1 ERROR APPORTIONMENT FOR ATMOSPHERIC CHEMISTRY-TRANSPORT

2 MODELS. A NEW APPROACH TO MODEL EVALUATION

3 E. Solazzo, S. Galmarini

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

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

9 **Abstract.** In this study, methods are proposed to diagnose the causes of errors in air quality
10 (AQ) modelling systems. We investigate the deviation between modelled and observed time
11 series of surface ozone through a revised formulation for breaking down the mean square
12 error (MSE) into bias, variance, and the minimum achievable MSE ($mMSE$). The bias
13 measures the accuracy and implies the existence of systematic errors and poor
14 representation of data complexity, the variance measures the precision and provides an
15 estimate of the variability of the modelling results in relation to the observed data, and the
16 $mMSE$ reflects unsystematic errors and provides a measure of the associativity between the
17 modelled and the observed fields through the correlation coefficient. Each of the error
18 components is analysed independently and apportioned to resolved processes based on the
19 corresponding timescale (long scale, synoptic, diurnal, and intra-day) and as a function of
20 model complexity.

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

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

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

29 1. INTRODUCTION

30 Due to their use for regulatory applications and to support legislation, air quality (AQ)
31 models must model correctly and be correctly applied, justifying the need for a thorough
32 evaluation. A framework for the operational and scientific evaluation of geophysical models
33 was already envisaged in the early '80s (Fox, 1981; Wilmott et al., 1985), the former being '*a*
34 *comparison with data exclusively within a particular application context*', and the latter
35 defined as '*some understanding of cause-and-effect relationship that relies on testing model*
36 *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:

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

73 evaluation activities (AQMEII 1, 2 and 3) using operational (Solazzo et al., 2012a,b; Solazzo
74 et al., 2013a; Im et al., 2015a,b), probabilistic (Solazzo et al., 2013b; Kioutsioukis et al.,
75 2014), and diagnostic (Hogrefe et al., 2014; Makar et al., 2015) evaluation frameworks.

76 The study we present here follows and complements the previous investigations based on
77 the AQMEII models collected in the first and second phases of the activity (AQMEII1 in 2006
78 and AQMEII2 in 2010). The main aim is to introduce a novel method that combines
79 operational and diagnostic evaluations. This method helps apportion the model error to its
80 components, thereby identifying the space/timescale at which it is most relevant and, when
81 possible, to infer which process/es could have generated it. This work is designed to support
82 the analysis of the currently ongoing third phase of the AQMEII activity (Galmarini et al.,
83 2015).

84 2. MEAN SQUARE ERROR AS A COMPREHENSIVE METRIC

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

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

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

104 2. 1. ERROR DECOMPOSITION
105 The MSE is the squared difference of the modelled (*mod*) and observed (*obs*) values:

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

EQ 1

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

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

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

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

108

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

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

111

112   the final formulation for the MSE components reads:

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

113

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

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

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

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

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

126

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

133   The bias measures the departure of the modelled from the observed results, and is a
134   measure of systematic error, since it measures the extent to which the average modelled
135   values deviate from the observed ones. The bias is commonly used to express the degree of
136   'trueness', i.e. "the closeness of agreement between the average value obtained from a

137   large series of measurements and the true value" (Johnson, 2008). The variance shows
138   whether the modelled variability is compatible with that observed. Finally, the covariance
139   term represents the unexplained proportion of the MSE due to the remaining unsystematic
140   errors, i.e. it represents the remaining error after deviations from the mean values have
141   been accounted for. This latter term is a measure of the lack of correlation of the model
142   with comparable observations, and is considered the least 'worrisome' portion of the error
143   (Pindick and Rubinfeld, 1998).

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

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

146   i.e. the best agreement between modelled and observed values is achieved by:
147

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

148

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

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

151

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

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

156

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

162   Ideally, the entire error should be attributable to unsystematic fluctuations. From a model
163   development perspective, the variance and covariance are possibly more revealing of model
164   deficiencies than is the bias term, as they are produced by the AQ model itself, while the
165   bias is also due to external sources (e.g. emissions, boundary conditions). From the

166 application viewpoint, however, it is the overall error that counts, which is mostly made up
167 of the bias.

168 ## 2.2. Spectral decomposition of modelled and observed time series
169 Hourly time series of (modelled and observed) ozone concentrations have been
170 decomposed using an iterative moving average approach known as the Kolmogorov-
171 Zurbenko (kz) low-pass filter (Zurbenko, 1986), whose applications to ozone are vastly
172 documented in the literature (Rao et al., 1997; Wise and Comrie, 2005; Hogrefe et al., 2000
173 and 2014; Galmarini et al., 2013; Kang et al., 2013; Solazzo and Galmarini, 2015). The kz
174 filter depends on two parameters: the length of the moving average window m and the
175 number of iterations k ($kz_{m,k}$). Since the kz is a low-pass filter, the filtered time series
176 consists of the low-frequency fluctuating component, while the difference between two
177 filtered time series provides a band-pass filter. This latter property is used to decompose the
178 ozone concentration time series as:

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

179

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

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

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

198 - The DU component accounts for more than half of the total variance, followed by
199   the LT and SY components;

200 - The ID component has the smallest influence due to the small amplitude of its
201   fluctuations;

202 - The variance of the spectral component is neither strongly nor systematically
203   associated with the area-type of the monitoring stations (i.e. rural, urban, suburban);

204 - Due to the bias, most of the error is accounted for by the LT component, followed by
205   the DU component. The ID contributes very little to the overall MSE.

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

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

213 3. DATA AND MODELS USED

214 The observational dataset derived from the surface AQ monitoring networks operating in
215 the EU and NA constitutes the same dataset used in the first and second phases of AQMEII
216 to support model evaluation. Only stations with over 75% valid records for the whole
217 periods and located at altitudes below 1 000 m have been used for this analysis. Details of
218 the modelled regions and number of receptor stations are reported in Table 1.

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

224 There are profound differences between the modelling systems that participated in
225 AQMEII1 and AQMEII2. The two sets of models have been applied to different years (2006
226 for phase 1 and 2010 for phase 2) and are therefore dissimilar with respect to the input data
227 of emissions and boundary conditions for chemistry. The AQ models of the second phase
228 are coupled (online chemistry feedbacks on meteorology), while those of the first phase are
229 not. The effect of using online models for simulating ozone accounts for the impact of
230 aerosols on radiation and therefore on temperature and photolysis rates (Baklanov et al.,
231 2014).

232 The model settings and input data for phase I are described in Solazzo et al. (2012a, b;
233 2013a), Schere et al. (2012), and Pouliot et al. (2012); for phase II, similar information is
234 presented in Im et al. (2015a, b), Brunner et al. (2015), and Pouliot et al. (2015).

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

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. The spatial average is carried out prior to the spectral decomposition and error apportionment.

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.

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

276 The results of Figure 1 help identify the time-scales and associated processes for which the
277 largest improvement in model accuracy can be achieved. The LT component has the largest
278 share of the error due to the bias (error breakdown is discussed in the next section), but
279 'internal' chemical processes, transport, and deposition also occur at this timescale. Diurnal
280 processes are the second largest source of error, including, among others, chemistry,
281 boundary layer dynamics, radiation forcing, and their interactions. The processes in the SY
282 band bridge meteorological and chemical processes, and discern between the fast-acting
283 diurnal processes and the baseline. As such, although the SY signal is not as strong as that of
284 the DU components (variance of SY is comparable to the variance of ID, see Hogrefe et al.,
285 2014), it accounts for a significant portion of the total error, as discussed next.

286 ## 4.2 THE QUALITY OF THE ERROR: ERROR APPORTIONMENT
287 The error breakdown (Eq 9) of each spectral component complements the analysis
288 presented in the previous section, and is reported in Figure 2 (please note that results in
289 Figure 2 are reported in ppb^2 for reason of clarity). The bias (only included in the LT
290 component) is the average amount by which the modelled time series is displaced with
291 respect to the observed time series, and is the main source of error. The bias can be either
292 due to 'internal' model errors, or inherited from external drivers (emissions, meteorology,
293 boundary conditions). Based on the experience matured within AQMEII, while the internal
294 model errors are of interest for model development because they are generated by
295 systematic modelling errors, the bias introduced by external drivers is responsible for the
296 largest share of modelling errors.

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

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

320 The $mMSE$ error of the LT component outweighs the variance error in most cases (in both
321 the EU and NA), and is due to the unexplained portion of observed variance.. The processes
322 responsible for the $mMSE$ error of the LT component (such as deposition, transport,
323 stratospheric mixing and photochemistry) act at timescales of more than 21 days.

324 The DU error (on average 3-4 ppb for AQMEII1 and 2-3 ppb for AQMEII2) makes up the
325 second highest contribution to the total error. The partitioning between variance and the
326 $mMSE$ error varies greatly from model to model. However, a comparison of the two AQMEII
327 phases shows that the $mMSE$ is predominant for AQMEII2, while the variance error
328 (typically due to model under-prediction of the observed variability) is most relevant in
329 several cases of AQMEII1. Therefore, at the DU scale, the 'quality' of the error of the
330 AQMEII2 phase is higher than that of its AQMEII1 counterpart. One possible explanation is
331 the fact that coupled models were used in AQMEII2, while AQMEII1 exclusively used non-
332 coupled models. As already mentioned (end of section 3), Makar et al. (2015) found that
333 AQMEII1 models performed better overall with respect to AQMEII2. An analysis of the LT
334 component showed that the bias in the AQMEII2 models is higher, possibly due to the 2010
335 emission inventory, while an analysis of the DU error found that the variance error in the
336 AQMEII2 models is significantly reduced with respect to the AQMEII1 models, and is almost
337 null. We postulate that the inclusion of feedback effects may have been beneficial, and that
338 the reduced performance of AQMEII2 models is likely due to external bias. The residual
339 $mMSE$ error of the DU component (~1-2 ppb on average for both continents) is mostly likely
340 generated by a number of processes, including chemistry, cloudiness, boundary layer
341 transition and vertical mixing. From Figure 2, the values of the correlation coefficient for the
342 DU component are very high (exceeding 0.8 in the majority of the cases). Such a high
343 performance can be misleadingly optimistic though, because it mostly reflects the 24-hour
344 and annual forcing embedded in both the observations and model values. Further analysis
345 on the amplitude and phase of the error can be more informative.

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

350   Finally, the error of the ID component is less than 1 ppb (on average ~0.2 ppb for AQMEII2)
351   and is generated by both variance (most commonly model over-prediction) and *mMSE*. The
352   fast-acting photochemical processes are, therefore, modelled with satisfactory precision, ,
353   although the small errors in the ID component can be quite large relative to the total
354   amount of ID variability. Furthermore, the spatial averaging of the time series prior to the
355   spectral decomposition might have suppressed the variance of the ID component which is
356   more uncorrelated in space than the other components and would therefore tend to
357   average out.

4.3. SPATIAL DISTRIBUTION OF THE SPECTRAL ERROR COMPONENTS

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

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

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

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

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

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 analyse 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 what type of error) and, if it exists, the time window where the error decreases. The information obtained is of immediate usefulness for model development, as it provides a clear temporal cut-off that distinguishes the dynamics of the error.

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).

425   In Figure 8, we report the spatially averaged curves of bias, variance, and covariance according

426   to Eq 6 as a function of model complexity. According to the regression analysis theories

427   outlined above, we would expect the variance to increase according to the complexity

428   ($\frac{d\sigma_m^2}{dcomplexity} > 0$), and the distance between the modelled and observed variance to

429   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

430   indeed turn downwards as predicted by the theory, while the curves of bias have a mixed

431   behaviour but are, basically, constant $\left(\frac{d(\overline{mod} - \overline{obs})^2}{dcomplexity} \approx 0\right)$.

432   More specifically:

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

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

448   Summarising, in most cases (both continents, both AQMEII phases), the $(\sigma_m - \sigma_o)^2$ term
449   decreases sharply after a timescale of resolved processes of ~1 day; the bias term is
450   surprisingly independent on complexity; the covariance is complementary to the variance.
451   Thus, the bias is the error term to which efforts should be expanded first, and current
452   studies are carried out to diagnose more precisely its origin within AQ modelling systems.

453   ## 5. CONCLUSIONS

454   This study presents a novel approach to model evaluation, and aims to combine standard

455   operational statistics with the time allocation of the component error. The methodology we

456   propose tackles the issue of diagnostic evaluation from the angle of the spectral

457   decomposition and error breakdown of model/data signals, introducing a compact operator

458   for the quantification of bias, variance, and the correlation coefficient.

459   When the analytical decomposition of the error into bias, variance and *mMSE* is applied to

460   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.

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

$$ID(t) = \mathbf{x}(t) - kz_{3,3}(\mathbf{x}(t))$$
$$DU(t) = kz_{3,3}(\mathbf{x}(t)) - kz_{13,5}(\mathbf{x}(t))$$
$$SY(t) = kz_{13,5}(\mathbf{x}(t)) - kz_{103,5}(\mathbf{x}(t))$$
$$LT(t) = kz_{103,5}(\mathbf{x}(t))$$
$$\mathbf{x}(t) = ID(t) + DU(t) + SY(t) + LT(t)$$

EQ. S.1

503 where $\mathbf{x}(t)$ is the time series vector.

