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's second comments (RC2)

We thank the reviewer for their additional comments, to which we respond below. The reviewer's comments are reproduced in italics.

1) There are indeed many statistics some of which are 'quite exotic' that can be used to assess model performance so one could argue that RMSE should not be included. However, in all articles quoted RMSE is included and this is not just because the authors had a big appetite for statistics but because this statistic 'completes the picture' when assessing model performance in conjunction with the bias and correlation. Why not extend your analysis with the RMSE?

Response: We do not dispute that RMSE is also an often-used model-measurement statistic, but we focused here on the correlation and bias statistics relevant for the health effects community using the output from these model simulations – as is stated in the abstract, with detailed epidemiological commentary on this in the Discussion section. (See also our response to the similar comment in our upload to the online Interactive Discussion section of this paper on 30/09/16.) Reviewer #2 accepts the appropriateness of our material. Our paper contains a large number of tables and graphics of model-measurement statistics already. All the raw model and measurement values used to calculate our statistics will be available to allow anyone to calculate any additional model-measurement statistics.

In the revised paper we have provided additional up-front confirmation of the model-measurement statistics computed in this work with the addition of the following sentence at the end of the Introduction: "Two important statistics for evaluation of air quality model output for health studies – correlation and bias (see Discussion) – were evaluated by type of monitor location, year, month and day-of-week." We have also changed the phrasing from "the two most important statistics" to "two important statistics" where similar text occurs elsewhere in the paper.

2) There was indeed a typo in my comment: this should off course have been 'UNcertainty'. However the response the authors provide to the comment I gave concerning model vs observation uncertainty rather me in my conviction that they did not fully understand the concept proposed by Thunis et al. Let me try to explain. The concept of a Model Quality Objective (MQO) presented by Thunis et al. is that statistics used to describe model performance (bias, R or RMSE, ...) in themselves do not allow an actual assessment of how good the model is performing. Thunis et al. therefore propose to use the observation uncertainty as a 'yard stick' by which model uncertainty can be assessed. This e.a. implies that if model uncertainty is smaller than observation uncertainty there is no statistical basis for trying to improve the model in the sense that you'll not be able to discern the improvement based on a comparison with measured values. This also means that if measurement uncertainty increases this does not 'degrade the values' but rather result in that 'poorer' model performance may still be acceptable.

Response: We thank the reviewer for providing further explanation of the concept of the model quality objective presented by Thunis and co-workers. We maintain that we understand the

concept, and that the issue is in the wording in some of the text we used to expresses our thoughts. We understand that the measurement uncertainty is used as a 'yard stick' against which model uncertainty can be assessed and that if model uncertainty is smaller than measurement uncertainty then it will not be possible to discern any improvement in model performance, when model performance is being assessed against measurements. We accept that the use of the phrasing "degrade the values" in our original statement – "The presence of measurement uncertainty degrades the values that can be expected from air quality model-measurement statistics" – is misleading. We have now amended this sentence to read: "The presence of measurement uncertainty constrains the extent to which model-measurement statistics can be used to evaluate the performance of a model." We have also amended the first sentence in the following paragraph from the original: "Table 3 shows that in the large majority of instances the values of model-measurement correlation and NMB from this EMEP4UK-WRF modelling exceed the threshold values described above for satisfactory model performance in the presence of measurement uncertainties at the levels assigned." to now read: "Table 3 shows that in the large majority of instances the values of model-measurement correlation and NMB from this EMEP4UK-WRF modelling satisfy the model performance criteria values derived for measurement uncertainties at the magnitudes discussed above." Phrasing in the relevant sentence in the abstract has also been modified to now read: "Model-measurement correlation and bias were generally better than values that incorporate realistic magnitudes of measurement uncertainties." We have thoroughly re-read all text and believe there is nothing that is incompatible with the work of Thunis and co-workers.