ECMWF COPERNICUS REPORT

Copernicus Atmosphere Monitoring Service

Annual report on the verification of validated re-analyses for 2017

VRA2017

Issued by: METEO-FRANCE / G. Collin

Date: 20/02/2020

Ref:

CAMS50_2018SC1_D5.4.1-2017_202002_Annual_verification_report_VRA2017_v1

Contributors

INERIS
F. Meleux
B. Raux
A. Ung
A. Colette

METEO-FRANCE
G. Collin
N. Assar

Table of Contents

Table of Figures

Executive summary

The present report presents an analysis of the performances of the annual air quality re-analyses throughout Europe, produced by the CAMS_50 service for the year 2017.

CAMS_50 services include the provision of ENSEMBLE air quality re-analyses resulting from the combination of seven well-validated and documented chemistry-transport models results. So-called 'validated' re-analyses are data assimilated fields of air pollutant concentrations, based on observation data rigorously validated according to the air quality reporting principles set in EU Decision 2011/850/EU *on reciprocal exchange of information and reporting on ambient air quality*. These data are reported by the Member States to the European Environment Agency (EEA) in autumn, every year for the previous year.

To run this evaluation process, INERIS computed a number of performance indicators and scores for ozone, nitrogen dioxide, PM_{10}, and $PM_{2.5}$ concentrations simulated by the seven individual models involved in the service and the resulting ENSEMBLE. They are presented in this report. Overall, the models performed as expected and the ENSEMBLE median model gives very often the best results. Consistency with previous validated assessment results from the former pre-operational MACC projects regarding models' behaviour is ensured.

We can highlight the following points:

- Even if model scores for ozone are good in a large part of Europe and consistent with the previous year, performances are better in Western and Central Europe than in Southern Europe (Mediterranean countries). This can be explained by uncertainties in meteorological fields and by the complexity of the photochemistry and deposition processes in those regions that may not be correctly approximated in the models. This is a well-known issue, but it should be carefully considered for future improvements of the CAMS_50 modelling system. Some difficulties are also noted in some part of Eastern Europe that should be further investigated, but the limited coverage of such region by stations makes it difficult to interpret the results.

- For ozone, a stability of the ENSEMBLE performances should be noted compared to last year. For the detection of the exceedances of the air quality objectives (information threshold in particular), a slight increase compared to last year was seen (30% of good detection of 180 µg/m^3 threshold), partly thanks to a more intensive summer in terms of ozone episodes than last year. The ENSEMBLE is among the best but not the best one for threshold detection. However, its behaviour produces a very limited number of false alarms.
- For nitrogen dioxide, the results are better than in 2016, with a general improvement for all re-analyses including the ENSEMBLE, which has very good scores. Nonetheless, observations are lacking in Eastern and Southern parts of Europe to draw robust conclusions for these parts of Europe.
- The same concern holds for PM_{10} and $PM_{2.5}$, too few observation data are available for evaluation in Eastern countries. Nevertheless, it may change in the coming years as noticed in 2017 in

Southern countries: indeed, many data for these countries were reported in 2017 for PM_{10}, allowing for a robust evaluation for this part of Europe. The results are very good for Western and Central Europe, with correlation coefficients higher than 0.85 and RMSE lower than 5 $\mu g/m^3$; for Southern Europe however, there is still some progress to be made to correct the regular underestimation of the PM_{10} concentrations, particularly over urban stations. The heterogeneity of the model responses for PM_{10} in Southern countries and especially Turkey should be investigated.

- For $PM_{2.5}$, the results are similar to PM_{10} in terms of re-analyses responses. The scores are good with low bias and RMSE values for most of the stations and correlation around 0.9. The number of stations is growing year after year, which improves the representativeness of this evaluation.
- Concerning the capability of the ENSEMBLE to detect threshold exceedances for O_3 and PM_{10}, it is worth noting that rates of false alarm are very low. This means that the representations of areas with exceedance are reliable and certainly underestimated, due to the tendency of the ENSEMBLE to have a significant number of missing exceedances.
- The overall performances of the ENSEMBLE re-analyses are quite satisfactory (and in line with the state of the art) for the set of pollutants analysed in this report (O_3, NO_2, PM_{10} and $PM_{2.5}$).
- This report provides insights of CAMS_50 re-analysis skills that may help users to rely on these data.

