Atmospheric Chemistry and Physics Discussions

Atmos. Chem. Phys. Discuss., referee comment RC1
https://doi.org/10.5194/acp-2021-864-RC1, 2022

Comment on acp-2021-864

Anonymous Referee #1

Referee comment on "Model output statistics (MOS) applied to Copernicus Atmospheric Monitoring Service (CAMS) O_3 forecasts: trade-offs between continuous and categorical skill scores" by Hervé Petetin et al., Atmos. Chem. Phys. Discuss., https://doi.org/10.5194/acp-2021-864-RC1, 2022

Review of

"Model Output Statistics (MOS) applied to CAMS O3 forecasts: trade-offs between continuous and categorical skill scores"

by Petetin et al.

Overview

The paper compares different MOS methods applied to the regional CAMS forecast at stations locations over the Iberian Peninsula for ozone in 2018-19. It uses an "operational scenario approach", namely that observations become only gradually available to learn the different methods. The paper finds that the MOS approaches have different strength and weaknesses with respect to improvements of the overall reduction of the forecast error and the error specifically for high pollution episodes and threshold exceedances.

General comments

The paper reports about a sound scientific effort, in particular the consideration of accuracy measures in general and for exceedances of threshold during episodes of high

pollutions is very welcome. But, some methodological aspects need to be re-considered. Also, the result, method and choice of accuracy measures are not explained with enough detail, which would be required to better understand the results and applicability of the different approaches. The paper does not present well the large amount of information coming from the combination of the many MOS approaches and accuracy measures. The authors need to introduce tables showing the accuracy measures for each MOS type, which allows the reader to digest the information. Tables could also help to substantially shorten the long narrative descriptions.

On the other hand, sensitivities to input parameters and variation of the methods are discussed with some detail, which make the paper somewhat unbalanced. Although interesting in itself, it is also not clear what the purpose of that section is. Are the results presented in 3.3 and 3.2 already carried out with the optimal choice of parameter setting or not ? The discussion in 3.4. should be shortened by focusing on application with a very high sensitivity to parameter choice.

It remains unsatisfactory to treat the persistence approach as a variant of MOS. As the author explain themselves, persistency is a reference forecast (to identify if a given forecast has skill compared to the reference) and both the RAW as well as the other MOS approaches should be more directly compared against it. An important question for all forecast application is, if RAW beats PERS (depending on the accuracy measure) and if and how MOS (using RAW) can improve the skill.

The AQ observations are used without discrimination of the representativeness for the scale of model grid boxes of the regional ensemble (10km). One would expect that some stations (i.e. rural, urban) are more representative than others (i.e. traffic). It is a missed opportunity of the paper to discuss the amount of correction by MOS for the different air quality observations stations based on the station type.

The assumption about an operational scenario (observations become gradually available after the start of the application on 1.1.2018) is in principle a welcome approach but several questions remain. It is unclear how different spin-up times (i.e. the time until further improvements by adding more previous data become very small) of the methods, which should also be stated more clearly, are taken into account in the evaluation. Second, it remains unclear what happens in the case, that observations are not available in near-real-time to be fed in to the MOS scheme. Consequently ,it is more important from an operational point of view to apply MOS approaches for the case that observations are always available in NRT or that they are not, which means that these MOS approaches could only be trained with past data. The latter is a typical cross-validation approach, which uses one data set to train and the other to evaluate the MOS. The impact of missing data needs to be discussed in more detail.

It does not make sense to use ER5 as a reference meteorological data set with respect to the HRES NWP forecast in this application. The HRES (IFS) forecast (9km) should be compared against HRES analysis that were the initial conditions of the forecast (step=0) (Both HRES and ER5 are produced with the IFS)

The graphical representation (Figures) needs to be improved. Choice of the colour range in maps and choice of colour in time series plot make it often impossible to discern the different data sets. Various aspects of Fig 3 remain unexplained.

Please summarise the result of 3.2, 3.3 and 3.4 in tables. That will shorten the paper and make it possible to compare the different results more easily.

Some specific comments:

Abstract

Please quantify the achieved improvements by MOS to replace or justify phrases such "can be substantially improved"

L67-71 A summary of the results of the paper is not required in the introduction

L 98 mention forecast start time

L 101 Flemming et al. 2015 is not a reference for the operational NWP forecast of ECMWF

L 120 Please consider the general comment about NRT availability of observations

L Please clarify, if model output is required for the PERS and MA approach. Please use the model independent methods as reference (see general comments) and not as an other MOS variant.

L 145 Can the choice of the length of the adjustment period (30 days) be substantiated ?

L 155 KF and other method are based on unbiased linear estimates (BLUE) So, the biases are not addressed in KF theory in general. Please clarify.

L 180 Please clarify “best analogue days”. How many days are required to get a spun-up AN (10) method.

L 209 Please motivate the choice of the 30 day training period.

L 233 Missing here is a skill score that assess the forecast skill against the persistency forecast

L 225-233 The amount of accuracy measures is overwhelming an the reader can not easily follow that. Please reduce the number of measures to a minimum and explain what specific characteristic of the forecast performance is quantified by that measure. Try to introduce a nomenclature (say upper case vs lower case, latin vs bold) for name of MOS methods and accuracy measures.

L 266 Please explain Fig 3 in more detail. What do the overlaying symbols mean (one per stations , forecast day ?).

L 277… Please provide the various accuracy measure in a table (also including the MOS results) for a better representation of the results.

L 335-445 Please see my general comment on section 3.4

L448 Using the ER5 data set as “truth” compared to the HRES NWP forecast does not make sense. The HRES analysis should be used for that. Because of different resolution and model cycle the two data sets are not consistent. Please avoid the term IFS for the forecast because both ER5 and HRES are produced with the IFS.

L 446 ER5 and HRES will not differ in the number of assimilated observations, if anything HRES will be better.

L 484 Please provide quantitative information about the improvements.

L 494 The skill of a forecast (in a scientific sense) is defined by the improvements w.r.t to a reference, which should be persistency in your case. How compares RAW and the MOS methods using RAW to PERS is a question that should be answered. See text book by D.S. Wilks, Statistical methods for atmospheric science.

L517 The finding that MOS results (using RAW) were more sensible to forecast lead time than PERS is interesting. One would expect a strong impact of the lead time for PERS. Please elaborate a bit more. Do the forecast show a drift perhaps introduced by the initialisation with analysis (assimilating AQ surface information)

L 575 After all this long discussions, it would be good to still make a recommendation. Which MOS scheme performed overall best and would be recommended for operational implementation.

L 575 Please provide reference for GHOST