We are thankful for the numerous constructive comments and suggestions provided by the two reviewers. Below, the comments of the reviewers are indicated in grey, our answers in black and the modifications of the text in blue.

Reviewer #1

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.

As discussed in more details below, we proposed in the revised version numerous modifications that will hopefully improve the manuscript along the lines suggested by reviewer #1. We added a new section to present the different metrics used (while at the same time reducing the corresponding section in the Appendix) and to introduce the corresponding skill scores to be discussed.

Regarding tables, in the initial version of the manuscript we decided on purpose to avoid including tables essentially because the results shown in Fig. 3 are multidimensional (for each statistical metric : MOS method, lead time, time scale). Thus, to our opinion, tables are not necessarily the most convenient way to present such a large quantity of results as they would not allow the different metrics to be easily compared among each other (as this would require many tables). However, we agree that some tables can facilitate the reading and shorten the discussion. Therefore, we selected a subset of these statistics and included them into two tables. We included additional tables in the Supplement (notably for the different sensitivity tests).

Also, as shown below, we shortened the narrative descriptions while analyzing in more detail the results shown in Fig. 3.

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.

In the revised version of the manuscript, we reorganized the discussion of the results in Sect. 3.2 and 3.3 (see below) and shortened the discussion on the sensitivity tests (section 3.4), which should improve the balance of the paper. Regarding this last section, although it might disrupt a bit the general narrative of the paper, we do think it is important to show the sensitivity of these different methods to their internal parameters. The methods used in Sect. 3.2 and 3.3 are not considering an optimal choice of internal parameters for the simple reason already raised in the discussion that there is no optimal MOS methods without a proper definition of user-specific needs and interests (in other words, a clear choice of the metric of strongest interest). We understand the frustration that the absence of clear recommendations might create but this is one of the conclusions of our study (as reflected in the title we choose): the behavior of the different MOS methods strongly varies with the choice of metric, for instance with some MOS methods showing the best continuous metrics and the worst categorical ones. Therefore, in an operational context, the choice of the MOS method (and its internal parameters) directly depends on the desired behavior of the forecast system (Is the user more interested in having forecasts with lowest bias and error or in predicting exceedances? For this latter category, is the user more interested in avoiding false alarms at the potential cost of missed episodes, or is he interested in a more precautionary approach where more false alarms are accepted?).
We applied the following modifications in the introduction of section 3.4 :

- L333 : "In the previous sections, we provided a first evaluation of the performance of a set of MOS methods. All methods rely on specific choices or parameters that can substantially influence the behavior of the MOS-corrected forecasts, and thus its general performance. In this section, we discuss some of these choices and investigate their impact on the performance through different sensitivity tests." ➔ "Each of the forecast methods considered in this study relies on a specific configuration, e.g. the time window of PERS or MA methods, the metric used internally in KF for optimizing the variance ratio, the number of analogs taken into account in AN, the choice of input features or metrics used internally for fitting the ML model in GBM. This configuration can substantially influence their general performance, although in a different way depending on the metric used. In the previous sections, we evaluated the performance of these different methods considering a relatively simple baseline configuration. In this section, we discuss some of these choices and investigate their impact on the performance through different sensitivity tests. Corresponding statistical results on continuous and categorical metrics are given in Tables in the Supplement."

And we slightly shortened some of subsections :

- L338 : “The persistence method essentially relies on the choice of the time window over which past observations are averaged to provide the O3 forecast. In the previous section, we used a window of 1 d. A sensitivity test is performed with windows ranging between 1 and 10 d (hereafter referred to as PERS(*n*) with *n* the window in days). Results are shown in Fig. G1 in Appendix G, and indicate that, while PERS(1) forecasts were unbiased (whatever the time scale), increasing the window leads to a growing negative bias on d1max and d8max scales. The bias is substantially reduced when working at dd1max and dd8max scales, i.e. when applying the PERS approach directly on daily 1-hour and 8-hour maximums rather than on the hourly time series. The differences between the two approaches originate from the day-to-day variability in the hour of the day when O3 mixing ratios peak. For illustration purposes, let's assume that O3 peaks between 15 and 17 h; on a given day, O3 mixing ratios at 15/16/17h reach 50/60/50 ppbv and on the following day 70/70/80 ppbv. Then, the $PERS(2)_{dd1max}$ O3 would be 70 ppbv (mean of 60 and 80 ppbv), while the $PERS(2)_{d1max}$ O3 would be only 65 ppbv (maximum of the mean diurnal profile of these two days, in this case 60/65/65).

  Conversely, both RMSE and PCC can be slightly improved with longer windows. However, averaging past observations over more days reduces the variability, which was unbiased in PERS(1)), thus introducing a substantial negative nMSDB. As a consequence, both H and F are slightly reduced, which means that PERS forecasts become more "conservative" with longer windows. The impact on SR for detecting exceedances of the target threshold is ambiguous for short lead times but positive for the longest ones. Interestingly, for information thresholds, the best SR are obtained around 4-7 d. However and more importantly, using longer windows deteriorates the general performance of the forecast, as shown by the decrease of both CSI and PSS. This deterioration is stronger in the first lead days, and softer during the last ones. Interestingly, there are also important differences in terms of AUC for detecting exceedances of the target threshold depending on the lead day, ranging from a decrease of AUC with longer windows at D+1 to an increase at D+4.”

  ➔ “The persistence method with a 1-d time window (PERS(1)) provides a reference forecast for assessing the skill scores on the different RAW and MOS-corrected forecasts. Here we explore how the time window, from 1 to 10 d (hereafter referred to as PERS(*n*) with *n* the window in days), impacts the performance of this PERS forecasts. Results are shown in Fig. G1 in Appendix G.

  Increasing the window leads to a growing negative bias on d1max and d8max scales that can be substantially reduced when working at dd1max and dd8max scales, i.e. when applying the PERS approach directly on daily 1-hour and 8-hour maxima rather than on the hourly time series. The differences between the two approaches originate from the day-to-day variability in the

hour of the day when O3 mixing ratios peak. For illustration purposes, let's assume that O3 peaks between 15 and 17 h; on a given day, O3 mixing ratios at 15/16/17h reach 50/60/50 ppbv and on the following day 70/70/80 ppbv. Then, the $PERS(2)_{dd1max}$ O3 would be 70 ppbv (mean of 60 and 80 ppbv), while the $PERS(2)_{d1max}$ O3 would be only 65 ppbv (maximum of the mean diurnal profile of these two days, in this case 60/65/65). Conversely, both nRMSE and PCC can be slightly improved with longer windows, but at the cost of a growing underestimation of the variability. As a consequence, both H and F are slightly reduced, which means that PERS forecasts become more "conservative" with longer windows. The impact on SR for detecting exceedances of the target threshold is low for short lead times but positive for the longest ones. Interestingly, for information thresholds, the best SR are obtained around 4-7 d. However and more importantly, using longer windows deteriorates the general performance of the forecast, as shown by the decrease of both CSI and PSS, especially at short lead times. Interestingly, there are also important differences in terms of AUC for detecting exceedances of the target threshold depending on the lead day, ranging from a decrease of AUC with longer windows at D+1 to an increase at D+4."

