18th International Conference on
Harmonisation within Atmospheric Dispersion Modelling for Regulatory Purposes
9-12 October 2017, Bologna, Italy

IN SEARCH OF A NEW PARADIGM FOR EVALUATING MODELS. LESSONS LEARNED AFTER THREE PHASES OF THE AQMEII ACTIVITY

Efisio Solazzo[1], Christian Hogrefe[2] and Stefano Galmarini[1]

[1] European Commission, Joint Research Centre (JRC), Ispra (VA)
[2] Atmospheric Model Application and Analysis Branch - Computational Exposure Division - NERL, ORD, U.S. EPA

Abstract: The Air Quality Model Evaluation International Initiative (AQMEII) has been active since 2010 with the aim of building a coordinated international effort on regional scale air quality modelling and evaluation, involving the modelling communities of North America and Europe. Over the years several dozen modelling groups from both continents have applied their modelling systems to common exercises, simulating air quality for a target year and delivered their results to a shared platform with a high level of harmonisation where they were evaluated against an extensive collection of available observations. The third and most recent phase of the activity was initiated in 2014 and is now nearing its completion.
Our experience suggests that the widespread practice of scoring the models using aggregate error metrics does not allow a comprehensive understating of error causes, and that the discussion about 'goodness' or 'badness' of a model based on such a practice can become sterile as it *i)* does not target the source of the error, *ii)* does not indicate if the model is doing the right thing for the right reason, and consequently *iii)* does not provide enough information for model development and improvement. Within AQMEII we have introduced the *error apportionment method*, where aggregated error metrics are used for time scale analysis and error qualification. Although this methodology provides a much clearer indication of the time scale and the type of model error with respect to conventional operational model evaluation, it still does not permit the unequivocal attribution of errors to specific processes.
We therefore argue that *evaluation needs to evolve from a practice into a discipline* designed to objectively and diagnostically develop and demonstrate viable performance evaluation techniques for regional air quality modelling systems.

Key words: *AQMEII, Model Evaluation*

INTRODUCTION

Evaluation of geophysical models is typically carried out under the theoretical umbrella proposed by Murphy in the early 1990s for assessing the dimensions of goodness of a forecast: consistency (*'the correspondence between forecasters' judgments and their forecasts*'), quality ('*the correspondence between the forecasts and the matching observations'*), and value ('*the incremental benefits realised by decision makers through the use of the forecasts'*) (Murphy, 1993). Since 2010, the Air Quality Model Evaluation International Initiative (AQMEII, Rao et al., 2011) has focused on the quality dimension of air quality model hindcast products, aiming at building an evalution strategy that is informative for modellers as well as for their users.

Our claim is that the *value* of a model's result depends strictly on the *quality* of the model that, in turn, depends on a sound evaluation. Operational metrics usually employed in air quality evaluation (e.g. error, bias, correlation) have several limitations: *interdependence* (they are related to each other and are redundant in the type of information they provide), *underdetermination* (they do not describe unique error features), and *incompleteness* (how many of these metrics are required to fully characterise the error?).

Over the three phases of AQMEII, the ozone error produced by the suites of modelling systems participating in the activity has not – on average– decreased (Figure 1) (although there are individual models that have improved their accuracy constantly). There is a need to understand what components of

the model require improvement, including the errors introduced in the models from input fields of emissions and boundary conditions.

