gmd-2016-183

"Spatiotemporal evaluation of EMEP4UK-WRF v4.3 atmospheric chemistry transport simulations of health-related metrics for NO_2, O_3, PM_{10} and $PM_{2.5}$ for 2001–2010" by C. Lin et al.

Responses to anonymous reviewer #1

We thank the reviewer for their time spent in reviewing our paper. Below, we respond to all comments made. The reviewer's comments are reproduced in their entirety, in italics.

The article presents a thorough evaluation of EMEP4UK model results against measurements of the AURN monitoring stations. While a thorough validation is a major and essential task when using an air quality model this article does not present any new insights or methodology on how such a validation should be done. Furthermore some of the presented validation work is IMO not complete and flawed to some extent. More specifically I have following remarks:

Response: We are pleased to read the reviewer's comment that our paper "presents a thorough evaluation of the EMEP4UK model results against measurements of the AURN monitoring stations." The reviewer then states that the article does not present any new insights or methodology on how such a validation should be done. In response we refer the reviewer and other readers of this discussion to the stated scope of Geoscientific Model Development, which encompasses articles reporting "full evaluations of previously published models" (http://www.geoscientific-model-development.net/about/aims_and_scope.html). Our article fits this scope: it reports, for the first time and for a temporally and spatially large dataset, the comparisons between output from the EMEP4UK model and observational data.

The criticisms in the latter part of the reviewer's comment above are repeated with more detail in their subsequent comments and we respond to them individually below.

1) For some unclear reason the authors have omitted the root mean square (RMSE) statistic from their analysis. They base this e.a. on the Thunis et al., 2012 paper. However in this paper even in the abstract the first statistic encountered is RMSE. In general there is agreement that a combination of bias, R and RMSE are best suited as each of these focuses on a different type of possible error in the model results when compared to observations.

Response: We are not clear why the reviewer thinks that our choice of the model-measurement statistics to present is based on the Thunis et al. (2012) paper. In our Introduction we cite several examples of the many studies that have discussed the choice of model-measurement statistic (for air quality studies), the work of Thunis and co-workers being amongst those we quote (P2, L5-9). We wrote: "Much has been written on air quality model evaluation (see, for example, Vautard et al., 2007; Dennis et al., 2010; Derwent et al., 2010; Rao et al., 2011; Thunis et al., 2012; Thunis et al., 2013; Pernigotti et al., 2013), including publications arising out of international collaborative programmes such as AQMEII (Air quality modelling evaluation international initiative, http://aqmeii-eu.wikidot.com) and FAIRMODE (Forum for air quality modelling in Europe, http://fairmode.jrc.ec.europa.eu)." These and other studies highlight the very wide suite of possible model-measurement statistics that can be used. We emphasise many times throughout our paper the basis of our selection of model-measurement comparison to publish in this paper (both the model-

measurement statistics and the air pollutant concentration averaging used in those statistics): namely that it was guided by the needs of the health burden and epidemiology community making first use of this large model dataset. The first two sentences and the fourth sentence of our Abstract make this clear: "This study was motivated by the use in air pollution epidemiology and health burden assessment of data simulated at 5 km × 5 km horizontal resolution by the EMEP4UK-WRF v4.3 atmospheric chemistry transport model. Thus the focus of the model-measurement comparison statistics presented here was on the health-relevant metrics of annual and daily means of NO_2, O_3, $PM_{2.5}$ and PM_{10} (daily maximum 8-hour running mean for O_3). "…"The two most important statistics highlighted in the literature for evaluation of air quality model output against policy (and hence health)-relevant standards – correlation and bias – were evaluated by site type, year, month and day-of-week." We do not dispute that RMSE is also a relevant model-measurement comparison statistic. But it is not practical to include results for all possible comparison statistics, which is why we focused on the correlation and bias statistics that are important for the health specialists. To further emphasise and justify this application of our evaluation we provided four paragraphs of discussion on the correlation and bias statistics in relation to health studies from P15, L1 to P15, L36. We will provide further emphasis and justification for our metrics in revised Introduction and Methods sections.

We refer to the work of Thunis et al. (2012) again in our Discussion section, in the context of commenting on the magnitudes of the model-measurement comparison statistics that may be expected for the type of air pollution model used in our work (see further comment on this below).

2) On p. 13 line 10 - 14 the authors blame deviations between modeled and observed data (almost) completely on the observed data's lack on representativenes and measurement error. Problems in representativenes are rather a problem of incompatibility: both model and observations are representative at a certain scale (neither of which is better than the other). However, these scales could (and are often) incompatible but this is neither the fault of model or observation. Measurement error is indeed a concern but in practice model error often by far exceeds the measurement error.