- L377 : "As explained in Sect. 2.3.3 (and Appendix B), the behaviour of the KF intrinsically depends on the $sigma_{eta}^2/sigma_{epsilon}^2$ ratio chosen. So far, this parameter has been adjusted dynamically (and updated regularly) to optimize the RMSE on past data. Here, a sensitivity test is performed with alternative strategies in which the variance ratio is chosen to optimize the SR, CSI, PSS or AUC with threshold values of 60 or 90 ppbv (hereafter referred to as SR-60, SR-90, CSI-60, CSI-90, PSS-60, PSS-90, AUC-60 and AUC-90). The objective is to investigate to what extent tuning the KF algorithm with appropriate categorical metrics allows improving the exceedance detection skills. Results (Fig. G3 in Appendix G) show that this tuning strategy barely impacts the performance obtained on continuous metrics, except for CSI-60 and PSS-60 that show slightly deteriorated RMSE and PCC. In return, the latter offer some PSS/CSI improvements compared to KF(RMSE) regarding the detection of target threshold exceedances, but these are mostly restricted to the first lead day. The improvement is stronger for the detection of the information threshold exceedances and extends further in lead time, especially for PSS-60. Surprisingly, a better performance on the detection of the 90 ppbv threshold is obtained with KF(PSS-60) compared to KF(PSS-90). The reasons for this unexpected result are not clear but may include the fact that optimizing KF based on the metric PSS-90 relies on much fewer events compared to PSS-60, which introduces more instability for rare events. Indeed, a common and well-known issue of PSS (as well as CSI and most other categorical metrics) is that it degenerates to trivial values (either 0 or 1) for rare events : as the frequency of the event decreases, the numbers of hits (a), false alarm (b) and missed exceedances (c) all decay toward zero but typically at different rates, which causes the metric to take meaningless values (either 0 or 1 in the case of PSS) (Jolliffe et al., 2011, Ferro et al., 2011). It is not entirely clear if we are already in a regime of rare events here but this potential issue may explain part of the results obtained here,

although further analysis are required to clarify this point. With KF(PSS-60), PSS at D+1/D+4 reaches about 0.17/0.05, against 0.02/0.01 for KF(RMSE). Therefore, the performance for detecting such high O3 concentrations remains very poor, especially far in time, but this sensitivity test demonstrates that choosing an appropriate tuning strategy can help slightly improving the detection skills at a potential cost in terms of continuous metrics."

➔ "As explained in Sect. 2.3.3 (and Appendix B), the behavior of the KF intrinsically depends on the $sigma_{eta}^2/sigma_{epsilon}^2$ ratio chosen. So far, this parameter has been adjusted dynamically (and updated regularly) to optimize the RMSE on past data. Here, a sensitivity test is performed with alternative strategies in which the variance ratio is chosen to optimize the SR, CSI, PSS or AUC with threshold values of 60 or 90 ppbv (hereafter referred to as SR-60, SR-90, CSI-60, CSI-90, PSS-60, PSS-90, AUC-60 and AUC-90). The objective is to investigate to what extent tuning the KF algorithm with appropriate categorical metrics allows improving the exceedance detection skills.

Results (Fig. G3 in Appendix G) show that this tuning strategy barely impacts the performance obtained on continuous metrics, except for CSI-60 and PSS-60 that show slightly deteriorated RMSE and PCC. Only small differences are also found on target threshold exceedances, except again with these two methods that show slightly improved CSI/PSS at short lead time. Results on information threshold exceedances show more variability depending on the time scale, but both CSI and PSS can typically be improved when used internally in the KF procedure, although often only at short lead times. The choice of the threshold in this optimizing metric leads to more ambiguous results. For instance, besides giving the best PSS on target threshold, KF(PSS-60) also gives better results than KF(PSS-90) on the information threshold. Reasons behind this behavior are not clear but may be due to some instabilities brought into PSS-90 by the rareness of such exceedances. Indeed, a common and well-known issue of PSS (as well as CSI and most other categorical metrics) is that it degenerates to trivial values (either 0 or 1) for rare events : as the frequency of the event decreases, the numbers of hits (a), false alarm (b) and missed exceedances (c) all decay toward zero but typically at different rates, which causes the metric to take meaningless values (either 0 or 1 in the case of PSS) (Jolliffe et al., 2011, Ferro et al., 2011). All in all, the performance for detecting such high O3 concentrations remains very poor, especially far in time, but this sensitivity test demonstrates that choosing an appropriate tuning strategy can help improving slightly the detection skills at a potential cost in terms of continuous metrics."

- L400 : "The AN method identifies the closest analog days to estimate the corresponding prediction, and thus depends on the number of analog days taken into account. We performed a sensitivity test with 1, 5, 10, 15, 20, 25 and 30 analog days (hereafter referred to as AN(N) with N the number of analogs). Results are shown in Fig. Fig. G4 in the Appendix G. Increasing the number of analog days up to 5 (AN(5)) positively impacts PCC but

deteriorates it when more days are included. It also increases the negative bias affecting the variability (nMSDB), which leads to a worse slope and intercept. Concerning the detection of target threshold exceedances, increasing the number of analog days logically makes the forecast more "conservative" (lower H and F), although the best SR are found with a number of analogs around 20. However, best CSI and PSS are obtained with lowest numbers of analogs (1 in this case). When focusing on information threshold exceedances, the AN forecasts based on 10 analogs or more never reach such high O3 values. Therefore, similarly to PERS and MA methods that reached their best skills for the shortest time windows, with AN the best CSI and PSS skills are obtained when using the lowest number of analogs (with a cost in the continuous metrics, as for PERS and MA). Computing the AN-corrected O3 mixing ratios based on a larger number of analogs gives smoother predictions, and our choice to weight the average by the distance to the different analogs is unable to substantially mitigate this issue."

➔ "The AN method identifies the closest analog days to estimate the corresponding prediction, and thus depends on the number of analog days taken into account. We performed a sensitivity test with 1, 5, 10, 15, 20, 25 and 30 analog days (hereafter referred to as AN(N) with N the number of analogs). Results are shown in Fig. G4 in the Appendix G.

Although the best slopes are found with smallest number of analogs, the best nRMSE and PCC are obtained using around 5-15 analogs. Using too numerous analogs increases the underestimation of the variability and deteriorates the slope. Regarding the detection of target thresholds, increasing the number of analogs makes the forecast more "conservative" (lower H and F, higher SR) and deteriorates the CSI and PSS. When focusing on information threshold exceedances, the AN forecasts based on 10 analogs or more never reach such high O3 values. Highest CSI and PSS are finally obtained with one single analog.

Therefore, similarly to PERS and MA methods that reached their best skills for the shortest time windows, with AN the best CSI and PSS skills are obtained when using the lowest number of analogs (with a cost in the continuous metrics, as for PERS and MA). Computing the AN-corrected O3 mixing ratios based on a larger number of analogs gives smoother predictions, and our choice to weight the average by the distance to the different analogs is unable to substantially mitigate this issue."

- L430 : "~~In this context, it appears interesting to evaluate to which extent the performance is altered when not relying on this specific information.~~ Results are shown in Fig. G5 in the Appendix G."
- L436 : "Regarding the skills for detecting d8max O3 above 60 ppbv, stronger weights typically increase both H and F, improve the (underestimated) FB, but deteriorate the SR and AUC (the forecasts become more liberal). Regarding the more balanced metrics (of strongest interest here), adding more weights on the tails of the O3 distribution has a positive although small impact on PSS. A minor positive impact is also found on CSI, but the best

results are obtained with GBM(W2), thus moderate weights. For both metrics, improvements are most obvious at the d8max scale, while changes at the dd8max scale are much smaller. Regarding the detection of d1max O3 above 90 ppbv, the influence of the weighting strategies is more ambiguous but the detection skills generally remain very poor. Again, the strongest CSI or PSS improvements are obtained at the d1max scale with much lower changes of the dd1max results." ➔ "Regarding the skills for detecting target threshold exceedances, stronger weights typically increase both H and F, improve the (underestimated) FB, but deteriorate the SR and AUC (the forecasts become more liberal). Regarding the more balanced metrics (of strongest interest here), adding more weights on the tails of the O3 distribution typically has a positive although small impact on CSI and PSS. Regarding the detection of information threshold exceedances, both CSI and PSS can also be slightly improved by adding some weight into the GBM, but the performance for detecting such high O3 values remain relatively low. The interest of using the O3 concentration observed one day before is here found to be limited."

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.