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

510
511 Assuming a spectral decomposition which is valid for the modelling and the observational
512 time series, the MSE formulation outlined in Galmarini et al. (2013) holds:

$$MSE(O_3) = MSE(LT + SY + DU + ID) = \sum MSE(spec\ comp) + \sum MSE\ (cc))$$

EQ. S.2

513

514 Where *spec comp* are the diagonal terms, and *LT, SY, DU, ID* and cc identifies the cross
515 components, i.e. the off-diagonal terms deriving from the squared nature of the MSE:
516 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,
517 DU_oID_o (o and m represent observed and modelled fields, respectively). For simplicity, the
518 cross-components are assumed to be symmetric, so the o and m subscripts are dropped.
519 This simplification has little impact on the MSE breakdown since, as shown by Galmarini et
520 al. (2013), the diagonal terms alone account for over 80% of the total variance.

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

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

EQ. S.3

523

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

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

525

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

528

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

529

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

532

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

536

537

538

539

540

541

542

543

544

545

546

547

548

549 REFERENCES

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

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

559 Dennis, R., Fox, T., Fuentes, M., Gilliland, A., Hanna, S., Hogrefe, C., Irwin, J., Rao, S.T., Scheffe, R., Schere, K.,
560 Steyn, D., Venkatram, A., 2010. A framework for evaluating regional-scale numerical photochemical
561 modeling systems. Environ. Fluid Mech. (Dordr.) 10, 471-489. http://dx.doi.org/10.1007/s10652-009-
562 9163e2.

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

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

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

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

573 Hastie, T., Tibshirani, R., Friedman, J., 2009. The elements of statistical learning (2nd edition). Springer-Verlag.
574 763 pages.

575 Hogrefe, C., Rao, S.T., Zurbenko, I.G., Porter, P.S., 2000. Interpreting the information in ozone observations and
576 model predictions relevant to regulatory policies in the Eastern United States. Bull. Am. Meteorol. Soc.
577 81, 2083e2106. http:// dx.doi.org/10.1175/1520-0477(2000)0812.3.CO;2.

578 Hogrefe, C., Roselle, S., Mathur, R., Rao, S.T., Galmarini, S., 2014. Space-time analysis of the Air Quality Model
579 Evaluation International Initiative (AQMEII) phase 1 air quality simulation. J. Air Waste Manag. Assoc. 64,
580 388-405.

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

586 Im, U., Bianconi, R., Solazzo, E., Kioutsioukis, I., Badia, A., Balzarini, A., Baro, R., Bellasio, R., Brunner, D.,
587 Chemel, C., Curci, G., Flemming, J., Forkel, R., Giordano, L., Jimenez-Guerrero, P., Hirtl, M., Hodzic, A.,
588 Honzak, L., Jorba, O., Knote, C., Kuenen, J. P., et al., 2015b. Evaluation of operational on-line-coupled

589   regional air quality models over Europe and North America in the context of AQMEII phase 2. Part I:
590   ozone. Atmos. Environ. 115, 404-420

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

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

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

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

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

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

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

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

610 Rao, K.S., 2005. Uncertainty analysis in atmospheric dispersion modelling. Pure and Applied Geophysics 162,
611   1893-1917.

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

614 Rao, S.T., Zurbenko, I.G., Neagu, R., Porter, P.S., Ku, J.Y., Henry, R.F., 1997. Space and time scales in ambient
615   ozone data. Bull. Am. Meteorol. Soc. 78, 2153e2166. http://dx.doi.org/10.1175/1520-
616   0477(1997)0782.0.CO;2.

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

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

626 Solazzo, E., Bianconi, R., Pirovano, G., Matthias, V., Vautard, R., Moran, M.D., Appel, K.W., Bessagnet, B.,
627   Brandt, J., Christensen, J.H., Chemel, C., Coll, I., Ferreira, J., Forkel, R., Francis, X.V., Grell, G., Grossi, P.,

628   Hansen, A.B., Hogrefe, C., Miranda, A.I., Nopmongco, U., Prank, M., Sartelet, K.N., Schaap, M., Silver, J.D.,
629   Sokhi, R.S., Vira, J., Werhahn, J., Wolke, R., Yarwood, G., Zhang, J., Rao, S.T., Galmarini, S., 2012b.
630   Operational model evaluation for particulate matter in Europe and North America. Atmos. Environ. 53,
631   75-92.

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

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

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

641 Stoeckenius, T.E., Hogrefe, C., Zagunis, J., Sturtz, T.M., Wells, B., Sakulyanontvittaya, T., 2015. A comparison
642   between 2010 and 2006 air quality and meteorological conditions, and emissions and boundary
643   conditions used in simulations of the AQMEII2 North American domain. Atmospheric Environment, 115,
644   389-403.

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

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

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

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

651

652

653

654

655

656

657

658

659

660

661

662

663   FIGURES

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

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

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

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

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

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

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

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

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

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

697   **Figure S3**. As in Figure S2 for the AQMEII2 models

698

699

700

701

702

703 TABLES

704 **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

705

706

707

708

709

710

711

712

713

714

715

716

717

718

719

720

721

722

723

724

725

726

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

730 § Standard anthropogenic emissions and biogenic emissions derived from meteorology (temperature and solar radiation) and land use
731 distribution implemented in the meteorological driver.
732 *Standard anthropogenic inventory but independent emission processing, exclusion of wildfires, and different versions of BEIS(v3.09)
733 used.
734 Refer to Solazzo et al. (2012a-b) and references therein for details.
735
736 *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

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

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

747

748

749

750

751

752

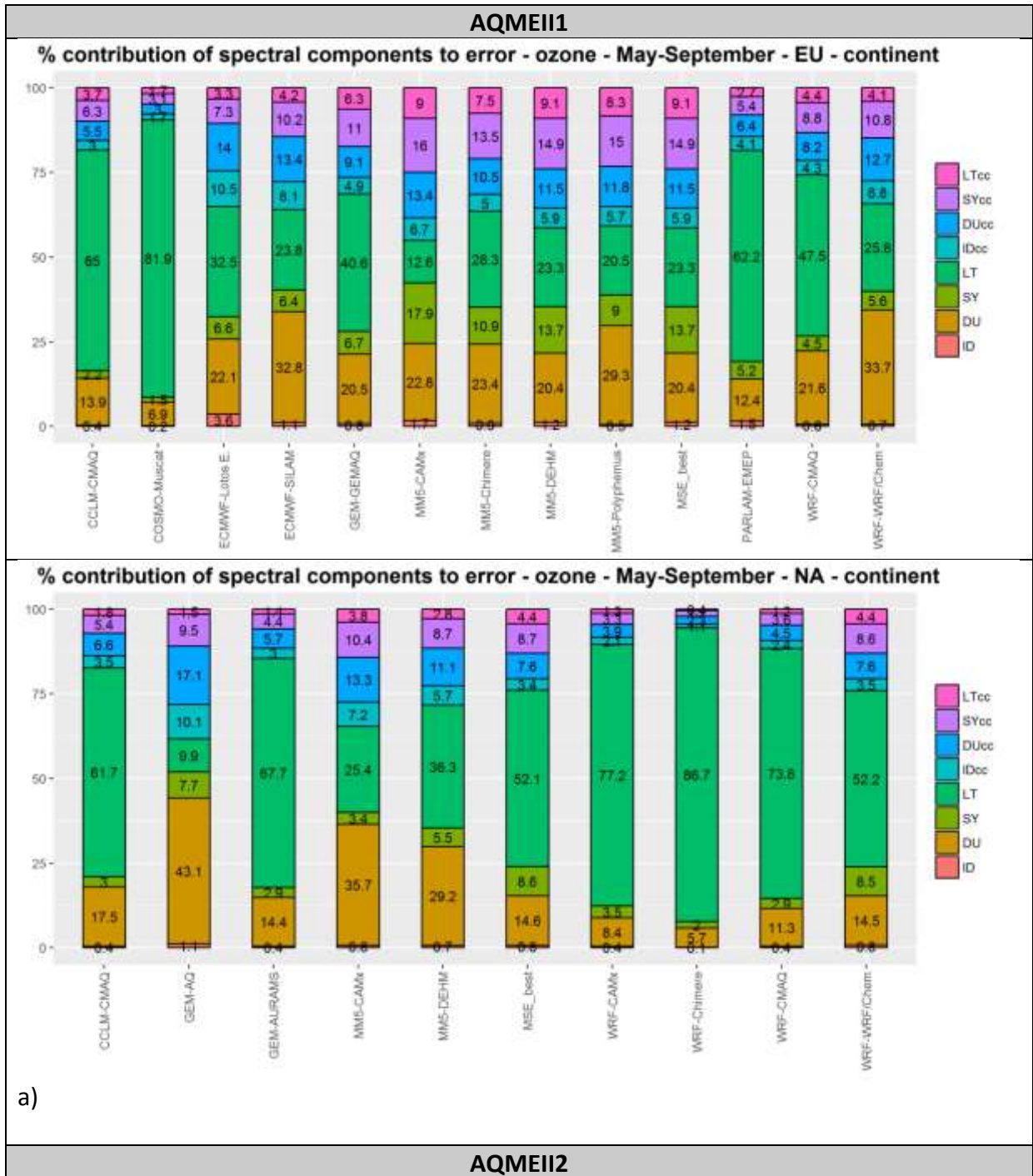
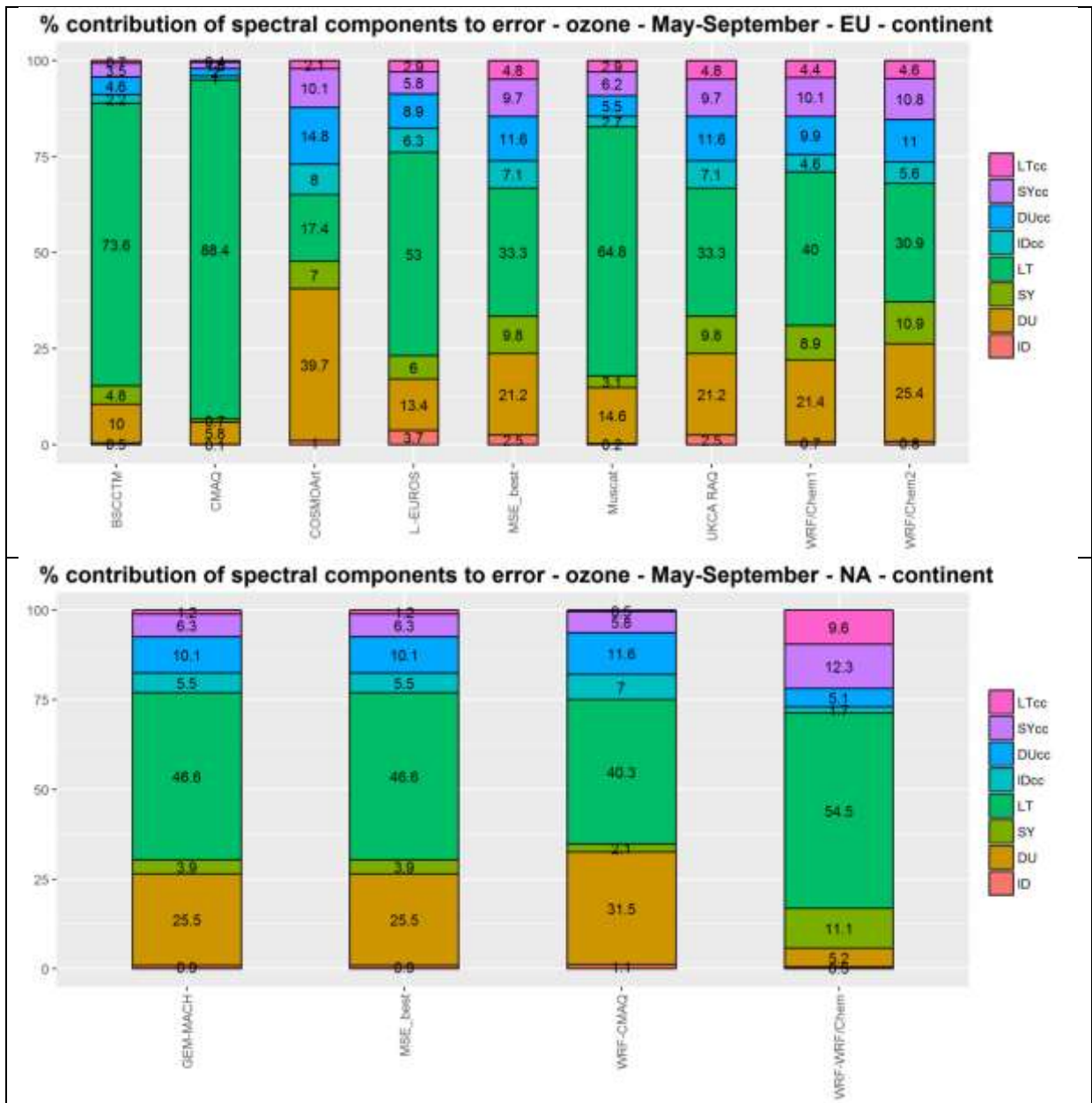
753

755

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.

