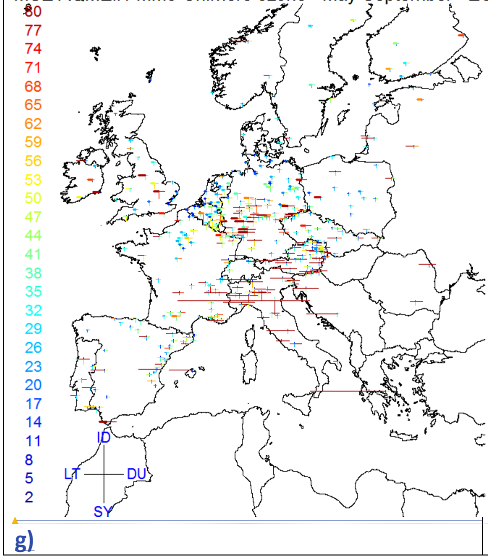
e)

MSE AQMEII1 MM5-CAMx ozone - May-September - EU



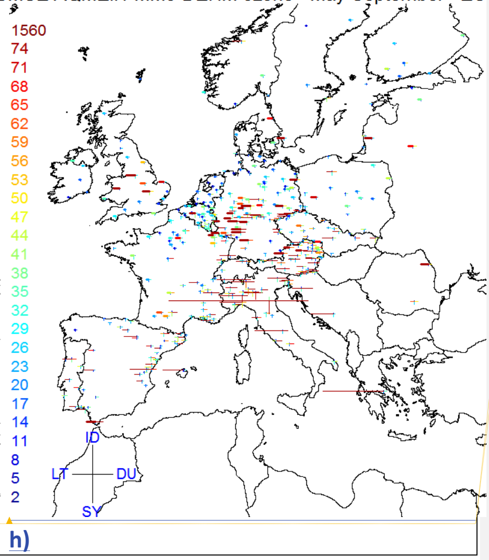
f)

MSE AQMEII1 MM5-Chimere ozone - May-September - EU



g)

MSE AQMEII1 MM5-DEHM ozone - May-September - EU



h)