We agree with the reviewer regarding the specificity of the PERS method, that cannot be considered as an additional MOS method. In the revised version of the manuscript, we modified the text to avoid confusion regarding this aspect. Concerning the presentation of the results, we consider it is useful to keep showing and discussing the different metrics for the RAW, the PERS and the different MOS-corrected forecast taken individually as done in the initial version. However, in the revised version, we added some extra discussion of the results in terms of skill scores, taking the PERS(1) forecast as a reference. We included a new figure equivalent to Fig. 3 showing the corresponding skills scores.

We applied the following modifications :

- L4 : "In this study, we investigate to what extent AQ forecasts can be improved using a variety of MOS methods, including persistence (PERS), moving average (MA), quantile mapping (QM), Kalman Filter (KF), analogs (AN), and gradient boosting machine (GBM)." ➔ "In this study, we investigate to what extent AQ forecasts can be improved using a variety of MOS methods, including moving average), quantile mapping, Kalman Filter, analogs, and gradient boosting machine, and consider as well the persistence method as a reference."
- L126 : "This section describes the different MOS methods implemented for correcting the raw forecasts (hereafter referred to as RAW), namely: persistence (PERS), moving average (MA), Kalman filter (KF), quantile mapping (QM), analogs (AN) and gradient boosting machine (GBM). All MOS

methods are applied independently on each monitoring station." ➔ "This section describes the different MOS methods implemented for correcting the raw forecasts (hereafter referred to as RAW), namely: moving average (MA), Kalman filter (KF), quantile mapping (QM), analogs (AN) and gradient boosting machine (GBM). All MOS methods are applied independently on each monitoring station. The skill of these different forecasts (including the RAW) is assessed relative to the Persistence (PERS) reference method, which uses the previously observed concentration values at a specific hour of the day (averaged over 1 or several days) as the predicted value. As a first approach, we use a time window of one single day (hereafter referred to as PERS(1))."

- L129 : "2.3.1 Persistence (PERS) and moving average (MA) methods" ➔ "2.3.1 Moving average (MA) methods"
- L130 : "We primarily consider two relatively simple MOS methods: the persistence (PERS) and the moving average (MA). The PERS method simply uses the previous observed concentrations values at a specific hour of the day (averaged over 1 or several days) as the predicted value for this specific hour. It is often used as a reference to measure the skill achieved by other methods, especially for very short-term forecasts. In the MA method, the forecast bias in the previous day or days is used to correct the forecast. As a first approach, we use a time window of one single day for both PERS and MA methods. The corresponding approaches are hereafter referred to as PERS(1) and MA(1). The sensitivity of both PERS and MA methods to the time window is discussed in Sect. 3.4.." ➔ "We primarily consider the Moving Average (MA) method, by which the raw CAMS forecast bias in the previous day(s) is used to correct the forecast. As a first approach, we use a time window of one single day (hereafter referred to as MA(1)). The sensitivity to the time window is discussed in Sect. 3.4".

We added a dedicated section to describe the evaluation metrics and the corresponding skill scores (and removed the corresponding text in L226-236) :

L221 : "**2.5 Evaluation metrics and skill scores**

In this study, O3 forecasts are evaluated using an extended panel of continuous and categorical metrics to provide a comprehensive view of the impact of the different MOS methods on the predictions. Continuous metrics used to evaluate the O3 concentrations include :

- nMB : normalized Mean Bias
- nRMSE : normalized Root Mean Square Error
- PCC : Pearson correlation coefficient
- Slope : slope of the predicted-versus-observed O3 mixing ratio, to quantify how well lowest and highest O3 concentrations are predicted
- nMSDB : normalized Mean Standard Deviation Bias, to investigate how well the O3 variability is reproduced by the forecast

Categorical metrics used to evaluate the O3 exceedances beyond certain thresholds include :

- H : Hit rate, to quantify the proportion of observed exceedances that are correctly detected
- F : False alarm rate, to quantify the proportion of observed non-exceedances erroneously forecast as exceedances
- FB : Frequency Bias, to investigate to which extent the forecast is predicting the same number of exceedances as observed (no matter if they are predicted on the correct days)
- SR : Success Ratio, to show how much of the predicted exceedances are indeed observed
- CSI : Critical Success Index, to quantify the proportion of correctly predicted exceedances when discarding all the corrected rejections
- PSS : Peirce Skill Score, to investigate to which extent the forecast is able to separate exceedances from non-exceedances
- AUC : Area Under the ROC Curve, to quantify the probability that the forecast predicts higher O3 concentrations during a situation of exceedance compared to a situation of non-exceedance

The formula of these different metrics can be found in the Appendix E. Each of them thus highlights a specific aspect of the performance. Regarding categorical metrics, Jolliffe et al. (2011) gave a detailed explanation of the different metric properties desirable for assessing the quality of a forecasting system (see Table 3.4 in Jolliffe et al. (2011)). In this framework, PSS can be considered as the one of the most interesting metric for assessing the accuracy of the different RAW and MOS-corrected forecasts, given that it gathers numerous valuable properties: (i) truly equitable (all random and fixed-value forecasting systems are awarded the same score, which provides a single no-skill baseline), (ii) not trivial to hedge (the forecaster cannot cheat on his forecast in order to increase PSS), (iii) base rate independent (PSS only depends on H and F, which makes it invariant to natural variations in climate, which is particularly interesting in the frame of AQ forecast where AQ standards and subsequently the base rate can also change) and (v) bounded (values are comprised within a fixed range). It is worth noting that no perfect metric exists, and PSS (as most other metrics) does not benefit from the properties of non-degeneracy (it tends to meaningless values for rare events).
In addition, results are also discussed in terms of skill scores, using the 1-d persistence (PERS(1)) as the reference forecast. Skill scores aim at measuring the accuracy of a forecast relatively to the accuracy of a chosen reference forecast (e.g. persistence, climatology, random choice). They can be computed as $S(X)=(X-X_{reference})/(X_{perfect}-X_{reference})$ with X the score of the forecast, $X_{reference}$ the score of the PERS(1) reference forecast and $X_{perfect}$ the score expected with a perfect forecast. Skill scores indicate if a given forecast has a perfect skill (value of 1), a better skill than the reference forecast (value between 0-1), an equivalent skill than the reference forecast (value of 0) or a worse skill than the reference (value below 0, unbounded). To be converted into skill scores, the aforementioned metrics of interest need to be transformed into scores following the rule "the higher the better" (to constrain the skill score to values below 1). For the different metrics M, the corresponding score X(M) is obtained applying the following transformations : $X(M)=-M$ for nRMSE and F and $X(M) = -|1-M|$ for slope; no transformation are

required for the other metrics (H, SR, CSI, PSS and AUC). Note that, as indicated by its name, PSS is already intrinsically defined as a skill score (where the reference corresponds to a climatology or random choice, both giving PSS values tending toward 0), but it does not prevent it to be converted into a skill score related to the persistence forecast.
In order to ensure fair comparisons between observations and all the different forecasts, O3 values at a given hour are discarded when at least one of these different dataset does not have data. Over the 2018-2019 period, the resulting data availability exceeds 94% whatever the time scale considered. Note that about 4% of the data is here missing due to the aforementioned minimum of 30 days (i.e. January 2018) of accumulated historical data requested to start computing the corrected forecasts with some MOS methods."

