https://doi.org/10.5194/acp-2021-864
Preprint. Discussion started: 1 December 2021

Atmospheric Chemistry and Physics
Discussions
Open Access
EGU

Model Output Statistics (MOS) applied to CAMS O_3 forecasts: trade-offs between continuous and categorical skill scores

Hervé Petetin[1], Dene Bowdalo[1], Pierre-Antoine Bretonnière[1], Marc Guevara[1], Oriol Jorba[1], Jan Mateu Armengol[1], Margarida Samso Cabre[1], Kim Serradell[1], Albert Soret[1], and Carlos Pérez Garcia-Pando[1,2]

[1]Barcelona Supercomputing Center, Barcelona, Spain
[2]ICREA, Catalan Institution for Research and Advanced Studies, Barcelona, Spain

Correspondence: Hervé Petetin (herve.petetin@bsc.es)

Abstract. Air quality (AQ) forecasting systems are usually built upon physics-based numerical models that are affected by a number of uncertainty sources. In order to reduce forecast errors, first and foremost the bias, they are often coupled with Model Output Statistics (MOS) modules. MOS methods are statistical techniques used to correct raw forecasts at surface monitoring station locations, where AQ observations are available. 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). We apply our analysis to the Copernicus Atmospheric Monitoring Service (CAMS) regional ensemble median O_3 forecasts over the Iberian Peninsula during 2018-2019. 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).

Our results show that O_3 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. Importantly, our results also clearly show the trade-offs between continuous and categorical skills and their dependencies on the MOS method. The most sophisticated MOS methods better reproduce O_3 mixing ratios overall, with lowest errors and highest correlations. However, they are not necessarily the best in predicting the highest O_3 episodes, for which simpler MOS methods can give better results. Although the complex impact of MOS methods on the distribution and variability of raw forecasts can only be comprehended through an extended set of complementary statistical metrics, our study shows that optimally implementing MOS in AQ forecast systems crucially requires selecting the appropriate skill score to be optimized for the forecast application of interest.

https://doi.org/10.5194/acp-2021-864
Preprint. Discussion started: 1 December 2021

1 Introduction

Air pollution is recognized as a major health and environmental issue (World Health Organization, 2016). Mitigating its negative impacts on health requires reducing both pollutant concentrations and population exposure. Air quality (AQ) forecasts can be used to warn the population on the potential occurrence of a pollution episode, while allowing the implementation of temporary emission reductions, including e.g. traffic restrictions, shutdown of industries and bans on the use of fertilizers in the agricultural sector).

AQ forecasting systems are typically based on regional chemistry-transport models (CTMs), which remain subject to numerous uncertainty sources, leading to persistent systematic and random errors, especially for ozone (O_3) and particulate matter (PM) (e.g. Im et al., 2015a, b). More importantly, they often largely underestimate the strongest episodes that exert the worst impacts upon health. In addition to the error sources related to the models themselves and the input data, part of the discrepancies between in-situ observations and geophysical forecasts are due to inherent representativeness issues, since concentrations measured at a specific location are not always comparable to the concentrations simulated over a relatively large volume.

To overcome these limitations, operational AQ forecasting systems based on geophysical models often rely on so-called Model Output Statistics (MOS) methods for correcting statistically the raw forecasts at monitoring stations. The basic idea of MOS methods is to combine raw forecasts with past observations, and eventually with other ancillary data, at a given station in order to produce a better forecast, preferably at a reasonable computational cost. 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.

Over the last decades, several MOS methods have been proposed for correcting weather forecasts, before their more recent application to AQ forecasts, essentially on O_3 and fine particulate matter ($PM_{2.5}$, with aerodynamic diameter lower than 2.5 µm). A very simple approach consists in subtracting the mean bias (or multiplying by a mean ratio to avoid negative values in the corrected forecasts) calculated from past data (McKeen et al., 2005). A more adaptive version consists in correcting the forecast by the model bias calculated over the previous days, which assumes some persistence in the errors (Djalalova et al., 2010). Other authors proposed fitting linear regression models between chemical concentration errors and meteorological parameters (e.g., Honoré et al., 2008; Struzewska et al., 2016). Liu et al. (2018) applied a set of autoregressive integrated moving average (ARIMA) models to improve Community Multiscale Air Quality (CMAQ) model forecasts. The Kalman Filter (KF) method is a more sophisticated approach, yet still relatively simple to implement, based on signal processing theory (e.g., Delle Monache et al., 2006; Kang et al., 2008, 2010; Borrego et al., 2011; Djalalova et al., 2010, 2015; Ma et al., 2018). Initially employed for correcting meteorological forecasts (Delle Monache et al., 2011; Hamill and Whitaker, 2006), the ANalogs (AN) method provides an observation-based forecast using historical forecasts and has recently provided encouraging results for correcting $PM_{2.5}$ CMAQ forecasts over the United States (Djalalova et al., 2015; Huang et al., 2017).

A common limitation in the aforementioned studies is that MOS corrections are assessed mainly in terms of continuous variables (i.e. pollutant mixing ratios), while typically less attention is put on the parallel impact in terms of categorical variables (i.e. exceedances of given thresholds), which is yet one of the primary goals of AQ forecasting systems. This can give a partial, if not misleading, view of the advantages and disadvantages of the different MOS approaches proposed in the literature.

The present study aims at providing a comprehensive assessment of the impact of different MOS approaches upon AQ forecasts. We consider a representative set of MOS methods, including some already proposed in the recent literature and another one based on machine learning (ML). These MOS corrective methods are applied to the Copernicus Atmospheric Monitoring Service (CAMS) regional ensemble O_3 forecasts, focusing on the Iberian Peninsula (Spain and Portugal) during the period 2018-2019. The MOS methods are evaluated for a comprehensive set of continuous and categorical metrics, at various time scales (hourly to daily), along different lead times (1 to 4 days), with different meteorological input data (forecast vs reanalyzed), in order to provide a more complete vision of their behaviour.

Our study unambiguously demonstrates the value of applying such MOS corrections to improve O_3 forecasts, while showing the trade-offs between continuous and categorical skills and their dependencies on the MOS method; the best method for reproducing O_3 mixing ratios does not always represent the best method for predicting the highest O_3 episodes. For instance, despite more sophisticated MOS methods achieve the best continuous skills, we show that simpler approaches can still provide better categorical skills for the highest O_3 episodes.

The paper is organized as follows: Sect. 2 first describes the data and MOS methods used in this study; Sect. 3 includes the evaluation of the raw (uncorrected) CAMS regional ensemble O_3 forecast over the Iberian Peninsula, along with a detailed assessment of the MOS results and some sensitivity analyses; a broader discussion and conclusion are provided in Sect. 4.

2 Data and methods

2.1 Data

2.1.1 Ozone observations

Hourly O_3 measurements over 2018-2019 are taken from the European Environmental Agency (EEA) AQ e-Reporting (EEA, 2020), and accessed through GHOST v3.2.2 (Globally Harmonised Observational Surface Treatment). GHOST is a project developed at the Earth Sciences Department of the Barcelona Supercomputing Center that aims at harmonizing global surface atmospheric observations and metadata, for the purpose of facilitating quality-assured comparisons between observations and models within the atmospheric chemistry community (Bowdalo, in preparation). On top of the public datasets it ingests, GHOST provides numerous data flags that are here used for quality assurance screening (see Appendix A). 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).

Our study focuses on the Iberian Peninsula, over a domain ranging from 10°W to 5°E longitude and from 35°N to 44°N latitude that includes Spain, Portugal and part of south-western France. In total, 455 O_3 monitoring stations are included, which represents an observational dataset of 7,437,862 hourly O_3 measurements with 93% of hourly data availability.

2.1.2 CAMS regional ensemble forecast

The benefit of MOS corrections is investigated on the CAMS regional ensemble forecasts. As one of the six Copernicus services, CAMS provides AQ forecast and reanalysis data at both regional and global scales (https://www.regional.atmosphere.copernicus.eu/). At regional scale, 9 state-of-the-art CTMs developed by European research institutions are currently participating in the operational ensemble AQ forecasts (CHIMERE from INERIS, EMEP from MET Norway, EURAD-IM from University of Cologne, LOTOS-EUROS from KNMI and TNO, MATCH from SMHI, MOCAGE from METEO-FRANCE, SILAM from FMI, DEHM from Aarhus University, GEM-AQ from IEP-NRI). In addition, MONARCH from BSC and MINNI from ENEA will join the ensemble soon. The ensemble forecast is computed as the median of all individual forecasts. Note that due to possible technical failures, all 9 forecasts are not always available for computing the full ensemble. The CAMS regional forecasts are provided over 4 lead days (hereafter referred to as D+1, D+2, D+3 and D+4).

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).

2.2 Applying MOS in a worse case scenario of operational-like conditions

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: (1) no past data, neither modeled nor observed, are available for training at the beginning of the period of study (here 2018/01/01), (2) the amount of modeled and observed data continuously grows with time along the period of study (here 2018-2019). Therefore, the approach consists in mimicking what could be possible in a worse case scenario where a new operational AQ forecasting system is implemented together with a MOS module, starting from scratch, i.e. without any hindcast data or past observations available. On a given day, all MOS methods can only rely on the historical data accumulated so far. We believe that such a strategy allows to compare the different MOS methods in a balanced way given the operational context. 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. In an operational context, the first category of methods might thus be advantaged at the beginning before being gradually supplanted with the second category.

2.3 Description of the Model Output Statistics (MOS) methods

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.

2.3.1 Persistence (PERS) and moving average (MA) methods

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.

2.3.2 Quantile mapping (QM) method

The quantile mapping (QM) method aims at adjusting the distribution of the forecast concentrations to the distribution of observed concentrations. For a given day, the QM method consists in (1) computing two cumulative distribution functions (CDF), corresponding to past modeled and observed O_3 mixing ratios, respectively, (2) locating the current O_3 forecast in the model CDF and (3) identifying the corresponding O_3 values in the observation CDF and using it as the QM-corrected O_3 forecast. For instance, if the current O_3 forecast gives a value corresponding to the 95th percentile, the QM-corrected O_3 forecast will correspond to the 95th percentile of the observed O_3 mixing ratios. This approach thus aims at correcting all quantiles of the distribution, and not only the mean.

In the operational-like context in which this study is conducted (Sect. 2.2), first QM corrections are computed when 30 days of data have been primarily accumulated, to ensure a minimum representativeness of the model and observation CDFs. 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).

2.3.3 Kalman filter (KF) method

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 (k_t) 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. A variance ratio close to zero induces a Kalman gain close to 0. In such situations, the estimated forecast bias corresponds to the estimated forecast bias of the previous day, independently from the forecast error. A very high (infinite) variance ratio gives a Kalman gain close to 1. In this case, the estimated forecast bias corresponds to the observed forecast bias of the previous day, which makes it thus equivalent to the MA(1) method.

In this study, the variance ratio is adjusted dynamically and updated regularly in order to optimize a specific statistical metric, in our case the RMSE (the corresponding approach being hereafter referred to as KF(RMSE)). The different steps are: (1) at a given day of update, the KF corrections over the entire historical dataset are computed considering different values of variance ratio, from 0.001 to 100 in a logarithmic progression; (2) the RMSE is computed for each of the corrected historical time series obtained; (3) the variance ratio associated to the best RMSE is retained and used until the next update. Other choices of metric to optimize are explored in Sect. 3.4.

As for QM, for computational reasons, the update frequency is set to 30 days in this study (although, again, an update frequency of one single day would be optimal).

2.3.4 Analogs (AN) method

The analogs method (AN) implemented here consists in (1) comparing the current forecast to all past forecasts available, (2) identifying the past days with the most similar forecast (hereafter referred to as analog days or analogs), and (3) using the corresponding past observed concentrations to estimate the AN-corrected O_3 forecast (Delle Monache et al., 2011, 2013; Djalalova et al., 2015; Huang et al., 2017, e.g.,). 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). From those best analog days, the MOS-corrected forecast is computed as the weighted average of the corresponding observed concentrations, where weights are taken as the inverse of the distance metric previously computed. In comparison to a normal average, introducing the weights is expected to slightly reduce the dependence upon the number of analog days chosen.

Therefore, in the analogs paradigm, the past days of similar chemical and/or meteorological conditions are identified in the forecast (i.e. model) space while the output (i.e. the AN-corrected forecast) is taken from the observation space. The AQ model thus only serves to identify the past observed situations that look similar to the current one.