Unknown
Formatted: Font:Bold

Unknown
Formatted: Font:Bold

Unknown
Formatted: Font:Bold

Unknown
Formatted: Font:Bold

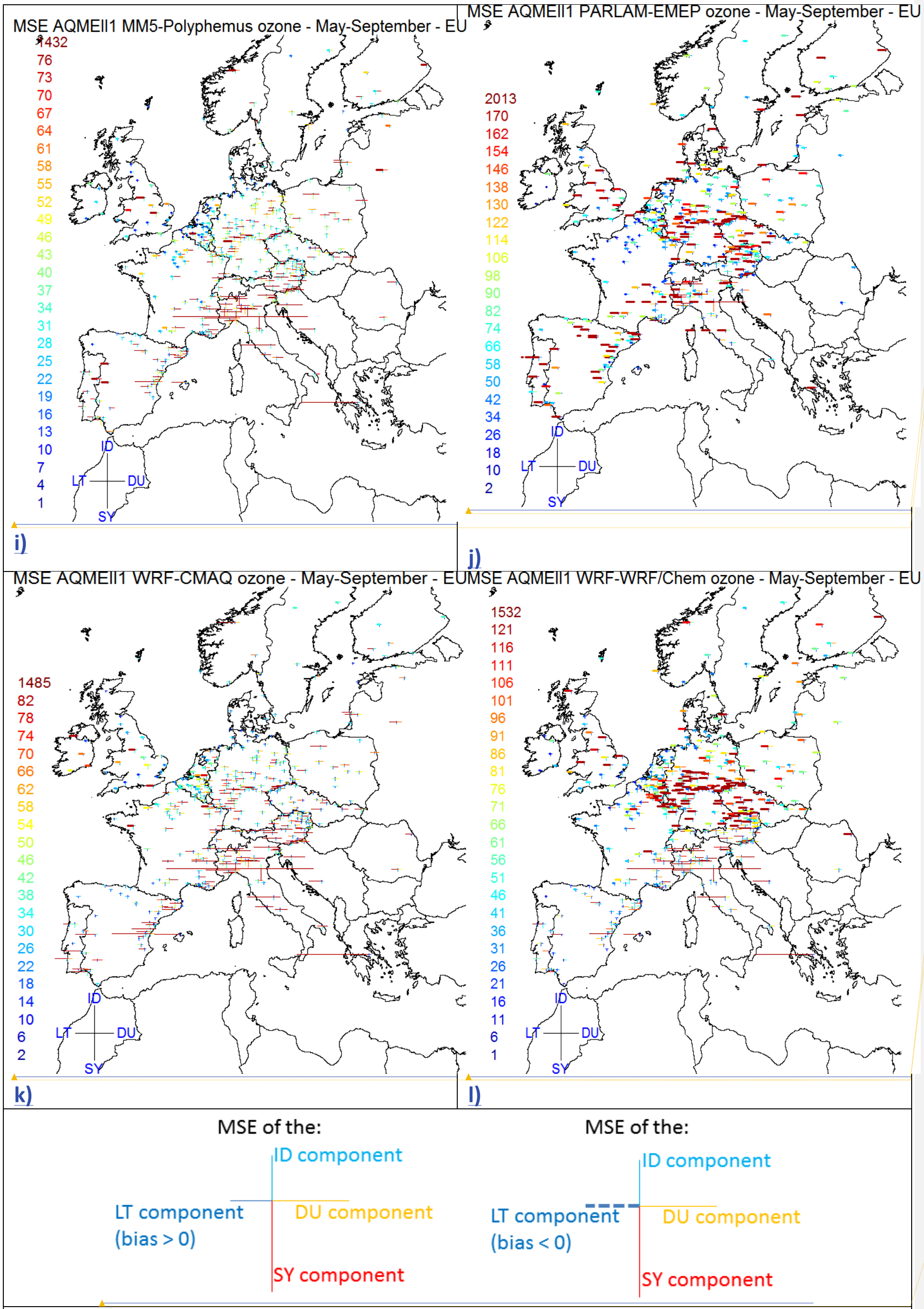


Figure 11. Spatial distribution of the MSE in the spectral components for the EU models of AQMEII1. The segments are centred at the rural receptors' position (clockwise from north:

Unknown
Formatted: Font:Bold

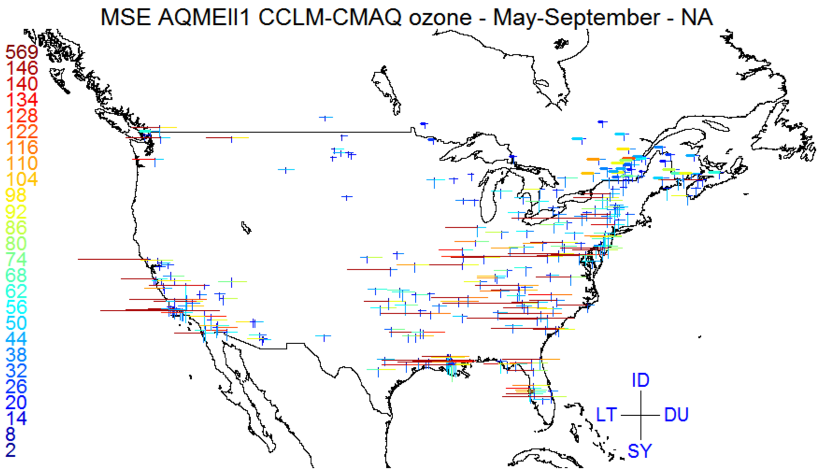
Unknown
Formatted: Font:Bold

Unknown
Formatted: Font:Bold

Unknown
Formatted: Font:Bold

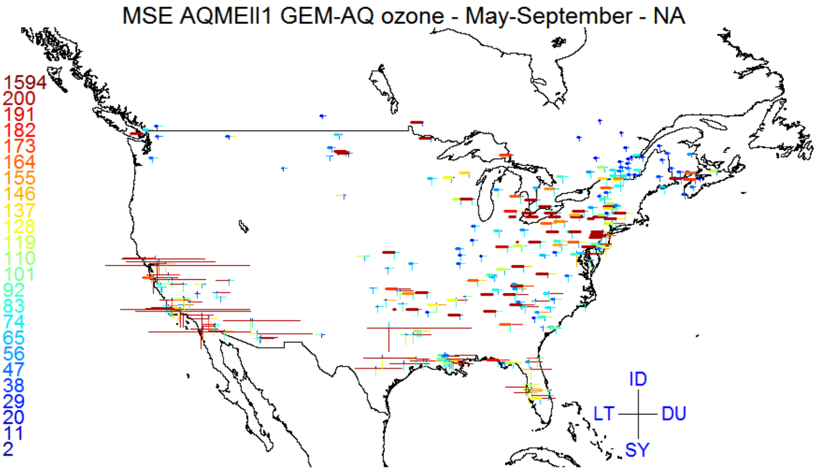
Unknown
Formatted: Font:Bold

MSE of ID, DU, SY, and LT). Their length is proportional to the MSE magnitude, coded according to the colour scale. For each model, the colour scale extends from zero up to the 75th percentile, and the last value of the scale is the maximum MSE. The colour of the MSE values above the 75th percentile represents the maximum value. The tick-dashed LT segment indicates model underestimation (low model bias), while thin continuous segment indicates model overestimation (high model bias). The example in the last panel indicates how the maps reports the error of the spectral components at each receptor (the colours are arbitrary). The example on the left represents the error at a receptor where the LT component is biased high, while the example on the right refers to a case where the bias is negative. The other components do not change.



a)

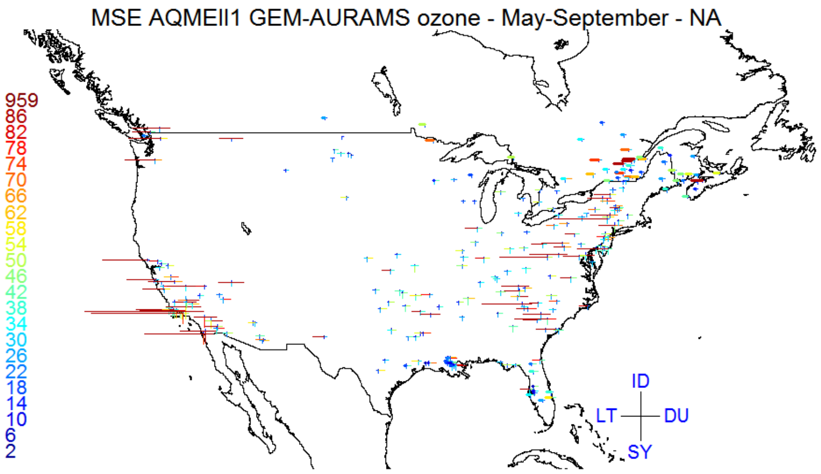
Unknown
Formatted: Font:Bold



b)