We re-organized the sections 3.1 and 3.2, but first focusing on the observations in a section entitled "Ozone pollution over Iberian Peninsula", and then on the "Performance on continuous forecasts" and "Performance on categorical forecasts", each of these two last sections being divided into a subsection "RAW forecasts" and "MOS-corrected forecasts" and including a discussion on both the evaluation metrics and their corresponding skill scores (for which a figure is added):
L266-305 (section 3.2) is replaced by :
"**3.2 Performance on continuous forecasts**
3.2.1 RAW forecasts
Considering the annual mean O3 mixing ratios at all 456 stations (Fig. 1), the raw CAMS ensemble forecast represents moderately well the spatial distribution of annual O3 over the Iberian Peninsula (PCC of 0.54 for D+1 forecasts) and strongly underestimates the spatial variability (nMSDB of -42%). At least part of these errors are due to the fact that all station types are taken into account here, including traffic stations where local road transport NOx emissions can strongly reduce the O3 levels (titration by NO), which cannot be properly represented by models at 10 km spatial resolution. In this study, all station types are included because we are ultimately interested in predicting O3 exceedances at all locations where they can be observed (and thus, where air quality standards apply). It is worth noting that the impact of the MOS methods on the different metrics might vary from one type of station to another, although this aspect is beyond the scope of our study. The raw CAMS ensemble forecast correctly identifies regions where most exceedances of the target threshold occur but often with underestimated frequency, especially around Madrid, in southern Spain (in-land part of Andalusia region) and along the Mediterranean coast. More severe deficiencies are found with the information threshold that is almost never reached by the CAMS ensemble (with one single exception around Porto).

The overall statistical results are shown in Fig. 3 for the different forecast methods, and a subset of these statistics is given in Table 1 (and in Table S1 in the Supplement for additional time scales). For a given lead day and time scale, statistics are here computed after aggregating data from all monitoring stations; therefore, statistics of D+1 O3 forecasts at hourly scale can be based on 730 d x 24 h x 455 stations = 7,971,600 points if there are no data gaps. The RAW forecast overestimates

moderately the O3 mixing ratios, especially at hourly and daily time scales, but shows a reasonable correlation at all time scales (above 0.75). However, its main deficiency lies in the underestimated variability (nMSDB around -30%), which is reflected in the low model-versus-observation linear slope obtained (around 0.5-0.6). The deterioration of the performance of the raw CAMS forecasts with lead time is very low, with hourly-scale nRMSE/PCC decreasing from 38%/0.75 at D+1 to 39%/0.72 at D+4, potentially due to their relatively coarse spatial resolution.

As expected (by construction), the PERS(1) reference forecast gives unbiased O3 forecasts. Due to the temporal auto-correlation of O3 concentrations, reasonable results are obtained at D+1 (nRMSE/PCC/slope of 36%/0.74/0.74) but quickly deteriorate with the lead time (down to 42%/0.65/0.64 at D+4). A subset of skill scores with PERS(1) as reference is shown in Fig. 4. Apart from the slope that is always better reproduced by PERS(1), the RAW forecast reaches better skill scores than PERS(1) on both the nRMSE and PCC but only beyond D+1 (with values typically ranging between 0-0.2), and not at all time scales (for instance, PERS(1) systematically shows better RMSE than RAW at daily scale).

3.2.2 MOS-corrected forecasts

The MA(1) method removes most of the bias of O3 concentrations and variability. Some residual biases appear when computing the daily 1-h maximum from the MOS-corrected hourly O3 concentrations (i.e. d1max scale), but can be removed by applying the MA(1) method directly at this time scale (i.e. dd1max scale). The MA(1) method substantially improves the other metrics for all lead days, with hourly-scale nRMSE/PCC/slope of 31%/0.81/0.82 at D+1 and 36%/0.74/0.75 at D+4. Thus, the performance still deteriorates with lead time, but slight less dramatically than with PERS(1). In terms of skill scores, such a simple approach as MA(1) is found to strongly improve the skills initially obtained with RAW alone, whatever the time scale or lead time. Skills scores range between 0.1-0.3 for nRMSE and 0.3-0.4 for PCC and slope, with slightly higher values at daily and d8max scales. The variations of skill along lead time differ between nRMSE/PCC (lowest and highest skills typically obtained at D+1 and D+2/D+3/D+4, respectively) and slope (skills tend to progressively decrease from D+1 to D+4, although slightly).

The QM method shows quite similar results than the MA(1) method, but usually with worse (better) performance at short (long) lead time. Thus, the deterioration of the performance with lead time tends to be slower in QM than in MA(1). Biases on O3 concentrations and O3 variability are often slightly higher with QM but remain relatively low (below ±5%). The strongest improvements of QM compared to MA(1) are found at hourly scale for longest lead times. On these continuous metrics, the skills of the QM method are only slightly positive or even negative at D+1 (except at hourly scale where skill scores are always positive) but are much higher between D+2 and D+4, and often slightly better than MA(1).

Compared to the previous MOS methods, the KF method provides a substantial improvement on both nRMSE and PCC, leading to skill scores of 0.3-0.4 and 0.4-0.6, respectively. However, this comes at the cost of an underestimation of the variability

(nMSDB around -10%, still much better than the -30% of nMSDB found in RAW). As for the previous methods, some small biases appear at d1max scale and to a lesser extent at d8max scale but applying this MOS method directly on d1max or d8max O3 mixing ratios rather than hourly data (i.e. dd1max and dd8max scales) mitigates the issue.

Overall, comparable results are found with AN and GBM methods, but the aforementioned issues are typically exacerbated. The negative biases at d1max and d8max time scales are much higher, especially for GBM, but can be removed at dd1max and dd8max scales. Similarly, the underestimation of the variability is much more pronounced, with nMSDB values around -15% and -10% for AN and GBM, respectively. These two MOS methods thus show a good performance for predicting the central part of the distribution of O3 mixing ratios, but have more difficulty in capturing the lowest and highest O3 concentrations observed on the tails of this distribution. Besides the negative nMSDB, this typically leads to lower slopes compared to the other MOS methods. Skill scores on nRMSE and PCC span over a relatively large range of values depending on the time scale and the lead time. They are typically the lowest at short lead times and/or at specific time scales (e.g. d1max) but can reach among the highest values (although slightly lower than KF), for instance with GBM, at hourly and daily scale at D+2/D+3/D+4. Concerning the slope, the aforementioned issues are here illustrated by the typically low skills of both AN and (to a slightly lesser extent) GBM methods, often worse than the other MOS methods.

Therefore, on this set of continuous metrics, the impact of the MOS corrections on the performance strongly varies with the method considered. Among the different MOS methods, KF seems to give the most balanced improvement with biases mostly removed, errors and correlation substantially improved and variability not too strongly underestimated. However, it is worth noting that since some MOS methods (namely QM, AN and GBM) can ingest increasing amounts of input data over time, we can expected their performance to change (increase) between the beginning of the period when very limited past data information is available and the end of the period when more past data have been accumulated. Investigating this aspect would ideally require a proper analysis, comparing the performance obtained over a given period using variable amount of past input data. Here, we simply provide some insights by comparing the relative difference of performance of these MOS methods against RAW, (1) when evaluated over the entire 2018-2019 period (i.e. including the beginning of the period of study when MOS methods can only rely on limited past data), and (2) when evaluated only over the year 2019 (i.e. when the first year is discarded). In the first case (evaluation over 2018-2019), the QM, AN and GBM show nRMSE 31, 41 and 44% lower than RAW, respectively. In the second case (evaluation over 2019), these MOS methods give nRMSE 33, 44 and 49% lower than RAW. Therefore, this basic comparison suggests that these MOS methods can indeed benefit from a larger amount of past data. Here, the change is more pronounced more GBM, which suggests that this MOS method is the one benefiting the most from more past training data. For GBM, this improvement is mainly due to the

relatively poor predictions made during the very first months of 2018 when the training dataset was the most limited (see time series in Fig. F1 in Appendix F).

3.3 Performance on categorical forecasts

3.3.1 RAW forecasts

Focusing now on the performance for detecting target and information thresholds, Fig. 3 (middle and bottom panels) shows a comprehensive set of metrics, where the most interesting ones are probably CSI and PSS, followed by SR and AUC.