756

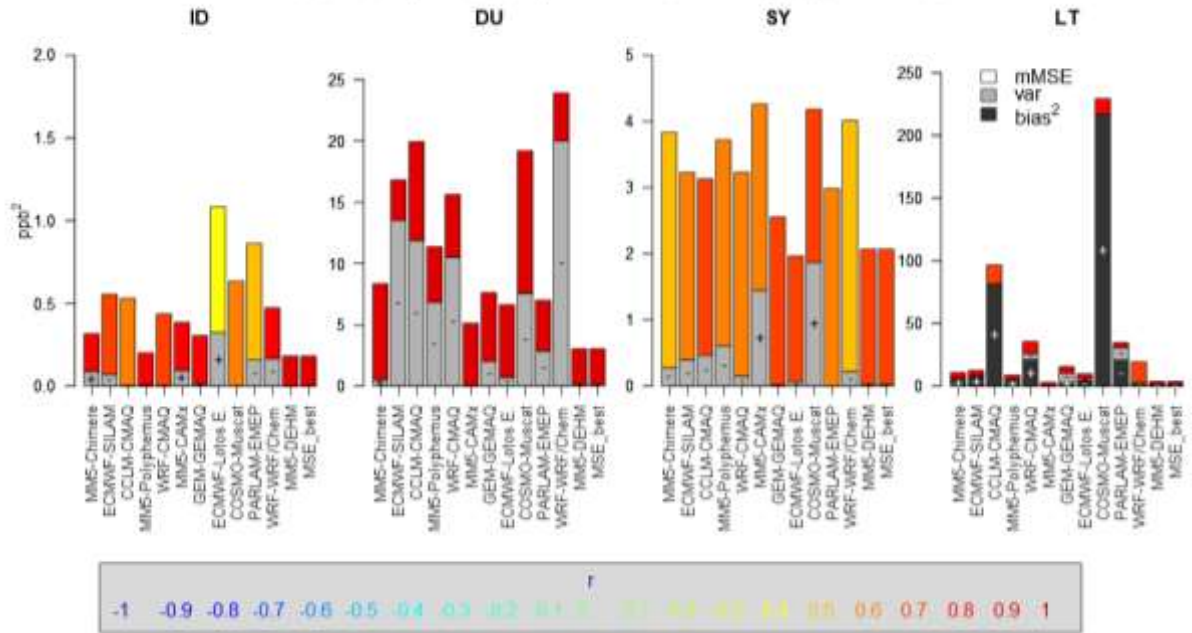
757

758

759

760

761

a)

AQMEII2

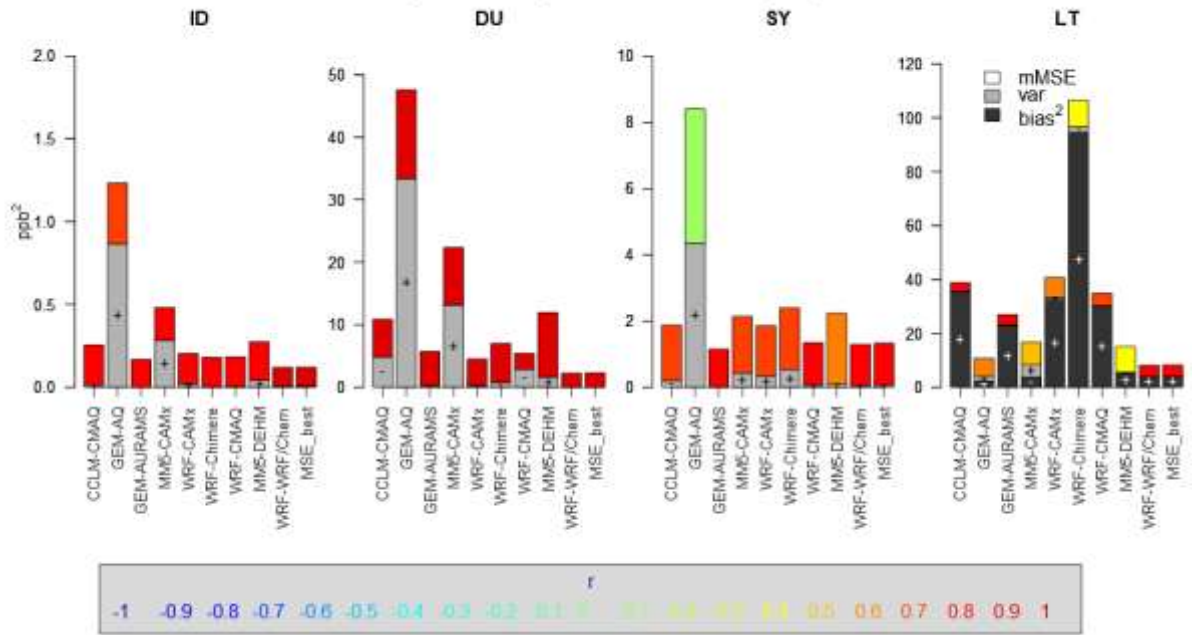
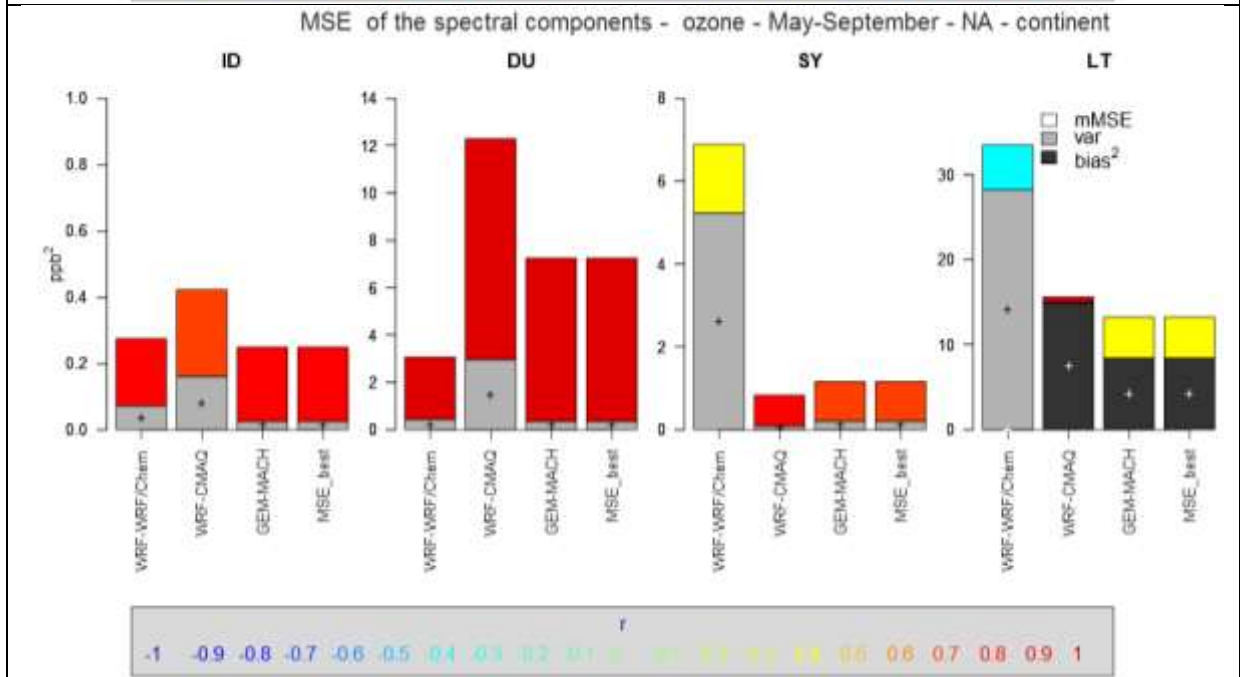