Unknown

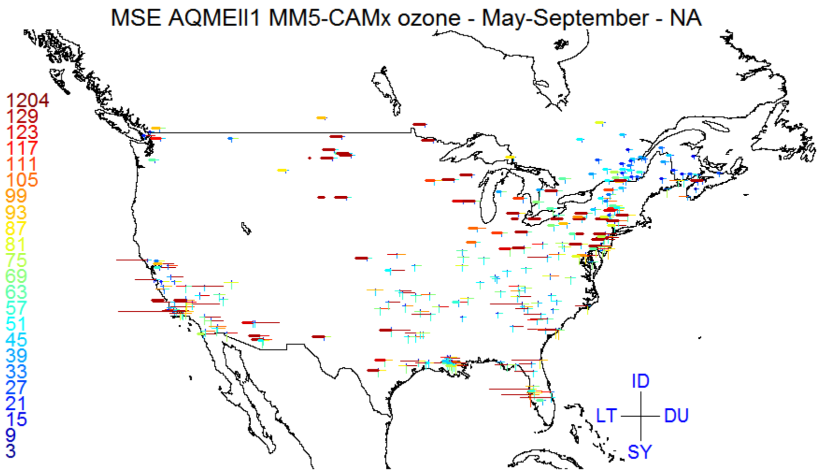
Formatted: Font:Bold



c)

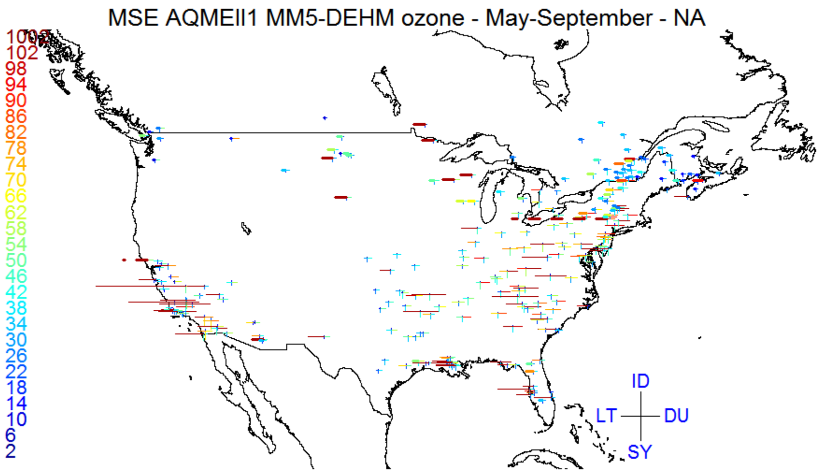
Unknown

Formatted: Font:Bold



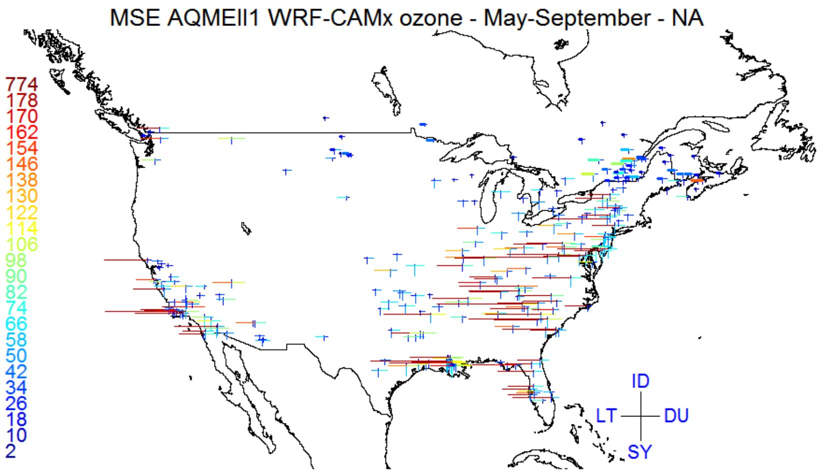
d)

Unknown
Formatted: Font:Bold



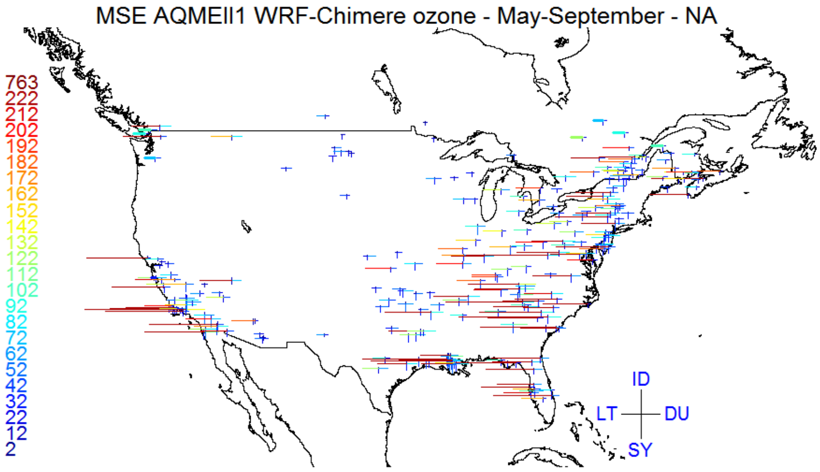
e)