Introduction

This report gives an overview of the performances of the European air quality **validated re-analysis** process developed by the CAMS Regional services and implemented to simulate air quality in Europe during the year 2017.

Air quality re-analyses result from a combination of chemistry-transport models' results that simulate the spatial and temporal evolution of regulatory air pollutant concentrations (according to the ambient Air quality Directive 2008/50/EC), and observations assimilated in each model to correct and improve its results. Each team providing air quality re-analyses developed appropriate and validated data assimilation chains, to provide best estimates of air pollution patterns according to available observation data.

The models implemented to calculate these re-analyses are the set of seven models run in other near-real-time CAMS regional services. The models are CHIMERE (INERIS, France), EMEP (MET Norway, Norway), EURAD-IM (RIU-UK, Germany), LOTOS-EUROS (KNMI-TNO, The Netherlands), MATCH (SMHI, Sweden), MOCAGE (METEO-FRANCE, France) and SILAM (FMI, Finland).

Observations that are assimilated in the chemistry transport models are issued from the regulatory air quality monitoring networks that report air quality data to the European Environment Agency (EEA), according to air quality Directive 2008/50/EC and Decision 2011/850/EU *on reciprocal exchange and reporting on ambient air quality*. Observation datasets have followed a stringent validation process by the national experts on the one side, and by the EEA on the other side. The 2011 Decision stipulates that the Member States must report monitoring data related to the year Y, the 30^{th} September of year Y+1. The data is gathered in the so-called AQ e-reporting database. Validated data should be released by the EEA in February-March of year Y+2.

The set of observation sites reported to the EEA is split in two subsets, one for data assimilation in re-analyses and the other for verification. Those datasets do not overlap, and verification cannot be biased by use of data for both assimilation and verification processes. Constraints have been set up in the splitting between assimilation and verification, to put co-located measurements of species from the same chemical family in the same group, either assimilation or evaluation. This is the case for the couple O_3 - NO_2 and for the couple PM_{10} - $PM_{2.5}$. It should be noted that the spatial density of observation sites can change significantly from a country to another. Logically, uneven distribution of monitoring stations impacts data assimilation and the evaluation processes. Consequently, data assimilation and evaluation cannot be performed in some geographical areas, and it will not be possible to draw some clear conclusions concerning the model capacities in those regions. Eastern Europe is more particularly concerned, and it should be mentioned that now Southern Europe is well covered with stations (which was not the case for the previous re-analysis).

The evaluation focuses on the seven individual models and the ENSEMBLE as well. The ENSEMBLE is the result of the median of the seven models. It is considered the best estimate of air pollution patterns and levels, since it combines the strengths of the other models. This will be checked in the present report.

Statistical indicators (bias, root mean square error, correlation coefficient) are presented to compare the models' results against observations. Maps, histograms and Taylor diagrams are proposed for a

better understanding and analysis of the performances. They are computed for the four regulatory pollutants targeted by the service: ozone (O_3), nitrogen dioxide (NO_2), particulate matter (PM_{10} and $PM_{2.5}$). Metrics relevant for policy purposes (regarding the content of the air quality directives) and for health impacts are considered for the evaluation.
All the results are presented below, after a short introduction on the computed performance indicators.

1. Performance indicators

The model performances are evaluated based on classical statistical indicators that measure objectively the gap between the model results and the observations at the available stations: bias, root mean square error (RMSE) and correlation coefficient are the most classical indicators. Comparison of observed and modelled averages is generally considered as well.
Obviously, the behaviour of performance indicators depends on the station typology and on the pollutant considered: the models used in the CAMS Regional service run at the European scale and their spatial resolution is 0.1° x 0.1°. Consequently, for pollutants which are largely influenced by local sources (NO_2, PM in some situations), these regional models are not able to reproduce hot spots monitored by traffic or industrial stations, and performance indicators will not be assessed. Difficulties can even be encountered at urban stations.

Conversely for pollutants characterized by long residence time in the atmosphere and large impacted areas (typically ozone and PM in some cases), performance indicators evaluated at all types of stations (except traffic and industrial sites) make sense.