2.3.5 Machine-learning-based MOS method

In this study, we also explore the use of ML algorithms as an innovative MOS approach for correcting AQ forecasts. In ML terms, it corresponds to a supervised regression problem where a ML model is trained to predict the observed concentrations, hereafter referred to as the target or output, based on multiple ancillary variables, hereafter referred to as the features or inputs, coming from meteorological and chemistry-transport geophysical models and/or past observations. In this context, the use of ML is of potential interest because (i) we suspect that some relationships exist between the target variable and at least some of these features, (ii) these relationships are likely too complex to be modeled in an analytical way, and (iii) data are available for extracting (learning) information about them. Over the last years, ML algorithms became very popular for many types of predictions, notably due to their ability to model complex (typically non-linear and multi-variable) relationships with good prediction skills. Among the myriad of ML algorithms developed so far, we focus on the decision tree-based ensemble methods, and more specifically on the gradient boosting machine (GBM), that often gives among the best prediction skills (as shown in various ML competitions and model intercomparisons, e.g., Caruana and Niculescu-Mizil, 2005).

At each monitoring station, one single ML model is trained to forecast O_3 concentrations at all lead hours (from 1 to 96) or days (from 1 to 4), depending on the time scale used (see Sect. 2.4). The features taken into account include a set of chemical features (raw forecast O_3 concentration, O_3 concentration observed one day before), meteorological features (2-m temperature, 10-m surface wind speed, normalized 10-m zonal and meridian wind speed components, surface pressure, total cloud cover, surface net solar radiation, surface solar radiation downwards, downward UV radiation at the surface, boundary layer height, and geopotential at 500 hPa; all forecast by the meteorological model) and time features (day of year, day of week, lead hour). Although the past O_3 observed concentration corresponds to recursive information that will not be available for all forecast lead days, we use here the same value for all lead days. The tuning of the GBM models is described in Appendix D.

As for QM, the GBM model is first trained (and tuned) only after 30 days to accumulate enough data, and then retrained every 30 days based on all historical data available.

2.4 Time scales of MOS corrections

Current AQ standards are defined according to pollutant-dependent time scales, e.g. daily 8-hour maximum (d8max) concentration in the case of O_3. In the literature, MOS corrections are typically applied to hourly concentrations, providing hourly corrected concentrations from which the value at the appropriate time scale can then be computed. Following this approach, for a given MOS method X, corrections in this study are first computed based on hourly time series (hereafter referred to as

https://doi.org/10.5194/acp-2021-864
Preprint. Discussion started: 1 December 2021

X_h), from which daily 24-hour average (X_d), daily 1-hour maximum (X_{d1max}) and daily 8-hour maximum (X_{d8max}) corrected concentrations are then deduced. In addition, MOS corrections are computed directly on daily 24-hour average (X_{dd}, the additional "d" indicating that the MOS method is applied directly on daily rather than hourly time series), daily 1-hour maximum (X_{dd1max}) and daily 8-hour maximum (X_{dd8max}) time series, respectively. When needed, meteorological features are used at the same time scale. This is done to investigate whether applying the MOS correction directly at the regulatory time scale can help achieving better performance.

3 Results

We first briefly describe the O_3 pollution over the Iberian Peninsula as observed by the monitoring stations and simulated by the CAMS regional ensemble forecast (Sect. 3.1). Then, we investigate the performance of the MOS methods on both continuous (Sect. 3.2) and categorical (Sect. 3.3) O_3 forecasts. Different sensitivity tests on the MOS methods are performed in Sect. 3.4. Finally, the impact of the input meteorological data on the MOS methods performance is discussed in Sect. 3.5.

The statistical performance of the forecasts is evaluated in terms of Mean Bias (MB), normalized Mean Bias (nMB), Root Mean Square Error (RMSE), normalized Root Mean Square Error (nRMSE), Pearson correlation coefficient (PCC), slope and intercepts (inter) computed by a linear regression applied to scatter plots of simulated versus observed O_3 mixing ratios (with observations in abscissa) and normalized Mean Standard Deviation bias (nMSDB) for the continuous forecasts. Hit rate (H), false alarm rate (F), frequency bias (FB), success ratio (SR), critical success index (CSI), Peirce skill score (PSS) and area under the ROC curve (AUC) are used the categorical forecasts. All categorical metrics are defined in Appendix E.

In order to ensure fair comparisons between observations and RAW/MOS forecasts, O_3 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.

3.1 Ozone pollution over Iberian Peninsula and raw CAMS forecasts

The European Union sets different standards regarding O_3 pollution, including (1) a target threshold of 60 ppbv for the daily 8-hour maximum, with 25 exceedances per year allowed on average over 3 years, (2) an information threshold of 90 ppbv for the daily 1-hour maximum, and (3) an alert threshold of 120 ppbv for the daily 1-hour maximum. In this study, we focus on the two first thresholds and exclude the last one mainly because exceedances of the alert threshold are extremely rare (only 13 exceedances over 314,005 points, i.e. 0.004%). With such a low frequency of occurrence, such events remain extremely difficult to predict (without predicting too many false alarms).

The mean O_3 mixing ratios, as well as the annual number of exceedances, are shown in Fig. 1, for both observations and raw CAMS ensemble forecasts. The time series at the different time scales are shown in Fig. 2. Over the Iberian Peninsula, O_3 mixing ratios range between 10 and 50 ppbv, depending on the type of monitoring station (urban traffic, urban background,

https://doi.org/10.5194/acp-2021-864
Preprint. Discussion started: 1 December 2021

rural background), with typically higher levels on the Mediterranean coast compared to the Atlantic one. Over the entire domain and time period, the target (d8max > 60 ppbv) and information (d1max > 90 ppbv) thresholds have been exceeded 13,221 and 274 times, respectively (i.e. 4 and 0.08% of the 314,005 points, respectively). These exceedances are well distributed in time along the 2018-2019 period, with 404/730 days (55%) with at least one station exceeding the target threshold, and 78/730 days (11%) with at least one station exceeding the information threshold. These exceedances are observed over a large part of the peninsula, but with a higher frequency in specific locations, including the surroundings (typically downwind) of the largest cities (e.g. Madrid, Barcelona, Valencia, Lisbon, Porto) and close to industrial areas (e.g. Puertollano, a major industrial hot spot at 200 km south of Madrid).

Considering the annual mean O_3 mixing ratios at all 456 stations, the CAMS ensemble forecast represents moderately well the spatial distribution of annual O_3 over the Iberian Peninsula (PCC of 0.54 for D+1 forecasts), and strongly underestimates the spatial variability (nMSDB of -42%), but bias and error remain reasonable (+19 and 25%, respectively). 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. Overall, considering all hourly O_3 forecasts at D+1, the CAMS ensemble shows a nMB/nRMSE/PCC of +19%/39%/0.75 (N=5,984,454) at the hourly scale. The 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).

3.2 Performance of MOS methods on continuous forecasts

The overall statistical results are shown in Fig. 3 for the different MOS methods. For a given lead day and time scale, statistics are computed after aggregating data at all monitoring stations. Therefore, statistics of D+1 O_3 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.

As mentioned in Sect. 3.1, the RAW O_3 D+1 forecasts over the Iberian Peninsula show a moderate overestimation with nMB around +18% at hourly and daily scales, reduced to +7 and +2% at d8max and d1max scales, respectively. Similarly, the nRMSE ranges between 38% at the hourly scale and 19% at the d1max scale. A reasonable correlation is obtained, around 0.75-0.79 depending on the time scale. The variability appears substantially underestimated, with a nMSDB between -28 and -30%, which is reflected in the low model-versus-observation linear slope obtained (between 0.53 and 0.57 depending on the time scale). 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. Such a slow decrease in performance might be due to the relatively coarse resolution of the CAMS forecasts.

The impact of the MOS corrections on the performance strongly varies with the method considered. As expected (by construction), the most basic PERS(1) method gives unbiased O_3 forecasts with unbiased variability (nMB and nMSDB of 0%).

https://doi.org/10.5194/acp-2021-864
Preprint. Discussion started: 1 December 2021

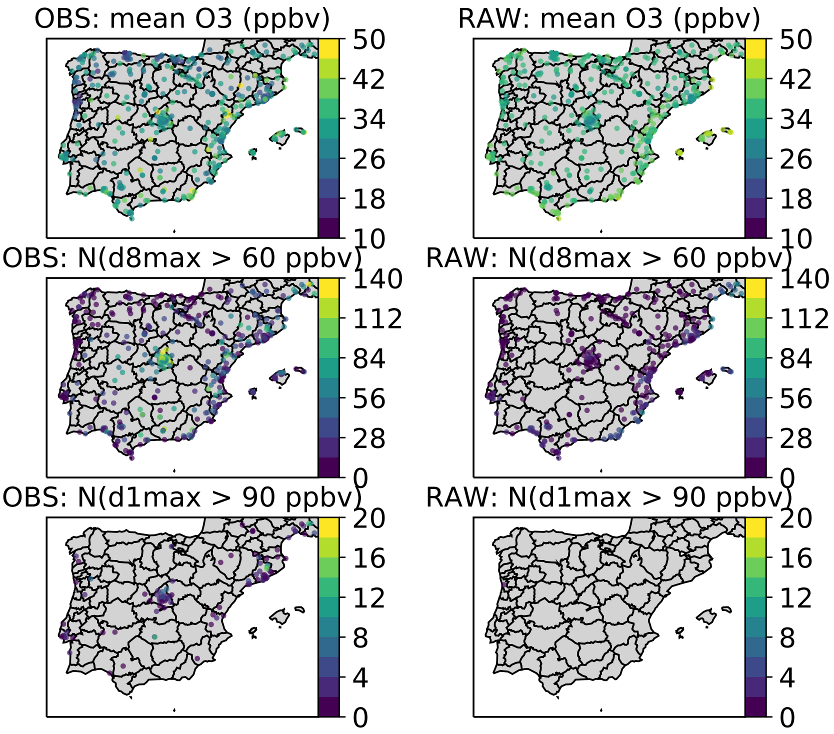

Figure 1. Overview of the O_3 pollution over the Iberian Peninsula, as observed by monitoring stations (left panels) and as simulated by the CAMS regional ensemble D+1 forecasts (right panels), showing the mean O_3 mixing ratios (top panels), and the number of exceedances of the standard (d8max > 60 ppbv; middle panels) and information threshold (d1max > 90 ppbv; bottom panels), over the period 2018-2019. For clarity, the stations without any observed or simulated exceedance are omitted.

Due to the temporal auto-correlation of O_3 concentrations, reasonable results are obtained at D+1. Compared to RAW, the PERS(1) method slightly reduces the nRMSE (36% at hourly scale), but does not improve the PCC (0.75). Although still too low, the slope is also greatly improved, with 0.75 at hourly scale (up to 0.84 at d8max scale). However, the performance of this simple method quickly deteriorates with lead time, down to nRMSE/PCC of 42%/0.65 at D+4, thus worst than the RAW forecast. The MA(1) method also allows to remove the bias and to correct most of the underestimated variability (absolute nMSDB below 2%). It 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 slightly deteriorates with lead time, but less dramatically than with PERS(1). The QM method shows quite similar results as the MA(1) method, but usually with slightly worse results at short lead time and better ones at longer lead time (thus with slower performance deterioration with lead time). When considering the hourly time scale, the KF, AN and GBM methods give relatively similar results on most of these continuous statistical metrics except the slope and nMSDB that are slightly better with KF (followed by GBM). Some negative biases are introduced by these MOS methods, essentially at d1max scale, as well as d8max scale for GBM specifically. Interestingly, applying these

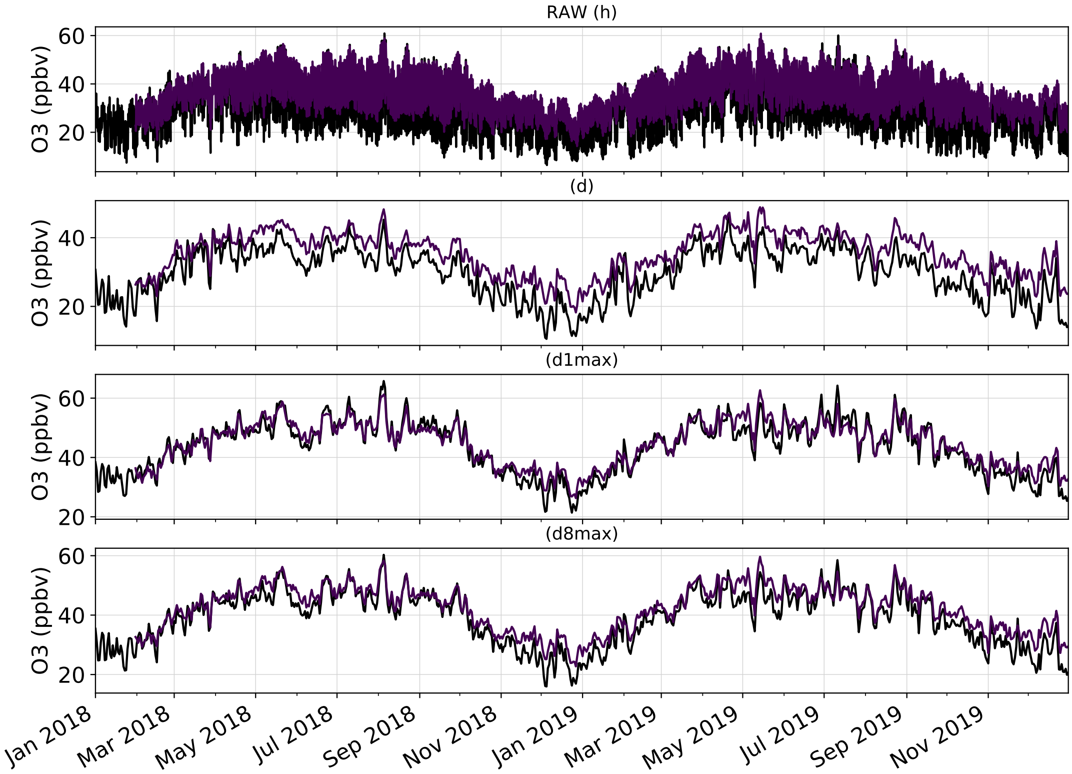

Figure 2. Time series of the mean O_3 mixing ratios over the Iberian Peninsula, as observed by monitoring stations (in black) and as simulated by the CAMS regional ensemble D+1 forecasts (in purple). Time series are shown at the hourly (h), daily mean (d), daily 1-hour maximum (d1max) and daily 8-hour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

MOS methods directly on d1max or d8max O_3 mixing ratios rather than hourly data (i.e. dd1max and dd8max scales) removes most of these biases. However, KF, AN and GBM all outperform the previous MOS methods in terms of nRMSE (about 25%) and PCC (about 0.86) that are substantially improved. Their main limitation lies in the variability that remains underestimated (nMSDB around -10%), although less than in RAW (-29%).