Unknown
Formatted: Font:Bold



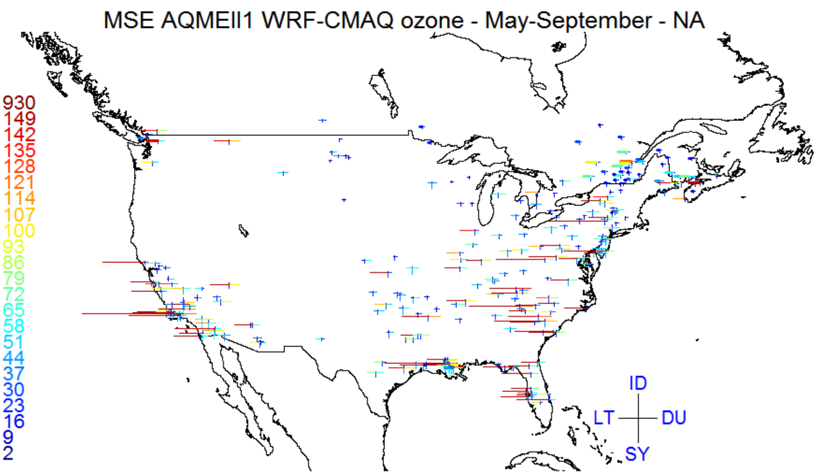
f)

Unknown
Formatted: Font:Bold



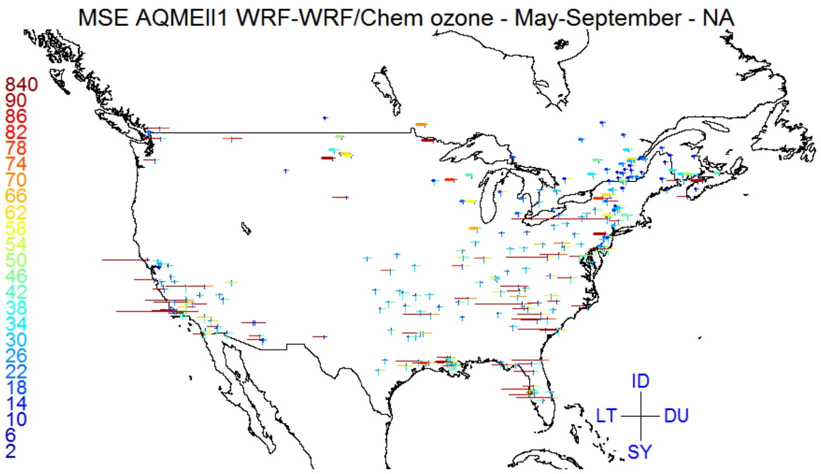
g)

Unknown
Formatted: Font:Bold



h)