The definitions of the various performance indicators used in the report are given below. They are very usual[1] in evaluation processes:

- Bias indicates, on average, if the simulations under or over-predicts the actual measured concentrations. In our case, negative values indicate under-prediction, whereas positive values indicate over-prediction; values close to 0 are the best ones:

  $$\frac{1}{N}\cdot\sum_{i=1}^{N}(P_i - O_i)$$

  Where N is the number of observations, P_i refers to the predictions and O_i to the observations. It is expressed in µg/m^3.
- Root Mean Square Error (RMSE) gives information about the skill of the model in predicting the overall magnitude of the observations. It should be as weak as possible:

  $$\sqrt{\frac{1}{N}\cdot\sum_{i=1}^{N}(P_i - O_i)^2}$$

  Where N is the number of observations, P_i refers to the predictions and O_i to the observations. It is expressed in µg/m^3.
- Correlation is a measure of whether predictions and observations change together in the same way (i.e. at the same time and/or place). The closer the correlation is to one, the better is the correspondence of extreme values of the two data sets.

  $$r = \frac{\mathrm{cov}(P_i, O_i)}{\sqrt{\mathrm{var}(P_i)}\cdot\sqrt{\mathrm{var}(O_i)}}$$

  Where N is the number of observations, P_i refers to the predictions and O_i to the observations. This is a non-dimensional number.

[1] Chang J.C. et Hanna S.R., 2004. Air quality model performance evaluation. *Meteorol. Atmos. Phys.* 87, 167–196.

Taylor diagrams synthesize, on a unique quadrant, various statistical indicators for different models: the radii correspond to the correlation coefficient values, the x-axis and the y-axis delimits arcs with bias values and the internal semi-circles correspond to the RMSE values. Therefore, this is a very pedagogical way of presenting an overview of the relative performances of a set of models and it is often used in model intercomparison exercises.

For indicators related to threshold values, for instance the number of days, hours when a certain concentration level is exceeded, some 'contingency tables' giving the percentages of correct predictions (GP), false alarms (FA), or missed events (ME) are estimated. These concepts come from the weather or air quality forecasting world. Although they are very severe and not objectively representative of the intrinsic model performance (because of the threshold cut-off effect, a result close to the threshold can fall arbitrarily in one or the other category), they can give useful information to compare various models' behaviours in different geographical regions. GP, FA and ME are expressed in percentage (%) and also referred sometimes to the total number of stations within each class (GP, FA and ME).

Several representations of the models' skills are proposed:

- Maps with colored patches at the location of the stations selected for the evaluation process. The color scale indicates how the model performs.
- Taylor diagrams provide a wider overview of the model performances.
- Histograms with model performances sorted by station typology and by European sub-region (Western, Northern, Southern, Central, Eastern) are proposed as well.

2. Performance indicators for ozone

In this evaluation, we focused on the ability of the model to correctly predict the ozone daily maximum (hourly average), which is the most relevant considering regulatory indicators like the number of exceedances of information and alert thresholds. The evaluation is performed over the "summer" period when ozone increases, reaching levels that may impact human health and ecosystems.

Error! Reference source not found. shows the Taylor diagram that synthesizes performances of individual CAMS re-analyses and the ENSEMBLE to reproduce hourly daily maximum of ozone in the summer period. The scores in this figure are computed with all typologies of background stations. The graph shows similar performances for all re-analyses (except for one production slightly aside from the others), highlighting slightly better scores for the ENSEMBLE with correlation closed to 0.9 and RMSE around 10 µg/m^3.

