Review of Adani et al.: Evaluation of air quality forecasting system FORAIR_IT over Europe and Italy at high resolution for year 2017, Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2020-54, in review, 2020.

Main comments

Adani and co-authors present the FORAIR_IT air quality forecasting system. They provide a detailed evaluation of the modelled concentrations of PM_{10}, $PM_{2.5}$, NO_2 and O_3 for the year 2017. For that purpose, they compare model output over a European (20 km x 20 km) domain and an Italian (4 km x 4 km) domain against surface observations. They also compare their results with those from the CAMS_50 interim ensemble reanalysis of the Copernicus Atmospheric Monitoring Service (CAMS). In addition, they investigate the skill of the model in reproducing the PM_{10} and $PM_{2.5}$ concentrations during a poor air quality episode that occurred in the Po Valley.

The manuscript is of broad interest for many readers of Geosci. Model Dev. Overall, the authors give proper credit to related work. The main structure of the paper (i.e. Section 2, description and set-up of the modelling system; Section 3, evaluation of European and Italian forecasts for 2017; Section 4, evaluation of a PM episode in the Po Valley) is appropriate. I like that in Section 4 they have used some model diagnostics (different contributions to the coarse fraction) and observations (chemical speciation) to try to understand the model performance for PM_{10} and $PM_{2.5}$, and to identify some potential deficiencies in the model.

However, I do not feel that the manuscript is well balanced and think it needs major revisions. These are the main issues I found:

1. The use of language is neither fluid nor precise. The manuscript should be corrected by a person with an advanced level of English. The authors must take this recommendation very seriously. Otherwise I could not recommend the manuscript for publication.

2. While the main structure is appropriate (as indicated above), the overall presentation within the different sections is not clear enough. See specific comments below.

3. In particular, Section 3 (evaluation of the European and national forecast) is long and repetitive. I understand that a first evaluation paper of a modelling system needs to be somewhat lengthy, but this text seems to belong in a technical report (for the authors' own use) rather than in a scientific paper (aimed at the readers). The authors repeat the same structures and analyses for four pollutants, but they hardly provide any insights into the reasons for the discrepancies of the model when compared with observations or CAMS.

4. The reproducibility of the results could be improved. For instance, there is no information on the model version used; the text only contains few details on the vertical and temporal disaggregation of emissions (see comments below), or on the chemistry scheme (are there any changes/updates with respect to the SPARC-99 reference they provide?). Furthermore, they could list the Italian stations used for the evaluation of the PM pollution episode. The paper is quite long already so some of this information could be in a supplement.

Specific comments

(1) The first paragraph of the introduction is very long. One could easily split it into shorter paragraphs.

(2) Some parts of Section 2 (Description and set-up of the forecasting system) should be better organised:

* I would break the paragraph on emissions around line 9 of page 6 (i.e. after "velocity scale, soil type and soil moisture by means of a climatic index").

* I would continue with another paragraph on the FORAIR_IT operational chain (line 15 of page 7)

* Then I would include all the information on model evaluation into one or two paragraphs (e.g. one for observations and another one for CAMS). Note that in the current version the authors start describing the observations used for evaluation in the same paragraph where they describe emissions (line 9 of page 6). Mixing different things within the same paragraph does not look good.

(3) Section 3 (evaluation of forecasts for 2017). I think it needs to be completely rewritten. It is very long and descriptive. I provide some comments below:

3.1. The text introducing Figures 2-5 is rather messy and hard to follow (pages 7-8).

3.2. Line 9 (or second line) of page 8: "Values are refferred to daily maximum (for O_3 and NO2) or daily mean (PM_{10} and PM2.5)". First, change "refferred" to "referred". The authors could justify why they have used the daily maximum instead of MDA8 in the case of O_3. Please also specify if you have used the daily maximum of the hourly values in the case of NO_2. One should indicate if the use of daily means, daily max and so on follows European or Italian recommendations / legislation.

3.3. The subsections on PM_{10} and $PM_{2.5}$ are very repetitive because both pollutants follow similar patterns. I feel it would be better to focus on the evaluation of one of them (e.g. PM_{10} in Figure 2 and Table 2). Then I would refer to Figure 3 and Table 3 only to highlight the main differences in the model performance for PM_{10} and $PM_{2.5}$. One could also do it the other way around, i.e. focus on $PM_{2.5}$ and then explain the differences with PM_{10}. With this one could lighten the text considerably, while clearly illustrating the main model deficiencies and the differences between both pollutants: e.g. better metrics for $PM_{2.5}$ than for PM_{10}, more negative bias for PM_{10} (attributable to difficulties in the representation of the coarse mode), more negative biases for urban than for rural sites (related to representativeness of the stations and model resolution), etc.