Note that some MOS methods (QM, AN and GBM) ingest increasing amounts of input data over time, and their performance is thus expected to be relatively lower at the beginning of the period, yet increase with time. Comparing the relative change of nRMSE and PCC obtained during the last year (2019) against those previously discussed (period 2018-2019), while RAW shows a slight relative deterioration of its performance (nRMSE increased by +2% and no change of PCC), all MOS methods depict a small relative improvement. Interestingly, the improvement for GBM is substantially larger than the other MOS methods, with nRMSE decreased by 5% (against +1 to +2% for the other methods) and PCC increased by +2% (against

+0 to +1% for the other methods). This improvement is mainly due to the relatively poor predictions made during the very first
305 months of 2018 when the training dataset was the most limited (see time series in Fig. F1 in Appendix F).

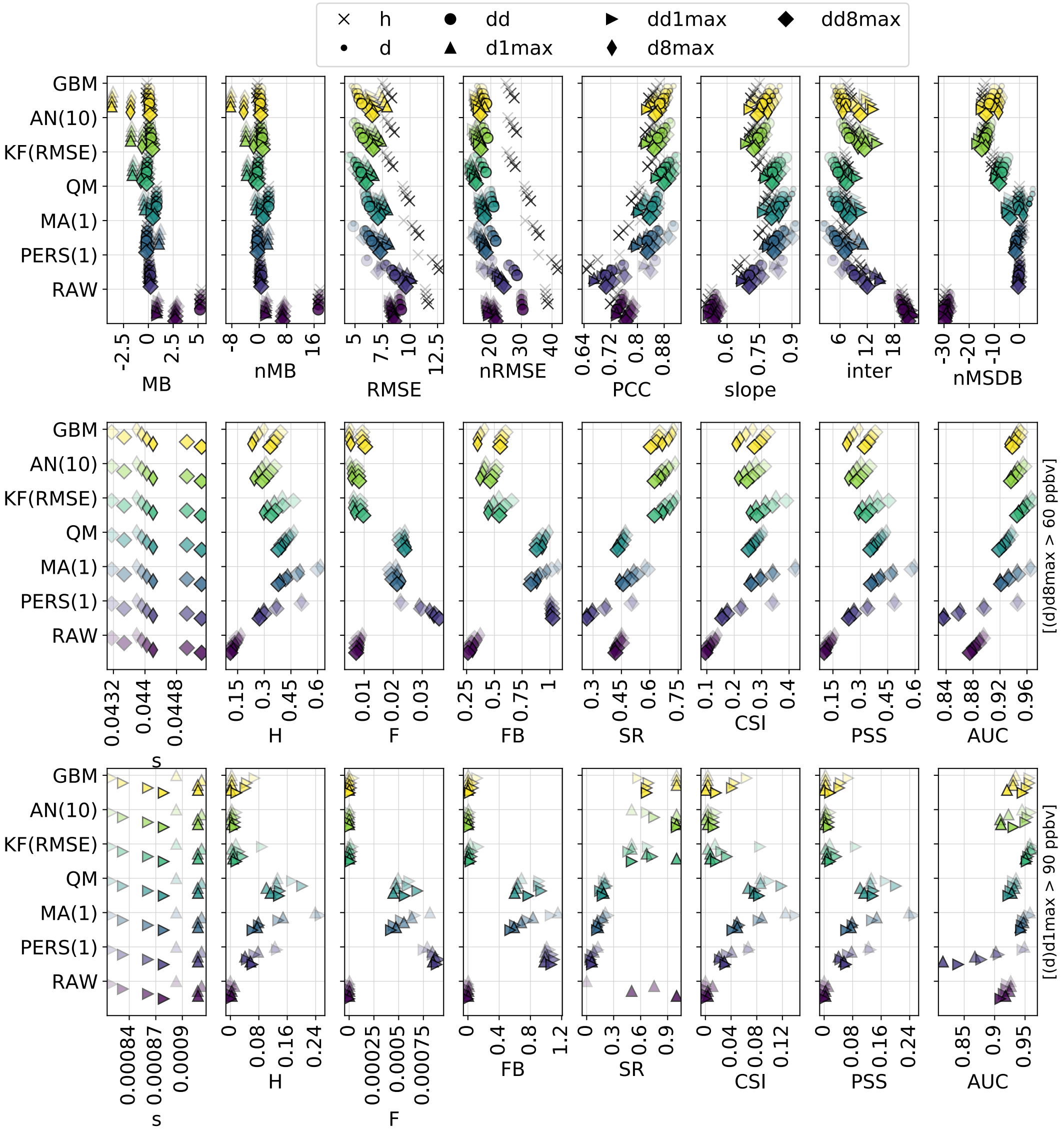

Figure 3. Statistical performance of RAW and MOS-corrected CAMS O_3 forecasts, for lead days D+1 to D+4 (ordered from top to bottom in each MOS method panel, with decreasing transparency) and different time scales (h: hourly; d: daily mean; d1max/dd1max: daily 1-hour maximum; d8max/dd8max: daily 8-hour maximum). See Sect. 2.4 for details on time scales and Appendix E for metrics definitions.

https://doi.org/10.5194/acp-2021-864
Preprint. Discussion started: 1 December 2021

3.3 Performance of MOS methods on categorical 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 relevant ones are probably CSI and PSS, followed by SR and AUC.

As previously foreseen, despite RAW is very "conservative" with low H and F (very few true positives and false negatives), it does not benefit from a strong SR (0.45), and finally shows the worst performance in terms of CSI (0.10) or PSS (0.15). The term "conservative" here refers to forecasting systems that predict exceedances only with strong evidence; it thus predicts very few exceedances but with higher confidence. It follows relatively well the variability of O_3 (as shown by a reasonably good AUC) but dramatically fails at reaching high O_3 mixing ratios, as illustrated by the low FB (0.25). Even a basic method like PERS(1) provides better detection skills regarding target thresholds. This is especially true during the first lead days, but the performance quickly decreases along 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. However, the usefulness of having geophysical O_3 forecasts is nicely illustrated by the results obtained with MA(1), QM and KF(RMSE), the MOS methods relying only on both RAW and observed O_3 data. Indeed, these methods show among the best CSI and/or PSS results, not so far from the two last methods, AN(10) and GBM. For short lead times (D+1), MA(1) clearly outperforms the other methods, especially for PSS. Differences of performance are reduced when considering longer lead times. At D+4, best CSI are obtained with KF(RMSE)$_{\mathrm{dd1max}}$ and GBM$_{\mathrm{dd1max}}$ (0.28), while best PSS are achieved by QM and MA(1). More generally, KF(RMSE), AN(10) and GBM appear as the most "conservative" MOS approaches here, with relatively low H and F, but a strong SR. In other terms, they predict fewer exceedances but with a higher reliability.

However, when considering the detection of the information threshold (d1max O_3 above 90 ppbv), the KF(RMSE), AN(10) and GBM methods still benefit from a strong SR but are missing too much the observed exceedances, which leads to a dramatic deterioration of both CSI and PSS. This means that there is a high change that an exceedance predicted by these methods indeed occurs but such exceedances are too rarely predicted. For detecting such high O_3 values, best methods are finally MA(1) for shortest lead times, and QM for longer ones. Both methods reproduce fairly well the geographical distribution of high O_3 episodes (PERS(1) reproduces it perfectly, by construction), as shown in Fig. 4, but still with very low SR (below 0.25 for exceedances of the information threshold). Note that the RAW model alone misses almost all exceedances of the information threshold.

3.4 Sensitivity tests

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 behaviour 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.

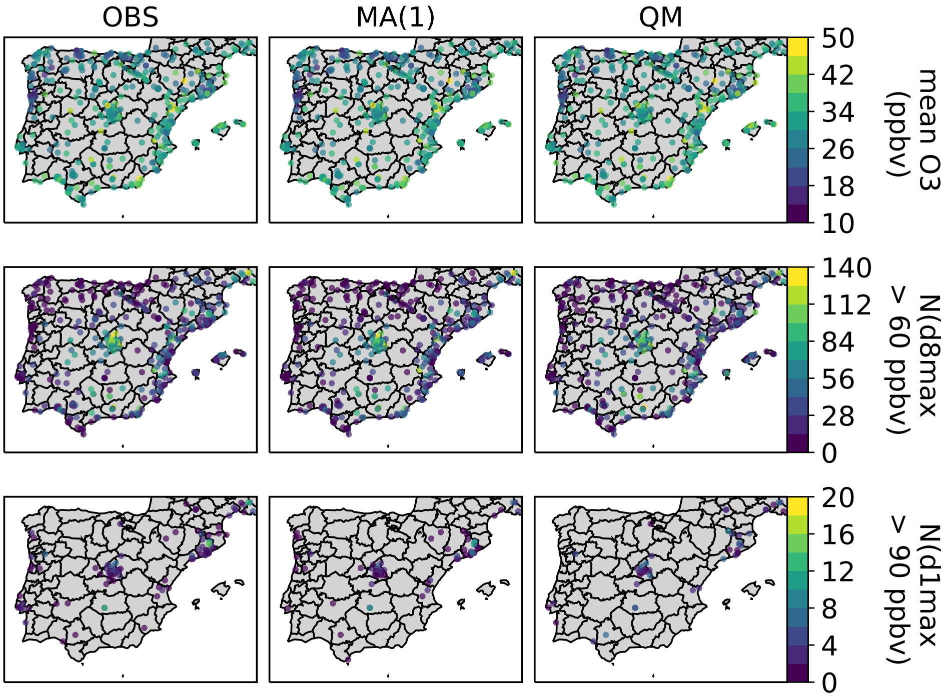

Figure 4. Similar to Fig. 1 but for observations, and D+4 O_3 forecasts corrected with MA(1) and QM methods.

3.4.1 Persistence method

The persistence method essentially relies on the choice of the time window over which past observations are averaged to provide the O_3 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 O_3 mixing ratios peak. For illustration purposes, let's assume that O_3 peaks between 15 and 17 h; on a given day, O_3 mixing ratios at 15/16/17h reach 50/60/50 ppbv and on the following day 70/70/80 ppbv. Then, the PERS(2)$_{dd1max}$ O_3 would be 70 ppbv (mean of 60 and 80 ppbv), while the PERS(2)$_{d1max}$ O_3 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

https://doi.org/10.5194/acp-2021-864
Preprint. Discussion started: 1 December 2021

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.

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. However, it gives very "liberal" O_3 forecasts with rather poor SR. The term "liberal" is here borrowed from (Fawcett, 2006) to designate forecasting systems that predict exceedances with weak evidence, in opposition with the aforementioned term "conservative". Longer time windows can improve SR, but result in an important deterioration of CSI and PSS, particularly for the shorter lead times (D+1/D+2).

3.4.2 Moving average method

Similarly to PERS, the MA method depends on the time window over which past model biases are averaged to correct the forecast. Similarly, a sensitivity test is performed with windows ranging between 1 and 10 d (hereafter referred to as MA(n) with n the window in days). Results are shown in Fig. G2 in Appendix G. Increasing the window length impacts the MA performance in a very similar way than for PERS, especially in terms of continuous metrics for which the sensitivity is almost exactly the same. Regarding the detection of the target threshold (d8max O_3 above 60 ppbv), the main noticeable difference is the absence of strong deterioration of some metrics like AUC, SR or CSI for shorter lead times. Regarding the detection of the information threshold (d1max O_3 above 90 ppbv), the clearest difference with PERS concerns the SR that substantially improves when considering longer windows. However, the deterioration of both CSI and PSS persists.