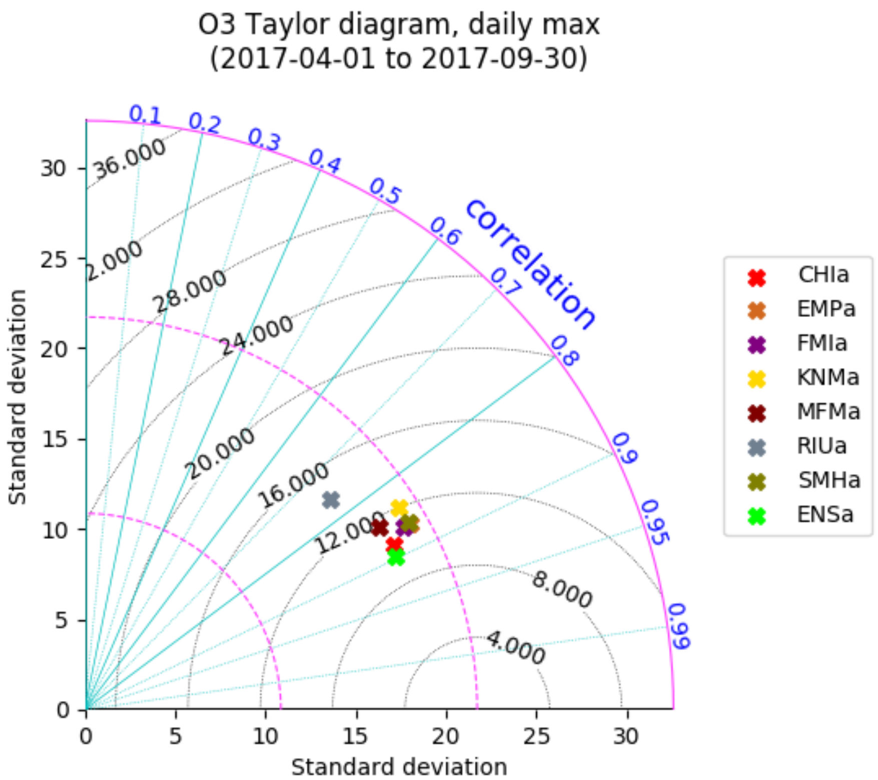
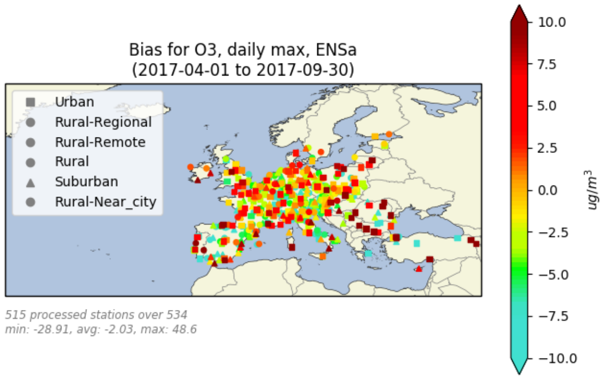


Figure 1 - Taylor diagram presenting performances of all CAMS regional models to simulate summer ozone daily maximum (hourly average), at all typologies of background stations.

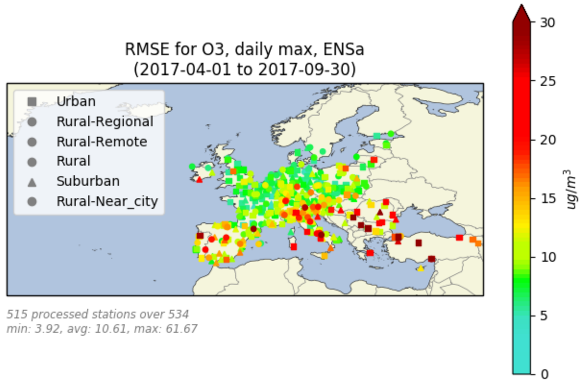
An in-depth analysis of the ENSEMBLE re-analyses can be elaborated considering the spatial distribution of the statistical indicators over Europe. 0 presents maps of bias, correlation coefficient and RMSE related to the ENSEMBLE, for daily maxima from April 1st to September 30th, 2017. Bias ranges in most parts of Europe between -5 and 5 μg/m^3. Higher bias values (overestimation) can be found in some specific locations in the Eastern part of Europe, and also for Mediterranean countries. Correlation coefficient is excellent with high values, most of them higher than 0.9, except for some locations in Eastern and Southern countries like Turkey and Romania. The same high quality can be seen with RMSE, although with a split between the North and the South - quite low RMSE in the North between 5 and 10 µg/m^3 and higher RMSE for many stations from the Mediterranean countries with values above 15 µg/m^3. Such performances are quite stable compared to the results of the previous year (2016).

To guide the interpretation of those maps, one can consider the same performance indicators for each individual model and the ENSEMBLE and various station typologies. 0, 4 and 5 present bias, RMSE, and correlation coefficient scores for all models at rural, suburban and urban stations respectively. The indicators are sorted per geographical region: Western Europe (EUW), Central Europe (EUC), Southern Europe (EUS), Northern Europe (EUN), Eastern Europe (EUE). The interpretation of the results is hampered by the low number of stations available for verification is some areas (especially for Eastern and Northern parts). The number of stations taken in consideration for computing the scores are mentioned on the figures. The situation is expected to improve in the coming years like it has improved in this report for Southern countries compared to previous years, with the integration of observations from Turkey. In Eastern and Northern Europe, the evaluation has been performed against a very little number of stations, which may be a problem regarding the representativeness of the obtained results.

