Comment on gmd-2020-433

Anonymous Referee #1

Referee comment on "Air Control Toolbox (ACT_v1.0): a flexible surrogate model to explore mitigation scenarios in air quality forecasts" by Augustin Colette et al., Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2020-433-RC1, 2021

The authors are presenting here the first version of the Air Control Toobox that allows flexible emission scenario analysis based on a sparse ensemble of CHIMERE emission scenario simulations. The approach is interesting and certainly useful for analysing pollution episodes, although it is true that having the emission reductions applied uniformly over the entire Europe remains a limitation for more practical uses. The paper also nicely illustrates how the ACT tool can be used for analysing the source allocation and chemical regimes. I think it should be accepted for publication after addressing the few minor comments below.

Main comments :

Abstract and introduction : Overall, the paper is well written but the authors should introduce more clearly the objective of their study and the general form of the tool they aimed to build. While reading the paper, it took me some time to understand precisely which type of tool they were developing. One reason may be that in the abstract and introduction, the authors are often talking about a "surrogate model" to tackle the complexity of geophysical chemistry-transport models (CTM). I was a bit confused at the beginning since this tends to suggest that they want to build a CTM surrogate model (able to predict concentrations based on emissions) while actually they want to develop an emission scenario surrogate model (able to predict the concentration changes in response to emission changes). Although some sentences indeed refer to the later, some others refer to the former, which maintains the confusion (e.g., P2L6 : "an emulator, or surrogate model, of a comprehensive air quality model"). While reading the paper, we progressively understand what the authors were referring to before, but the first part of the paper could be clarified so that things are clear earlier in the text. I would personally understand better and easier the goal of their study by introducing it in such a way :

The overall objective of the authors is to build a tool for conducting flexible emission scenario analysis on an every-day basis, for exploring the impact of different mitigation strategies on the current forecast. More specifically, they want the tool to estimate the impact of a given set of emission reductions (in this case, looking at the options available for the user on the ACT web interface, 21 emission scenarios of emission reduction bins of 5%, namely 0%, -5%, -10%..., -95%, -100%) applied to a given set of emission sectors (in this case, 4 emission sectors). Simulating explicitly all possible combinations would require a total of $21^4=194,481$ emission scenario simulations (or $101^4=104,060,401$ simulations with emission reduction bins of 1%), which is obviously not feasible for computational reasons. This motivates the need for developing, in each cell of the domain and for each day of forecast, (1) a regression model able to estimate the pollutant concentration responses to different emission reductions based on (2) a reasonable number of emission scenario simulations that could be run in parallel to the reference CTM simulation on an every-day basis.

Then, this naturally justifies the type of regression model chosen. Indeed, although more complex regression models can eventually achieve better accuracy, they require larger training datasets (in this case, more emission scenario simulations). As a compromise between these two aspects, this motivates the choice of multivariate second-order polynomial models with interaction terms able to capture some of the non-linearities existing between concentration changes and emission change (at least in some emission sectors) while remaining simple enough, in the sense that only a few coefficients need to be fitted, which allows relying on a limited number of emission scenario simulations.

Section3 : I agree with the first reviewer that section 3 could be shorter. Also, it is not very clear to which extent the step-by-step approach - (1) univariate first/second/third order model, (2) bivariate second order model with interaction terms, (3) quadrivariate first/second order model with interaction terms - is followed only for pedagogical purposes or if (at least part of) the methodology is really sequential (it seems so). Please clarify. If I understand correctly, pairs of emission sectors with most important interactions are chosen in section 3.2 based on different two-variable polynomial regression models, but how this ensures that it remains then the most appropriate choice in the four-variable polynomial models developed in section 3.3? The authors acknowledge that "by doing [such a step-by-step approach], [they] assume that the optimal structure and training scenario remains valid when including interactions whereas there is a possibility that the addition of an interaction term could change the selection of univariate terms." However, in the type of model proposed in section 3.3, the authors keep only the interaction terms selected before. Also, to test this final model, they retain only the subset of emission scenario selected before. These different methodological aspects should be clarified.

I wonder to which extent not considering any emission scenario with reductions applied to 3 or 4 sectors simultaneously (besides the scenario with all 4 sectors reduced by 100%), for the training or at least the testing is a strong limitation in this study. This should be discussed (and ideally tested).

Additional comments :

P1L13 : "flexibility" : please be more specific on which type of flexibility you are talking about (here, flexibility on the emission forcing)

P4L11 : "Note however that those periods do not constitute a specific training period for the surrogate model which is intended to be re-fitted to new CTM simulations every forecasted day in an automated machine learning approach." Do the authors mean that all regression models in this study are trained only based on the current forecast day? Please clarify. In total for this study, the authors have performed 46 CHIMERE simulations for each of the 4 months mentioned in Section 2.1, is it correct?

P4L25 : "operational analyses of the IFS" : not IFS forecasts?

P5L20 : "we conclude that the surrogate model will be at most a third order polynomial, less if interaction terms are accounted for." Please clarify that this is motivated by the number of coefficients to be fitted.

P6L4 : "emissions are applied uniformly over Europe" should be "emission reductions…"

P7L16 : "both dates" : which dates are you referring to?

P9L1 : "testing and validation" should be "training and testing"

FIG3 : Please describe in the legend the meaning of the numbers into brackets (I guess they correspond to the minimum and maximum error over the domain?)

P10L1 : "perform" should be "performs"; "largest emission reduction" should be "largest emission reduction"

P11L6 : "yield" should be "yields"

Sect. 3.1 : Results are discussed for PM10 and briefly mentioned for O3max but not for the other pollutants

Table 2 : a comma is missing between IND30 and IND60

P12L10 : It seems to me the authors are actually talking about lines 3 and 4 where no interactions are included in the training set, please correct if needed.

All formulas : In order to make formulas easier to read (especially the longest ones), I suggest to introduce another variable such as $DELTA^{sector}=EPSILON^{sector}-EPSILON^{ref}$.

P14L13 : "This final model is tested the" should be "This final model is tested against the"

P17L6 : "ammonia emissions December is" should be "ammonia emissions in December are"

P18L4 : "the allocation we display here" : where?

P18L14 : "assess" should be "assessed"

ACT web interface : Why only O3 is proposed and not O3max or O3avg? Also, the source allocation model is useful for analysing pollution episodes and I would suggest to include it on the ACT web interface in the future.