Reply to referee #2

1. The title of this manuscript emphasizes the application of machine learning. However, the methods described in this manuscript, such as calibration of the ensemble in the first stage and statistical modelling of the space-time process in the second stage, appear to be statistical methods instead of machine learning.

We thank the referee for the suggestion and agree with its comment. We suggest changing the title to: "Accurate, reliable and high-resolution air quality predictions by improving the Copernicus Atmosphere Monitoring Service using a novel statistical post-processing method"

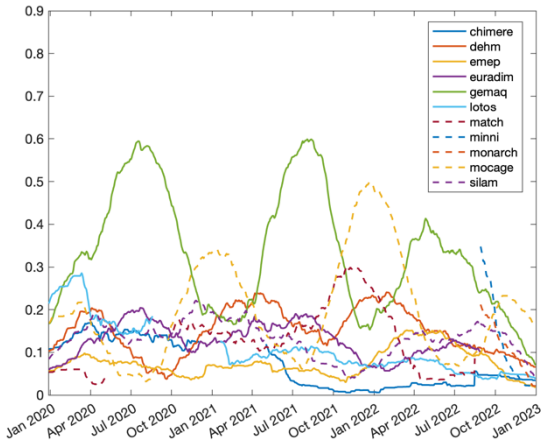
2. calibration of the ensemble: here the objective is to produce calibrated concentrations based on the linear regressions of multi-models and observations with parameters b1, … bm. I am a little surprised to see that this process needs to be repeated every day. I am wondering whether there are temporal and spatial changes in the parameters b1, … bm in the model training and what these changes in the parameters b1, … bm represent.

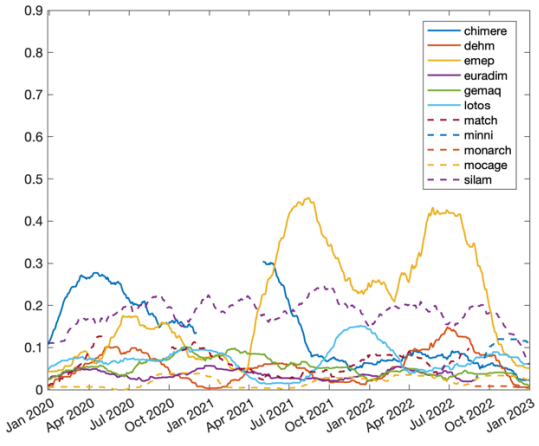
For brevity, in the first version of the paper, we did not include details about the b1, … bm properties. However, you raise an important point. We reply to your question under answer '3' (see below)

3. statistical modelling of the space-time process: similar to the above question, does this process need to be repeated every day?

The b1, …, bm weights (stage 1) can be interpreted as a measure of the overall performance of each ensemble member, over the training period, relative to the other members. These weights are constructed using monitoring data located throughout the spatial domain. Due to different meteorological conditions and seasonal variabilities, their estimation is repeated regularly using a predefined time window. 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 next day. We did not find large differences in using different training windows, so we chose the less resource-consuming scenario (a 3-day training period, as described in the paper). A short training window has also the advantage of adapting 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. The same consideration also applies for the space-time process. This process was repeated every day to mimic an operational system running during the 2022-year period, and a new model was trained regularly with the most recent data to stay close to new forecasting situations.

In the revised version, we include a section dedicated to the aspects related to the interpretations that coefficients might have. To give you an idea of the results obtained, we extended our analysis to the whole period from 2020 to 2022 (three years of continuous update of b1, …, bm weights). The following figure (left panel) shows the temporal dependence of these weights for PM_{10} (left panel) and O_3 (right panel) for each model. For many models, the weights range from 0.05 to 0.3, reflecting their relative performance. Some models (GEMAQ, MATCH, MOCAGE) show a marked seasonal dependence, with the weight of the GEMAQ model increasing significantly during the summer period, while the weight of the MATCH and MOCAGE models increases during the winter period, indicating both a different performance, depending on the season, and complementarity of these models. It is also interesting to note that for ozone, a pollutant with a marked seasonal cycle, most models perform equally well in both the winter and summer seasons. This analysis is reported in the revised paper (see new Figure 2).





4. Section 4.2-4.3: I suggest shortening these two sections and perhaps moving some figures into supplement because the parameters in both Stage 1 and Stage 2 are trained with observations, and it is thus expected to see some improvements after these two Stages.

We partially agree with the comment of the referee. As explained in Section 3.2 (Validation), we split the whole dataset (about 700 monitoring stations) in two sets. ≈90% were used to train the model, the remaining set (the validation dataset) was used for validation purposes (not used during the training phase for the first and second stage). Comparison with the validation data set is an integral part of the validation process since it represents a measure of the expected error in forecasting independent data. We consider this information as a significant result of applying our post-processing method. However, we followed your suggestion and shortened Sections 4.2 and 4.3 as much as possible (for example, we discard the old Figure 2, which is a replica of the same information reported in Table 2).

5. Section 4.4: In contrast, it could be better to extend this section as it is most interesting to the readers. For example, Fig. 7 is not convincing as these three stations may not provide a good representation of the whole domain. It could be better to provide scatter plots to show the overall performances of the predictions; While the high-resolution PM concentrations in Fig. 8 are interesting, it is useful to show the map of the differences between predictions and observations to demonstrate the spatial performance of the predictions; in addition, as the model needs to be trained every day, I am wondering whether the performance of the predictions has seasonal variabilities.

We expanded the analysis, 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) and season (see new Figure 3) and geographical region (see new Figure 4).*
2. *a comparison extended to all pollutants (not only PM10, see the new section 4.3)*
3. *box-whiskers plots for the bias of all pollutants (see Figures 3 and 4), equivalent to scatterplots.*

The statistical prost-processing method works equally well, independently from summer and winter seasons (see new Figure 3). The first stage is able to modulate the seasonal dependency of models (see new Figure 2), and the second stage, starting from the calibrated and more accurate results, improves the forecasts, further reducing the bias (see new Figures 3 and 4)

6. is it possible that this figure overestimates PM concentrations over rural areas because most stations in Table 1 are higher polluted urban and suburban stations?

As already stated in the previous answer, we conducted a thorough comparison between observations and model values, highlighting the dependence on the type of monitoring station (urban, suburban and rural). We anticipate that the post-processing method can manage very well the spatial dependence (several spatio-

temporal predictors were used precisely because they are able to consider the spatial dependence in the surroundings of the measurement station). In the revised version we include a new figure and a new table (see Table 1 and new Figure 3) showing the distribution of the average bias (distinguished by urban, suburban and rural stations) for all pollutants.

Technical Comments:

A flow chart is suggested to provide a clearer description of the methods.
It would be helpful to provide a List to show the variables which were used in the model training including their temporal and spatial resolutions.

We inserted a flow chart in the revised version (see new Figure 1). Table A1 includes the temporal and spatial resolutions

Why the Section number of Stage 1 is 3.1.1 but 3.2 for Stage 2? I assume the description of these two stages is parallel.

Corrected

Lines 130-131: why most met predictors are selected at 12 UTC?

Some met variables, i.e. boundary layer, were selected at 00 and 12 UTC to take in account the daily cycle. The other variables were selected at 12:00 UTC. We considered that day to day variations may be represented by this time level.