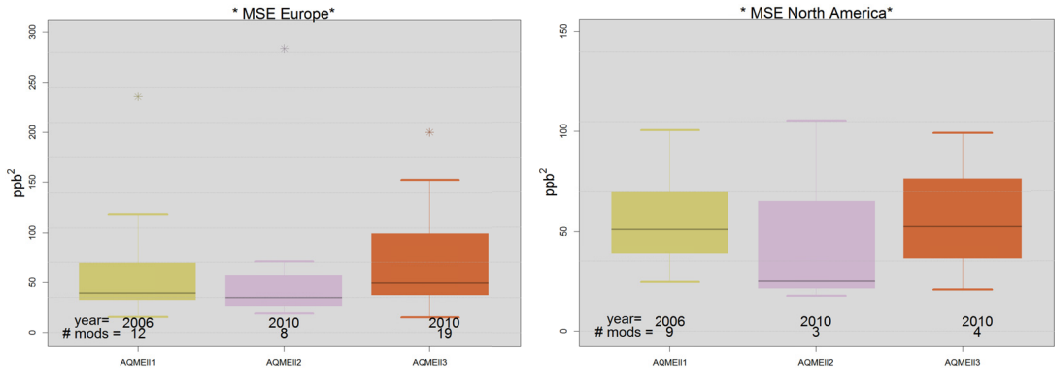


Figure 1. Range of variability of the Mean Square Error (MSE) of the suites of AQMEII models for surface ozone in Europe (left) and North America (right). The composition of the ensemble, the modelled year and the features of the models (on-line coupled models were used for AQMEII2) vary substantially among the different editions of AQMEII.

Following the requests from modellers to help diagnose the source of modelling error, the main aims of this study are to move towards tools devised to enable diagnostic interpretation of model errors and to advance the evaluation strategy outlined in the course of the three phases of AQMEII. This study attempts to:

- Attribute, where possible, the type of error to processes by utilizing modelling runs with modified fluxes at the boundaries (anthropogenic emissions and deposition at the surface, and boundary conditions at the bounding planes of the domain) and breaking down the mean square error (MSE) into bias, variance and covariance;
- Identify the time scales (or frequencies) of the error of modelled ozone and investigate the periodicity of the ozone error which can be symptomatic of recursive (either casual or systematic) model deficiencies.

METHODOLOGY AND RESULTS

We use the suits of regional scale models participating in the three phases of AQMEII and extensively described in a number of publications (http://aqmeii.jrc.ec.europa.eu/publications.html). The models are applied to the continental domains of Europe and North America and evaluated using surface network measurements. In this study the focus is on surface ozone.

To aid diagnostic interpretation, the mean square (or quadratic) error MSE (MSE = $E[mod\text{-}obs]^2$) is decomposed according to

$$MSE = \left(\overline{mod} - \overline{obs}\right)^2 + (\sigma_m - \sigma_o)^2 + 2\sigma_m\sigma_o(1-r) = bias^2 + var + covar \qquad \textbf{Eq 1}$$

where σ_m and σ_o are the modelled and observed standard deviation, *var* and *covar* are the variance and covariance operators, *r* is the linear correlation coefficient, and *bias* is the time averaged offset between the mean modelled and observed ozone concentration.

The error breakdown reported in **Figure 2** shows that, to the extent that outputs from the three phases are comparable, the median values as well as the overall distribution of the error components have not improved significantly over the years but have rather stayed constant or have changed only slightly. Very little information about the actual causes of modelling error can be gained from the error breakdown, and it actually raises more questions than it helps to answer: Why is the bias of the AQMEII3 models in North America higher than then the bias of the previous phases? Is the covariance showing a decreasing trend over both continents due to model improvement or due to casual error compensation?