The RAW forecast shows low H and F (very few true positives and false negatives). With an intermediate SR (0.45, i.e. only 45% of the exceedances predicted by RAW indeed occur), it can be seen as a moderately "conservative" forecast for target thresholds (d8max O3 above 60 ppbv); the term "conservative" here refers to forecasting systems that predict exceedances only with strong evidence (it thus predicts very few exceedances but with a moderate confidence). Despite showing a reasonably good AUC, the RAW forecast strongly fails at reproducing high O3 mixing ratios, as illustrated by the low FB (0.25, i.e. RAW predicts 4 times less exceedances than the observations), and finally shows the worst performance in terms of CSI (0.10) or PSS (0.15). In comparison, the PERS(1) reference forecast provides better detection skills regarding target thresholds. This is especially true at short lead days, but the performance then quickly decreases with the lead time, with CSI/PSS reduced from about 0.27/0.42 at D+1 to about 0.14/0.23 at D+4. Except FB, all categorical metrics show a similarly strong sensitivity to the lead time. With PERS(1) taken as a reference, the skill scores of RAW clearly show negative and positive values for H and F, respectively (i.e. it predicts less true exceedances but produces less false alarms). The consequence in terms of SR skills is positive but only beyond D+1. With positive skills on AUC, RAW is able to discriminate exceedances and non-exceedances slightly better than PERS(1), but only beyond D+2. However, its skills on the important CSI and PSS metrics are strongly negative at all lead times, which highlights its overall deficiency for predicting correctly the exceedances of the target threshold (i.e. without too many false alarms).

Exceedances of the information threshold (d1max O3 above 90 ppbv) appear even more difficult to capture for the RAW forecast with CSI and PSS typically below 0.02. However, given that it is also more difficult for PERS(1) to capture these exceedances, the skills of RAW on these two metrics are substantially better (although still negative) on this information threshold compared to the target threshold. Results also show much better SR, especially at longest lead times (i.e. most of the predicted exceedances indeed occur), but this apparently good result has to be put in front of the extremely low H (i.e. RAW almost never predict exceedances).

3.3.2 MOS-corrected forecasts

Although the RAW forecast alone shows quite limited skills for predicting high O3 exceedances, its potential usefulness is nicely illustrated by the results obtained when it is combined with observations, such as in MA(1), QM or KF(RMSE). When considering the target threshold exceedances, CSI and PSS are indeed greatly

improved with these last MOS methods, and to a lesser extent by the two other methods, AN(10) and GBM. KF(RMSE), AN(10) and GBM clearly appear as the most "conservative" MOS approaches here, with relatively low H and F, but strong SR. In other terms, they predict fewer exceedances but with a higher reliability. In terms of skill scores, all these MOS-corrected forecasts always have better skills than RAW. However, only MA(1) always beats PERS(1) at all lead times, while the other MOS methods provide positive skills only beyond D+1/D+2. This MA(1) method thus clearly outperforms the other methods at D+1, while differences of performance are reduced when considering longer lead times. At longer lead times, the ranking between these different MOS methods varies substantially depending on the considered metric, with MA(1), KF(RMSE) and GBM showing best skills on CSI, and MA(1) and QM showing best skills on PSS.

However, when considering the detection of the information threshold, the KF(RMSE), AN(10) and GBM methods still benefit from a strong SR but are missing too many of the observed exceedances, which leads to a dramatic deterioration of both CSI and PSS. As for RAW, this means that there is a high change that an exceedance predicted by these methods indeed occurs but such exceedances are too rarely predicted. Most of their skill scores on PSI are found to be negative, while only a few positive skills are obtained on CSI for specific time scales in KF and GBM methods. For detecting such high O3 values, best methods are finally MA(1) for shortest lead times. At longer lead times, the skills of MA(1) quickly deteriorate and best skills are finally obtained for QM. Both methods reproduce fairly well the geographical distribution of high O3 episodes (PERS(1) reproduces it perfectly, by construction), as shown in Fig. 5, but still with very low SR (below 0.25 for exceedances of the information threshold)."

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.

Indeed, such relatively coarse spatial resolution does not allow to properly represent the pollutant concentrations observed at stations close to strong emission sources (e.g. urban traffic, industrial), as already mentioned in the text (L257 : "Part of this positive nMB and negative nMSDB is expected since this broad comparison includes all station types, including traffic stations where local road transport NOx emissions can strongly reduce the O_3 levels (titration by NO), which cannot be fully represented by models at 10 km spatial resolution."). In this study, we chose to consider all stations because we are ultimately interested in predicting O3 exceedances at all locations where observations are available and therefore where air quality standards apply. Given the numerous aspects already covered, we consider that it is beyond the scope of this study to explore in more detail the impact of the station types. We added a brief comment :

L260 : "In this study, all station types are included because we are ultimately interested in predicting O3 exceedances at all locations where they can be observed (and thus, where the air quality standards apply). It is worth noting that the impact

of the MOS methods on the different skills might vary from one type of station to another, although this aspect is beyond the scope of our study."

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.

Again, the reviewer is raising here interesting questions but the application of MOS methods to operational air quality forecasts is a vast topic of research and we do not claim in this first study to cover all its relevant aspects. To our opinion, regarding the application of MOS methods to air quality forecasts, this work already goes beyond most of the other studies available in the literature, and therefore we consider the points raised by the reviewer as important aspects but beyond the scope of the present study. Nonetheless, according to the reviewer comments we discussed these different points in the revised version :

L124 : "We note, however, that methods relying on limited past data may respond better to an abrupt change in environmental conditions, as experienced for instance during the COVID-19 lockdowns. Although not covered by the present study, we acknowledge here that in an operational context, the relationship between the length of past training data and the performance of the corresponding MOS prediction is an interesting aspect to investigate, as is the quantification of the spin-up time beyond which the MOS method might not significantly improve. Only some insights will be given by comparing the performance obtained in 2019 with and without using the data available in 2018. Similarly, our study does not investigates how potential issues (delays) in the near-real time availability of the observations can impact the performance of the MOS methods, although this might be another important aspect to take into account in operational conditions; to the best of our knowledge, EEA observations are typically available with a 2-h lag but some sporadic technical failures can induce extended delays."

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)

We thank the reviewer for clarifying this point. However, to the best of our knowledge, only a limited subset of the meteorological variables used in this paper are available in the HRES analysis, which thus prevents a fully consistent comparison between the two meteorological products (HRES forecast and analysis).

Nonetheless, in order to avoid the confusion existing in the first version of the manuscript, we modified the corresponding sections by presenting this test HRES versus ERA5 as a simple sensitivity test on the meteorological input data :

L99 : "**2.1.3 IFS and ERA5 meteorological data**

Some MOS methods rely on meteorological data. In this study, meteorological data are taken from the European Centre for Medium-Range Weather Forecasts (ECMWF) Integrated Forecast System (IFS) (Flemming et al., 2015). IFS has a native spatial resolution of about 9 km and 137 vertical levels. In addition, to investigate to which extent the quality of the meteorological input data impacts the performance of the meteorology-dependent MOS methods (Sect. 3.5), we replicated all our experiments with the ERA5 reanalysis dataset (Copernicus Climate Change Service (C3S), 2017). ERA5 data have a native spatial resolution of about 31 km and 137 vertical levels, although data were downloaded on a 0.25°x0.25° regular longitude-latitude grid from the Climate Data Store. Although reanalysis meteorological data would obviously not be available in an operational context, testing the MOS methods with this reference dataset allows to estimate the upper range of performance that could be expected.

At all surface O_3 monitoring stations, for both IFS and ERA5, we extracted the following variables at the hourly scale: 2-m temperature (code 167), 10-m surface wind speed (207), normalized 10-m zonal and meridian wind speed components (165 and 166), surface pressure (134), total cloud cover (164), surface net solar radiation (176), surface solar radiation downwards (169), downward UV radiation at the surface (57), boundary layer height (159), and geopotential at 500 hPa (129)."