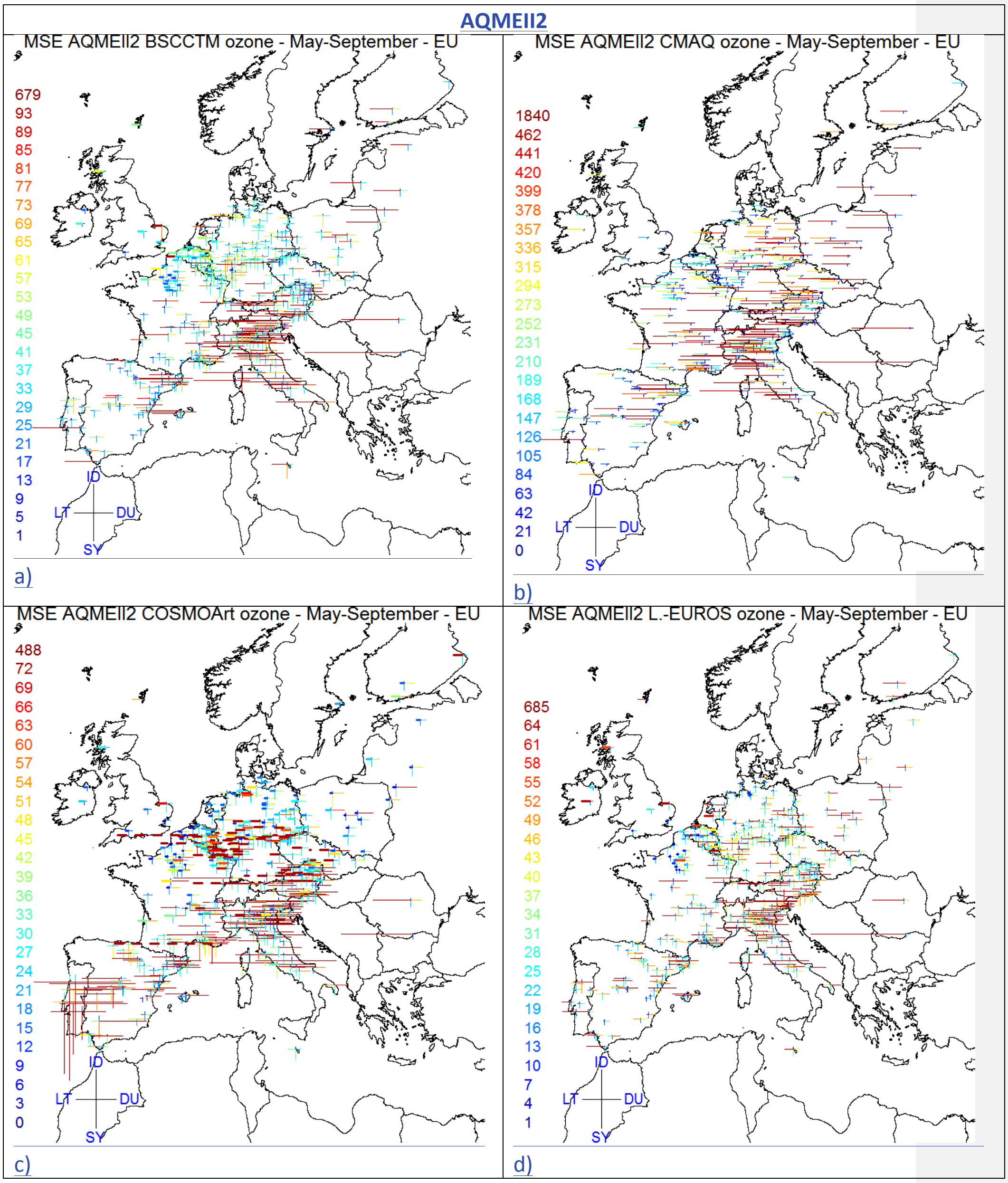
Unknown
Formatted: Font:Bold



i)

Figure 12. As in **Figure 11** but for the NA models of AQMEII1

Unknown
Formatted: Font:Bold


AQMEII2
MSE AQMEII2 BSCCTM ozone - May-September - EU
679
93
89
85
81
77
73
69
65
61
57
53
49
45
41
37
33
29
25
21
17
13
9
5
1
ID
LT
DU
SY
a)
MSE AQMEII2 CMAQ ozone - May-September - EU
1840
462
441
420
399
378
357
336
315
294
273
252
231
210
189
168
147
126
105
84
63
42
21
0
ID
LT
DU
SY
b)
MSE AQMEII2 COSMOArt ozone - May-September - EU
488
72
69
66
63
60
57
54
51
48
45
42
39
36
33
30
27
24
21
18
15
12
9
6
3
0
ID
LT
DU
SY
c)
MSE AQMEII2 L.-EUROS ozone - May-September - EU
685
64
61
58
55
52
49
46
43
40
37
34
31
28
25
22
19
16
13
10
7
4
1
ID
LT
DU
SY
d)