Therefore, the detection of O_3 exceedances with the MA method shows its best skills with shortest windows (1 d). As for PERS, the corresponding forecasts are quite liberal with low SR. However, in contrast to PERS, the SR associated to strong thresholds (d1max above 90 ppbv) can be substantially improved when using longer windows, which may be an interesting option if the corresponding deterioration of CSI/PSS is seen as acceptable.

3.4.3 Kalman filter method

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 and Stephenson, 2011; Ferro and Stephenson, 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 O_3 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.

3.4.4 Analog method

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. 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 O_3 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 O_3 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.

3.4.5 Gradient boosting machine method

Although GBM gives among the best RMSE and PCC, it strongly underestimates the variability of O_3 mixing ratios, with critical consequences in terms of detection skills, especially for the highest thresholds (e.g. d1max > 90 ppbv). This is at least partly due to the low frequency of occurrence of such episodes, and their corresponding low weight in the entire population of points used for the training. One way of mitigating this issue consists in specifying different weights to the different training instances. This aims at forcing the GBM model to better predict the instances of higher weight, at the cost of a potential deterioration of the performance on the instances of lower weight.

In order to assess to which extent it may improve the performance of the GBM MOS method, we tested different weighting strategies. At each training phase, we compute the absolute distance D between all observed O_3 mixing ratio instances and the mean O_3 mixing ratio (averaged over the entire training dataset). Then several sensitivity tests are performed, weighting the training data by D, D^2 and D^3, respectively (hereafter referred to as GBM(W), GBM(W2), GBM(W3), respectively). Using such weights, we want the GBM model to better predict the lower and upper tails of the O_3 distribution in order to better represent the variability of the O_3 mixing ratios. Given that the O_3 mixing ratio distribution is typically positively skewed, the highest weights are put on the strongest positive deviations from the mean.

As a parallel sensitivity test, we explore the performance of these different ML models but removing the input feature corresponding to the previous (one day before) observed O_3 mixing ratio (hereafter referred to as GBM(noO), GBM(noO,W), GBM(noO,W2) and GBM(noO,W3)). This additional test is of interest for operational purposes since O_3 observations are not always available in near real-time. 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.

As expected, results highlight a deterioration of the RMSE and PCC combined with an improvement of the slope, intercept and nMSDB. The negative bias affecting the variability with the unweighted GBM is substantially reduced when using weights, although too strong weights (as in GBM(W3) for instance) can lead to a slight overestimation of the variability at specific time scales.

Regarding the skills for detecting d8max O_3 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 O_3 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 O_3 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.

Therefore, adopting an appropriate weighting strategy is simple yet effective for achieving slightly better O_3 exceedance detection skills in exchange of a reasonable deterioration in RMSE and PCC. Overall, the improvements are relatively small, but still valuable given the initially very low detection skills for the strongest O_3 episodes.

3.5 Influence of the meteorological input data in AN and GBM methods

In the previous sections, O_3 corrections with AN and GBM methods relied on IFS meteorological forecasts. Here, we investigate the impact of using ERA5 reanalysis. Generally, the hourly O_3 predictions with both meteorological datasets are consistent. Assuming ERA5-based O_3 mixing ratios as the truth, the IFS-based O_3 predictions with AN(10) method show a nRMSE/PCC of 8%/0.98 at D+1 (N=7,067,085), slowly deteriorating up to 10%/0.97 at D+4 (N=6,960,524). Similarly, nRMSE/PCC with GBM method evolves from 12%/0.96 to 13%/0.95. Whatever the lead day or the MOS method, no differences are found between the ERA5-based and IFS-based predictions.

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 O_3 predictions are close, the statistical performance against observations is also very consistent between both meteorological datasets. As expected, the performance is slightly lower with IFS data and the discrepancies increase with lead time. Using IFS rather than ERA5 data increases the nRMSE of AN(10) by 1% at D+1 and by 5% at D+4. This relative deterioration at D+4 depends upon the MOS method, with 5, 4, 4 and 8% for AN(1), AN(5), AN(10) and GBM, respectively. Similarly, the PCC is slightly reduced when using IFS data, by only 1% at D+1 whatever the MOS method, and up to 3, 2, 2 and 3% for AN(1), AN(5), AN(10) and GBM, respectively at D+4. Therefore, the sensitivity to the quality of the meteorological input data varies with the MOS method and the metric considered, and GBM is the most sensitive to this aspect.

Overall, similar conclusions can be drawn for categorical metrics. GBM shows a relative deterioration of CSI/PSS from -7 to -9%. Again, this deterioration of the performance is also observed with the AN method, with up to -5, -8 and -8% for AN(1), AN(5) and AN(10), respectively.

Compared to IFS, the ERA5 reanalysis undoubtedly benefits from the assimilation of many meteorological observations but has conversely a coarser spatial resolution (about 31 versus 9 km), which may have a negative impact on its reliability, especially in specific areas (e.g. complex orography, urban areas). All in all, ERA5 likely gives better meteorological information, in particular for longer lead times. Here, our results show that a better performance of the MOS correction can be obtained using such higher quality meteorological inputs. At the same time, the deterioration introduced by the use of IFS forecasts remains relatively small (at lead times below 4 d). 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 O_3 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 O_3, the GBM models can also learn indirectly at least part of the potential errors affecting some meteorological variables and how they relate to O_3 mixing ratios. Therefore, the presence of systematic biases in some of the ancillary features is not expected to strongly impact the performance of the predictions. However, results at longer lead times are still clearly better with ERA5 than with IFS, because of the chaotic nature of weather and the unavoidable increase of errors with lead time.

4 Discussion and conclusions

We demonstrate the strong impact of MOS methods to enhance raw CAMS O_3 forecasts, not only by removing potential systematic biases but also for correcting other issues related to the distribution and/or variability of O_3 mixing ratios. Apart from the PERS method, all MOS approaches were indeed able to substantially improve at least some aspects of the RAW O_3 forecasts, first and foremost the RMSE and PCC, for which the strongest improvements are obtained with most sophisticated MOS methods like KF, AN or GBM. However, although all MOS methods were able to increase the underestimated variability of O_3 mixing ratios of RAW, the strongest improvements of slope and nMSDB were obtained with more simple MOS methods like MA or QM. O_3 mixing ratios corrected with AN, GBM and to a lesser extent KF, remained too smooth, and such a deficiency has a major impact on the detection skills for high O_3 thresholds. All in all, the best PSS or CSI are usually obtained with the more simple MOS methods. Therefore, there is a clear trade-off between the continuous and categorical skills scores, as also shown by the different sensitivity tests. The quality of a MOS-corrected forecast assessed solely based on metrics like RMSE or PCC thus tells little about the forecast value, here understood as an information a user can benefit from to make better decisions, notably for mitigating O_3 short-term episodes.

More generally, our study highlights the complexity of identifying the "best" MOS method given the multiple dimensions of the problem. The relative performance of the MOS methods can vary depending on the metric used, the threshold considered in the case of categorical metrics (or more specifically the base rate), the time scale at which MOS corrections are computed and/or evaluated, or the lead time. Other dimensions not covered by this study, like the seasonality of the performance, are also susceptible of shedding a different light on the inter-comparison.
Among the continuous metrics, both RMSE and PCC provide a first valuable information on the performance of a MOS method. However, a MOS method can give the best RMSE and PCC, yet the poorest high O_3 detection skills. This was the case of the unweighted GBM method. Continuous metrics like the model-versus-observation linear slope or nMSDB provide important complementary information, potentially less misleading, especially in a context where the final objective is to predict episodes of strong O_3. Among the categorical metrics, although results were presented on a relatively large set of metrics, all metrics do not benefit from the same properties. PSS may be considered as one of the most valuable, notably due to its independence from the base rate, in contrast to CSI. Such a property is particularly useful when comparing scores over different regions and/or time periods where the frequency of observed exceedances might vary, for instance due to different emission forcing and/or meteorological conditions. In an operational context where statistical metrics are continuously monitored, the independence from the base rate is an interesting property because it may change with time, which prevents from a consistent comparison between different periods. However, a well-known issue of both PSS and CSI (as well as many other categorical metrics) is that they degenerate to trivial values (either 0 or 1) as events become rarer (Jolliffe and Stephenson, 2011; Ferro and Stephenson, 2011), which should restrict their use to the detection of not too rare (and therefore not too high) O_3 episodes. In this study, the base rate of the target threshold was likely sufficiently high (s around 5%), but we were probably already at the limit regarding the information threshold (s around 0.1%). All in all, the selection of the evaluation metrics depends on the subjective choices and intended use, and is fundamentally a cost-loss problem where the user should arbitrate between the cost of missing exceedances and predicting false alarms.

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. This aspect is important, although different users may have different needs in terms of lead time, depending on the intended use of the AQ forecast. Forecasts at D+1 may already be useful for some applications like warning in advance the vulnerable population so that they could adapt their outdoor activities. However, implementing short-term emission reduction measures at local scale usually goes through decisions taken at different administrative and political levels, and thus typically requires forecasts at least at D+2. If such measures would have to be taken at larger scale, the occurrence of O_3 episodes would probably need to be forecasted even more in advance.

We saw that some MOS methods like PERS can provide a reasonable performance at D+1 but quickly deteriorate when looking further in the future. Actually, the performance of a basic method like PERS(1) obviously depends on the typical duration of O_3 episodes over the region of study; one (single) episode is defined here as a suite of successive days showing an exceedance of a given threshold at a given station. Over the Iberian Peninsula domain in 2018-2019, considering the target threshold (d8max > 60 ppbv), a total of 6,540 such O_3 episodes were observed on the O_3 monitoring network with min/mean/max duration of 1/2/27 d (and 5th/25th/50th/75th/95th percentiles of 1.0/1.0/1.0/2.0/5.0 d). Note the 27-d-long O_3 exceedance occurred in June-July 2019 at about 30 km north of Madrid (station code *ES1802A*). Considering the information threshold, 240 episodes were observed, with min/mean/max duration of 1/1.1/5 d (and 5th/25th/50th/75th/95th percentiles of 1.0/1.0/1.0/1.0/2.0 d). This may partly explain why the deterioration of performance with lead time was stronger for target thresholds compared to information thresholds.

The performance of the MA(1) method also substantially depends on the lead time, although less than PERS(1). Conversely, some MOS methods like GBM, AN or QM were less impacted by the increasing lead time.

By comparing the MOS results obtained with ERA5 reanalysis data rather than IFS forecasts, we demonstrated that higher-quality meteorological input data helps improving the performance of the prediction. However, the improvement obtained with ERA5 was relatively small, which is an important result for the use of MOS in an operational context where only meteorological forecasts can be used. Although data are so far only available until July 2019, it would be interesting in the near-future to extend the present analysis using the UERRA regional reanalysis for Europe that provides meteorological information at a refined spatial resolution of 5.5x5.5 km^2 (https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-uerra-europe-single-levels?tab=overview).

For operational purposes, other important aspects are to be taken into account. A first aspect concerns the input data required by the MOS method. Does the MOS method rely on observations, models or a combination of both? When the method relies on observations, are they needed in near real-time? How much historical data are required? When the method relies on historical data, to which extent the length of the historical dataset impacts the performance? Related to this last point, another essential aspect concerns the ability of the MOS method to handle progressive and/or abrupt changes in the AQ forecasting system (e.g. configuration, parameterizations, input data like emissions) and/or in the Earth's atmosphere (long-term trends, anomalous events like the COVID-19-related emission reduction, climate change). In this frame, the year 2020 obviously offers a unique large-scale case study to investigate the behaviour of the different MOS methods.

MOS methods relying only on very recent data (namely PERS, MA and KF methods) are evidently more adaptable to rapid changes, which is a clear asset under changing atmospheric conditions or modeling system configurations. On the other hand, they naturally discard all the potentially useful information available within the historical dataset. Methods like QM, AN or GBM aim at extracting such information to produce better forecasts, but implicitly rely on the assumption that these historical data are still up-to-date and thus representative of the current conditions, which can be a too strong hypothesis when the historical dataset is long or the emission forcing and/or meteorological conditions are changing rapidly. 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 implemented a relatively simple ML-based MOS method. Although the performance on categorical metrics was found limited despite encouraging results on continuous metrics, there is likely room for improvements in near-future developments. In order to improve the high O_3 detection skills, potential interesting aspects to explore include testing other types of ML models, customizing loss function and/or cross-validation scores, designing specific weighting strategies and/or re-sampling approaches or comparing regression and classification ML models for the detection of exceedances. Along the preparation of this study, some of them have been investigated but more efforts are required to draw firm conclusions regarding their potential for better predicting O_3 episodes. Finally, we focused here on the CAMS regional ensemble but including the individual CAMS models in the set of ML input features may help achieving better performance if the ML model is somehow able to learn the variability (in time and space, or during specific meteorological conditions) of strengths and weaknesses of each model and build its predictions based on the most appropriate sub-set of individual models. More generally, the performance of the different MOS methods is expected to vary from one raw model to another. Investigating the performance and behavior of these methods on the different individual models might shed an interesting light on the results obtained here with the ensemble, and eventually allow generalizing some of our conclusions.