➔ **"2.1.3 HRES and ERA5 meteorological data**

Some MOS methods rely on meteorological data. In this study, meteorological data are taken from the Atmospheric Model high resolution 10-days forecast (HRES) (https://www.ecmwf.int/en/forecasts/datasets/set-i) provided by the European Centre for Medium-Range Weather Forecasts (ECMWF). HRES has a native spatial resolution of about 9 km and 137 vertical levels. In addition, to investigate the sensitivity to the meteorological input data, we replicated all our experiments with the ERA5 reanalysis dataset (Copernicus Climate Change Service (C3S), 2017) (https://www.ecmwf.int/en/forecasts/datasets/reanalysis-datasets/era5). ERA5 data have a native spatial resolution of about 31 km and 137 vertical levels, although data were downloaded on a 0.25°x0.25° regular longitude-latitude grid from the Climate Data Store. At all surface O3 monitoring stations, for both HRES and ERA5, we extracted the following variables at the hourly scale: 2-m temperature (code 167), 10-m surface wind speed (207), normalized 10-m zonal and meridian wind speed components (165 and 166), surface pressure (134), total cloud cover (164), surface net solar radiation (176), surface solar radiation downwards (169), downward UV radiation at the surface (57), boundary layer height (159), and geopotential at 500 hPa (129)."

And we substantially reduced the size of section 3.5 and included it as a subsection in section 3.4 (i.e. as an additional sensitivity tests) :

"3.4.6 Influence of the meteorological input data in AN and GBM methods

In the previous sections, O3 corrections with AN and GBM methods relied on HRES meteorological forecasts. Here, we investigate the impact of using an alternative meteorological data, namely the ERA5 meteorological reanalysis. For both AN and

GBM methods, the MOS-corrected O3 mixing ratios obtained with these two meteorological dataset are very similar, with PCC above 0.95. The results obtained against observations are shown in Fig. G6 in the Appendix G, for the AN(1), AN(5), AN(10) and GBM methods. Since O3 predictions are close, the statistical performance against observations is also very consistent between both meteorological datasets. For both continuous and categorical metrics, the performance obtained with HRES data is found to be slightly lower than with ERA5. Discrepancies between both meteorological dataset tend to increase with lead time, with GBM being slightly more sensitive to the meteorological input data than AN.
Therefore, this experiment highlights a relatively low sensitivity of both AN and GBM methods to the two meteorological datasets tested here. The very similar results obtained with IFS and ERA5 meteorological input data are likely not explained by the fact that both datasets give very similar values for the different meteorological variables, but rather by the intrinsic characteristics of both AN and GBM methods. The AN method make use of the meteorological data only to identify past days with more or less similar meteorological conditions, and can thus handle to some extent the presence of biases in meteorological variables as far as they are systematic (and thus do not impact the identification of the analogs). On the other side, the GBM method uses past information to learn the complex relationship between O3 mixing ratios and the other ancillary features. Although the better the input data, the higher the chances are to fit a reliable model for predicting O3, the GBM models can also learn indirectly at least part of the potential errors affecting some meteorological variables and how they relate to O3 mixing ratios. Therefore, the presence of biases in some of the ancillary features is not expected to strongly impact the performance of the predictions."

We also modified the abstract :

L13 : "When considering MOS methods relying on meteorological information and comparing the results obtained with IFS forecasts and ERA5 reanalysis, the relative deterioration brought by the use of IFS is minor, which paves the way for their use in operational MOS applications." ➔ "The MOS methods relying on meteorological data were found to provide relatively similar performance with two different meteorological inputs."

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.

Regarding the maps (Figs. 1 and 4 in the first version), the color range corresponds to the range of values obtained at the different stations (for both OBS and RAW, in order to allow a direct comparison). We suspect that the reviewer's comment is more directly related to the bottom panels (about exceedances of d1max above 90 ppbv) in which the color bar values range between 0 and 20 while most of the points shown on the corresponding maps are purple/blue. The reason is that some high values were hidden behind other overlapping points. We modified the plot to make it easier to read, and updated accordingly the legend by adding : "In order to limit the overlap, stations are here plotted by decreasing value and with decreasing size (lowest values with largest symbols but in background, highest values with smallest symbols but in foreground).". We also modified the color of the time series (Fig. 2 in

the first version) to make easier to read. Regarding Fig. 3, we greatly modified the discussion in the revised version.

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.
We included in the Supplement several tables to show a subset of the evaluation results on continuous and categorical metrics for all these sensitivity tests. We also modified the corresponding figures in the Appendix so that to be consistent with the new version of Fig. 3 (fewer metrics). As described in another answer, we greatly shortened the discussion on the sensitivity tests (section 3.4).

Some specific comments:
Abstract: Please quantify the achieved improvements by MOS to replace or justify phrases such "can be substantially improved"
We added information on the improvement in terms of RMSE and PCC :
L11 : "Our results show that O3 forecasts can be substantially improved using such MOS corrections and that this improvement goes much beyond the correction of the systematic bias. Although it typically affects all lead times, some MOS methods appear more adversely impacted by the lead time. When considering MOS methods relying on meteorological information and comparing the results obtained with IFS forecasts and ERA5 reanalysis, the relative deterioration brought by the use of IFS is minor, which paves the way for their use in operational MOS applications. " ➔ "Our results show that O3 forecasts can be substantially improved using such MOS corrections and that improvements go well beyond the correction of the systematic bias. Depending on the time scale and lead time, root mean square errors decreased from 20-40% to 10-30%, while Pearson Correlation coefficients increased from 0.7-0.8 to 0.8-0.9. Although the improvement typically affects all lead times, some MOS methods appear more adversely impacted by the lead time. The MOS methods relying on meteorological data were found to provide relatively similar performance with two different meteorological inputs."

L67-71 A summary of the results of the paper is not required in the introduction
We removed these lines.

L 98 mention forecast start time
We modified the sentence L97 : "The CAMS regional forecasts are provided over 4 lead days, hereafter referred to as D+1, D+2, D+3 and D+4 (starting at 0 UTC)."

L 101 Flemming et al. 2015 is not a reference for the operational NWP forecast of ECMWF
We replaced this reference by the following link : https://www.ecmwf.int/en/forecasts/datasets/set-i

L 120 Please consider the general comment about NRT availability of observations
As far as we know, EEA observations are usually available with a 2-hours lag. We included this information, as described in a previous answer.

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.

The persistence (PERS) approach only depends on observations, and is now considered apart from the MOS methods (and used as a reference forecast for computing skill scores). The moving average (MA) method is a MOS method where the raw CAMS predictions are used together with the observations : the raw forecast on a given day is corrected by the mean difference between raw and observed concentrations one or several days before. The modifications of this part of the text are described in another answer (related to the reviewer's comment on the persistence method).

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

The choice of 30 days is arbitrary and motivated only by computational reasons, we extended the discussion related to this aspect :

L146 : "For computational reasons, both CDFs are updated every 30 days (although an update frequency of one single day would be optimal in a real operational context)." ➔ "For computational reasons, both CDFs are updated every 30 days (although an update frequency of one single day would be optimal in a real operational context). The choice of a 30-day update frequency only aims at reducing the computational cost of running all MOS methods at all stations during the 2-year period. In a real operational context, only one day would have to be run, which would allow increasing the update frequency up to 1 day, i.e., the CDFs would be updated every day ensuring that we are taking benefit from the entire observational dataset available at a given time."

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.

To the best of our understanding, although biases are indeed not addressed in the KF theory, the application of KF as a MOS approach is specific in the sense that it takes the forecast bias itself as the state variable of interest. We reformulated part of this section :

L150 : "Over the last decades, the Kalman filter (KF) theory has found numerous applications in problems with different levels of complexity. In atmospheric sciences, it offers a popular frame for sophisticated data assimilation applications (e.g., Gaubert et al., 2014, Di Tomaso et al. 2017), but can also be used as a simple yet powerful MOS method for correcting forecasts (e.g., Delle Monache et al., 2006, Kang et al., 2008, De Ridder et al., 2012). A detailed description of the KF algorithm can be found in Appendix B (as well as in Delle Monache et al., 2006).

