Reply to referee #1

1) Section 2.2: It is suggested to provide a table of all the data and their sources in the supplement or other appropriate locations of the paper. Too many URLs are present in this section.

We thank the referee for the suggestion. We provide a new table in the supplement with information on the data source (see Table A1).

2) Line 132: Temperature is a very important parameter relevant with photochemistry and lifetime of air pollutants. Why is it not included in the predictor list?

We thank the referee for pointing out this feature. Temperature was considered as a predictor, but we forgot to list this parameter in the predictor list in Section 2.2. We are sorry for this oversight; in the revised version we specified the use of temperature.

3) Line 200: Since the sites are unevenly distributed and many non-urban areas are not monitored. I would anticipate significantly reduced site density in Central and South Italy. So, should we also consider more even inclusion/representation of sites in different part of Italy?

The density of monitoring sites in Central and Southern Italy is reduced compared to that in Northern Italy. In the revised version we included more details on the performance of the model as a function of the characteristics of the measurement stations (urban, urban background, or rural/remote, as classified in the EEA database). Each type of station is represented in each Italian region, and, according to the results shown in the revised paper, we found no differences in performance depending on geographical location. Finally, we underline that stage 2 of our approach is based on the use of different types of information, mainly linked to geographical characteristics and meteorological conditions. This additional information is uniformly distributed throughout the Italian territory and as highlighted in the work, allows an excellent correction of the systematic bias and an adaptation to local conditions. We included a new figure and a new table (see Table 1 and new Figure 3) showing the distribution of the average bias (distinguished by north, center and south stations) for all pollutants.

4) Section 4.1: I think it is well anticipated that the 11 models will have varying biases and precision. Maybe this section can be moved to the supplement?

We agree with your suggestion and move Section 4.1 to the appendix.

5) Table 2: For all the four pollutants and in the "training" and "prediction" rows, the absolute biases are amplified from Step 1 to Step 2. It appears unusual and not found in previous studies. Why and does it matter?

The first stage is a bas correction step, ie. the b1,…,bm parameters are constructed so as to remove the 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 □g/m^3, which is below 1 □g/m^3 in both cases. For $PM_{2.5}$ changes from 0.37 to -0.58 □g/m^3. Similar considerations are also valid for NO_2 and O_3. These 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.

6) If Figure 2 is only briefly discussed and Table 2 is mainly used in Section 4.2.1, maybe Figure 2 should also be moved to the supplement?

Figure 2 shows the Taylor diagrams for the validation and prediction dataset, but they are a replica of the same information reported in Table 2 (this is why we did not re-discuss these diagrams in detail). For conciseness, we decided to discard Figure 2 in the revised version.

7) Section 4.4: NO2 has the strongest spatiotemporal variability due to its short lifetime. I believe case studies using NO2 can provide the most relevant information about model capability. Why is PM10 discussed here? Should similar results for the other pollutants be included in the supplement?

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. *Predictions on the regular grid is shown for all pollutants*

8) Figure 8: Please 1) add a map of median of raw predictions and 2) add observed values on the maps. Also, how to assess if the predicted values over unmonitored areas are accurate? Line 346 discussed "extrapolation ability", but quantitative evaluation of such ability is missing. Some "spatial-clustered" cross-validation idea (e.g., doi: s41467-020-18321-y) might be useful.

We included the box-wisher plots for all pollutants in the revised paper (see new Figures 3 and 4). This analysis has been made to highlight the performance over the dependence on the type of monitoring station (urban, suburban and rural) and season. 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). The comparison with the validation data set represents a measure of the expected error in forecasting independent data at "unmonitored" locations.
We are sorry but we are unable to resolve the doi number you provided. Missing digits?