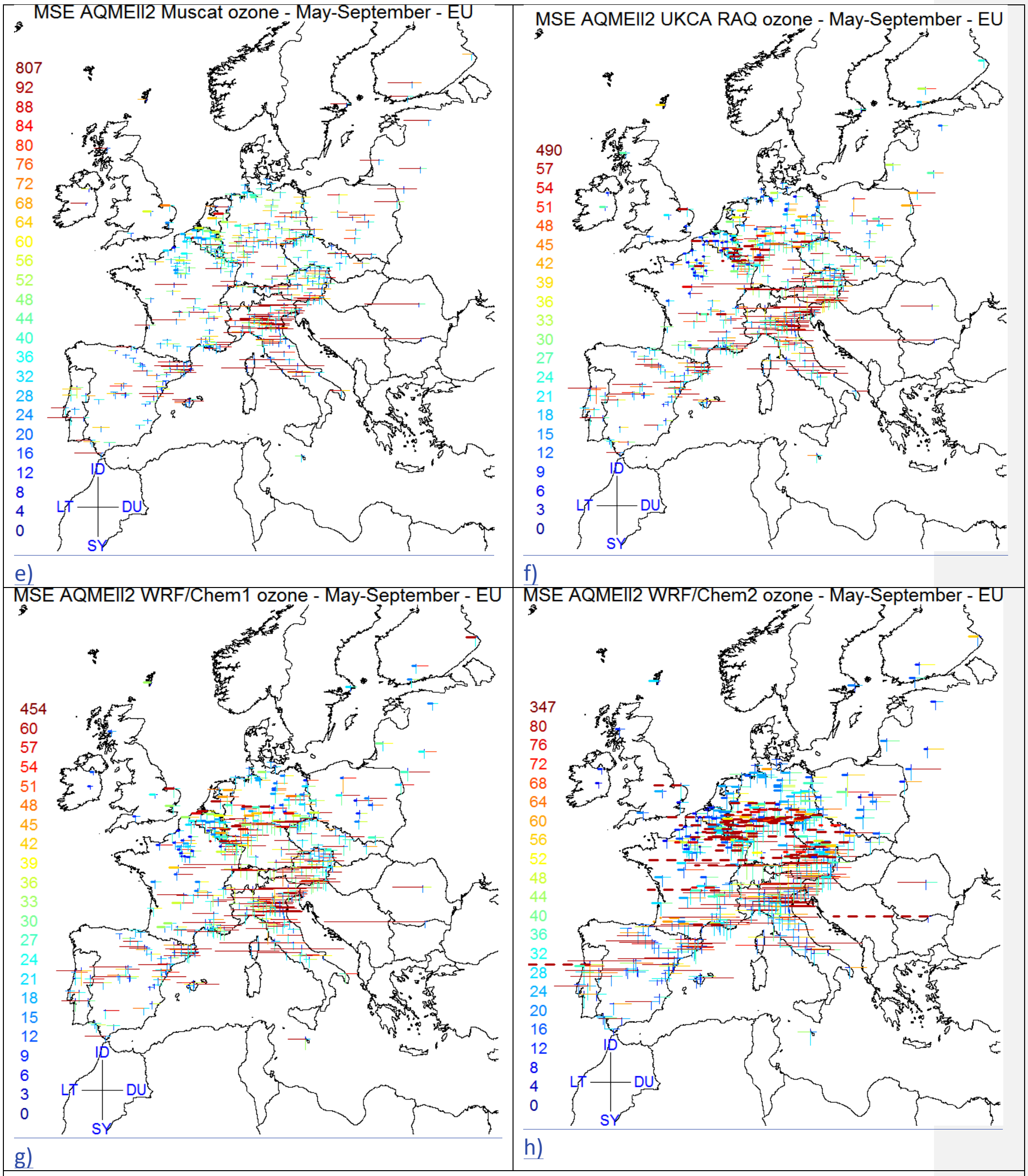
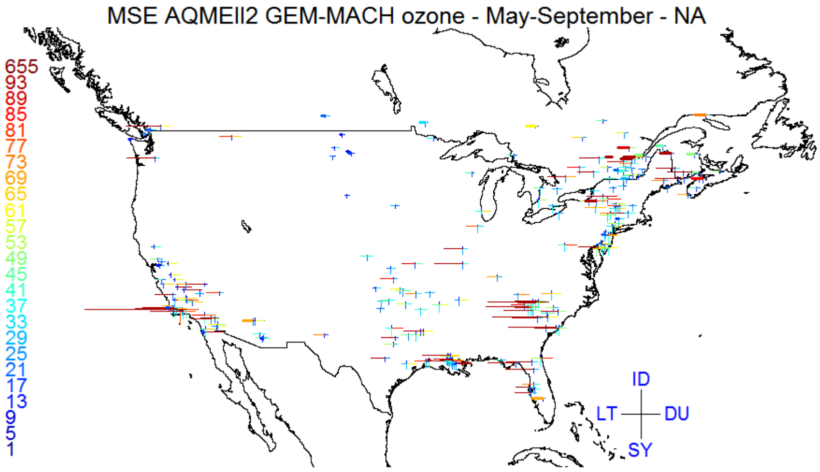
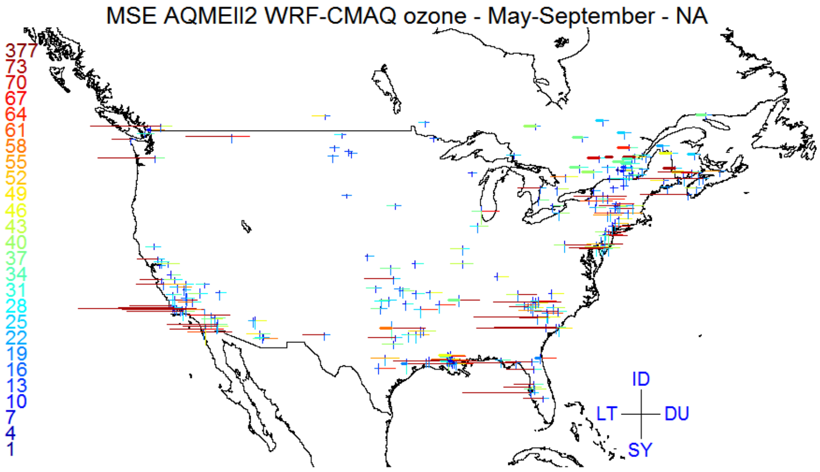


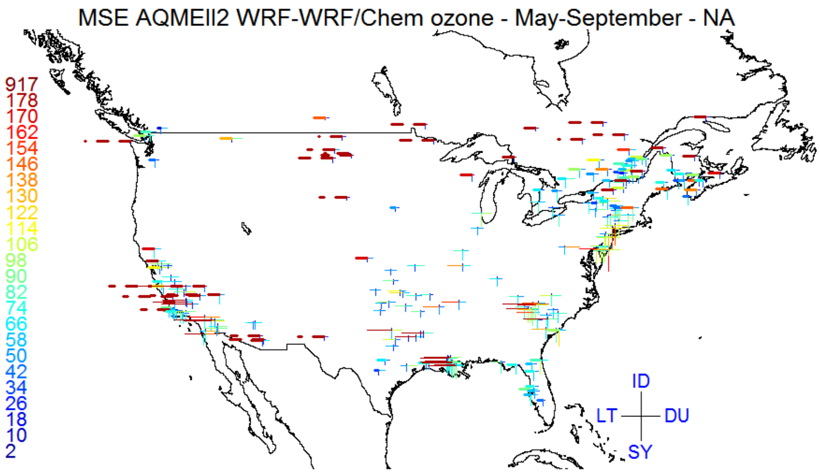
Figure 13. As in **Figure 11** but for the EU models of AQMEII2


MSE AQMEII2 GEM-MACH ozone - May-September - NA
655
93
89
85
81
77
73
69
65
61
57
53
49
45
41
37
33
29
25
21
17
13
9
5
1
ID
LT
DU
SY

a)