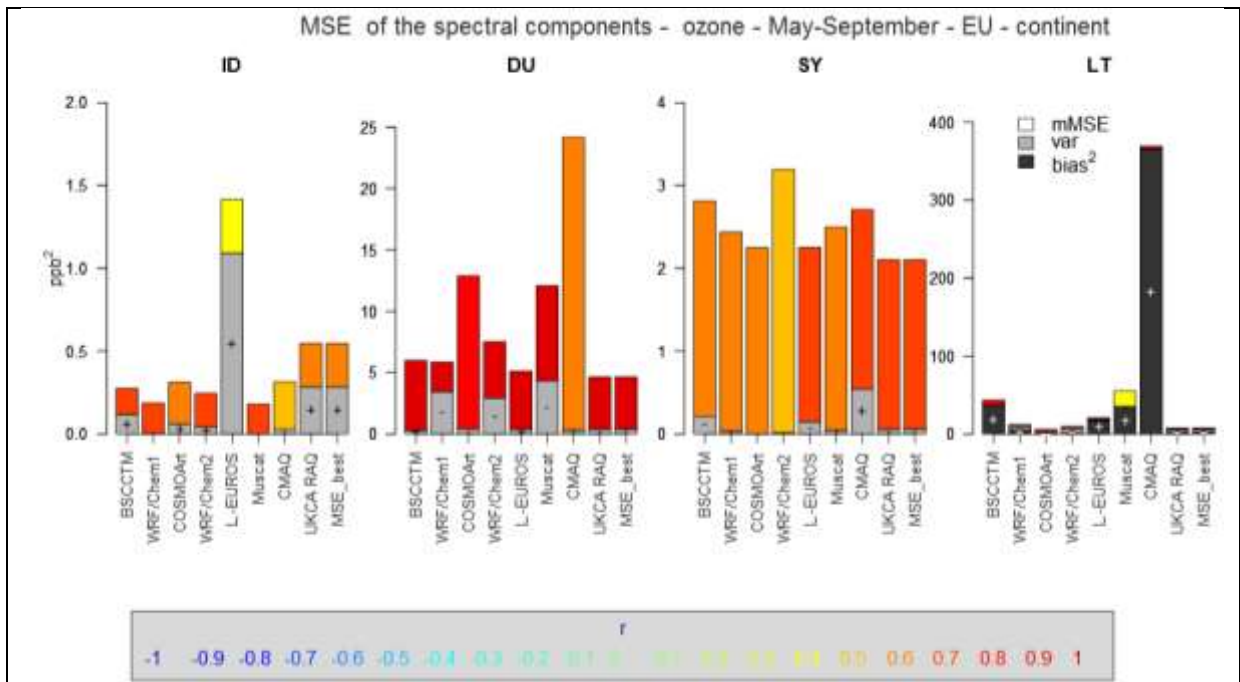
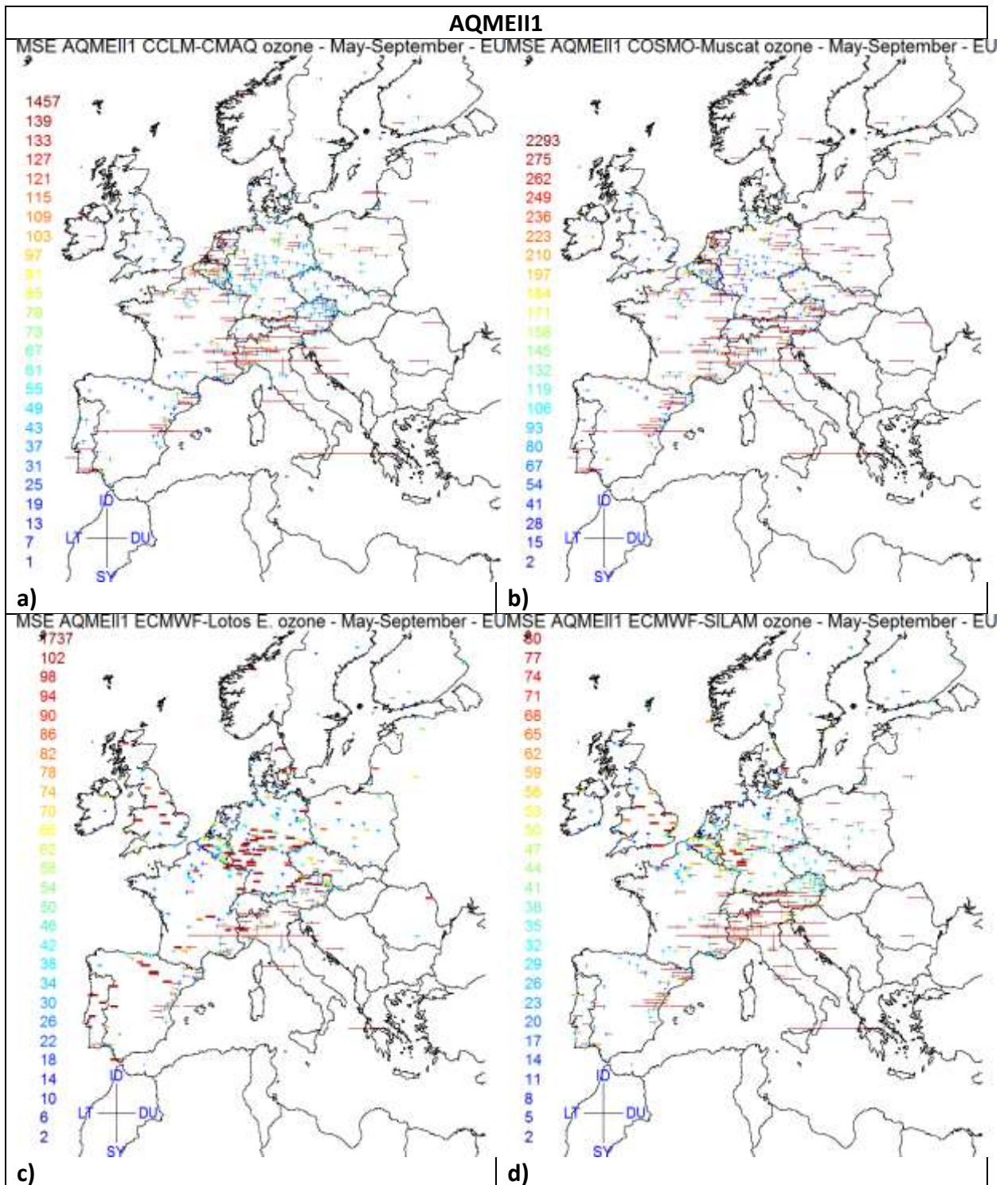
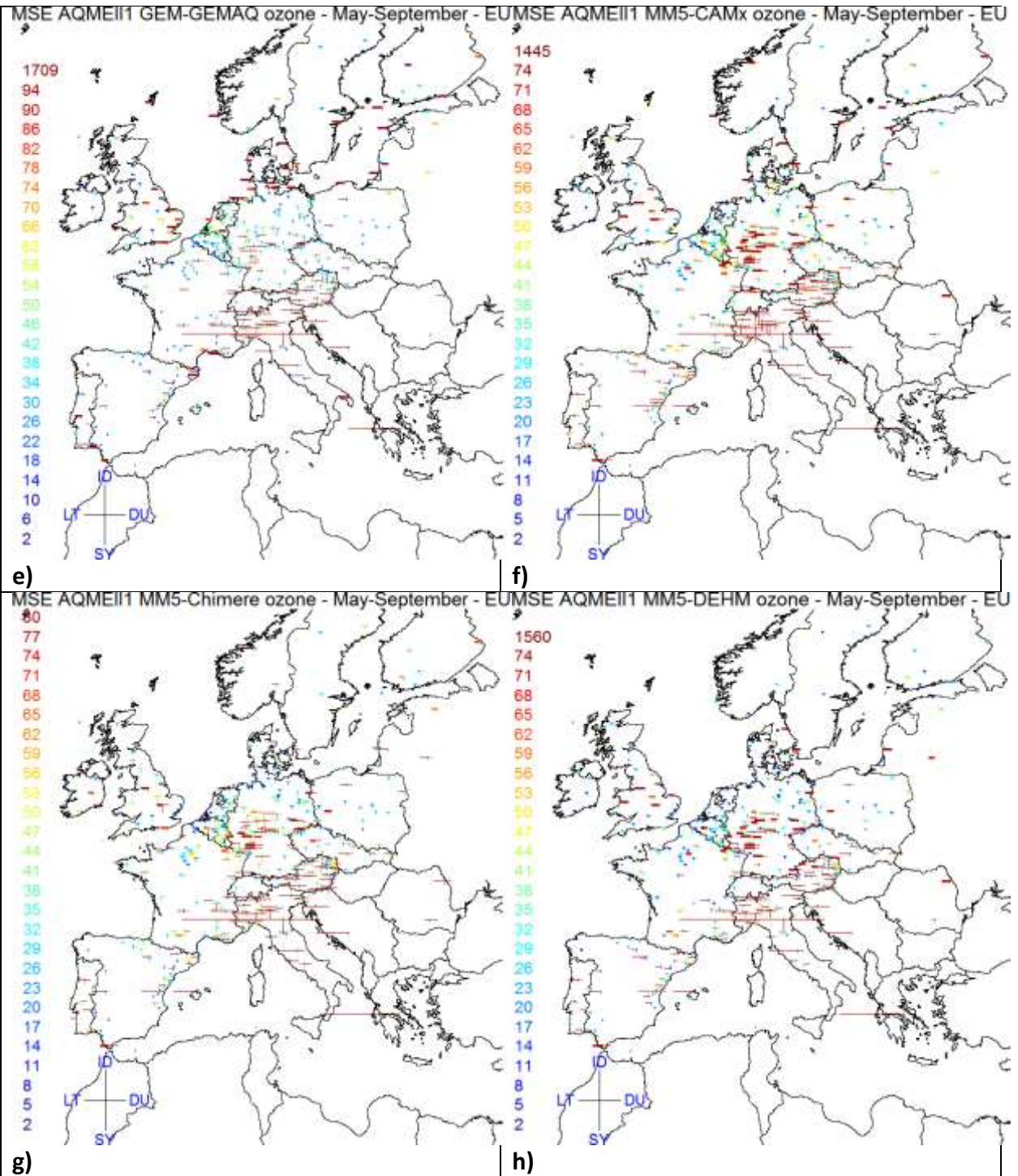
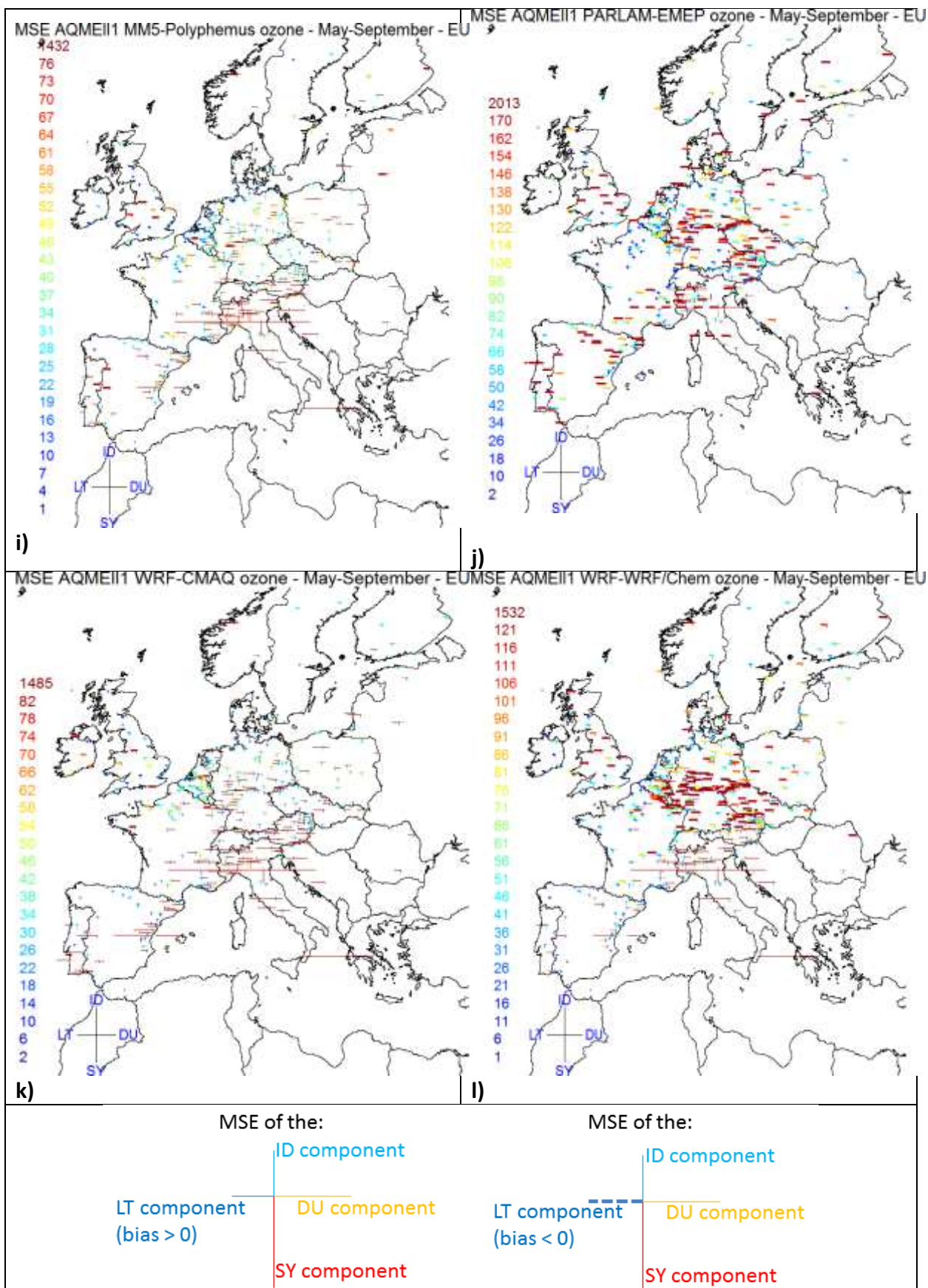