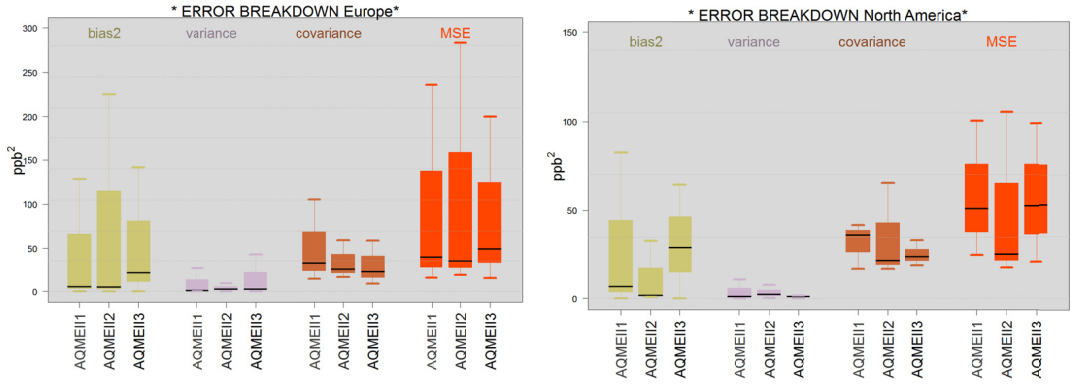
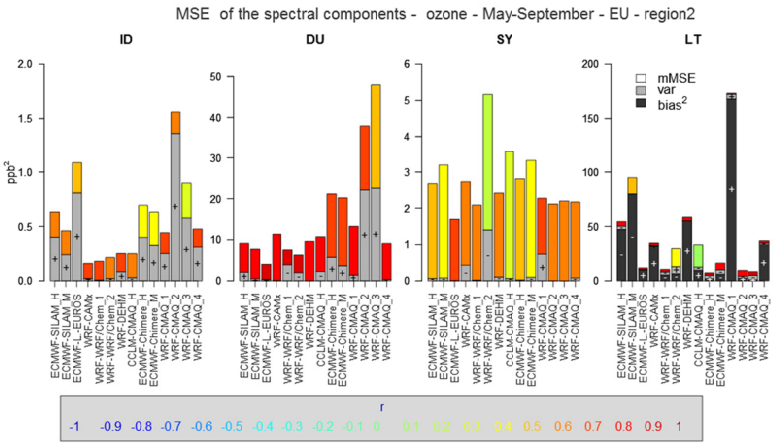


Figure 2. MSE breakdown for the three phases of AQMEII. Europe on the left and North America on the right

Aiming to identify the time scales (or frequencies) of the error we have advanced the analysis of the error components based on **Eq 1** by apportioning them to four relevant time scales: the base line (LT), synoptic (SY), diurnal (DU), and intra-day (ID) components, each representing a range of processes in a specific spectral range (details are given in Solazzo and Galmarini, 2016 and Solazzo et al, 2017a). The deviation of the modelled from the observed spectral component is informative about the time scale of the process(es) causing the error.

According to the results of **Figure 3**, the bias accounts for the largest share of the error, followed by the covariance of the diurnal fluctuations (having a periodicity of 0.5 to 1.5 days). Since the bias reflects systematic errors while the covariance is due to timing error, we can now advance informed hypotheses about the possible causes of the models error. For instance, the systematic model over-prediction can be due to error in deposition, missing processes (such as forest canopy removal) and an excess of precursor emissions (all of these causes are backed by evidence, summarised in Solazzo et al., 2017b). The error due to timing of the ozone signal can be introduced by errors in the radiative energy balance causing a too early (or late) boundary layer growth and/or collapse (Solazzo et al., 2017b).

Aiming to associate the error to the process, in the third phase of AQMEII two modelling systems (Chimere and CMAQ, operated respectively by the INERIS institute in France and by the US EPA in the US) have been used to perform a series of sensitivity simulations aimed at a better understanding of the causes of differences between the base model simulations and observed data, including : i) one annual run with zeroed anthropogenic emissions (referred to as '*zero Emi*'); ii) one annual run with a constant value of ozone at the lateral boundaries of the model domain (referred to as '*zero BC*' and '*const BC*', and iii) one run with ozone dry deposition velocity set to zero (referred to as '*zero Dep*') (**Figure 4**). Full details are provided in in Solazzo et al. (2017b).