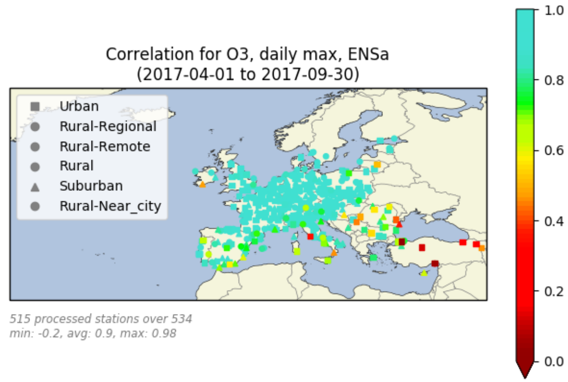
When observation data is available, the panel of re-analyses shows consistency in the scores from one region to another. The scores of the re-analyses are very similar for EUW, EUC and EUS, where most of the productions tends to underestimate the ozone concentrations at rural sites with underestimations more intensive for Southern Europe.
The responses of the models for EUE and EUN are more disperse, but this cannot be taken into consideration due to the low number of stations used for this evaluation.
The bias for suburban stations is very similar to rural stations scores with underestimations less important for EUW, EUC and EUS, while it becomes slightly positive for urban sites.
Regarding RMSE, we can note once again good consistency between results, even if they are slightly better for suburban stations. For the whole of Europe, the RMSE obtained by the models varies from 10 to 15 µg/m^3 for rural and suburban sites. It is worth noting that the better results are for the ENSEMBLE (this also holds whatever the region considered).
Correlation coefficient is quite high, ranging from 0.8 to 0.9 in the best cases among which is the ENSEMBLE. For some models and regions, significant differences appear, with correlations closed to 0.6 far from the best ones.
Obviously, there are more uncertainties in the re-analyses in Eastern and Northern regions due to uncertainties in emissions and to the complexity of the photochemical processes and meteorology. Yet, there are also much fewer stations than in other regions, which makes the scores very sensitive and less robust. For this reason, conclusions should be established with care.
Nevertheless, overall performances of the models to simulate ozone daily maxima are satisfactory and consistent with previous results obtained in the past and with the state of the art.
A slight improvement has been noticed compared to scores of the ENSEMBLE re-analyses for 2016.



(a) ozone bias (daily maximum)

