Atmos. Chem. Phys. Discuss., 10, C10094–C10103, 2010
www.atmos-chem-phys-discuss.net/10/C10094/2010/

Atmospheric Chemistry and Physics Discussions

Interactive comment on "Using measurements for evaluation of black carbon modeling" *by* S. Gilardoni et al.

S. Gilardoni et al.

sgilardoni@gmail.com

Received and published: 22 November 2010

We thank both the referees for their suggestions that we believe improved the readability of the paper.

Referee #2

1. This paper describes and quantifies some issues in the use of measured, surface black carbon concentrations to evaluate the fidelity of transport model simulations. It analyzes model and observational statistics in light of measurement uncertainty, and this approach should be more widely applied. These issues are frequently discussed when such evaluations are published, but in most papers they are only mentioned and not quantified. This paper raises the bar for these comparisons and I think it

is an important contribution. I would like to see some of the points sharpened and synthesized before publication. My comments are given below.

The authors are grateful to the referee for his effort to understand the purpose and the potential applications of this work.

2. General comment. I realize that it is common to divide a paper into "Method" and "Results" sections but I think this paper should be an exception. Section 2 describes the basic data used for the analysis. Statistical methods are discussed in the Model/observation comparison section. The sections could be renamed to reflect their content.

We agree with the referee's suggestions; we renamed section "Methods" as "Observations and model data", section "Spatial representativeness of experimental measurements" as "Variability within a grid box", and section "Model/observation comparison" as "Statistical methods".

3. p 11319 "Systematic model evaluation has received an increasing attention..." I should hope that this is not true. There has always been attention to model evaluation but it gets more refined as time goes on.

The sentence at page 11319 is rephrased as:

Model evaluation is crucial because it helps understanding pollutant dynamics and makes it possible to use model output to quantify the effect of environmental policies on regional and global scale (Koch et al. 2009).

4 Page 11319 line 18 and on. Representativeness is important, and interferences are important, but interferences don't make spatial and temporal representativeness more important. These are separate issues.

We agree with the referee that sentence at page 11319 line 18 could be misleading. For this reason we reword the sentence as follow:

First, we recognize that the model/observation comparison is only meaningful when the temporal and spatial representativeness of observations are tested. In addition, light absorbing interfering species and uncertainties in the BC optical properties might strongly affect the interpretation of the observations.

5. Authors should discuss why these particular stations were chosen for analysis. The selection is not obvious.

The authors explain the site choice in the following lead-in paragraph of section 2:

In the following sections we describe the measurement sites used for the model / observation comparison, the physical meaning of equivalent black carbon based on light attenuation measurements, and the model details. The choice of the measurement sites was based on the availability of long time series of observations. Unfortunately, the number of locations that offer public data access and long time coverage is limited and this study could not be extended to more than six sites.

6. Section 2.1.1. There is a lengthy discussion about choices of correction for optical measurements. This is necessary as most data users don't recognize the importance of these assumptions, which are often made by the manufacturer or the measurement operator. However, I find it hard to synthesize and compare the data. I would like to see a table containing all these corrections: C, R(ATN) and sigma for each station. Perhaps this could be added to Table 1.

The authors add correction parameters and σ^* to Table 1 and add the following paragraph at page 11324 line 2:

A summary of σ^* values used at each site is reported in Table 1, together with C and R correction factors when σ^* was calculated from the theoretical σ value of 14625/λ.

7. Page 11324 line 20. "Further validation was performed" Of what?

The further validation refers to the model transport. The sentence is rephrased as:

The model transport has been extensively validated using 222Rn and SF6 (Krol et al. 2005, Peters et al., 2004) and its further validation was performed within the EVERGREEN Project (Bergamaschi et al. 2006).

The model ability to simulate BC concentration has been evaluated in Vignati et al. (2010b), and this reference is added at page 11325 line 6.

8. Page 11325, discussion of removal. Is information available on dominant removal (convective or stratiform)?

On annual global scale the removal was about 8 Tg C year-1, with 5.8 Tg C year-1 due to large scale precipitation and 2.1 Tg C year-1 due to convective precipitation. The authors add these information to the text.

9. Page 11326: temporal representativeness. Good idea. I wish more models did this.

The authors appreciate the referee's comment.

10. Page 11327: spatial representativeness. I don't think this section can really be interpreted as spatial representativeness. Two sets of observations within a single grid box may agree reasonably well. This is a necessary, but not a sufficient condition for an observation point to be representative. I suggest the section be retitled "Variability within a grid box" or something similar and the issue could be discussed.

We re-title the section as suggested and replace the term "spatial representativeness" with "variability within the grid box" throughout the text.