Figure 2. MSE (ppb^2) breakdown in bias squared, variance and *mMSE* of the spectral components ID, DU, SY, LT, based on Eq 9. The bias is entirely accounted for by the LT component. The sign within the share of bias and variance indicates model overestimation (+) or underestimation (-) of mean concentration (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.

762

763

764

AQMEII1

MSE AQMEII1 CCLM-CMAQ ozone - May-September - EU   MSE AQMEII1 COSMO-Muscat ozone - May-September - EU

a)   b)

MSE AQMEII1 ECMWF-Lotos E. ozone - May-September - EU   MSE AQMEII1 ECMWF-SILAM ozone - May-September - EU

c)   d)

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.

766

a)

b)

c)

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

d)

e)

g)

h)

i)

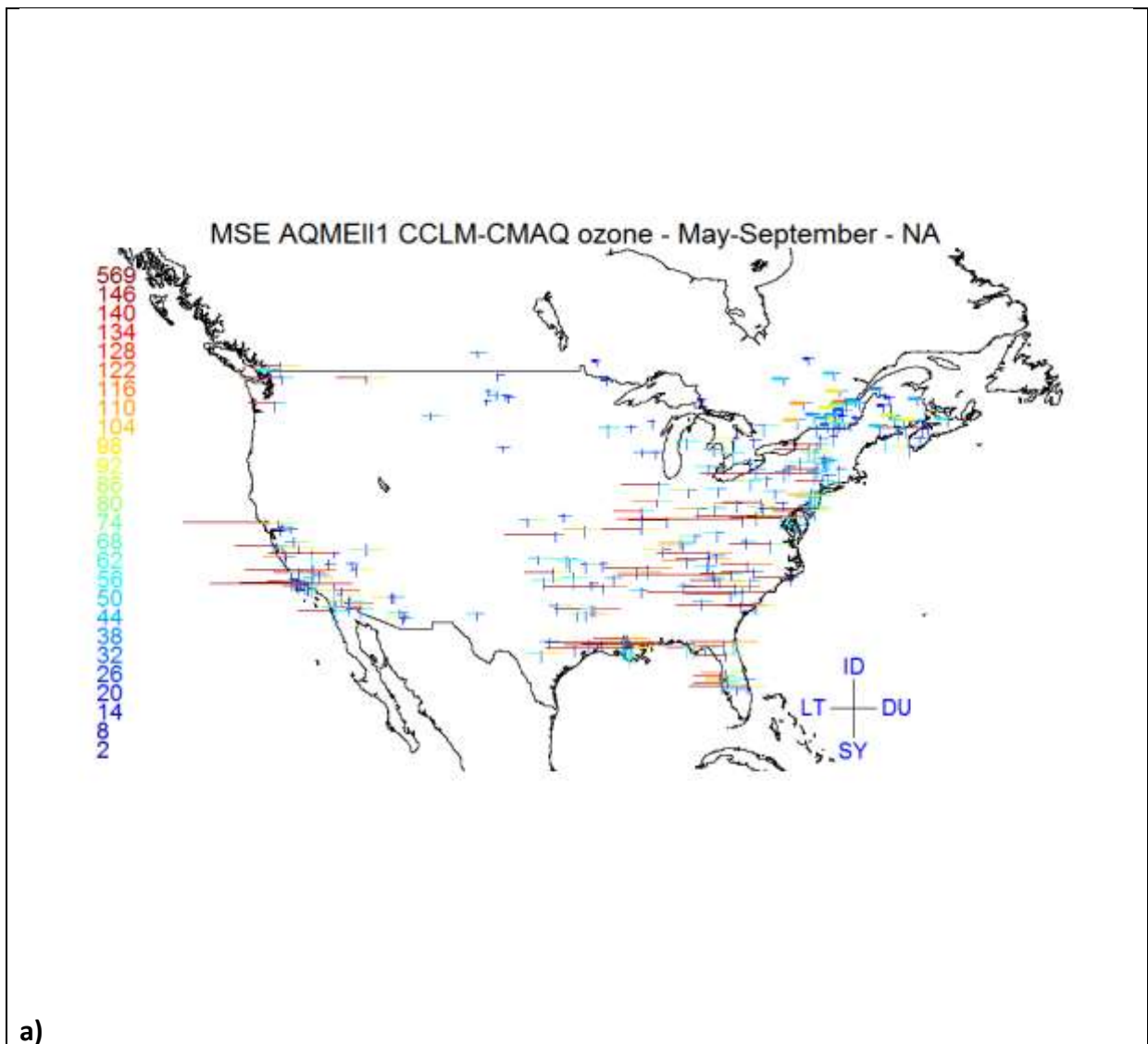
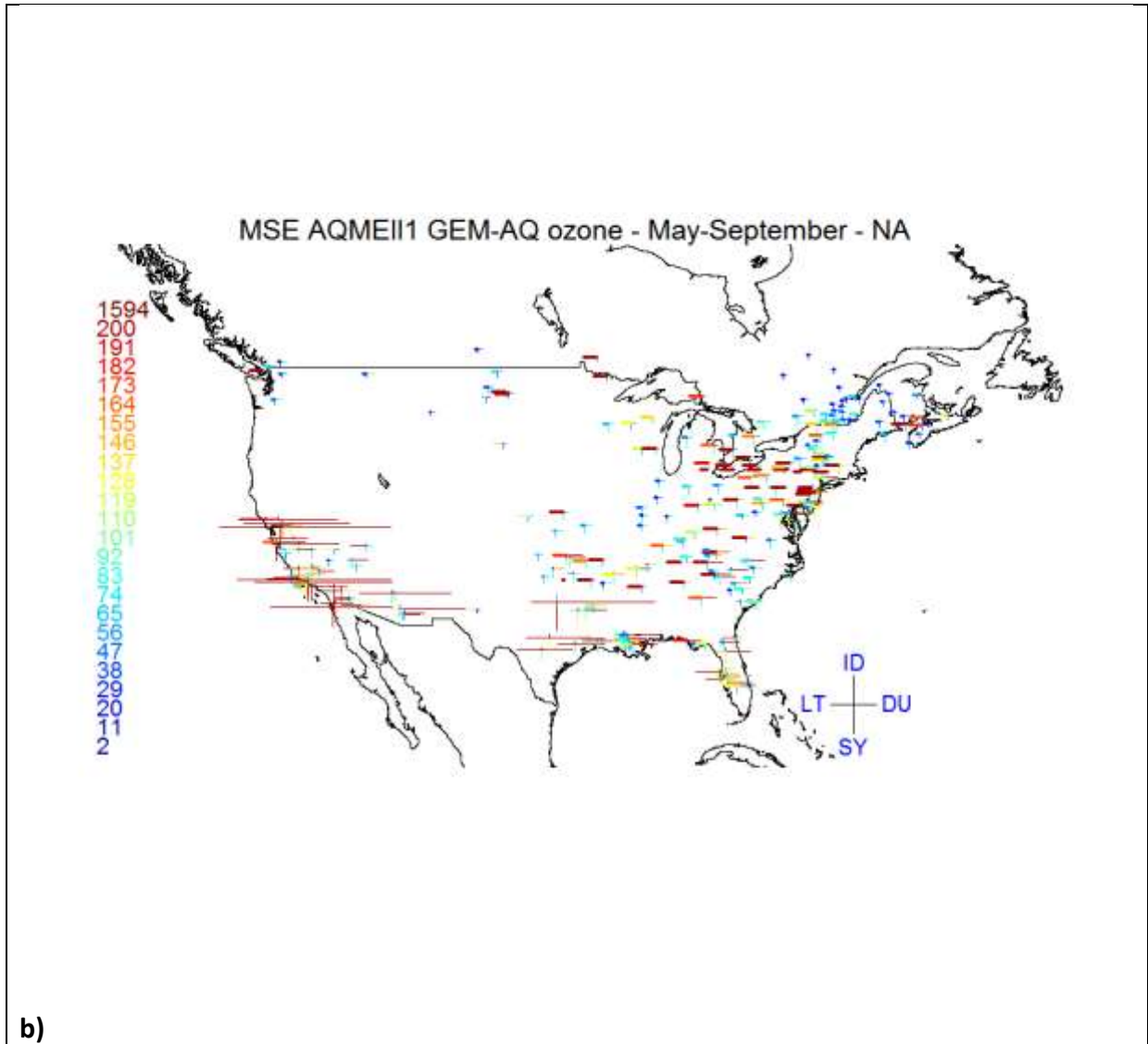
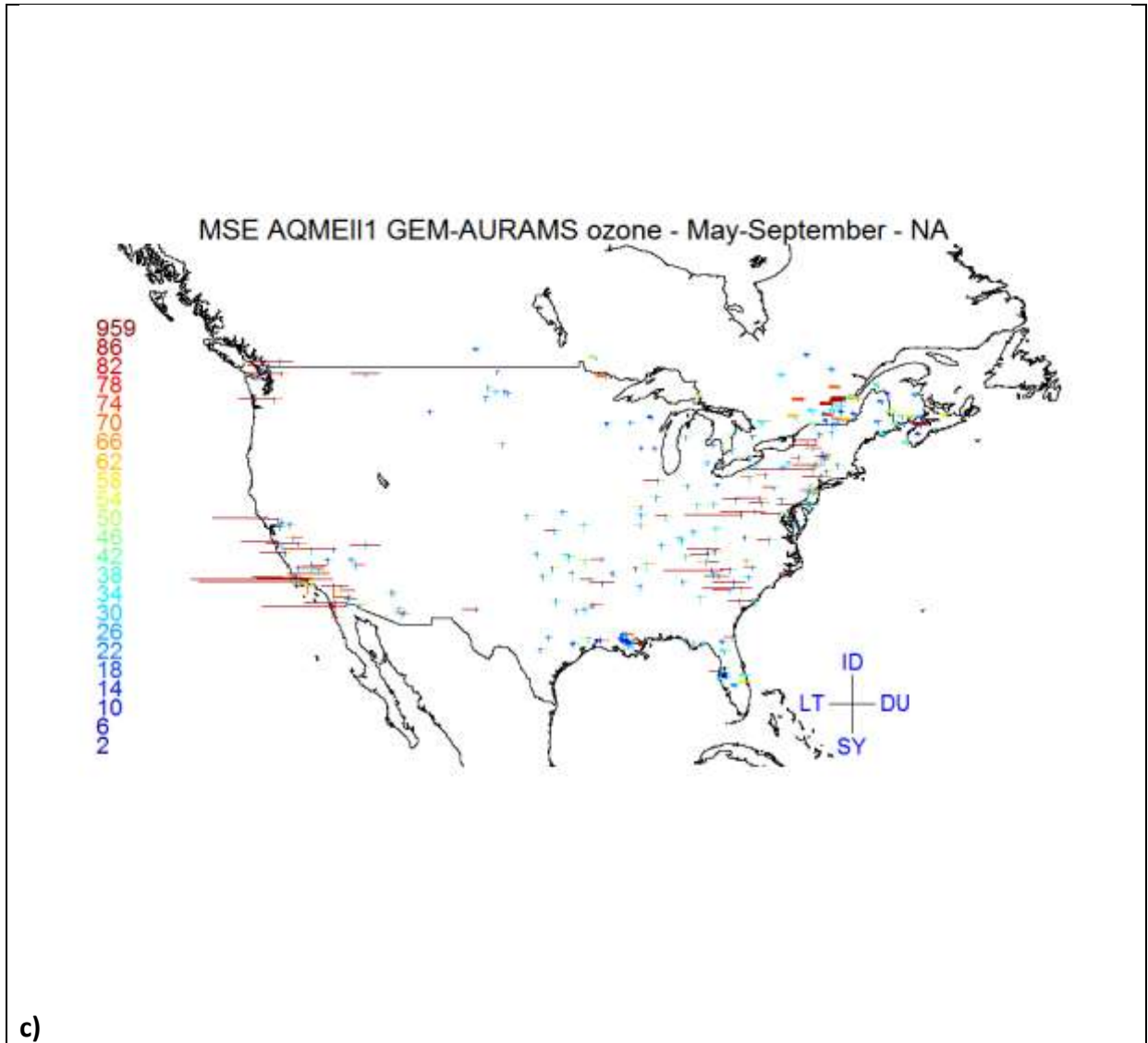
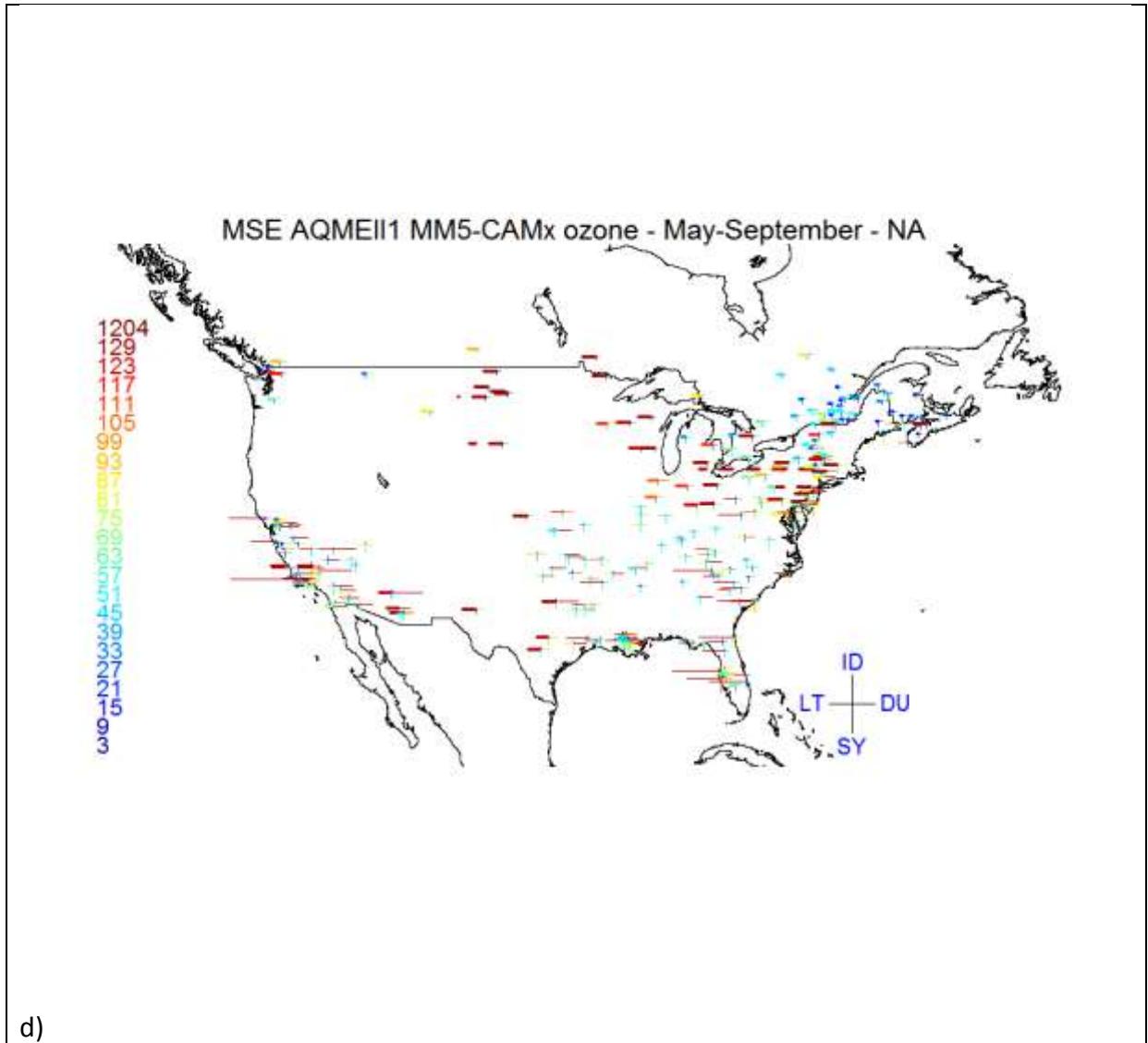
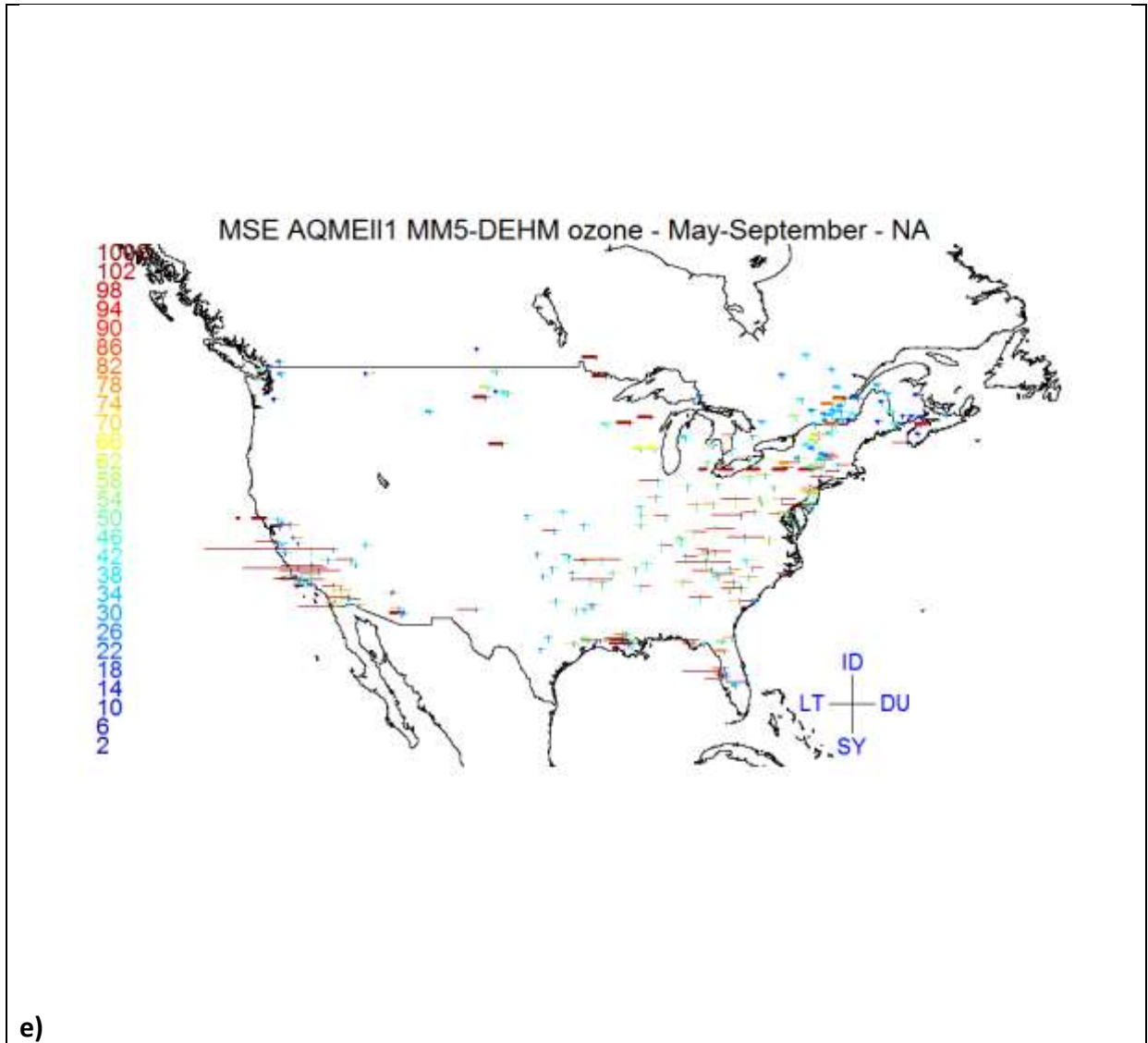
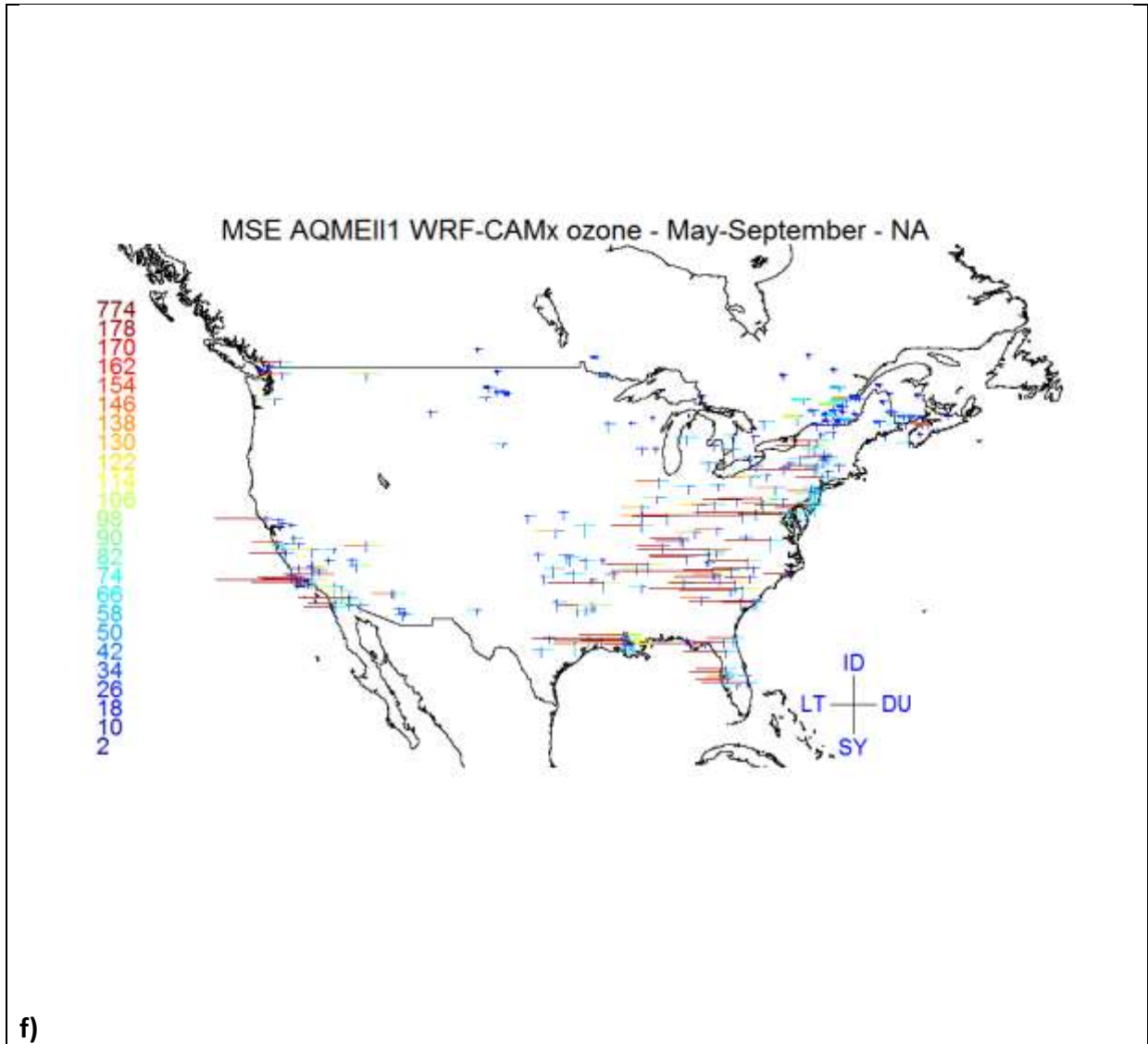
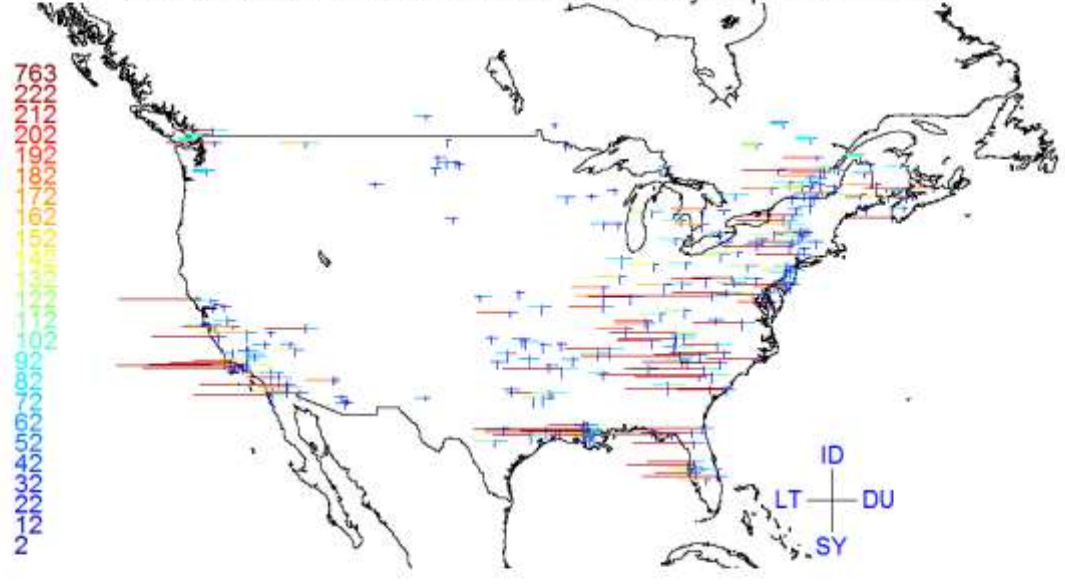
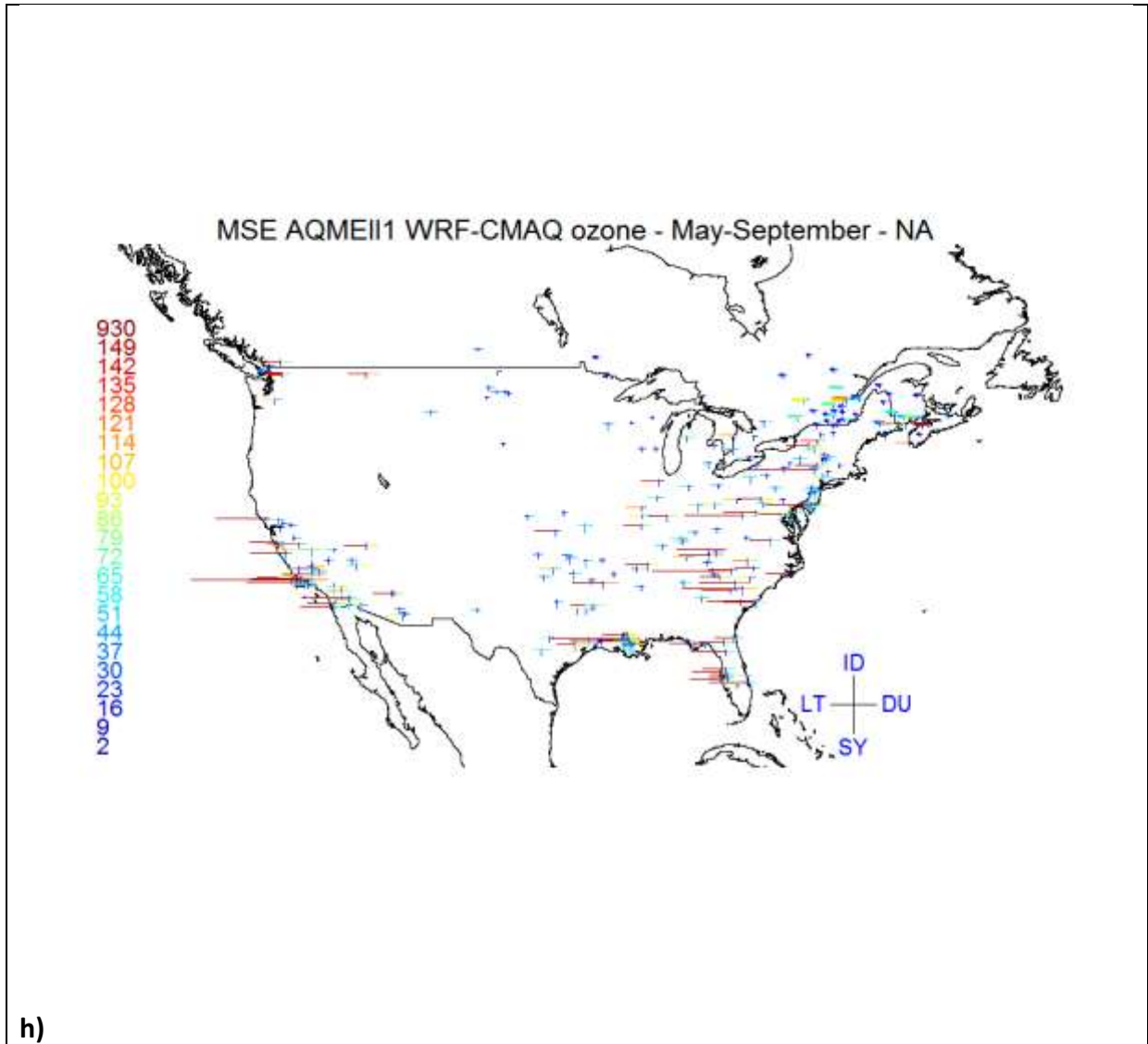
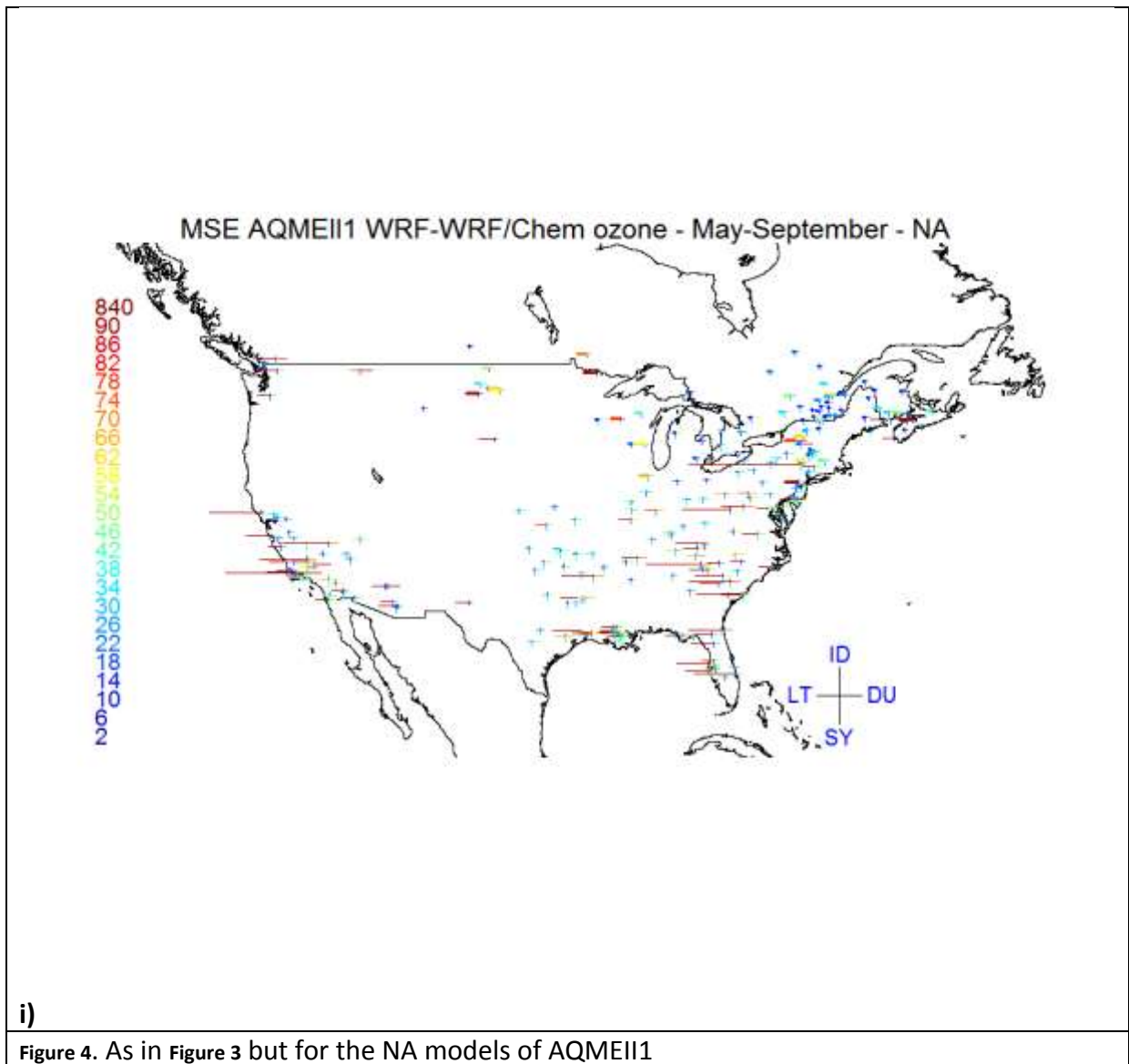