We believe the reviewer puts an interpretation on our text here that is not what we state, and at the same time ignores one of the key messages we promote from our model-measurement comparison. The specific text to which the reviewer refers above reads: "Even for a well-specified Eulerian model (in terms of input data, transport, chemistry, etc.), model-measurement agreement may not be perfect for (at least) the following two reasons: first, the model simulates a volume-averaged concentration whereas the monitor records the composition of the air in one part of that volume, which may or may not reflect the average concentration for the whole volume over the relevant time-averaging period; and, secondly, the measurement may be in error." So we and the reviewer are in agreement that there is an intrinsic incompatibility in the spatial scale of model and measurement. At no point here, or elsewhere in the paper, do we claim that one is better than the other, or 'blame' deviations between modelled and observed data "(almost) completely on measurements." We are simply reminding readers of this intrinsic incompatibility in scales, together with the reminder that measurements have an associated uncertainty. In fact, we do fully acknowledge model error at several points in our presentation and discussion of results, including in both the conclusions and in the abstract. We specifically emphasise (i.e. 'blame') shortcomings in emissions input into the model as being the dominant driver for the model-measurement

deviations (shortcomings in absolute magnitudes in emissions, in their temporal disaggregation and in the averaging of emissions across a model grid). For example, this is the text we write in the Abstract: "The directions of these biases are consistent with expectations of the effects of averaging primary emissions across the 5 km × 5 km model grid in urban areas, compared with monitor locations that are more influenced by these emissions than the grid average. …The biases are also indicative of potential underestimations of primary NOx and PM emissions in the model, and, for PM, with known omissions in the model of some PM components, e.g. wind-blown dust."; and, as further example, this is the text we write in the Conclusions "….is strongly indicative that the main driver of model shortcoming is inaccuracy of emissions (totals and the monthly and day-of-week temporal factors applied in the model to the totals)."

2) In line with the previous remark, after reading the text I have some doubts on whether the authors have understood the full extent of the methodology presented to the FAIRMODE community and outlined in the articles by Thunis et al. (2012) to which they refer. A sentence like p13 line 21" The presence of measurement certainty degrades the values that can be expected from air quality model measurement statistics" is a case in point: in the methodology proposed by Thunis et al. measurement uncertainty is used as the 'ruler' by which model uncertainty is judged: more measurement uncertainty then effectively means that model results can also be more uncertain!

Response: (We presume the reviewer intended to quote our text in their comment as "The presence of measurement UNcertainty degrades the values that can be expected from air quality model-measurement statistics", which is what we wrote, rather than "The presence of measurement certainty degrades the values…" which is what the reviewer writes that we wrote.) We don't understand why the reviewer thinks that we don't understand the concept that the greater the uncertainty that may exist in measurements the poorer the model-measurement comparison statistics may be. We think our sentence fully encapsulates this concept. We refer to the work of Thunis and co-workers at this point in the Discussion as a very useful previously-published 'yard stick' for the magnitudes of correlation coefficients and bias that might be expected for atmospheric chemistry transport model output vs. measurement (which are of similar construct to our model-measurement comparisons) when allowing for the possibility that there may be uncertainty in the measurement up to the level permitted under EU directives for reporting air pollutant measurements. We do not claim that these levels of uncertainties are the actual uncertainties in our particular set of measurements, but that if they were then these are the sorts of magnitudes of model-measurement statistics that might be expected.

In the end I was therefore left somewhat disconcerted by the text. Amassing all these results in a, admittedly, clear form must have been a major undertaking but there is not really anything new here. Worse yet, the authors seem to have missed some of the points made in the articles that they refer. I therefore recommend not publishing this article.

Response: We hope that our extensive responses above have addressed the reviewer's concerns. In summary, the novelty of work is the publication of new model evaluation statistics derived from an extensive set of simulations from the EMEP4UK model, with deliberate focus on the model-measurement comparison needs of the health burden and epidemiology community users of these simulations.

References cited in this response

Dennis, R., Fox, T., Fuentes, M., Gilliland, A., Hanna, S., Hogrefe, C., Irwin, J., Rao, S. T., Scheffe, R., Schere, K., Steyn, D. and Venkatram, A.: A framework for evaluating regional-scale numerical photochemical modeling systems, Environmental Fluid Mechanics, 10, 471-489, 2010.

Derwent, D., Fraser, A., Abbott, J., Jenkin, M. E., Willis, P. and Murrells, T.: Evaluating the performance of air quality models, A report for Defra and the Devolved Administrations, http://www.airquality.co.uk/reports/cat05/1006241607_100608_MIP_Final_Version.pdf, 2010.

Pernigotti, D., Gerboles, M., Belis, C. A. and Thunis, P.: Model quality objectives based on measurement uncertainty. Part II: NO2 and PM10, Atmos. Environ., 79, 869-878, 2013.

Rao, S. T., Galmarini, S. and Puckett, K.: Air Quality Model Evaluation International Initiative (AQMEII) Advancing the State of the Science in Regional Photochemical Modeling and Its Applications, Bulletin of the American Meteorological Society, 92, 23-30, 2011.

Thunis, P., Pederzoli, A. and Pernigotti, D.: Performance criteria to evaluate air quality modeling applications, Atmos. Environ., 59, 476-482, 2012.

Thunis, P., Pernigotti, D. and Gerboles, M.: Model quality objectives based on measurement uncertainty. Part I: Ozone, Atmos. Environ., 79, 861-868, 2013.

Vautard, R., Builtjes, P. H. J., Thunis, P., Cuvelier, C., Bedogni, M., Bessagnet, B., Honore, C., Moussiopoulos, N., Pirovano, G., Schaap, M., Stern, R., Tarrason, L. and Wind, P.: Evaluation and intercomparison of Ozone and PM10 simulations by several chemistry transport models over four European cities within the CityDelta project, Atmos. Environ., 41, 173-188, 2007.