MSE AQMEII2 WRF-CMAQ ozone - May-September - NA
377
73
70
67
64
61
58
55
52
49
46
43
40
37
34
31
28
25
22
19
16
13
10
7
4
1
ID
LT
DU
SY

b)



c)

Figure 14. As in **Figure 11** but for the NA models of AQMEII2

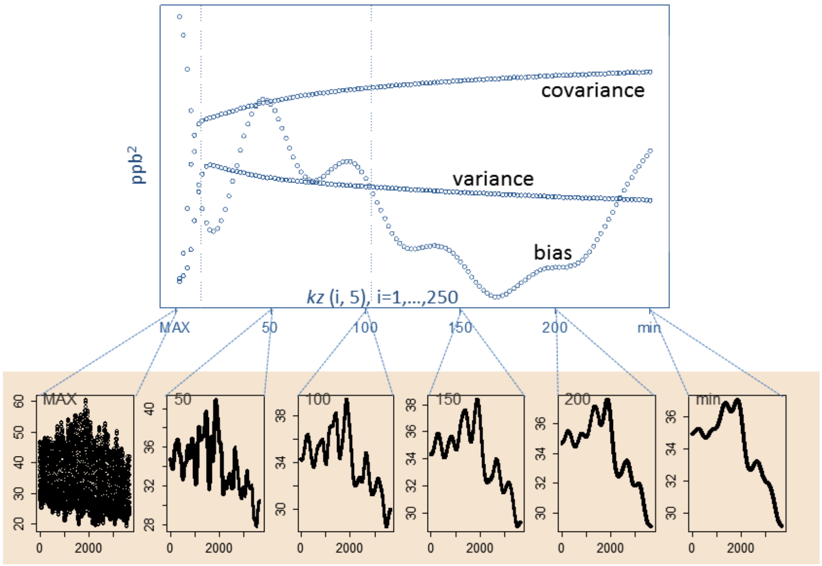
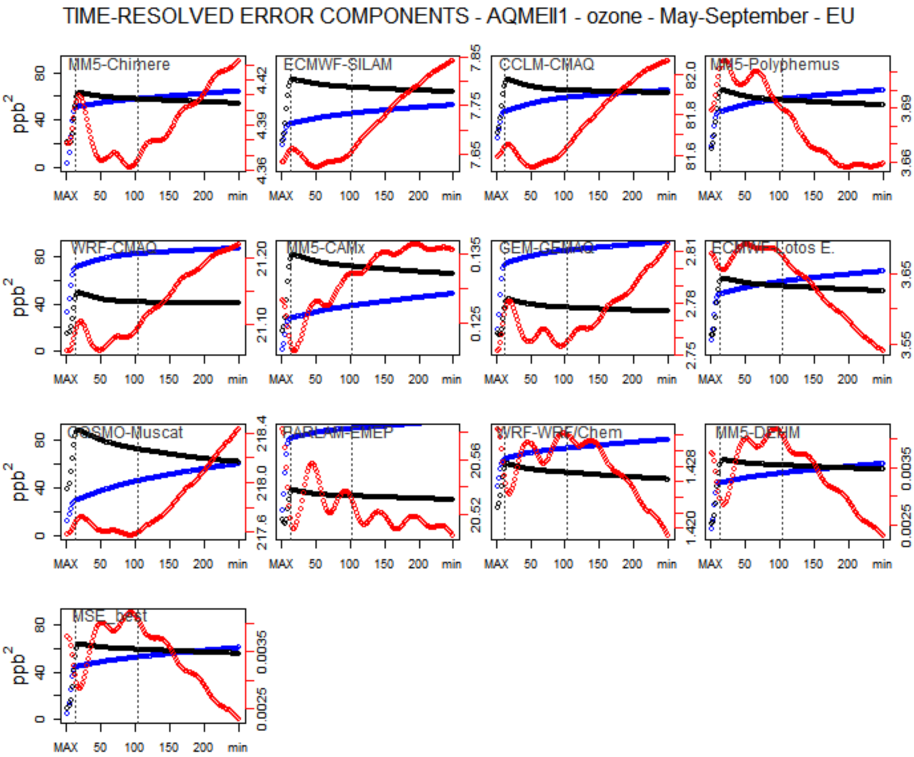
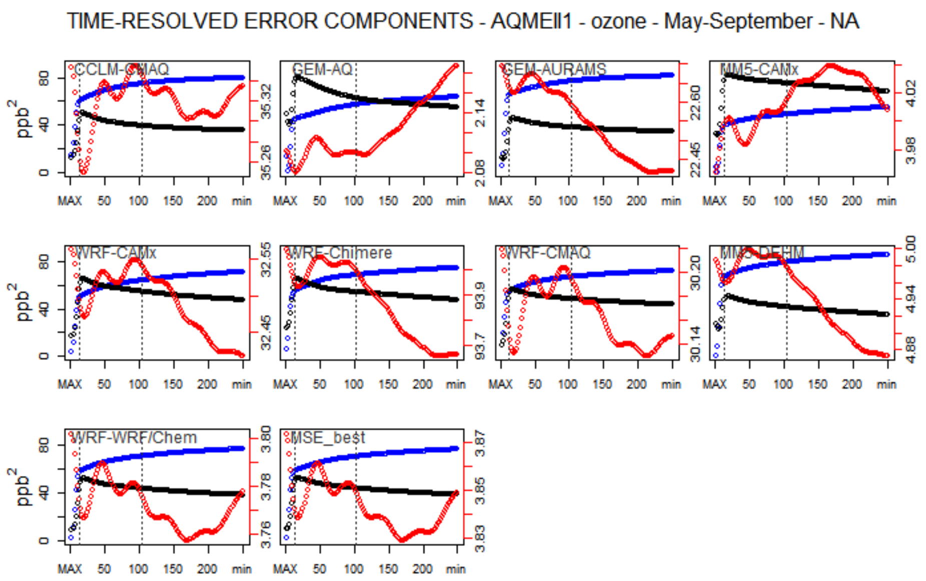