Figure 4. As in **Figure 3** but for the NA models of AQMEII1

767

768

769

770

771

772

773

774

775

776

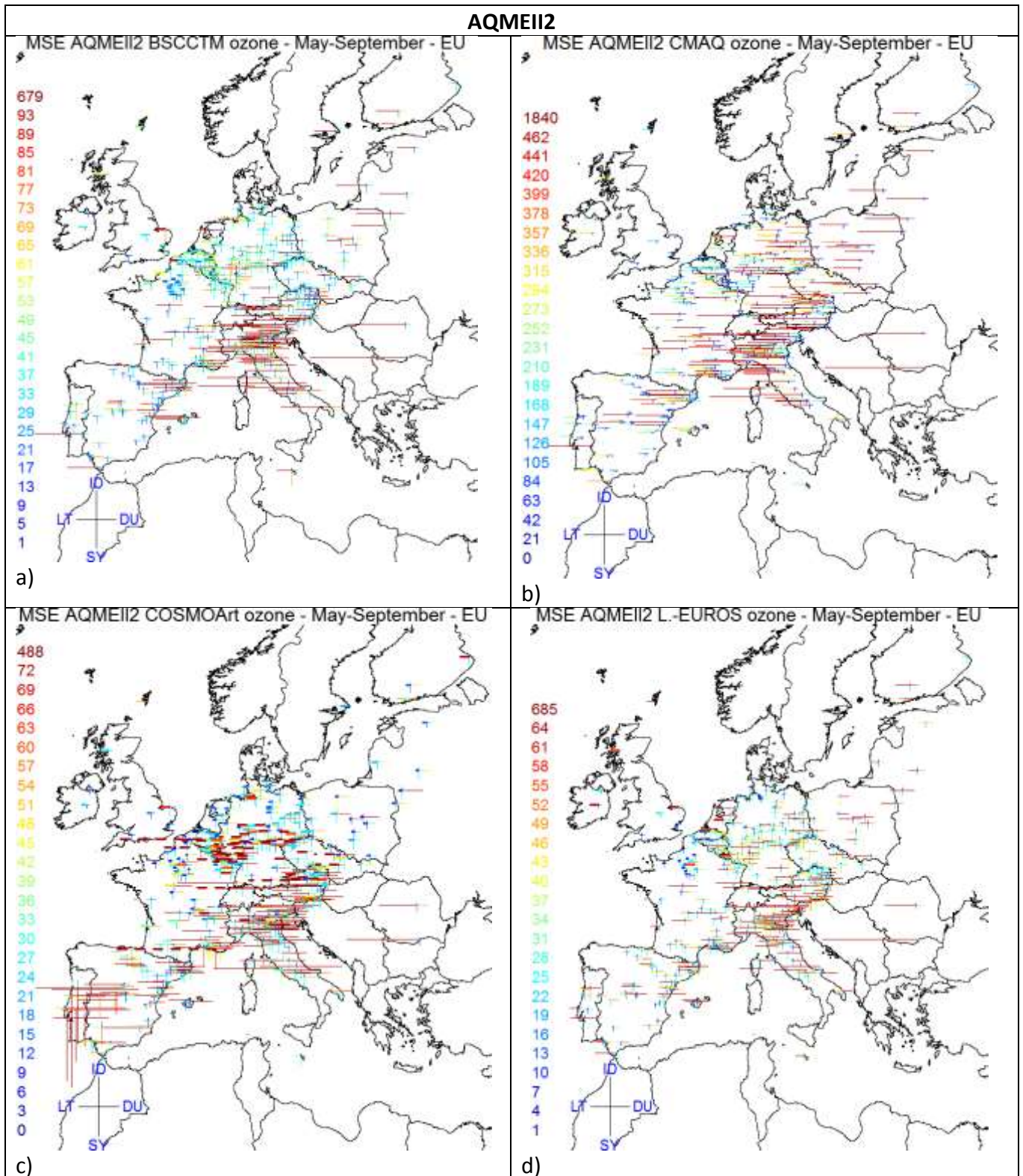
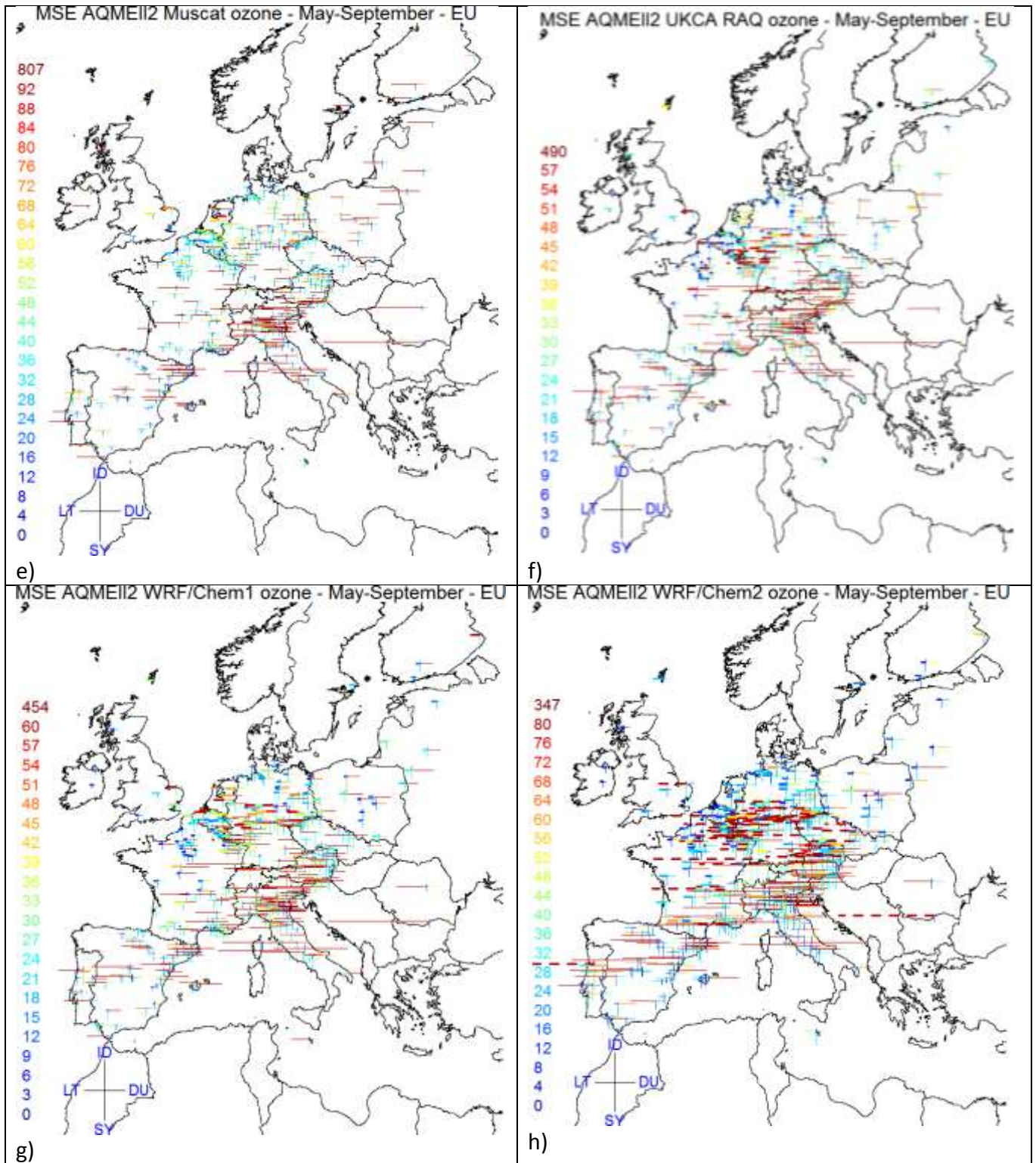



Figure 5. As in **Figure 3** but for the EU models of AQMEII2

778

779

a)

b)

