Reply to referee #3

This paper developed a statistical two stage method to better forecast the pollutant levels. The methods are evaluated with respect to key statistical metrics of both deterministic and probabilistic nature. The idea presented in this work is interesting and is of practical importance. The paper overall has good technical quality, although improvements can be made to further improve the manuscript. I suggest publication of this work after the following comments are addressed.

Major:

Line 173: Why are three days' data used to train the coefficients in the first stage? Are the coefficients sensitive to the number of days used for training.

The b1, …, bm weights (stage 1) can be interpreted as a measure of the overall performance of each member of the ensemble, over the training period, relative to the other members. We tested the training of our global model using the last 3 days, the last 7 days, or the last 14 days, and applied it to predict the concentrations of the upcoming day. We did not find significant differences in using different training windows, so that we chose the less resource consuming scenario (a 3-day training period, as described in the paper). Moreover, a short training window also has the advantage to adapt the bias correction rapidly (in case of rapid changes in meteorological conditions or pollutant emissions, as, for example, experimented during the COVID crisis) and is less computing intensive.

Table 2: From this table it seems that stage 2 worsens the prediction in terms of bias as well the RMSE of PM2.5. What is the reason for this?

The first stage is a bas correction step, ie. the b1,…,bm parameters are constructed so as to remove the bias The expected result is a very low bias. The second stage introduces new information, i.e. the spatio-temporal covariates. This new information is always beneficial, as shown by the further reduction of the mean squared error (reported in the same table), at the small cost of increasing the mean bias in some cases. In any case, we are talking about rather small deviations. For example, for PM_{10} the bias changes from 0.20 to -0.97 $\mu g/m^3$, which is below 1 $\mu g/m^3$ in both cases. For $PM_{2.5}$ changes from 0.37 to -0.58 $\mu g/m^3$. Similar considerations are also valid for NO_2 and O_3. Differences are less than one or two orders of magnitude of the typical average concentration values of these pollutants, in line with those reported in other studies. We believe that these errors do not detract from the significance of our statistical treatment, even if at present we cannot exclude a deepening of the nature of this behavior in a future work.

Line 349: More technical description can be provided, e.g., how are the coefficients used in stage 1 obtained? Are they the same as those trained in previous sections? The application in 4.4 is quite interesting and this section could be expanded to include more details.
Figure 8 and related text: Please provide comparison with observations.
I'd like to see some comments on the computational cost of the current method. Low computational cost indicates sensitivity studies (e.g., with respect to spatiotemporal predictors) can be easily performed to potentially improve the current method.

The weights estimated in stage 1 were used to obtain an ensemble average, and this average was interpolated onto the new 4x4 km grid (using a bi-linear interpolation). In the second stage, the spatio-temporal predictors are estimated at the cell centers of the 4x4 km grid and then the statistical post-processing is applied. In the revised version, we describe this process more precisely (see section 4.5). Moreover, we expand the analysis of the properties of the post-processing approach, including:

1. *a map for the comparison between observations and model values, highlighting the dependence on the type of monitoring station (urban, suburban and rural, see new Figure 3) and geographical regions (stations in north, center and south regions), see new Figure 4).*

2. *a comparison extended to all pollutants (not only PM10, see the new section 4.3)*
3. *scatterplots for all pollutants*

Computational costs are an important point of our approach. In the literature, the problem of building spatially continuous concentrations maps over large domains has been approached by different perspectives. Most of the studies use hierarchical models based on the Markov chain Monte Carlo (MCMC) approach; despite the existence of user-friendly programming tools, the application of a MCMC approach is rather cumbersome, requiring a lot of CPU-time as well as tweaking of simulation and model parameters' specifications. Some strategies have been proposed to alleviate the computational burden of fitting complex spatio-temporal hierarchical models. One of such strategies is that applied in INLA. Computationally, INLA is much more efficient as it is based on the use of sparse matrices. Moreover, INLA is based on approximating the marginal posterior distributions (by using Laplace and other numerical approximations and numerical integration schemes) and is usually faster and more accurate than MCMC alternatives. We include these details on the computational costs in the conclusion section of revised paper.

Minor:

Line 134: Add references for the relation between temperature, wind speed, RH and ozone, PM and NO2.

References have been included for these parameters.

Technical:

Line 179: 'given in Appendix A'.

corrected

Can the authors also add legends to Figure 7 instead of only describing them in the text?

We will substitute Figure 7 with a more detailed comparison between the results of the statistical model and the observations, extended to all pollutants (not only PM_{10}). Anyway, the paper has been revised to include a better description of all figures.