Data availability. The EEA AQ e-Reporting and ERA5 dataset used in this study are publicly available.

Appendix A: Quality assurance with GHOST

Using the metadata available in GHOST (Globally Harmonised Observational Surface Treatment), a quality assurance screening is applied to O_3 hourly observations, in which the following data are removed : missing measurements (GHOST's flag 0), infinite values (flag 1), negative measurements (flag 2), zero measurements (flag 4), measurements associated with data quality flags given by the data provider which have been decreed by the GHOST project architects to suggest the measurements are associated with substantial uncertainty or bias (flag 6), measurements for which no valid data remains to average in temporal window after screening by key QA flags (flag 8), measurements showing persistently recurring values (rolling 7 out of 9 data points; flag 10), concentrations greater than a scientifically feasible limit (above 5000 ppbv) (flag 12), measurements detected as distributional outliers using adjusted boxplot analysis (flag 13), measurements manually flagged as too extreme (flag 14), data with too coarse reported measurement resolution (above 1.0 ppbv) (flag 17), data with too coarse empirically derived measurement resolution (above 1.0 ppbv) (flag 18), measurements below the reported lower limit of detection (flag

22), measurements above the reported upper limit of detection (flag 25), measurements with inappropriate primary sampling for preparing NO_2 for subsequent measurement (flag 40), measurements with inappropriate sample preparation for preparing NO_2 for subsequent measurement (flag 41) and measurements with erroneous measurement methodology (flag 42).

Appendix B: Kalman filter

CAMS forecasts are available over 4 lead days, from D+1 to D+4. We define here the time t as the day D at a given hour of the day ($t+1$ thus corresponds to D+1 at this specific hour of the day). In an operational context, observations at this hour of the day are available only until time t (included). In this frame, the primary objective of the Kalman filter is to estimate the so-called $x_{t+1|t}$, that designates here the true (unknown) forecast bias at time $t+1$ using the information available until t (included). We distinguish estimated values from true values using an hat (ˆ) ($\hat{x}_{t+1|t}$ therefore corresponds to the estimated value of $x_{t+1|t}$). In its application as a MOS method, the Kalman filter considers the following *system equation* for describing the time evolution of the true forecast bias :

$$x_{t+1|t} = x_{t|t-1} + \eta_t \tag{B1}$$

where η_t is assumed to be a white noise term with normal distribution, zero-mean, variance σ_η^2 and uncorrelated in time.
It also assumes that the forecast error (forecast minus observation) y_t observed at time t does not represents the true $x_{t|t-1}$ due to the presence of some random error ϵ_t:

$$y_t = x_{t|t-1} + \epsilon_t \tag{B2}$$

where ϵ_t is assumed to be a white noise term with normal distribution, zero-mean, variance σ_ϵ^2 and uncorrelated in time, and independent from η_t.
Using the Kalman filter theory, it can be demonstrated that the optimal estimate for the true forecast bias $x_{t+1|t}$ can be obtained from the following equations :

$$K_t = (p_{t-1|t-2} + \sigma_\eta^2)/(p_{t-1|t-2} + \sigma_\eta^2 + \sigma_\epsilon^2) \tag{B3}$$

$$\hat{x}_{t+1|t} = \hat{x}_{t|t-1} + K_t(y_t - \hat{x}_{t|t-1}) \tag{B4}$$

$$\hat{p}_{t+1|t} = (\hat{p}_{t|t-1} + \sigma_\eta^2)(1 - K_t) \tag{B5}$$

where K_t corresponds to the so-called Kalman gain used to weight the respective importance of the previous forecast bias estimate ($\hat{x}_{t|t-1}$) and its observed value (y_t), and $\hat{p}_{t+1|t}$ the expected error of the forecast bias estimate (i.e. the variance of the forecast bias error : $\hat{p}_{t+1|t} = Var(x_{t+1|t} - \hat{x}_{t+1|t})$).

In practise, the KF algorithm first requires initializing the $\hat{x}_{0|-1}$ and $\hat{p}_{0|-1}$ values (any reasonable value can be chosen, given that the KF quickly converges). Then the algorithm starts its first iteration for $t = 0$, which includes the sequential calculation of : (1) the forecast error y_0, (2) the Kalman gain K_0 (using y_0, $\hat{p}_{0|-1}$, σ_η^2 and σ_ϵ^2 in Eq. B3), (3) both the $\hat{x}_{1|0}$ and $\hat{p}_{1|0}$ (using

https://doi.org/10.5194/acp-2021-864
Preprint. Discussion started: 1 December 2021

K_0, $\hat{x}_{0|-1}$ and y_0 in Eq. B4, and K_0, $\hat{p}_{0|-1}$ and σ_η^2 in Eq. B5, respectively).
Solving these equations requires assigning values to both σ_η^2 and σ_ϵ^2. It can be demonstrated that, once σ_ϵ^2 is set to a fixed value (any reasonable value can be chosen, for instance $\sigma_\epsilon^2 = 1$), the KF results solely depend on the $\sigma_\eta^2/\sigma_\epsilon^2$ variance ratio. Various strategies can be used to choose an appropriate value for this variance ratio. This aspect is discussed in Sect. 2.3.3.

Appendix C: Analogs norm

The analogs (AN) method requires to identify which past forecast days are the most similar to the current one. Given a set of features to take into account, this similarity is computed using the norm introduced by (Delle Monache et al., 2006) :

$$\|F_t, A_{t'}\| = \sum_{i=1}^{N} \frac{w_i}{\sigma_i} \sqrt{\sum_{t=-T}^{T} \left(F_{i,t+k} - A_{i,t'+k}\right)^2} \tag{C1}$$

with F_t the raw forecast at time t, $A_{t'}$ an analog forecast at time t', N the number of features taken into account, w_i the weight of the feature i, σ_i its standard deviation calculated over past forecasts, T the half the width of the time window over which to compute the metric (i.e. a value $T = 2$ means that the squared difference between the forecast and the analog will be computed over a ±2 hours time window). In our study, we used weights of 1 for all features (wind speed, wind direction, temperature, surface pressure) and $T = 1$.

Appendix D: Tuning of the GBM models

The GBM models are tuned using a so-called *randomized search* in which a range of values is given for each hyperparameter of interest and a total number of hyperparameters combinations to test. After fixing the learning rate to 0.05 (*learning_rate* in the *scikit-learn* Python package), the tuning of the GBM model was done over the following set of hyperparameters: the tree maximum depth (*max_depth* : from 1 to 5 by 1), the subsample (*subsample* : from 0.3 to 1.0 by 0.1), the number of trees (*n_estimators*: from 50 to 1000 by 50) and the minimum number of samples required to be at a leaf node (*min_samples_leaf*: from 1 to 50). As we are dealing here with time series, this tuning is conducted through a rolling-origin cross-validation in which validation data are always posterior to train data.

Appendix E: Evaluation metrics

Continuous forecasts of hourly pollutant concentrations are evaluated in terms of Mean Bias (MB), normalized Mean Bias (nMB), Root Mean Square Error (RMSE), normalized Root Mean Square Error (nRMSE), and Pearson correlation coefficient (PCC).

The performance of the categorical forecasts of exceedances of the AQ standard can primarily be described through a contingency table (Tab. E1). Based on these individual numbers a (hits), b (false alarms), c (misses) and d (correct rejections),

https://doi.org/10.5194/acp-2021-864
Preprint. Discussion started: 1 December 2021

Table E1. Schematic contingency table for deterministic forecasts of binary exceedances of the regulatory limit values.

	Exceedance observed		
Exceedance forecast	Yes	No	Total
Yes	a (hits)	b (false alarms)	$a+b$
No	c (misses)	d (correct rejections)	$c+d$
Total	$a+c$	$b+d$	$a+b+c+d=n$

a wide number of verification metrics have been proposed in the literature, often with inconsistent nomenclature. In order to avoid confusions, all metrics used in this paper systematically follow the nomenclature given in the reference book of Jolliffe and Stephenson (2011).

For a given total number of data n $(= a+b+c+d)$, the 2x2 contingency table can be fully described by three independent measures, namely the base rate s independent from the forecasting system (total proportion of observed exceedances, also known as the climatological probability of an exceedance), the hit rate H (proportion of the observed exceedances that are correctly detected) and the false alarm rate F (proportion of the observed non-exceedances erroneously forecast as exceedances, to be distinguished from the false alarm ratio) defined as follows:

$$s = (a+c)/n \tag{E1a}$$

$$H = a/(a+c) \tag{E1b}$$

$$F = b/(b+d) \tag{E1c}$$

Any categorical metric initially function of a, b, c and d can be expressed in terms of s, H and F. One interest of considering this s-H-F framework (so-called likelihood-base rate factorization, see chapter 3 of Jolliffe and Stephenson (2011) for a detailed description) lies in the fact that, since the forecaster does not have any influence on s, the tri-dimensional problem is reduced to bi-dimensional (H and F). Since it is easily possible to maximize H (by always predicting an exceedance) or F (by always predicting a non-exceedance), none of these two metrics taken individually is a good and balanced metric for assessing the quality of a forecasting system; only some combinations of both (eventually with s) can eventually provide a good way to assess this detection skills. Some common examples of metrics are the Proportion of Correct (PC), the Frequency Bias (FB), the False Alarm Ratio (FAR), the Success Ratio (SR), the Critical Success Index (CSI), the Gilbert Skill Score (GSS) or the Peirce Skill Score (PSS), defined as follows:

$$\mathrm{PC} = (a+d)/n = (1-s)(1-F) + sH \tag{E2a}$$

$$\mathrm{FB} = (a+b)/(a+c) = (1-s)F/s + H \tag{E2b}$$

$$\mathrm{FAR} = b/(a+b) = \left[1 + \left(\frac{s}{1-s}\right)\frac{H}{F}\right]^{-1} \tag{E2c}$$

$$\mathrm{SR} = (a)/(a+b) = 1 - \left[1 + \left(\frac{s}{1-s}\right)\frac{H}{F}\right]^{-1} \tag{E2d}$$

$$\mathrm{CSI} = a/(a+b+c) = \frac{H}{1+F(1-s)s} \tag{E2e}$$

$$\mathrm{GSS} = \frac{a-a_r}{a+b+c-a_r} = \frac{H-F}{(1-sH)/(1-s)+F(1-s)/s} \tag{E2f}$$

$$\mathrm{PSS} = \frac{ad-bc}{(b+d)(a+c)} = H-F \tag{E2g}$$

with $a_r = (a+b)(a+c)/n$ the expected number of hits (a) for a random forecast with the same s (meaning that GSS is an equivalent of CSI where the number of hits is adjusted for the hits that are associated to random chance, due to the climatology frequency of the event).

Jolliffe and Stephenson (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 and Stephenson, 2011). In this framework, among the previous metrics, we retained PSS as the best choice for assessing the skills of our MOS methods, given that it gathers numerous interesting 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 with a fixed range). Note also 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).
Finally, another useful metric is the Area under the ROC curve (AUC).