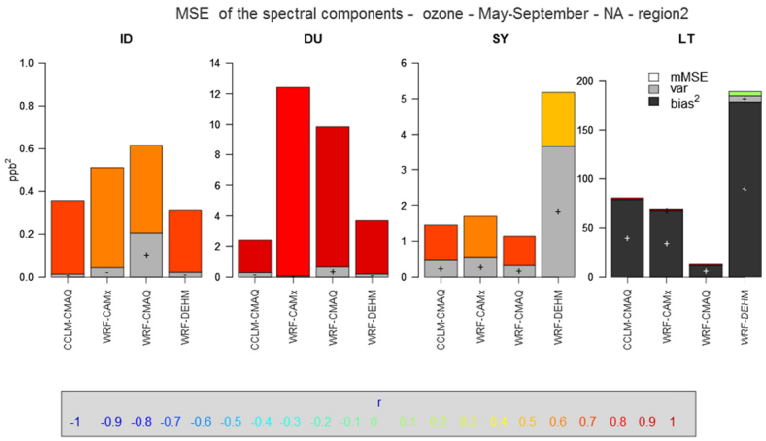


Figure 3. MSE (ppb^2) breakdown into bias squared, variance and covariance for the spectral components of the spatial average time series of ozone during the months of May to September 2010. By construction, the bias is entirely accounted for by the base line (LT) component. The signs within the bias and variance portion of the bars indicate model overestimation (+) or underestimation (-) of the bias and variance. The colour of the covariance share (*mMSE* in the legend is a proxy for the covariance) 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. Top panel: continental Europe; lower panel: south-east US.

The results of the error breakdown for the sensitivity runs are reported in **Figure 4**. While the zeroing/modification of input of ozone from the lateral boundaries causes a shift of the ozone diurnal cycle in both CMAQ and Chimere, the response of the two models to a modification of anthropogenic emission and deposition fluxes is very different. For CMAQ, the effect of removing anthropogenic emissions causes a shift and a flattening of the diurnal curve (bias and variance error), while for Chimere the effect is restricted to a shift. In contrast, setting the ozone dry deposition velocity to zero causes a shift (bias error) for CMAQ, while a profound change of the error structure occurs for Chimere with significant impacts not only on the bias but also the variance and covariance terms. Furthermore, several investigations indicate that the dynamics of the boundary layer is responsible for a recursive (systematic) daily error. The most revealing indicator is the analysis of the ACF and PACF of the time series of ozone residuals (Figure 5) that shows a marked daily periodicity: the 24-hour errors are highly associated throughout the year, i.e. the error repeats itself with daily regularity. Analyses of the error periodicity of primary species (to exclude the role of chemical transformations) and of the scenario with zeroed anthropogenic emissions (to exclude the role of emissions) have shown the same error structure, pointing to boundary layer processes as the main cause of daily error.

CONCLUSIONS

Based on the lesson learned after almost 10 years of model evaluation within AQMEII, we suggest that evaluation methods aiming to improve the models need to be diagnostic in focus. Air quality models have grown in complexity beyond the capacity of developers to control each process in isolation and so should evaluation techniques. Continuous improvements in process physics have the advantage of enhancing the variability, but improving the representation of variance can inflate the bias (and often the opposite is true as well). Bias correction methods, although successful in removing the offsetting error, have the considerable shortcoming of removing potential systematic, cumulative errors, thus masking the nature and source of the error. Although having exploited several evaluation frameworks over the past ten years within AQMEII (operational, diagnostic, and probabilistic) the goal of clearly associating errors to processes has not yet been achieved. As already suggested in the conclusions of the collective analysis of the AQMEII3 suite of model runs summarised by Solazzo et al. (2017), future model evaluation activities would benefit from incorporating sensitivity simulations and process specific analyses that help to disentangle the non-linearity of the many model variables, possibly by focusing on smaller modelling communities.