3.4. A few things about Figures 2-5:

* First of all, please increase the font size in all panels. Indicate whether the pollutant is "daily PM_{10}", "maximum daily NO_2" and so on in the figure captions.

* Extend the caption of some of the Figures (e.g. Fig 2) to indicate what you show and for which species (daily PM_{10} in that case). Then, shorten other captions like this: "As Fig 2 but for daily $PM_{2.5}$", and so on.

* I do not find it very useful to show the time series of some metrics considering D0, D1 and D2. First, with this everything looks quite small in those panels, which makes it very hard to follow the text about the diurnal cycles. Second, the differences for D0-D1-D2 are perfectly summarised in Tables 2-5. In those panels of Figures 2-5 I would simply show the daily cycles of those metrics for D0 together with those of CAMS.

3.5. There are many tables: Tables 2-5 for "EU+IT", and Tables A1-A8 for either "EU" or "IT". That is too much information to digest. I would never use more than 8 tables: 4 to summarise the model performance over the European domain and other 4 over the Italian domain. All captions should be improved. For instance, the text in Table 2 reads "PM10 EU+IT" while it should be something like "Evaluation of model performance metrics for daily PM_{10} over the European domain". Need to indicate units for some statistics (RMSE and mean of observations).

3.6. Need to remove Appendix 4 (Evaluation in reproducing the exceedances) because the authors do not provide any explanations in the text. That section could be useful with some appropriate explanations, but I believe the text is already quite long.

3.7. As indicated above, large parts of text are neither appealing nor useful for the readers. Overall, the text is extremely descriptive and quite often does not help understand the model deficiencies. Here are just some examples about some parts of the text that could be improved:

* Description of diurnal cycles of RMSE and NMB for PM_{10} (lines 15-31 of page 10). The last sentence there reads: "This difference between the daily cycles of MMB may be due to differences in daily disaggregation of the emissions used for CAMS_50 ensemble models and FORAIR_IT, which is based on EMEP temporal profiles". Please simplify the descriptions and provide some information about the temporal profiles of FORAIR_T as well as those of CAMS (if available). The authors need to discuss the injection of surface emissions (only from lowest model level?) and the diurnal emission profiles applied, to understand the average hourly evolution of RMSE and bias.

* Need to consider the comment above (i.e. provide information about vertical and time profiles of emissions) to try to document the reasons for the diurnal cycles of RMSE and bias in the case of NO_2 and O_3. It is not clear what one can learn from all the long descriptions provided in Sections 3.3 and 3.4, because there is no clear attribution of model biases to the treatment of processes in the model (i.e. emissions, vertical mixing, chemistry). Note also that the diurnal cycle of each pollutant also needs to be considered to understand the diurnal evolution of some metrics (e.g. RMSE) and that the O_3 biases could be put in the context of the NO_2 biases.

* Beginning of page 14: Discussion on the monthly RMSE values for $PM_{2.5}$ What do we learn from that seasonal cycle? I wonder myself whether the high RMSE values in winter are simply due to higher $PM_{2.5}$ concentrations during those months than in the rest of the year.

* Line 15 (comment about the $PM_{2.5}$ biases): "Table 3 shows that model performances don't change so much as function of forecast day. Correlation skills slightly deteriorate along with the forecast time. On the contrary, BIAS tends to be reduced from D0 to D2". Rewrite. The bias becomes less positive for rural sites and more negative for both suburban and urban sites. The absolute value of the bias increases for the latter types of sites.

* The evolution of monthly RMSE values in the case of NO$_2$ is discussed at the beginning of page 18. I would expect the high values in winter to be at least partly due to the high concentrations during that season, as indicated above for PM$_{2.5}$, but that is not mentioned there. Then, one has to wait to read lines 32-33, where the authors indicate that the errors are lower in summer because of low concentrations. The authors need to be more focused and reorganize the text.

* As indicated above for PM, please do brief comparisons of the model performance from rural to urban sites for both NO$_2$ and O$_3$ (Tables 4-5), keeping in mind the grid size.