Appendix F: Time series

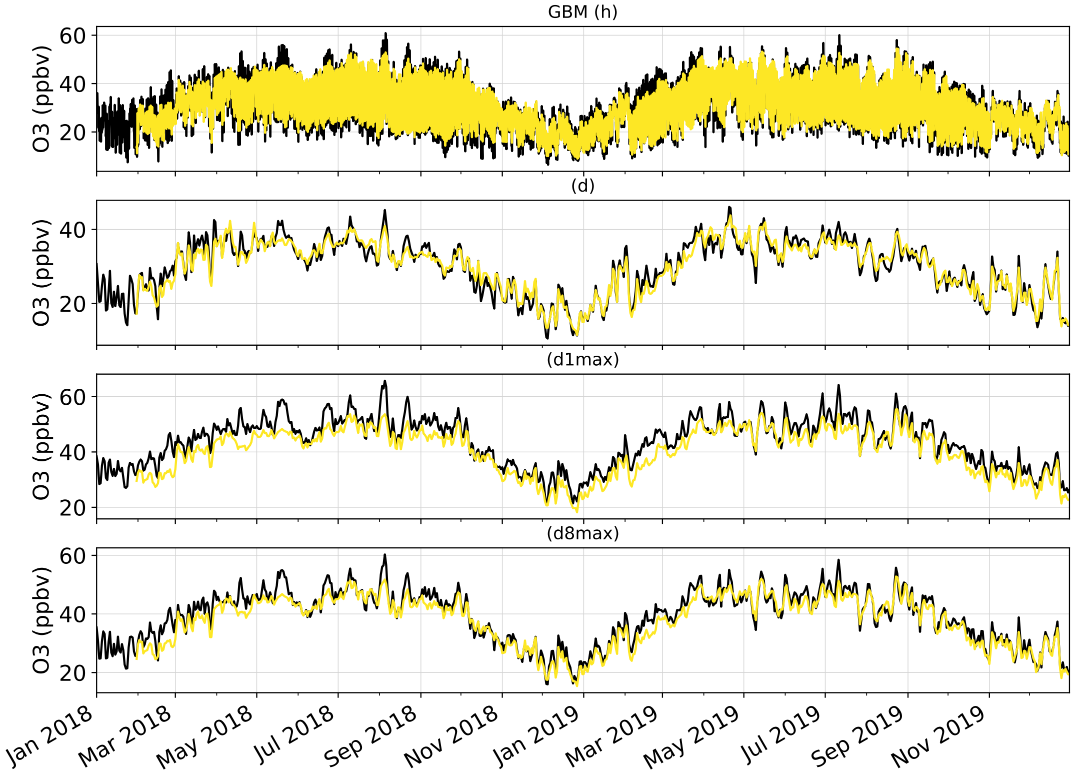

Figure F1. Time series of the mean O_3 mixing ratios over the Iberian Peninsula, as observed by monitoring stations (in black) and as simulated by CAMS D+1 forecasts corrected with the GBM MOS method (in yellow). Time series are shown at the hourly (h), daily mean (d), daily 1-hour maximum (d1max) and daily 8-hour maximum (d8max) time scales. O_3 mixing ratios are averaged over all surface stations of the domain.

Appendix G: Sensitivity tests on MOS methods

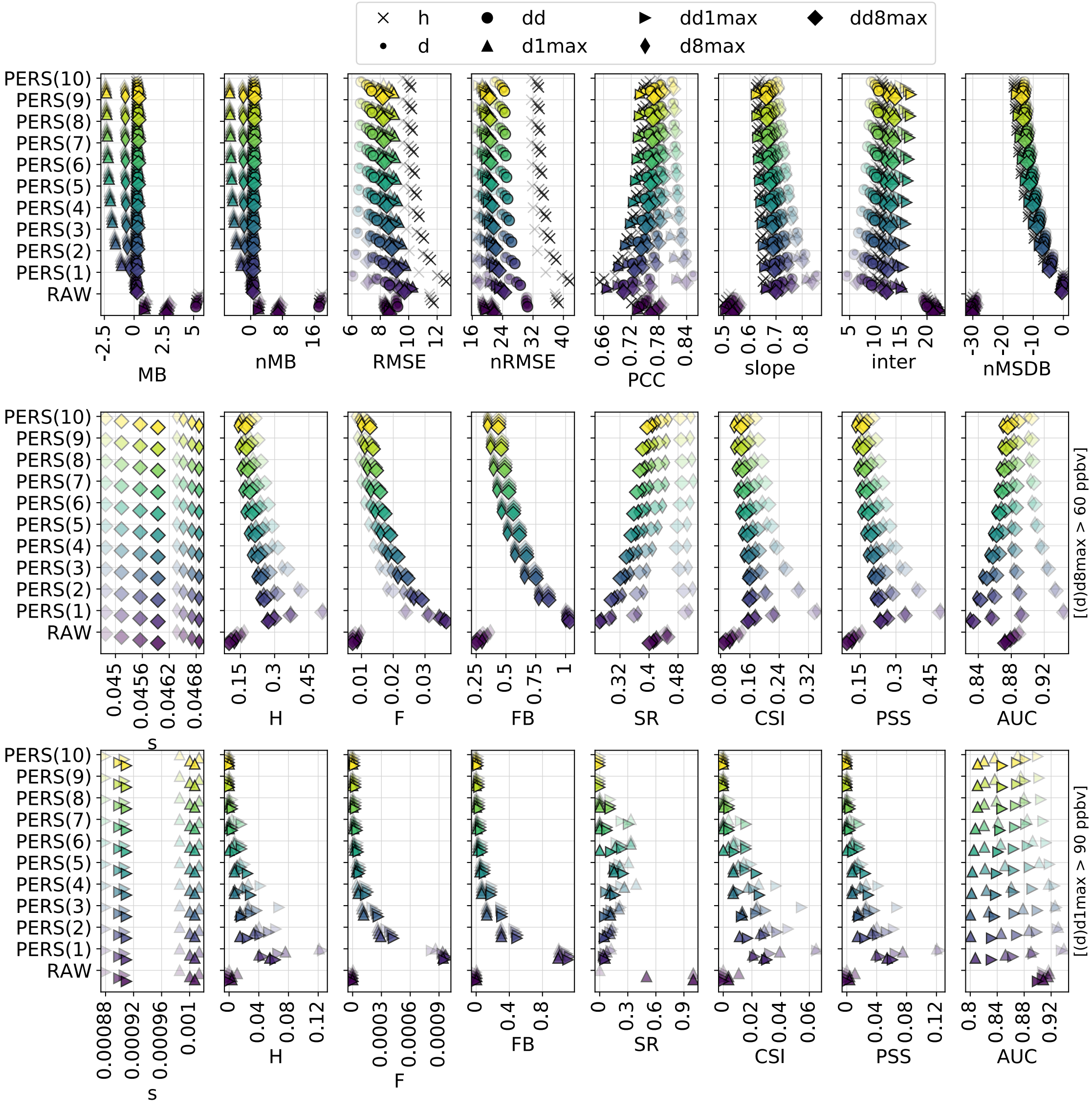

Figure G1. Similar to Fig. 3 for sensitivity tests on the PERS method.

https://doi.org/10.5194/acp-2021-864
Preprint. Discussion started: 1 December 2021

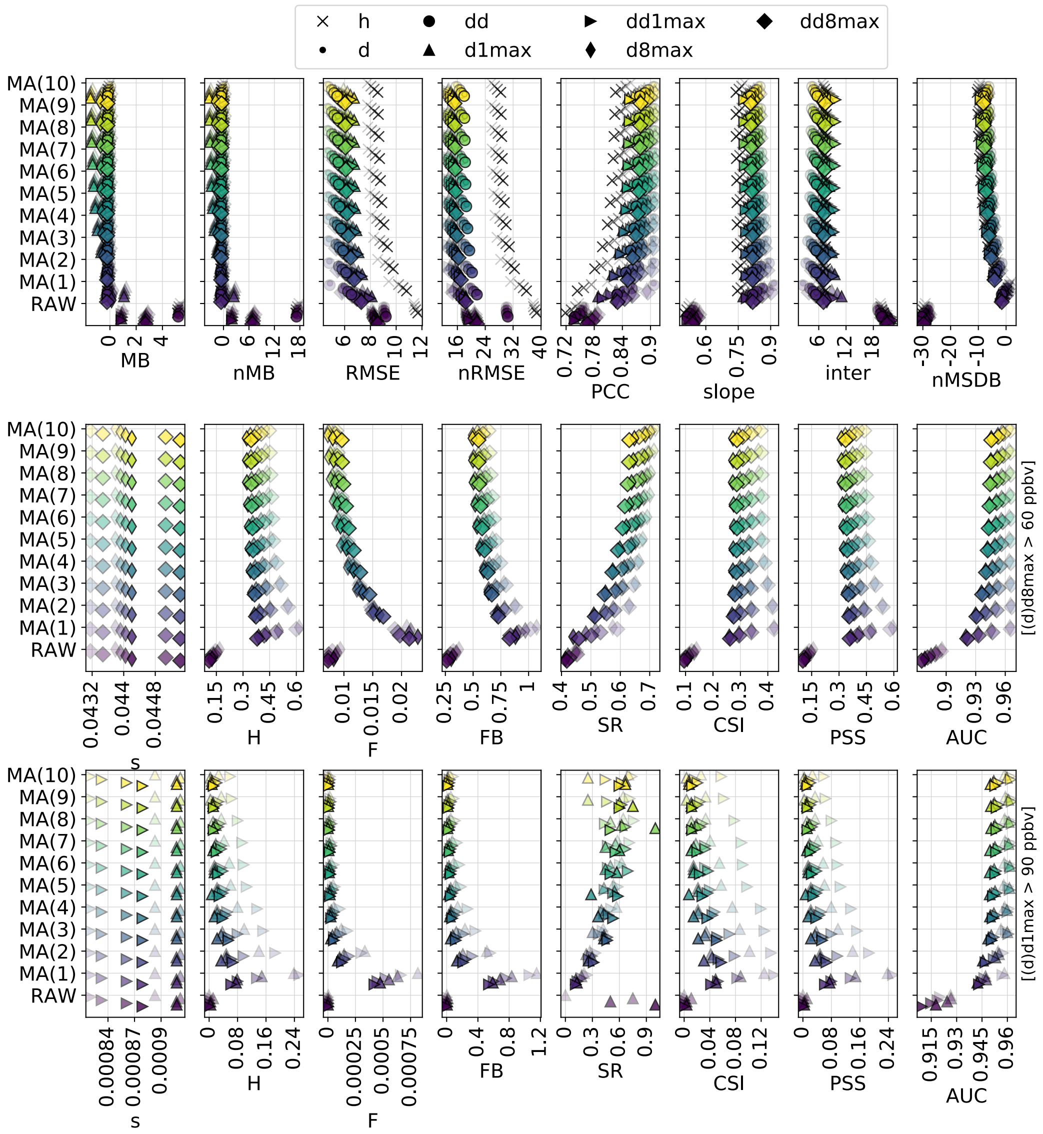

Figure G2. Similar to Fig. 3 for sensitivity tests on the MA method.

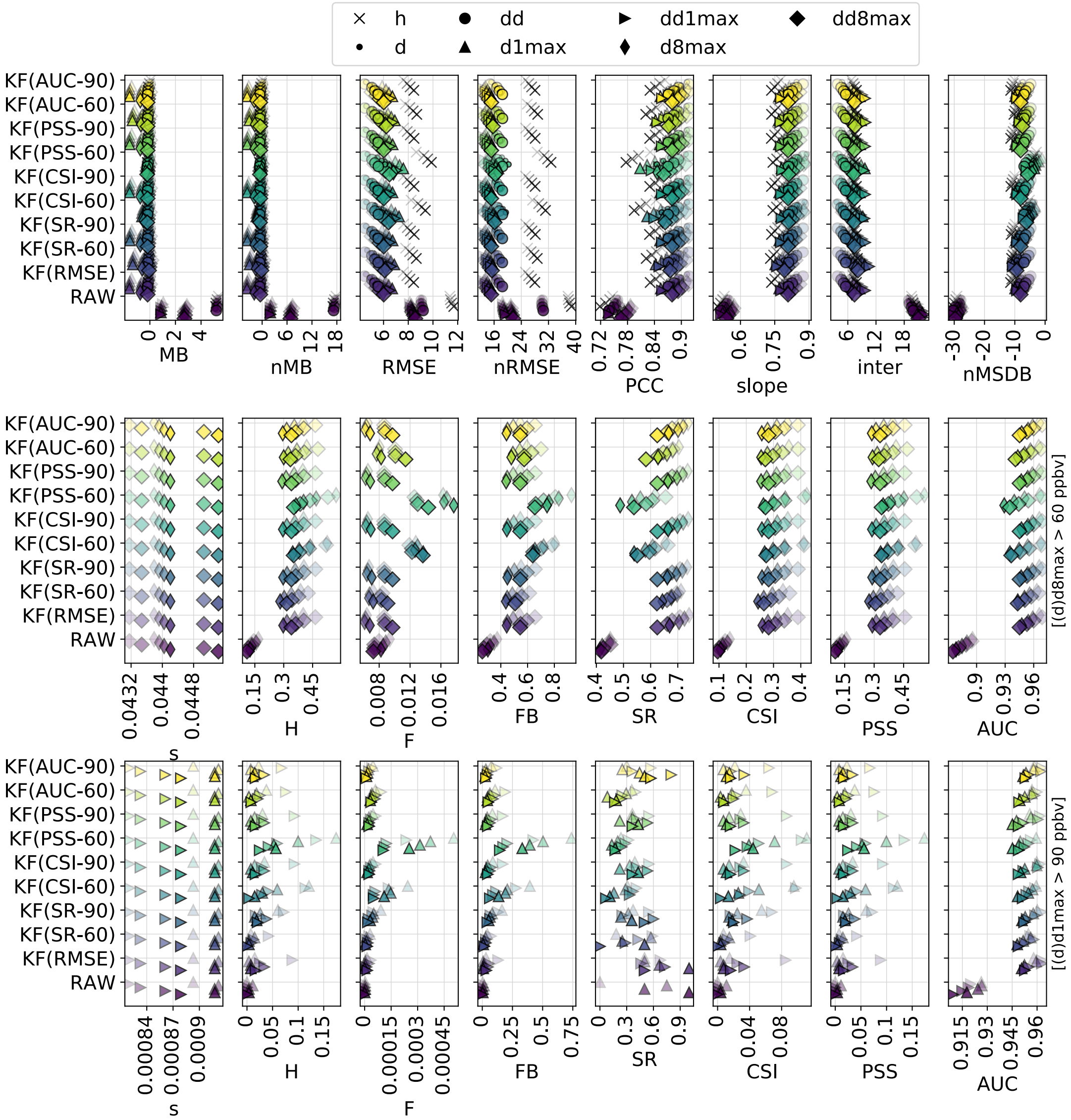

Figure G3. Similar to Fig. 3 for sensitivity tests on the KF method.