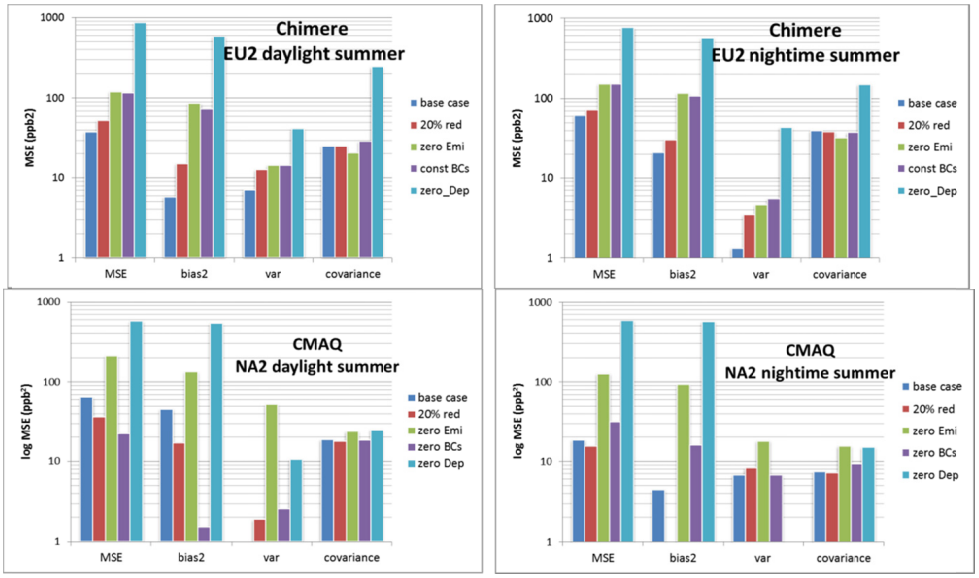


Figure 4. MSE decomposition for June – August hourly ozone into bias^2, variance and covariance for Europe (top) and North America (bottom). Results are presented separately for daylight hours (left) and night-time hours (right)

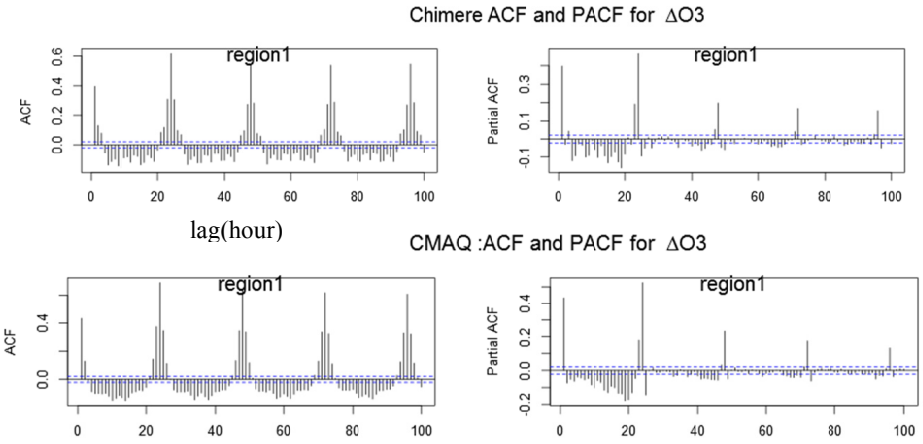


Figure 5. Auto and partial autocorrelation of the deviation of the Chimere (top) and CMAQ (bottom) models form the observation for two sub-regions of Europe and North America, respectively (the x-axis reports the lag in hours)

REFERENCES

Murphy, A. H., 1993: What is a good forecast?: An essay on the nature of goodness in weather forecasting. *Weather Forecast* , **8**, 281-293.

Rao, S. T., Galmarini, S. and K. Puckett, 2011: Air quality model evaluation international initiative (AQMEII), *B. Am. Meteorol. Soc*., **92**, 23–30

Solazzo, E. and S. Galmarini, 2016: Error Apportionment for atmospheric chemistry transport models: a new approach to model evaluation. *Atmospeheric Chemistry and Physics,* **16**, 6263-6283.

Solazzo, E. and et al., 2017a. Evaluation and error apportionment of an ensemble of atmospheric chemistry transport modelling systems: multi variable temporal and spatial breakdown. *Atmospheric Chemistry and Physic* **17***,* 3001-3054

Solazzo, E., Hogrefe, C., Colette, A., Garcia-Vivanco, M., and S. Galmarini, 2017b: Advanced error diagnostics of the CMAQ and Chimere modelling systems within the AQMEII3 model evaluation framework. *Atmospheric Chemistry and Physic,* in press