KF provides an efficient way of estimating the forecast bias based on past model and observation information. For a given day at a given hour, the forecast bias is computed as a weighted average of (1) the forecast bias estimated one day before and (2) the corresponding observed forecast bias. Each of these two terms is weighted according to the value of the so-called Kalman gain (Kt) that intrinsically depends on the so-called variance ratio (see Appendix B for more details). The value chosen for this internal parameter substantially affects the behaviour of the KF, and

thus the obtained MOS corrections." ➔ "The Kalman Filter (KF) is an optimal recursive data processing algorithm with numerous science and engineering applications (see Pei et al., 2017 for an introduction). In atmospheric sciences, it offers a popular frame for sophisticated data assimilation applications (e.g., Gaubert et al., 2014, Di Tomaso et al. 2017), but can also be used as a simple yet powerful MOS method for correcting forecasts (e.g., Delle Monache et al., 2006, Kang et al., 2008, De Ridder et al., 2012). The KF-based MOS method aims at estimating recursively the unknown forecast bias (here taken as the state variable of interest) combining previous forecast bias estimates with forecast bias observations. The updated forecast bias estimate is computed as a weighted average of these two terms, both being considered as uncertain, i.e. affected by a noise with zero-mean and a given variance. A detailed description of the KF algorithm can be found in Appendix B but an important aspect to be mentioned here is that each of these two terms is weighted according to the value of the so-called Kalman gain that intrinsically depends on the ratio of both variances (hereafter referred to as the variance ratio). The value chosen for this internal parameter substantially affects the behavior of the KF, and thus the obtained MOS corrections."

Also, to make it easier to follow, we modified the corresponding Appendix and adopted notations more consistent with those used by Pei et al. (2017) in their gentle introduction to KF.

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

The aforementioned distance metric is used to compute the distance between the current forecast day and each of the past days, this distance representing how similar or different are the current forecast day and one given past day (a small distance means that both are very similar). In this frame, best analog days refers to the most similar days (i.e. the days with smallest distance to the current forecast). We clarified the text :

L176 : "The current forecast is compared to past forecasts based on a set of features including the raw O_3 mixing ratio forecast from the AQ model and the 10-meter wind speed, 2-meter temperature, surface pressure and boundary layer height forecast from the meteorological model. The similarity of each day of forecast is assessed using the distance metric proposed by Delle Monache et al. (2011) and previously used in Djalalova et al. (2015) (see the formula in Appendix C). As a first approach, we consider the 10 best analog days, hereafter referred to as AN(10); other values are tested in Sect. 3.4)." ➔ "The current forecast is compared to each individual past forecasts in order to identify which ones are the most similar. Based on a set of features including the raw O_3 mixing ratio forecast from the AQ model and the 10-meter wind speed, 2-meter temperature, surface pressure and boundary layer height forecast from the meteorological model, the distance metric proposed by Delle Monache et al. (2011) and previously used in Djalalova et al. (2015) (see the formula in Appendix C) is used to compute the distance (i.e., to quantify the similarity) of each individual past forecast with respect to the current forecast. Then, as a first approach, the 10 best analog days that correspond here to the 10 most similar past forecasts are identified (hereafter referred to as AN(10); other values are tested in Sect. 3.4)."

L 209 Please motivate the choice of the 30 day training period.
See previous answer on the topic.

L 233 Missing here is a skill score that assess the forecast skill against the persistency forecast
See previous answer on the topic.

L 225-233 The amount of accuracy measures is overwhelming and 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.
Although we understand it might appear overwhelming at first read, we do think it is useful to show such a comprehensive set of metrics to highlight different aspects of the forecast performance (while MOS results in the literature are often shown only with a very limited number of metric, typically only continuous). In the revised version, we simplified Fig. 3 (and the figures in the Appendix) by removing a few metrics, namely MB and RMSE (we only kept nMB and nRMSE), as well as the intercept and the base rate, which should make it slightly easier to follow. More importantly, we added a section where metrics are more clearly introduced (as previous mentioned in another answer), and we tried to interpret in more detail the discussion of these different metrics in the discussion of the results.

L 266 Please explain Fig 3 in more detail. What do the overlaying symbols mean (one per stations , forecast day ?).
We clarified the legend of Fig. 3 : "Figure 3. Statistical performance of RAW and MOS-corrected CAMS O3 forecasts for continuous metrics (top panels) and categorical metrics related to the exceedance of the target (intermediate panels) and information threshold (bottom panels). The different symbols depict results obtained at different time scales (h: hourly; d: daily mean; d1max/dd1max: daily 1-hour maximum; d8max/dd8max: daily 8-hour maximum). In each panel, results are shown for the different methods (each with a given color). The overlaying symbols of decreasing transparency show the results at the different lead days from D+1 (most transparent) to D+4 (most opaque). […]"
We also extended the discussion of Fig. 3, as described in another answer.

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

L 335-445 Please see my general comment on section 3.4
As described in another answer, we shortened this section.

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.; L446 ER5 and HRES will not differ in the number of assimilated observations, if anything HRES will be better.

As described in another answer, we modified this section to take into account the comments of the reviewer.

L 484 Please provide quantitative information about the improvements.

Given that numerous quantitative information was provided in the previous section, we do not think it is useful to provide again some quantitative information here. We rather prefer to keep this discussion for a general discussion around the use of MOS.

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.

As described in another answer, we take into account the comments of the reviewer regarding this aspect, and greatly modified the manuscript.

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)

We are not sure to follow the reviewer on this comment. Among all the forecast methods, PERS is clearly the most strongly impacted by the lead time, while the impact of the lead time on RAW is found to relatively small (and we indicated in the text that it could be "potentially due to their relatively coarse spatial resolution"). The MOS methods are typically moderately impacted by the lead time, likely simply because they typically rely on both raw forecast and past recent observations. We added L517 : "The performance of the RAW forecasts was found to be only slightly sensitive to the lead day, but this sensitivity was substantially stronger with some MOS methods (although lower than for the persistence method)."

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.

One key message of our study lies in the large variability of performance of the different MOS methods from one metric to another. Thus, it is not possible to conclude with a clear recommendation as it directly depends on what the user is most interested in, and in the specific case of O3 exceedances forecasts, the respective cost of false positive and false negative predictions. However, through its results and sensitivity tests, our study provides a rich and useful material to help users to make their decision (although any MOS implementation requires testing different MOS methods and/or configurations as results obtained here with CAMS ensemble forecasts over Spain might evidently differ with another model and/or in a different region).

L 575 Please provide reference for GHOST

Although it is still in preparation, the reference for GHOST is already provided in the list of references (Bowdalo, D.: Globally Harmonised Observational Surface Treatment: Database of global surface gas observations, in preparation).

Reviewer #3

The paper entitled "Model Output Statistics (MOS) applied to CAMS O3 forecasts: trade-offs between continuous and categorical skill scores" by Petetin et al. is well written and provides a very interesting perspective on different statistical tools and machine learning approaches that can be used to improve air quality forecasts. It falls within the scope of ACP and I truly enjoyed reading it. The analysis is sound and truly comprehensive. I have only a few minor comments, that the authors may want to consider to improve the manuscript further.