https://doi.org/10.5194/acp-2021-864
Preprint. Discussion started: 1 December 2021

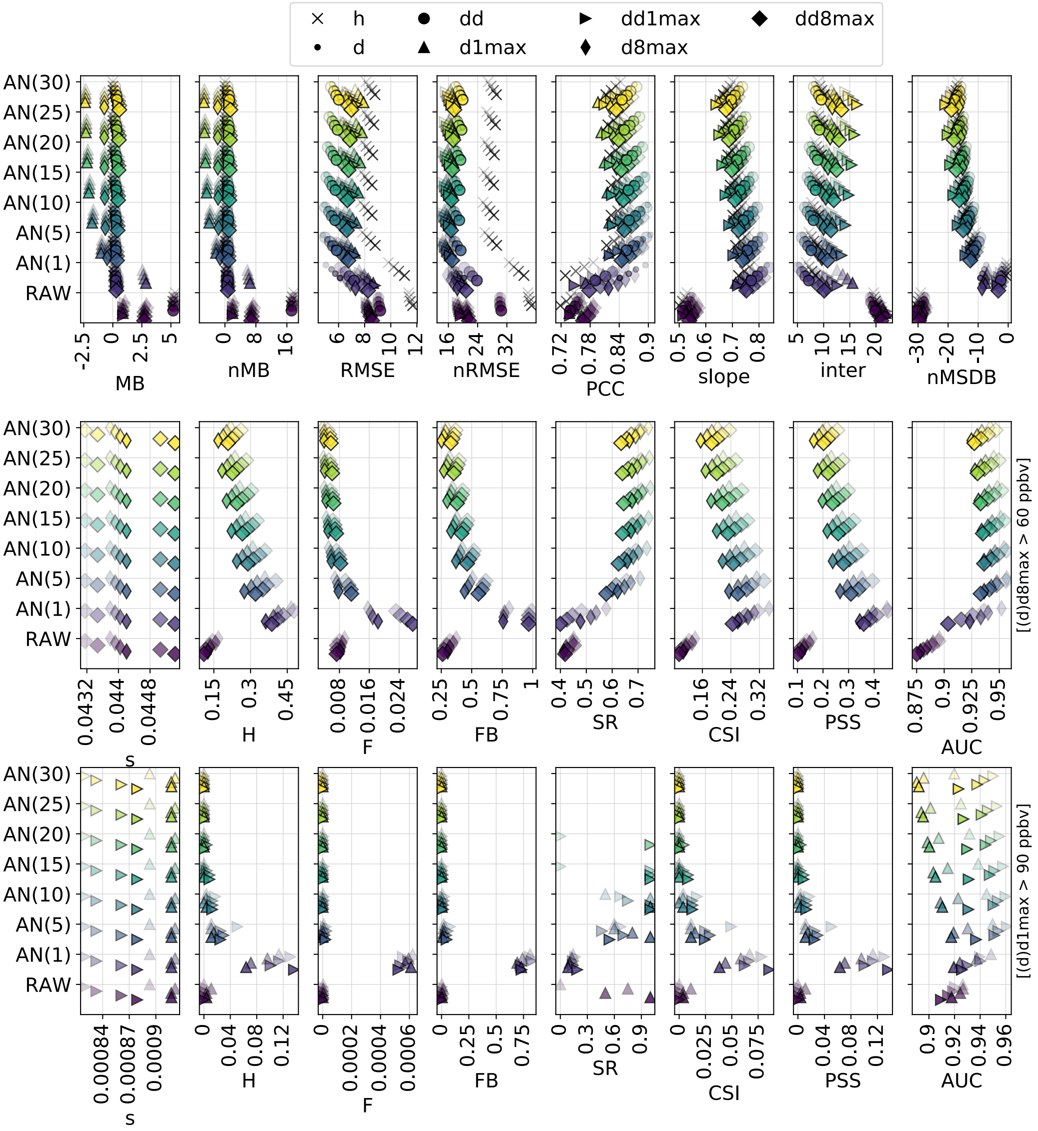

Figure G4. Similar to Fig. 3 for sensitivity tests on the AN method.

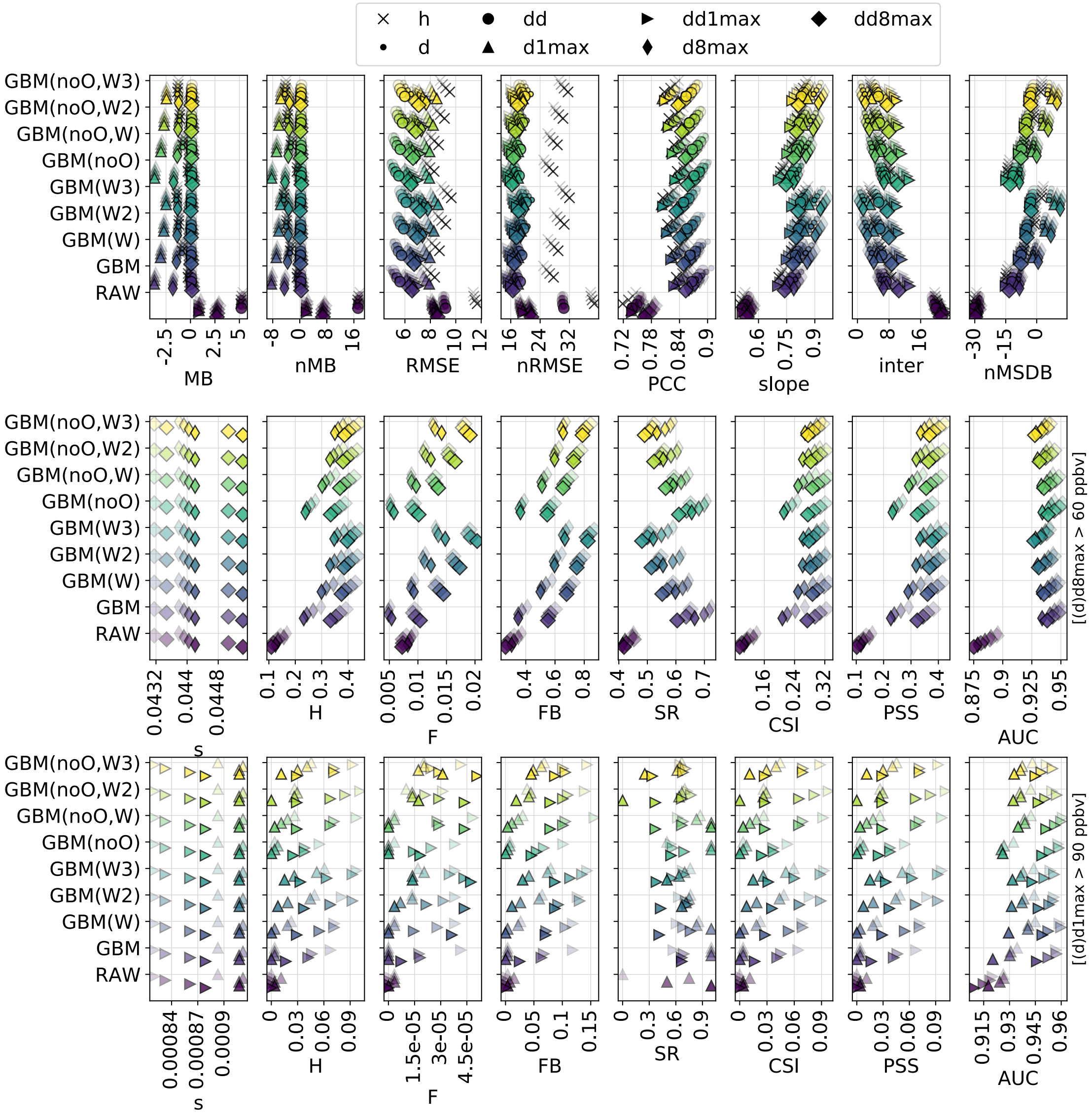

Figure G5. Similar to Fig. 3 for sensitivity tests on the GBM method.

https://doi.org/10.5194/acp-2021-864
Preprint. Discussion started: 1 December 2021

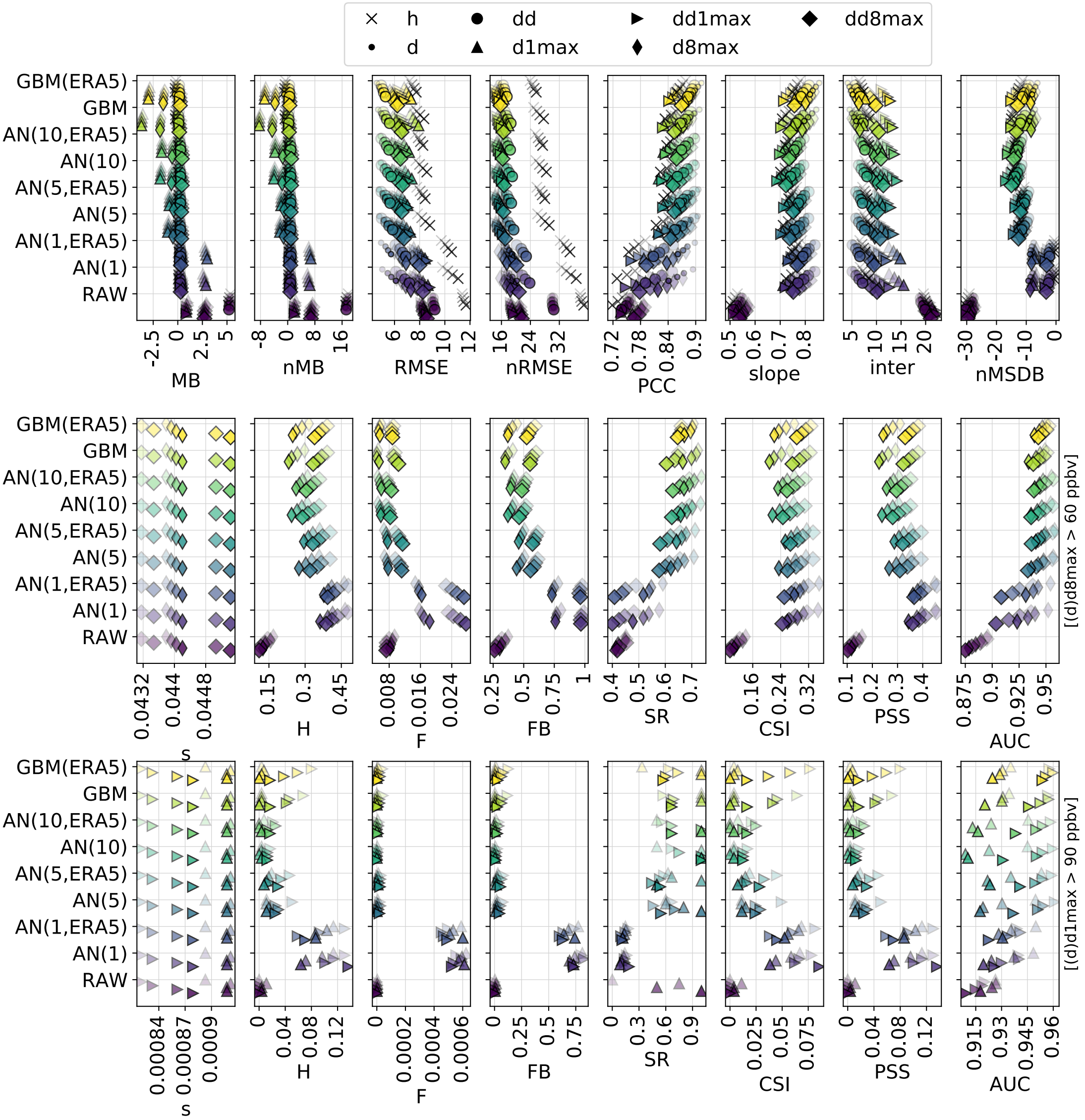

Figure G6. Similar to Fig. 3 for sensitivity tests on the meteorological data (IFS versus ERA5) used in AN and GBM methods.