11. End of section 3: It would be useful to have a summary paragraph of the EBC and EC issue. It seems that EBC = BC in some regions but not where dust is present in large quantities. Please indicate which stations could be affected by dust and how you could determine this.

We add the following summary paragraph to section 3.2.

EBC at Bondville compares well with EC at the same site and at the nearest IMPROVE

sites, indicating both similarity among rural - background sites, and an adequate choice of σ^* to estimate EBC from optical measurements. On the contrary, the Trinidad Head grid cell is characterized by a larger variability of EC and EC is systematically smaller than EBC; the discrepancy is likely due to an incorrect σ^* rather than dust interference on optical measurements; at Trinidad Head the uncertainty of black carbon optical properties and the variability of the grid cell compromise the use of EBC observations for model evaluation.

12. Section 4 General comment. This section discusses a few statistical tests. It would be nice to lead this section with an introduction to the tests that will be used, their value in analysis, and what the reader should look for in each test. Also compare these tests to each other during the discussion throughout. For example, what information do the results of the Mann-Whitney test provide, in contrast with the PD overlap? Authors may have become familiar with the physical interpretation of these measures through the course of their analysis, but most readers (and perhaps modelers) may not be so comfortable. Finally, finish the section with a concluding paragraph about what has been learned about the comparison at the different stations.

The authors add the following lead-in and concluding paragraphs to section 4:

Page 11330 line 1

The statistical tools used in the following paragraphs include Fast Fuorier Transformation (FFT) analysis, skewness analysis, median comparison, variability analysis, probability distribution curve comparison, Student's t-test, and Mann-Whitney test. The FFT analysis is used to identify the frequency and amplitude of data modulation. The skewness analysis investigates the symmetry of data distribution (of model or observations); when skewness is zero the data are symmetrically distributed around their average value. Median comparison and variability analysis are used to test the similarity of two data distributions: the median comparison focuses on the similarity of central values of each distribution, while the variability analysis investigates how data

are spread around the central values. The similarity of two data distributions can further be investigated with the overlap of the probability distribution curves, the Student's t-test, and the Mann-Whitney test; in the first two tests, higher value of the output (overlap area and test probability) corresponds to higher similarity of data distributions, while the Mann-Whitney test shows lower output parameter Z for higher similarity of distributions.

In the following sections we will use the different statistical tools within three tests with three different objectives. The comparison of monthly medians is used to evaluate the ability of the model to simulate the time trend of measurements over a longer time period. . .

Page 11332 line 16

In this section we illustrate how to investigate the ability of model to simulate observation variability, and to verify if the model fails in simulating the daily modulation. First we compare 1–hour and 24–hour resolution data to identify cases when variability is dominated by daily modulation, and then we use the FFT analysis to verify how model describes this modulation.

Page 11334 line 1

Model overestimates observation variability at Bondville and Ispra, and underestimates it at Alert. At Ispra the overestimation is limited to the summer months, when the model variability is dominated by the daily modulation and the model overestimates daily modulation of observations. At Bondville and Alert the model fails in reproducing both daily modulations and 1-hour resolution data variability.

Page 11334 line 5

We compare here four different statistical tools to quantify the agreement between model and observations. The most accurate way to quantify this agreement is the overlap of the probability distribution curves. Considering that this method is elaborate,

we discuss limitations and advantages of alternative and more common statistical tools: comparison of medians, Student's t-test, and Mann-Whitney test.

Page 11337 line 22

. . .although Mann–Whitney test does not assume that the two populations of data are normally distributed, it requires that the shapes of their distributions are identical. For this reason we re-applied the Mann-Whitney test only to those months for whom model does not overestimate observation variability (see section 4.2). Considering only months when the model to observed variability ratio is smaller than 1, the correlation of Z values and PD overlap area is improved (r2=0.79).

Page 11338 line 3

Monthly medians, Student's t test, and Mann-Whitney test are employed to quantify the agreement between model and observation data distributions and compared to the probability distribution curve overlap method. The three tools are based on different sets of assumptions that are not always encountered in real population samples. If these assumptions are not discussed and the methods are applied without discrimination to the entire dataset, the result can be misleading.

In summary, the comparison of median is inaccurate and it should be discouraged because ignores the data distribution around the medians. The Student's t-test is a good choice only when the data distribution is similar to a normal distribution; this assumption is rarely verified and its testing can be time consuming. The Mann-Whitney test is a satisfactory alternative to the comparison of PD curves, but it requires that observation and model data distribution have the same shape; this can be easily verified by comparison of observation and model variability.

13. Section 4.2: variability. I like the idea of this section but found it hard to understand. The name "overall" variability for 1-hour data is confusing to me. Do the terms daily, diurnal, day-to-day mean the same thing? Perhaps frame this in terms of the Fast

Fourier transform first so that you can discuss the frequency components.