In line 85 the authors say that in this study, daily mean, daily 1-hour maximum and daily 8-hour maximum are computed only when at least 75% of the hourly data are available (i.e. 18 over 24 hours). Theoretically speaking a day during which the data from 9 am to 4 pm is missing could qualify this criterion, yet both the computed d1max and d8max of such a day would be far off. Instrumental interventions such as service visits, purging with zero air to get moisture out of the system and calibrations usually occur during working hours. Hence the authors may want to consider applying a filter directed at daytime rather than night-time observations next time. This may remove a few extra data points but would have been preferable considering the target of the paper. The current choice is, however, hardly going to impact the results pertaining to the model evaluation of the various techniques with respect to d8max and d1max. The days analysed are driven mostly by observational stations reporting d8max >60 ppb or D1max>90. Hence most days included in the analysis would have daytime data. Nigh time events with d8max >60 ppb and d1max >90 ppb are relatively rare, although they do occur occasionally at high altitude stations. So, in my opinion the best way out without redoing the analysis would be to run a quick check on the data for the following two parameters
How many days with large data gaps during the day (9 am to 4 pm) were included in the analysis?
How many of the observed d8max and d1max events are night time events?
Both numbers would be small and can be reported and discussed as limitation.
Considering all Spanish stations and all days in 2018-2019 with at least 18 over 24 hourly values available, we checked how many had at least 6-hour data gaps occurring between 8 and 15 UTC. In total, the frequency of such large daytime data gaps is only 167/314,005 (0.05%), of which only 12 are exceedances of the target threshold (over a total of 13,221), and 0 are exceedances of the information threshold. Checking how many days and stations have at least 4-hour data gaps occurring between 8 and 15 UTC, the total frequency increases to 1854/314,005 (0.6%), of which only 77 are exceedances of the target threshold and 0 are exceedances of the information threshold. Therefore, the situation of large data gaps during daytime indeed occurs, but very rarely, and thus should not impact significantly our results. We added some elements of information regarding this

point : L83 : “In this study, daily mean, daily 1-hour maximum and daily 8-hour maximum (hereafter respectively referred to as d, d1max and d8max) are computed only when at least 75% of the hourly data are available (i.e. 18 over 24 hours). Note that despite such data availability criteria, large data gaps at some stations and during some days might occur mainly during daytime (for instance due to maintenance operations that typically occur during working hours). Considering all stations and days with at least 18 hours of data, the frequency of data gaps exceeding 4 hours between 8 and 15 UTC was found to be only 0.6% (1854/314,005). Such situation occurs with a similarly low frequency on days exceeding the target threshold (77/13,221 or 0.6%) and never occurs on days exceeding the information threshold.”

Figure 2: I find it hard to see the colour difference between the purple and black line. In particular where they are not superimposed. The colour contrast in Figure F1 which is similar is much better

We modified the color of this plot, and updated the caption :

“Figure 2. Time series of the mean O_3 mixing ratios over the Iberian Peninsula, as observed by monitoring stations (in black) and as simulated by the CAMS regional ensemble D+1 forecasts (in ~~purple~~ yellow). Time series are shown at the hourly (h), daily mean (d), daily 1-hour maximum (d1max) and daily 8-hour maximum (d8max) time scales. O_3 mixing ratios are averaged over all surface stations of the domain.”

Figure 3: Some people have bad memory for abbreviations or the habit of skipping to the figures. Just like the authors gave the full form for (h: hourly; d: daily mean; d1max/dd1max: daily 1-hour maximum; d8max/dd8max: daily 8-hour maximum) which is much appreciated can they please give the full form of the abbreviations S, H, F, FB, SR, CSI, PSS, AUC, PCC (which people may be more familiar with a R) in the figure caption. It will save a lot of readers from having to scroll back to the method section where these are defined.

We added more information in the caption (note that following the recommendation of the other reviewer, we removed some of the metrics to make the plot easier to follow) :

“Figure 3. Statistical performance of RAW and MOS-corrected CAMS O3 forecasts for continuous metrics (top panels) and categorical metrics related to the exceedance of the target (intermediate panels) and information threshold (bottom panels). The different symbols depict results obtained at different time scales (h: hourly; d: daily mean; d1max/dd1max: daily 1-hour maximum; d8max/dd8max: daily 8-hour maximum). In each panel, results are shown for the different methods (each with a given color). The overlaying symbols of decreasing transparency show the results at the different lead days from D+1 (most transparent) to D+4 (most opaque). Metrics : normalized Mean Bias (nMB in %), normalized Root Mean Square Error (nRMSE in %), Pearson correlation coefficient (PCC), slope (unitless), normalized Mean Standard Deviation bias (nMSDB in %), Hit rate (H), False alarm rate (F), Frequency Bias (FB), Success Ratio (SR), Critical Success Index (CSI), Peirce Skill Score (PSS), Area Under the ROC Curve (AUC). See Sect. 2.4 and 2.5 for details on time scales and metrics, respectively.”

Other modifications

- We harmonized the manuscript to use exclusively American English (e.g. "behavior")
- We added at L245 : "Over the Iberian Peninsula, annual mean O3 mixing ratios"
- L237 : "Ozone pollution over Iberian Peninsula and raw CAMS forecasts" ➔ "Ozone pollution over Iberian Peninsula"
- Abstract L7 : "A key aspect of our study is the evaluation, which is performed using a very comprehensive set of continuous and categorical metrics at various time scales (hourly to daily), along different lead times (1 to 4 days), and using different meteorological input data (forecast vs reanalyzed)." ➔ "A key aspect of our study is the evaluation, which is performed using a comprehensive set of continuous and categorical metrics at various time scales, along different lead times, and using different meteorological input datasets."
- Abstract L16 : "However, they are not necessarily the best in predicting the highest O3 episodes, for which simpler MOS methods can give better results." ➔ "However, they are not necessarily the best in predicting the peak O3 episodes, for which simpler MOS methods can achieve better results.
- L37 : "As these MOS methods often significantly reduce systematic errors, bringing mean biases close to zero, they are also commonly referred to as bias-correction or bias-adjustment methods, although they may not aimed at reducing directly this specific metric. MOS methods relying on local data (first and foremost the local observations) can also be seen as so-called downscaling methods as they allow capturing some of the local features that cannot be reproduced at typical CTM spatial resolution." ➔ "As these MOS methods often significantly reduce systematic errors, bringing mean biases close to zero, they are also commonly referred to as bias-correction or bias-adjustment methods, although they may not be aimed at reducing directly this specific metric. MOS methods relying on local data (first and foremost the local observations) can also be seen as so-called downscaling methods since they allow capturing some of the local features that cannot be reproduced at typical CTM spatial resolution."
- L113 : "Applying MOS in a worse case scenario of operational-like conditions" ➔ "Applying MOS under restrictive operational conditions"
- L114 : "A novel aspect of this study is that it provides a comparison of a set of MOS methods under a worse case scenario of operational-like conditions, which can be described through two assumptions:" ➔ "A novel aspect of this study is that we provide a comparison of a set of MOS methods under potentially restrictive training conditions in operational context. To mimic such restrictions we assume that"
- L119 : "On a given day, all MOS methods can only rely on the historical data accumulated so far." ➔ "On a given day, the MOS methods can therefore only rely on the historical data accumulated since the beginning of the period. Our approach consists in understanding the behaviour of the

different MOS methods in a worse case scenario where a new or upgraded operational AQ forecasting system is implemented together with a MOS module for which there is little or no hindcast data."

- L121 : "As it will be described in more detail in the next section, some MOS methods require very limited prior information to achieve their optimal performance, while other need a larger amount of training data." ➔ "As described in detail in the next section, some MOS methods require very limited prior information to achieve their optimal performance, while others need a larger amount of training data."
- L189 : "In this study, we also explore the use of ML algorithms as an innovative MOS approach for correcting AQ forecasts." ➔ "We also explore the use of ML algorithms as an innovative MOS approach for correcting AQ forecasts."
- L358 : "Therefore, for detecting exceedances, considering PSS and/or CSI as the most relevant metrics ~~(Appendix E)~~, the PERS method shows its best performance for a time window of 1 d."
- L557 : "In this study, we considered a relatively short 2-year dataset but using a longer training dataset would likely require to build specific methodologies to tackle this issue, either by identifying and discarding the potentially outdated data, or by giving them a lower weight in the procedure." ➔ "In this study, we considered a relatively short 2-year dataset but using a longer training dataset would likely require building specific methodologies to tackle this issue, either by identifying and discarding the potentially outdated data, or by giving them a lower weight in the procedure."