* Section 3.4. Why not just focusing on the evaluation of ozone during an extended summer period (e.g. April - September) rather than during the whole year. I think that producing statistics during summer would be useful, while combining cold and warm periods could lead to results which may be hard to interpret.

* First paragraph of page 22 (discussion of the months with high RMSE values for O$_3$): "Conversely the winter one is quite unusual; anyway it is present both in FORAIR_IT and CAMS_50 outcomes and it may be related to high concentrations of chemical reactive compounds in Eastern Europe that caused anomalous high concentrations of NO$_2$ and particulate matter, in the first months of the year". The authors are speculating. I think they could easily check that. Anyway, that might not be needed if they focus on the evaluation of O$_3$ during the extended summer.

* Around line 27 of page 22: "Since RMSE is composed of two terms (the BIAS component and the Centered Root Mean Square Error, CRMSE) ... This latter metric summarizes in turn the model capacity of capturing the observed standard deviation and the correlation ...". Further below, the authors attribute most of the error associated with CRMSE to correlation, but it is not clear to me how that is done. CRMSE is not a very common metric. Shouldn't the formulation and definition of CRMSE be included in appendix A1? Or maybe just provide a reference in that part of the text?

* Just within the same paragraph of page 22: "... the amplitude of the standard deviations is quite well captured since the median of the skill variance (ratio between the observed and the simulated standard deviations) ...". Again, shouldn't this metric be included in section A1 and applied to other pollutants apart from O$_3$? Please see also comment 3.9 on the potential importance of this metric in the comparison of the performance of FORAIR_IT to that of CAMS.

3.8. Need to use different paragraphs to clearly separate the results for the European and Italian domains (see e.g. page 10, line 38: "Concerning ...").

3.9. It might be useful to summarise some aspects of the FORAIR_IT modelling system (over the Italian domain) that tend to outperform CAMS. For instance, is the ratio of modelled-to-observed standard deviations (based on daily concentrations) closer to 1 in FORAIR_IT than in the CAMS ensemble? I would expect that to be the case if the latter smooths data from different models, but I do not know if the term "model ensemble" used here refers to model mean or to model median.

3.10. The authors could summarise their results by indicating whether the skill scores – e.g. RMSE, NMB, correlation and even FAC2 (fraction of predictors within a factor of two of the observations) – are within the performance criteria recommended by the literature for regional models. Not sure if this would help or make the text too long …

(4) Section 4 (PM episode in the Po Valley).

4.1. Around line 15 of page 24: "During October 2017, a 10 day outbreak ...". Here the authors should give some details about the typical PM levels during the episode. For instance, one could refer to the top panel of Figure 6 to indicate the average PM_{10} concentrations in the model and to Figure A1 to briefly describe the daily evolution of PM_{10}. From those descriptions it should be clear for the reader that the PM_{10} plume developed in Lombardy and mainly affected the north of Emilia Romagna. Then I would focus on the evaluation of PM_{10} and $PM_{2.5}$.

4.2. There are three types of symbols on the top panel of Figure 6: filled squares, filled circles and open squares. From the explanations provided in the figure caption and on page 24 it is not clear what each symbol represents.

4.3. There is a very long paragraph on page 26. Split into several paragraphs.

4.4. Lines 15-20 of page 26: "The model average forecast .... This high gradient area may be particularly challenging for the model and specific analysis were then conducted to analyze its variability over the period under consideration". Why not overplotting circles with the average PM_{10} concentrations at the locations of the sites in Figure 6 (same colour scale as modelled PM_{10}) to get an overall impression of whether the model captures the observed concentration gradients? One could also do the same for the daily concentrations (Figure A1), even if not shown because the circles would look too small in those panels.

4.5. Around line 26 of page 26: "The peculiar topographical conformation of the Po Valley, surrounded by complex orography, could have lead the meteorological model to forecast too intense zonal wind ...". Have the authors carried out any evaluation of the meteorology in the model? By the way, change "lead" to "led".

4.6. Towards the end of page 26: "Relative contribution of boundary intrusion along with soil and sea salt resuspension is also shown with black, blue and cyan dots, respectively". Need to make it clear that these are model diagnostics. In addition, it is not clear what the term "boundary intrusion" means; change it to something like "intrusion of PM_{10} from the model boundaries".