Author contributions. HP contributed to the conception and design of the study. PAB and MSC were responsible for downloading the CAMS and meteorological data. KS was responsible for installing the python packages and other useful modules on the Mare Nostrum supercomputer. DB was responsible for the acquisition and preprocessing of the air quality data through the GHOST project. HP carried out the analysis. HP, CPGP, OJ, AS, MG, JMA and DB contributed to the interpretation of results. HP was responsible for writing the article, with a careful review from CPGP and JAM.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. This research has been funded by the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement H2020-MSCA-COFUND-2016-754433, as well as the MITIGATE project (PID2020-116324RA-I00 / AEI / 10.13039/501100011033) from the Agencia Estatal de Investigacion (AEI). We also acknowledge support by the the AXA Research Fund and Red Temática ACTRIS España (CGL2017-90884-REDT), the BSC-CNS "Centro de Excelencia Severo Ochoa 2015-2019" Program (SEV-2015-0493), PRACE and RES for awarding us access to Marenostrum Supercomputer in the Barcelona Supercomputing Center, and H2020 ACTRIS IMP (#871115).

References

Borrego, C., Monteiro, A., Pay, M., Ribeiro, I., Miranda, A., Basart, S., and Baldasano, J.: How bias-correction can improve air quality forecasts over Portugal, Atmospheric Environment, 45, 6629–6641, https://doi.org/10.1016/j.atmosenv.2011.09.006, https://linkinghub.elsevier.com/retrieve/pii/S1352231011009381, 2011.

Bowdalo, D.: Globally Harmonised Observational Surface Treatment: Database of global surface gas observations, in preparation.

Caruana, R. and Niculescu-Mizil, A.: An empirical comparison of supervised learning algorithms using different performance metrics, Tech. rep., Technical Report TR2005-1973, Cornell University, 2005.

Copernicus Climate Change Service (C3S): ERA5: Fifth generation of ECMWF atmospheric reanalyses of the global climate, 2017.

De Ridder, K., Kumar, U., Lauwaet, D., Blyth, L., and Lefebvre, W.: Kalman filter-based air quality forecast adjustment, Atmospheric Environment, 50, 381–384, https://doi.org/10.1016/j.atmosenv.2012.01.032, https://linkinghub.elsevier.com/retrieve/pii/S1352231012000532, 2012.

Delle Monache, L., Nipen, T., Deng, X., Zhou, Y., and Stull, R.: Ozone ensemble forecasts: 2. A Kalman filter predictor bias correction, Journal of Geophysical Research, 111, D05 308, https://doi.org/10.1029/2005JD006311, http://doi.wiley.com/10.1029/2005JD006311, 2006.

Delle Monache, L., Nipen, T., Liu, Y., Roux, G., and Stull, R.: Kalman Filter and Analog Schemes to Postprocess Numerical Weather Predictions, Monthly Weather Review, 139, 3554–3570, https://doi.org/10.1175/2011MWR3653.1, http://journals.ametsoc.org/doi/abs/10.1175/2011MWR3653.1, 2011.

Delle Monache, L., Eckel, F. A., Rife, D. L., Nagarajan, B., and Searight, K.: Probabilistic Weather Prediction with an Analog Ensemble, Monthly Weather Review, 141, 3498–3516, https://doi.org/10.1175/MWR-D-12-00281.1, http://journals.ametsoc.org/doi/abs/10.1175/MWR-D-12-00281.1, 2013.

Di Tomaso, E., Schutgens, N. A. J., Jorba, O., and Pérez García-Pando, C.: Assimilation of MODIS Dark Target and Deep Blue observations in the dust aerosol component of NMMB-MONARCH version 1.0, Geoscientific Model Development, 10, 1107–1129, https://doi.org/10.5194/gmd-10-1107-2017, https://www.geosci-model-dev.net/10/1107/2017/, 2017.

Djalalova, I., Wilczak, J., McKeen, S., Grell, G., Peckham, S., Pagowski, M., DelleMonache, L., McQueen, J., Tang, Y., and Lee, P.: Ensemble and bias-correction techniques for air quality model forecasts of surface O3 and PM2.5 during the TEXAQS-II experiment of 2006, Atmospheric Environment, 44, 455–467, https://doi.org/10.1016/j.atmosenv.2009.11.007, https://linkinghub.elsevier.com/retrieve/pii/S1352231009009510, 2010.

Djalalova, I., Delle Monache, L., and Wilczak, J.: PM2.5 analog forecast and Kalman filter post-processing for the Community Multiscale Air Quality (CMAQ) model, Atmospheric Environment, 108, 76–87, https://doi.org/10.1016/j.atmosenv.2015.02.021, https://linkinghub.elsevier.com/retrieve/pii/S1352231015001405, 2015.

EEA: Air Quality e-Reporting Database, European Environment Agency (http://www.eea.europa.eu/data- and-maps/data/aqereporting-8) (accessed 1 May 2020), 2020.

Fawcett, T.: An introduction to ROC analysis, Pattern Recognition Letters, 27, 861–874, https://doi.org/10.1016/j.patrec.2005.10.010, https://linkinghub.elsevier.com/retrieve/pii/S016786550500303X, 2006.

Ferro, C. A. T. and Stephenson, D. B.: Extremal Dependence Indices: Improved Verification Measures for Deterministic Forecasts of Rare Binary Events, Weather and Forecasting, 26, 699–713, 2011.

Flemming, J., Huijnen, V., Arteta, J., Bechtold, P., Beljaars, A., Blechschmidt, A.-M., Diamantakis, M., Engelen, R. J., Gaudel, A., Inness, A., Jones, L., Josse, B., Katragkou, E., Marecal, V., Peuch, V.-H., Richter, A., Schultz, M. G., Stein, O., and Tsikerdekis, A.: Tropospheric chemistry in the Integrated Forecasting System of ECMWF, Geoscientific Model Development, 8, 975–1003, https://doi.org/10.5194/gmd-8-975-2015, https://gmd.copernicus.org/articles/8/975/2015/, 2015.

Gaubert, B., Coman, A., Foret, G., Meleux, F., Ung, A., Rouil, L., Ionescu, A., Candau, Y., and Beekmann, M.: Regional scale ozone data assimilation using an ensemble Kalman filter and the CHIMERE chemical transport model, Geoscientific Model Development, 7, 283–302, https://doi.org/10.5194/gmd-7-283-2014, https://www.geosci-model-dev.net/7/283/2014/, 2014.

Hamill, T. M. and Whitaker, J. S.: Probabilistic Quantitative Precipitation Forecasts Based on Reforecast Analogs: Theory and Application, Monthly Weather Review, 134, 3209–3229, https://doi.org/10.1175/MWR3237.1, http://journals.ametsoc.org/doi/abs/10.1175/MWR3237.1, 2006.

Honoré, C., Rouïl, L., Vautard, R., Beekmann, M., Bessagnet, B., Dufour, A., Elichegaray, C., Flaud, J.-M., Malherbe, L., Meleux, F., Menut, L., Martin, D., Peuch, A., Peuch, V.-H., and Poisson, N.: Predictability of European air quality: Assessment of 3 years of operational forecasts and analyses by the PREV'AIR system, Journal of Geophysical Research, 113, D04 301, https://doi.org/10.1029/2007JD008761, http://doi.wiley.com/10.1029/2007JD008761, 2008.

Huang, J., McQueen, J., Wilczak, J., Djalalova, I., Stajner, I., Shafran, P., Allured, D., Lee, P., Pan, L., Tong, D., Huang, H.-C., DiMego, G., Upadhayay, S., and Delle Monache, L.: Improving NOAA NAQFC PM 2.5 Predictions with a Bias Correction Approach, Weather and Forecasting, 32, 407–421, https://doi.org/10.1175/WAF-D-16-0118.1, http://journals.ametsoc.org/doi/10.1175/WAF-D-16-0118.1, 2017.

Im, U., Bianconi, R., Solazzo, E., Kioutsioukis, I., Badia, A., Balzarini, A., Baró, R., Bellasio, R., Brunner, D., Chemel, C., Curci, G., Denier van der Gon, H., Flemming, J., Forkel, R., Giordano, L., Jiménez-Guerrero, P., Hirtl, M., Hodzic, A., Honzak, L., Jorba, O., Knote, C., Makar, P. A., Manders-Groot, A., Neal, L., Pérez, J. L., Pirovano, G., Pouliot, G., San Jose, R., Savage, N., Schroder, W., Sokhi, R. S., Syrakov, D., Torian, A., Tuccella, P., Wang, K., Werhahn, J., Wolke, R., Zabkar, R., Zhang, Y., Zhang, J., Hogrefe, C., and Galmarini, S.: Evaluation of operational online-coupled regional air quality models over Europe and North America in the context of AQMEII phase 2. Part II: Particulate matter, Atmospheric Environment, 115, 421–441, https://doi.org/10.1016/j.atmosenv.2014.08.072, http://linkinghub.elsevier.com/retrieve/pii/S1352231014006839, 2015a.

Im, U., Bianconi, R., Solazzo, E., Kioutsioukis, I., Badia, A., Balzarini, A., Baró, R., Bellasio, R., Brunner, D., Chemel, C., Curci, G., Flemming, J., Forkel, R., Giordano, L., Jiménez-Guerrero, P., Hirtl, M., Hodzic, A., Honzak, L., Jorba, O., Knote, C., Kuenen, J. J., Makar, P. A., Manders-Groot, A., Neal, L., Pérez, J. L., Pirovano, G., Pouliot, G., San Jose, R., Savage, N., Schroder, W., Sokhi, R. S., Syrakov, D., Torian, A., Tuccella, P., Werhahn, J., Wolke, R., Yahya, K., Zabkar, R., Zhang, Y., Zhang, J., Hogrefe, C., and Galmarini, S.: Evaluation of operational on-line-coupled regional air quality models over Europe and North America in the context of AQMEII phase 2. Part I: Ozone, Atmospheric Environment, 115, 404–420, https://doi.org/10.1016/j.atmosenv.2014.09.042, http://linkinghub.elsevier.com/retrieve/pii/S1352231014007353, 2015b.

Jolliffe, I. T. and Stephenson, D. B.: Forecast Verification: A Practitioner's Guide in Atmospheric Science, 2nd Edition, Chichester, United Kingdom: J. Wiley, 2011.

Kang, D., Mathur, R., Rao, S. T., and Yu, S.: Bias adjustment techniques for improving ozone air quality forecasts, Journal of Geophysical Research, 113, D23 308, https://doi.org/10.1029/2008JD010151, http://doi.wiley.com/10.1029/2008JD010151, 2008.

Kang, D., Mathur, R., and Trivikrama Rao, S.: Real-time bias-adjusted O3 and PM2.5 air quality index forecasts and their performance evaluations over the continental United States, Atmospheric Environment, 44, 2203–2212, https://doi.org/10.1016/j.atmosenv.2010.03.017, https://linkinghub.elsevier.com/retrieve/pii/S1352231010002128, 2010.

Liu, T., Lau, A. K. H., Sandbrink, K., and Fung, J. C. H.: Time Series Forecasting of Air Quality Based On Regional Numerical Modeling in Hong Kong, Journal of Geophysical Research: Atmospheres, 123, 4175–4196, https://doi.org/10.1002/2017JD028052, http://doi.wiley.com/10.1002/2017JD028052, 2018.

Ma, C., Wang, T., Zang, Z., and Li, Z.: Comparisons of Three-Dimensional Variational Data Assimilation and Model Output Statistics in Improving Atmospheric Chemistry Forecasts, Advances in Atmospheric Sciences, 35, 813–825, https://doi.org/10.1007/s00376-017-7179-y, http://link.springer.com/10.1007/s00376-017-7179-y, 2018.

McKeen, S., Wilczak, J., Grell, G., Djalalova, I., Peckham, S., Hsie, E.-Y., Gong, W., Bouchet, V., Menard, S., Moffet, R., McHenry, J., McQueen, J., Tang, Y., Carmichael, G. R., Pagowski, M., Chan, A., Dye, T., Frost, G., Lee, P., and Mathur, R.: Assessment of an ensemble of seven real-time ozone forecasts over eastern North America during the summer of 2004, Journal of Geophysical Research, 110, D21 307, https://doi.org/10.1029/2005JD005858, http://doi.wiley.com/10.1029/2005JD005858, 2005.

Struzewska, J., Kaminski, J., and Jefimow, M.: Application of model output statistics to the GEM-AQ high resolution air quality forecast, Atmospheric Research, 181, 186–199, https://doi.org/10.1016/j.atmosres.2016.06.012, https://linkinghub.elsevier.com/retrieve/pii/S016980951630165X, 2016.

World Health Organization: Ambient air pollution: a global assessment of exposure and burden of disease, Tech. rep., https://apps.who.int/iris/bitstream/handle/10665/250141/9789241511353-eng.pdf?sequence=1, 2016.