FIGURE 15 Example of the model complexity as time-resolved scale of the transport and dispersion processes: the minimum complexity (far right) is a poor time-resolving time series obtained as *kz(250,5)* (> 1 month). The complexity increases towards the left, with the scale of resolved processes becoming finer up to the maximum complexity (far left), which represents the full time series. The upper panel shows an example of how the curves of the error for covariance, variance and bias vary according to complexity.


TIME-RESOLVED ERROR COMPONENTS - AQMEII1 - ozone - May-September - EU
MM5-Chimere
ECMWF-SILAM
CCLM-CMAQ
MM5-Polyphemus
WRF-CMAQ
MM5-CAMx
GEM-CEMAQ
ECMWF-Lotos E.
COSMO-Muscat
HARLAM-EMEP
WRF-WRF/Chem
MM5-DEHM
MSE_best
ppb^2
0
40
80
MAX
50
100
150
200
min
4.36
4.39
4.42
7.65
7.75
7.85
81.6
81.8
82.0
3.66
3.69
21.10
21.20
0.125
0.135
2.75
2.78
2.81
3.55
3.65
217.6
218.0
218.4
20.52
20.56
1.420
1.428
0.0025
0.0035


TIME-RESOLVED ERROR COMPONENTS - AQMEII1 - ozone - May-September - NA
CCLM-CMAQ
GEM-AQ
GEM-AURAMS
MM5-CAMx
WRF-CAMx
WRF-Chimere
WRF-CMAQ
MM5-DEHM
WRF-WRF/Chem
MSE_best
ppb^2
0
40
80
MAX
50
100
150
200
min
35.26
35.32
2.08
2.14
22.45
22.60
3.98
4.02
32.45
32.55
93.7
93.9
30.14
30.20
4.88
4.94
5.00
3.76
3.78
3.80
3.83
3.85
3.87

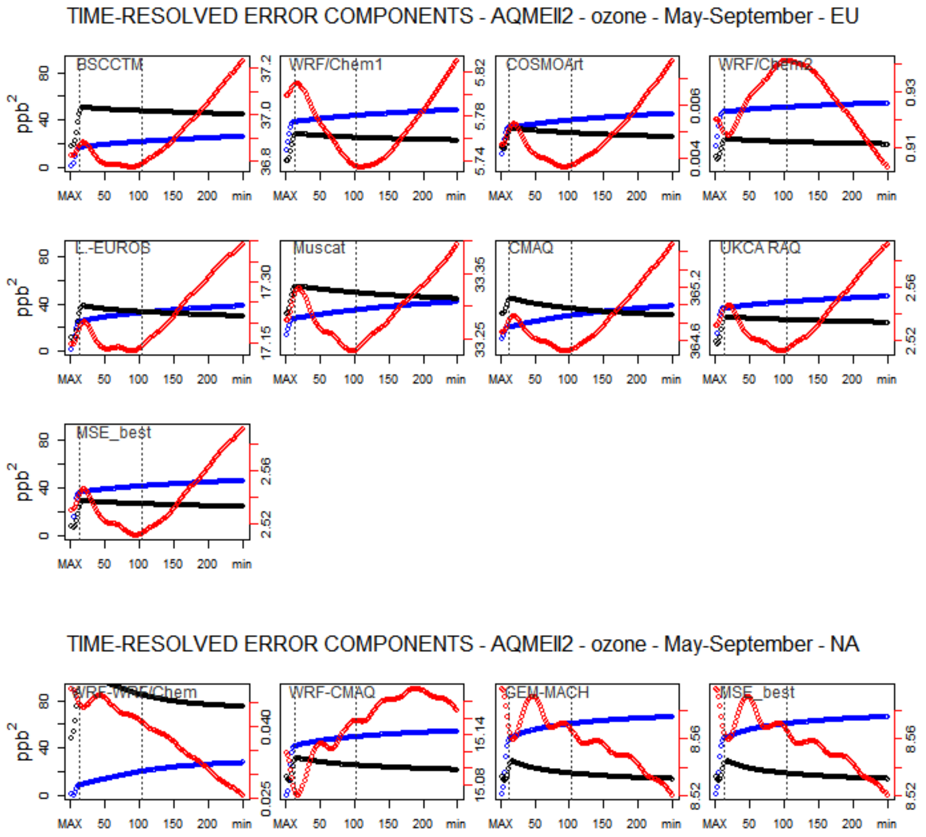


FIGURE 16 Evolution of error components (red: bias; Blue: variance; Black: covariance) as a function of model complexity. Complexity increases from right (min) to left (MAX) and is calculated as the temporal scale of the resolved process using the kz filter on the modelled signal: kz(i,5), i=2,…,250.