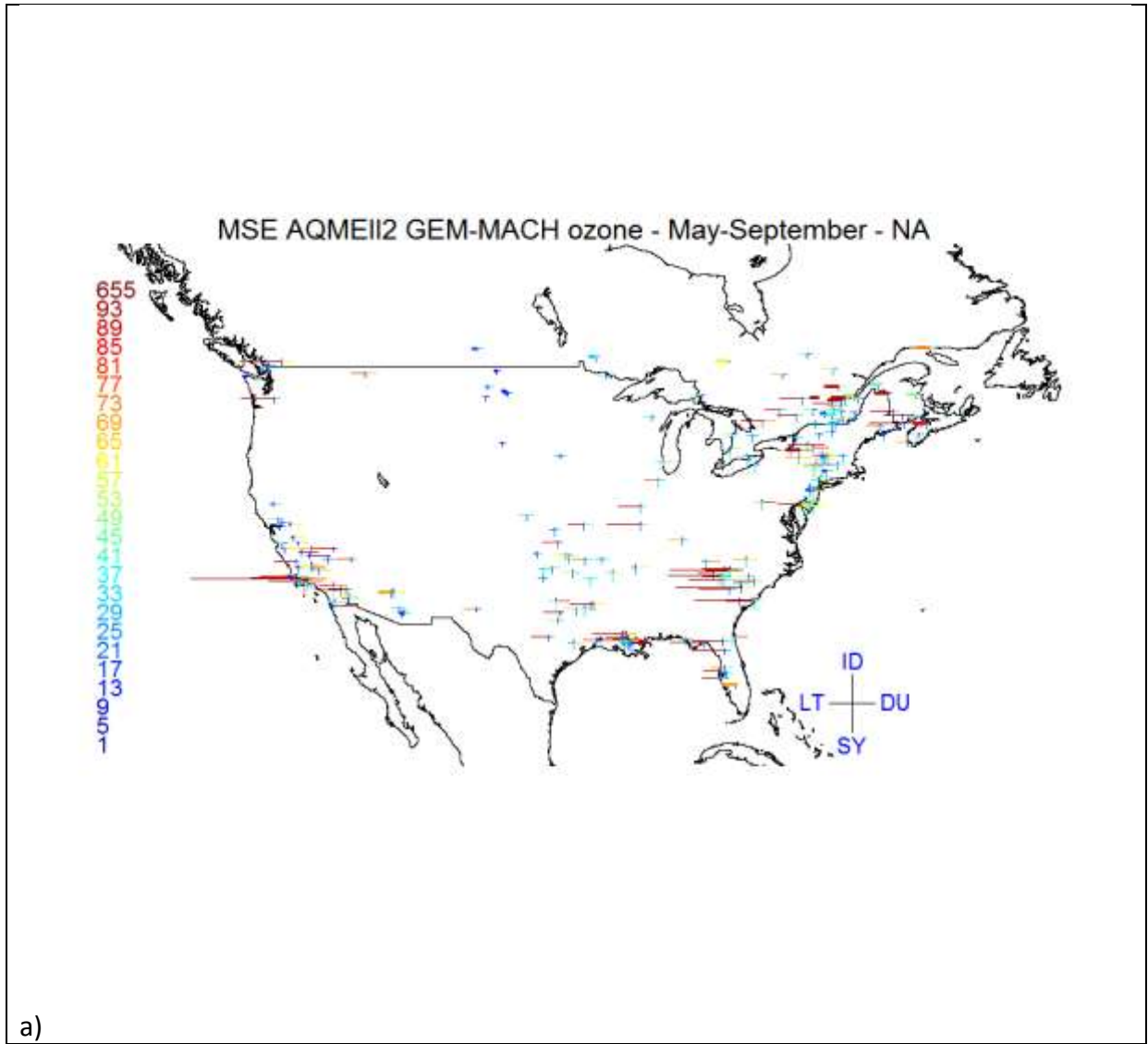
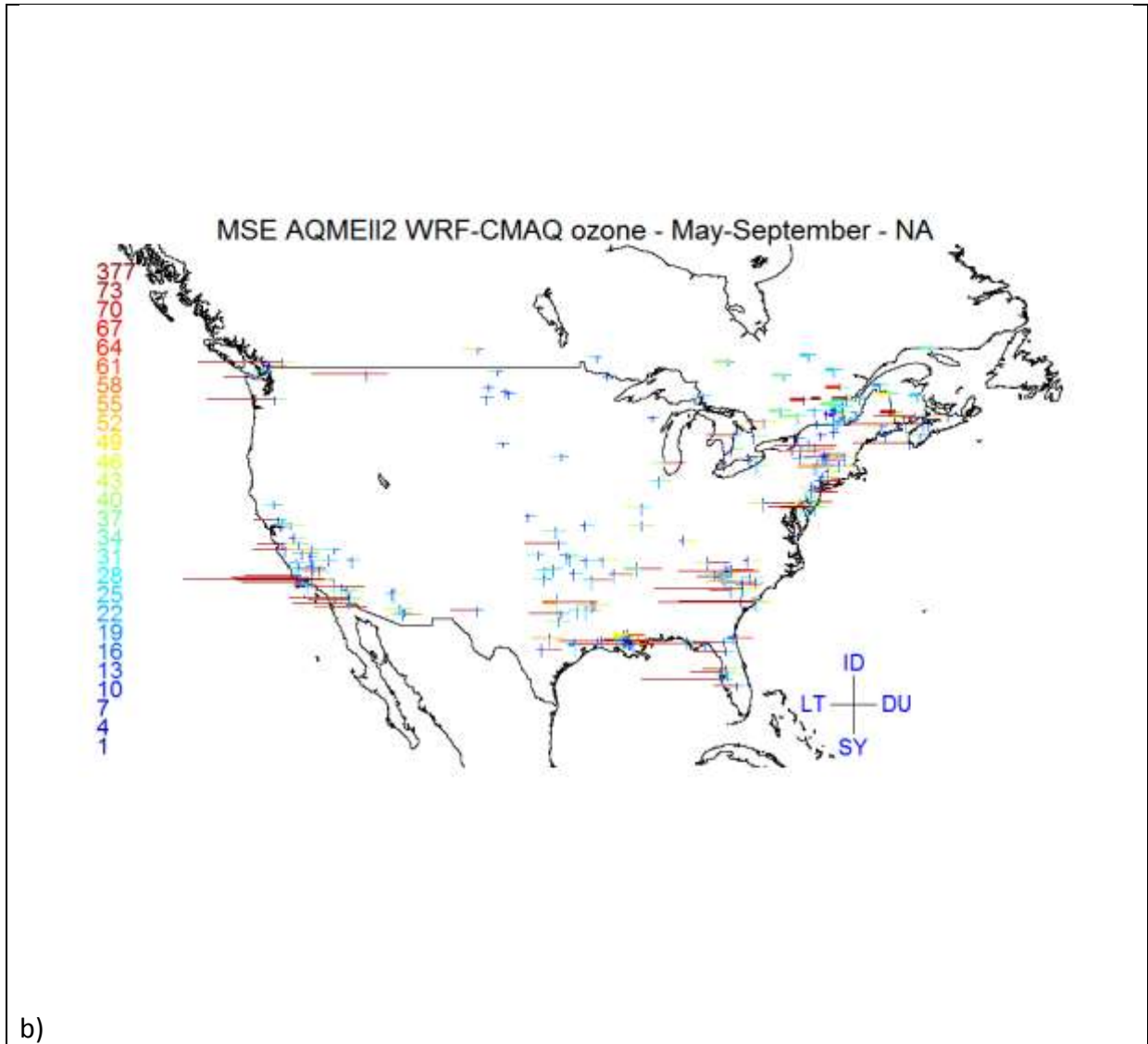
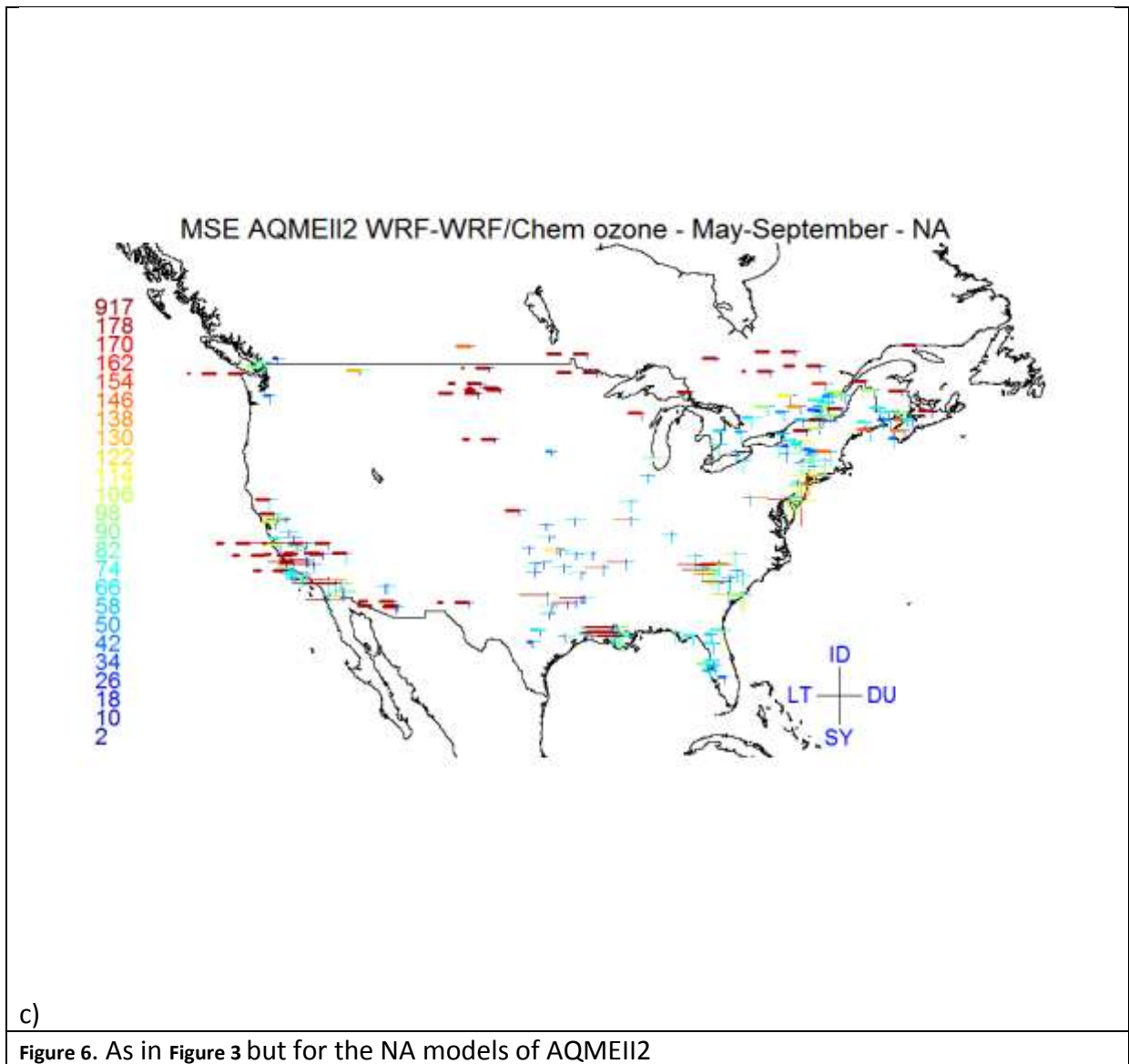
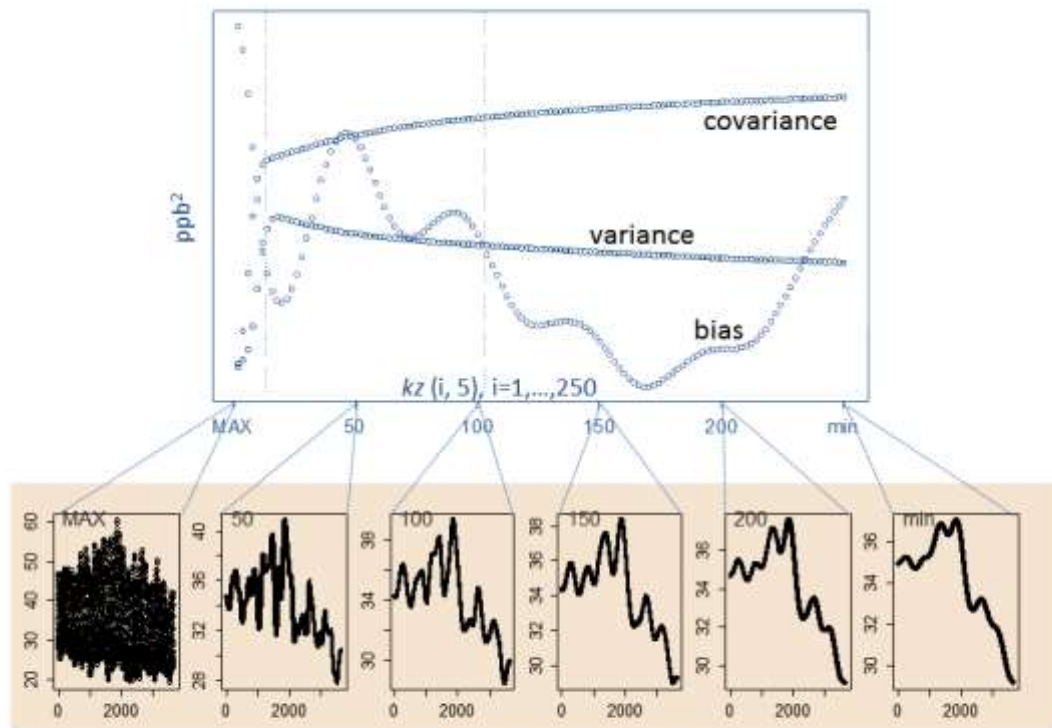
c)

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

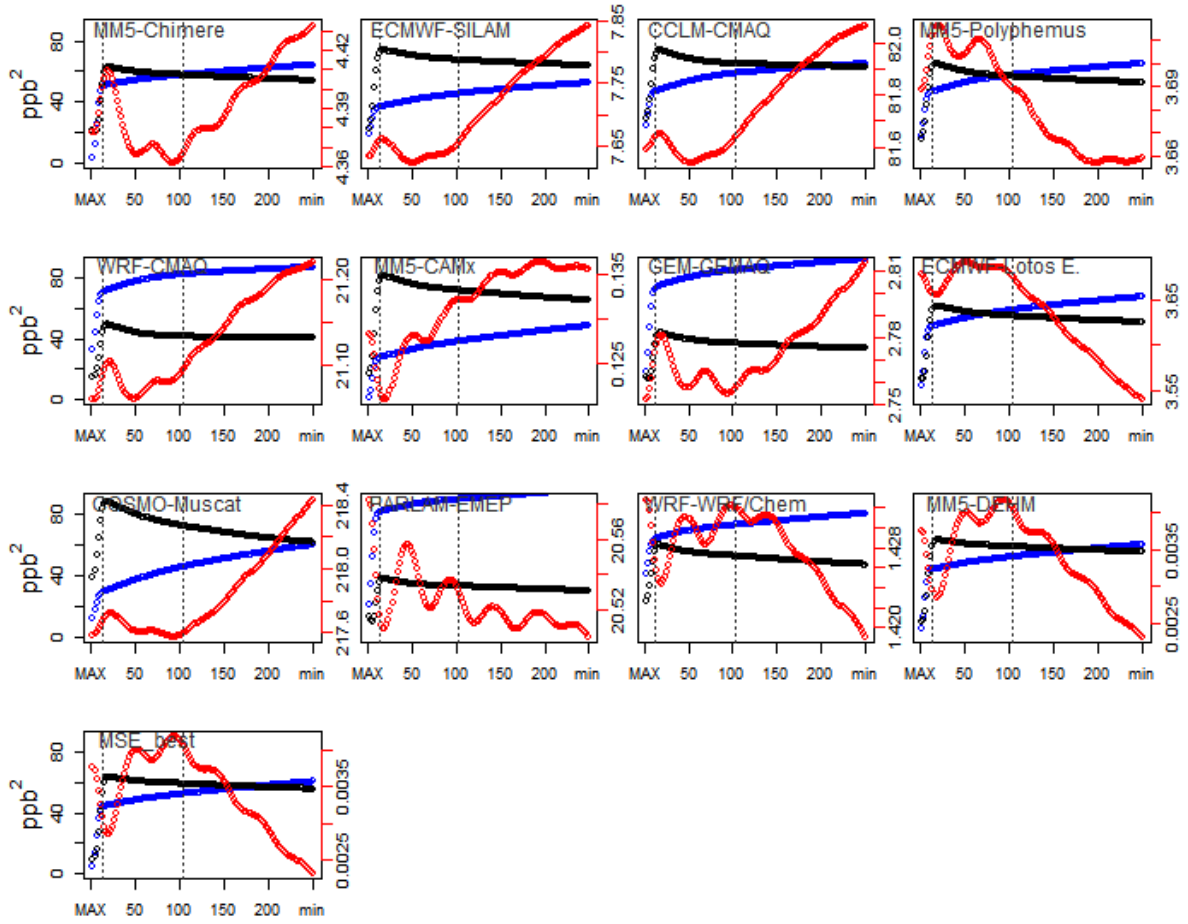
780

781

782

783

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

790

TIME-RESOLVED ERROR COMPONENTS - AQMEII1 - ozone - May-September - EU

791

TIME-RESOLVED ERROR COMPONENTS - AQMEII1 - ozone - May-September - NA

792

793

794

795

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.