4.7. What are the main conclusions that can be drawn from all the analyses regarding the underestimation of the coarse mode? Low emissions in the model over Emilia Romagna or in other areas of the Po Valley? Underestimation of dust resuspension or other processes? What about pollution entering through the boundaries? Some notes on this:

* Figure 7 (on page 27) is very useful to understand how good the model is at representing processes related to the coarse aerosol mode, but I don't know how it helps understand the October episode. For instance, the coarse fraction of PM_{10} coming from the boundaries does not seem to be high during October (no black circles in the figure in that month), when the coarse fraction is clearly underestimated; however, this might not be very relevant because during the episode PM_{10} pollution seems to come mostly from Lombardy (panels in Figure A1), which is within the inner model domain. Some discussion on this could help.

* Page 28: "Trying to quantify the underestimation of the coarse fraction of PM_{10} in the anthropogenic emission inventory, among the different causes, if we consider a closed box of the

same surface of the ER region (where this emission is about 3300 ton/y), with the same thickness of the first model layer (40 m) within which no removal, advection and transformation processes happen, the ER coarse fraction of PM_{10} will contribute with a concentration of 0.4 μg/m3". This sentence is very long, please rewrite. In addition, as I indicate above, most of the pollution seems to come from Lombardy during the episode. Shouldn't the authors comment on the importance of Lombardy emissions and advection to Emilia Romagna?

4.8. Towards the end of this section or in the following section (Sect 5, Summary and conclusions) one could summarise the main aerosol processes that could be improved in the model: e.g. insufficient emissions in Northern Italy?, issues with dust resuspension or other processes contributing to the coarse mode?, improve processes related to nitrate and ammonium chemistry during PM episodes?

(5) Section 5 (Summary and conclusions). As indicated in the previous comment, one could summarise some of the processes (and potential model developments needed) that could be improved in the model: e.g. as in many other modelling systems, there are obvious issues with the coarse mode (what could be improved?); modelling of nitrate during episodes; emissions and vertical mixing for O_3 and NO_2?, ...

Technical corrections

As indicated above, a person with an advanced level of English must revise the use of the language throughout the manuscript.

Some minor technical corrections are listed here:

* Use subscript for 10 and 2.5 in PM_{10} and $PM_{2.5}$.

* Change "station area" to "station type" everywhere.

* Page 1, line 21: Pope et al., 2019 --> Pope III et al., 2019

* Page 3, line 1: D'Allura et al. --> D'Allura et al., 2018

* Page 4, caption of Figure 1: "2017 valid observations for each evaluated pollutant" could be changed to something like "observations used for the evaluation of four pollutants in 2017".

* Page 5, line 38: "stack and flue gas-PM parameters". This could be hard to follow because you have not spelled out PM before.

* Towards the end of page 7: "The group of panels a of Figs. from 2 to 5 and group of panels b of Figs. 4 and 5 present the same skills for each pollutant, within a common layout". This is hard to follow. Rewrite.

* Last lines of page 7. This is a good example of a long sentence that can be simplified: "The skill scores considered are RMSE, Modified Mean Bias and Correlation Coefficient (CORR) in panels a2(b2), a3(b3) and a4(b4) as a function of forecast day (D0, D1, D2),respectively, whose definition is

reported in Appendix A1 of the Supplementary Material". Rewrite and split into a couple of sentences. This can be done in other parts of the text too.

* Last lines of page 24: "During this period, PM10 concentration underestimation is partly compensated by an overestimation of PM2.5". It should be "in contrast with", not "partly compensated by". I know what the authors mean (i.e. that the overestimation of $PM_{2.5}$ compensates for the underestimation of the coarse mode), but they should rewrite the whole sentence to make their point clear.

* Page 31, line 23: "forecast performances are measured" --> "forecast performance is measured". There ae other parts of the text where plural can be changed to singular, but I am not supposed to rewrite the manuscript …

* Page 31, lines 23-24: "CAMS_50 ensemble members in the quarterly evaluation report". Can you provide a reference?

* Page 31, line 25: Change "lack" to "deficiencies in "the lack of the modelling system"

* Figure A1: Please extend the text to describe what you show there. "Daily mean concentration of PM10" is not very informative.

* Reference list: Provide web address for all technical documents and reports when possible.

* Reference list: Remove Nabielek et al. (2016). It is not mentioned in the main text.