To avoid potential confusion in section 4.2, we replace the terms “overall variability” and “day-to-day variability” with “1-hour resolution data variability” and “24-hour resolution data variability”. Figure 6 was modified accordingly. To improve the reading we also explain the sequence of techniques employed in the lead-in paragraph (see point 12).

14. Section 4.3: probability distributions and overlap. Again I like this idea but I am not sure what the presentation means. I do understand the statistical principle, but the physical implication is unclear. Since there are many reasons for model disagreement and agreement, I hope the authors can explore what a PD overlap means physically. If they can use it to explain the differences in model-measurement comparison at the various stations, it could be quite powerful.

The authors add the following section “PD overlap summary” at the end of “Model / observation agreement”. Figure 9 is here attached.

We explore here the physical mean of PD overlap and how PD overlap analysis can be used to discuss model performance. For this purpose we focus on four cases that were already discussed in the previous sections.

At Alert the model is not able to reproduce the transport of pollutants from lower latitudes, which leads to PD overlap ranging between 0 and 25%. In March and April model distribution data is narrow and centered around BC concentrations smaller than 50 ng m-3, while observations are shifted towards higher concentrations; from December to February observations distribution is bimodal and the model simulates only the lower concentration mode, again below 50 ng m-3.(Fig. 9a) The model does not show higher BC concentrations since it fails in describing pollutant transport taking place at the end of winter and in spring.

A similar effect is observed at Jungfraujoch in summer, when local sources might affect the sampling site. The model fails in describing the transport of emissions from lower

altitudes (Fig. 9b), and thus it does not simulate the occasionally higher BC concentrations in May – August. This results in PD overlap slightly smaller during summer than in winter.

Figure 9c shows that at Ispra in summer the model data distribution has a bimodal shape and the mode at concentrations larger than 2000 ng m-3 is not present in the observations. This discrepancy is likely due to the model underestimation of dilution processes, which also leads to the overestimation of daily cycle amplitude, as seen during variability analysis.

At Bondville the model does not simulate the lower BC concentrations (Fig. 9d). This could be due to underestimation of dilution, but more likely by the large model grid cell that contains industrial, urban, and background areas.

15. Figures 5 and 7. Suggest you use different symbols in addition to colors, for readers who print in black and white or who have colorblindness.

We modify the color codes used in Figure 5 and 7, and change the corresponding legends.

Vignati, E., Karl, M., Krol, M., Wilson, J., Stier, P., and Cavalli, F., Sources of uncertainties in modeling black carbon at the global scale, Atmospheric Chemistry and Physics, 10, 2010b.

Interactive comment on Atmos. Chem. Phys. Discuss., 10, 11315, 2010.

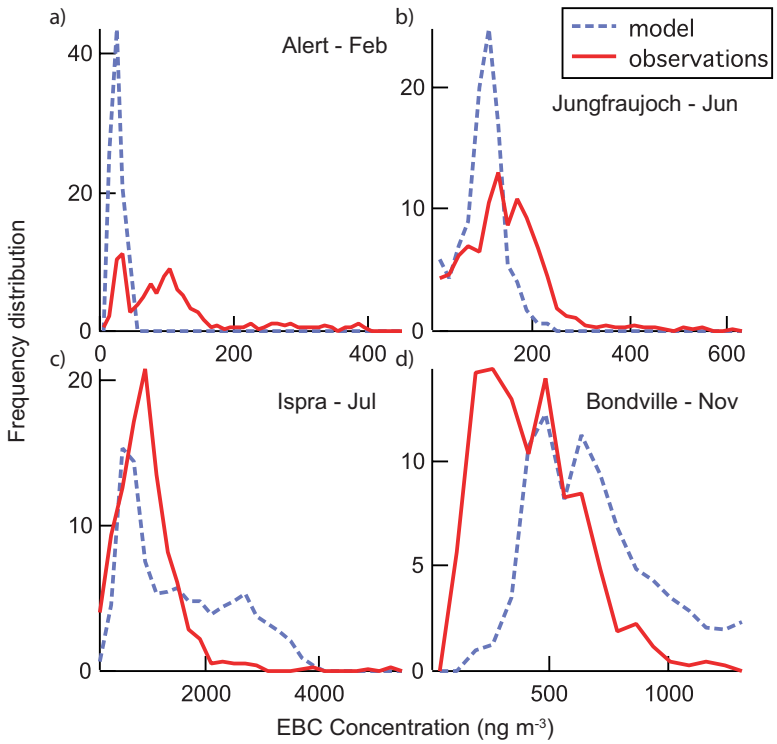

Fig. 1. Fig 9. PD curves of model (blue) and observations (red) at Alert in February (a), at Jungfraujoch in June (b), at Ispra in July (c), and at Bondville in November (d).