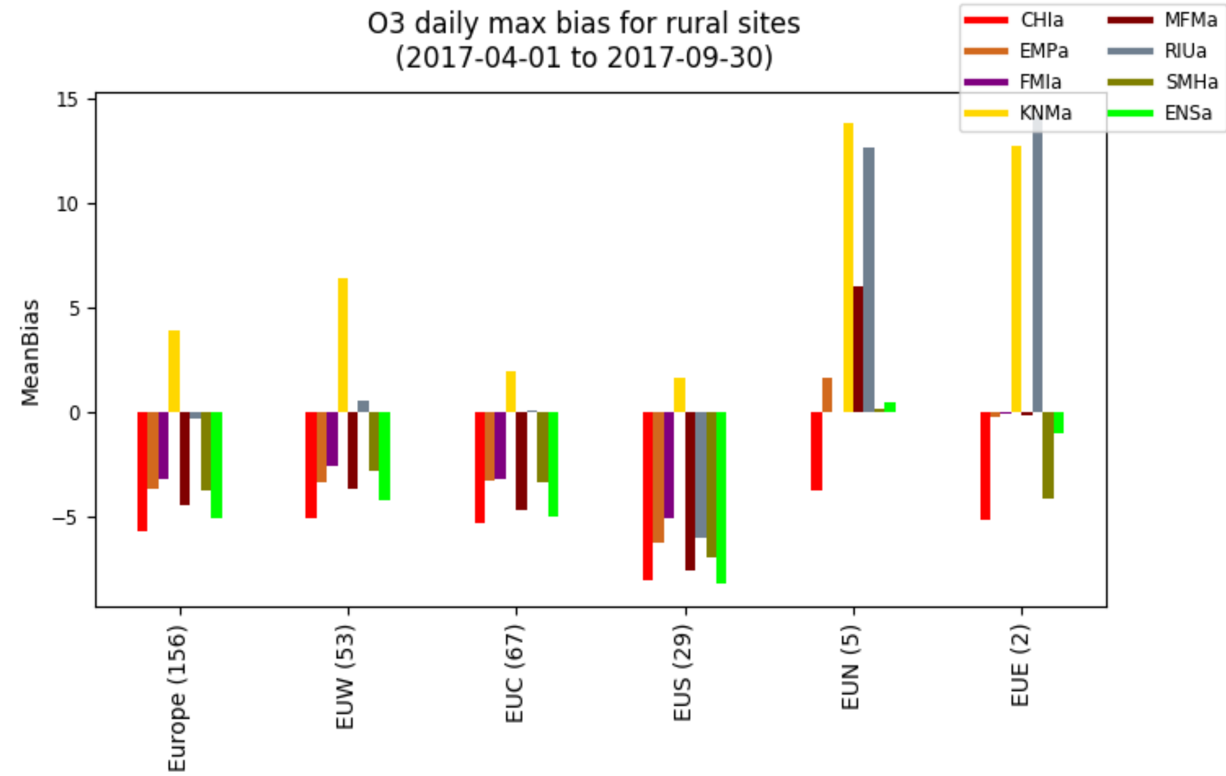


(b) ozone RMSE (daily maximum)

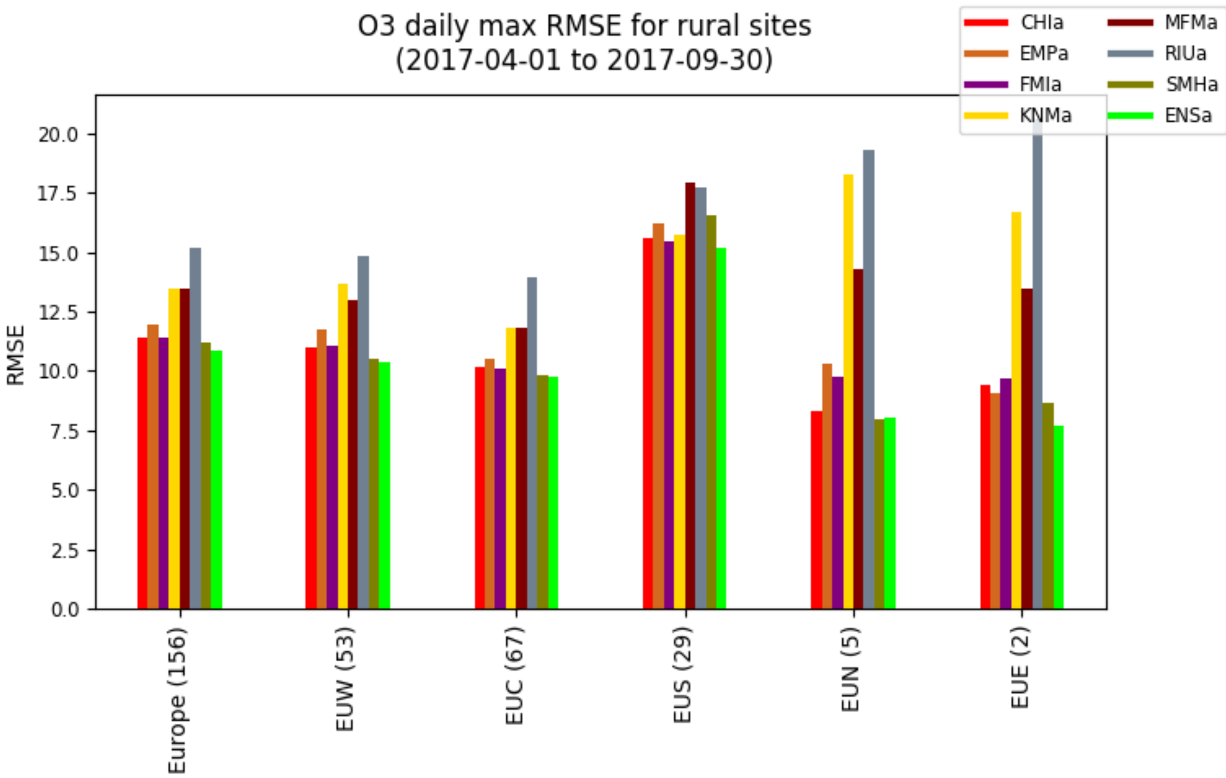


(c) ozone correlation coefficient (daily maximum)

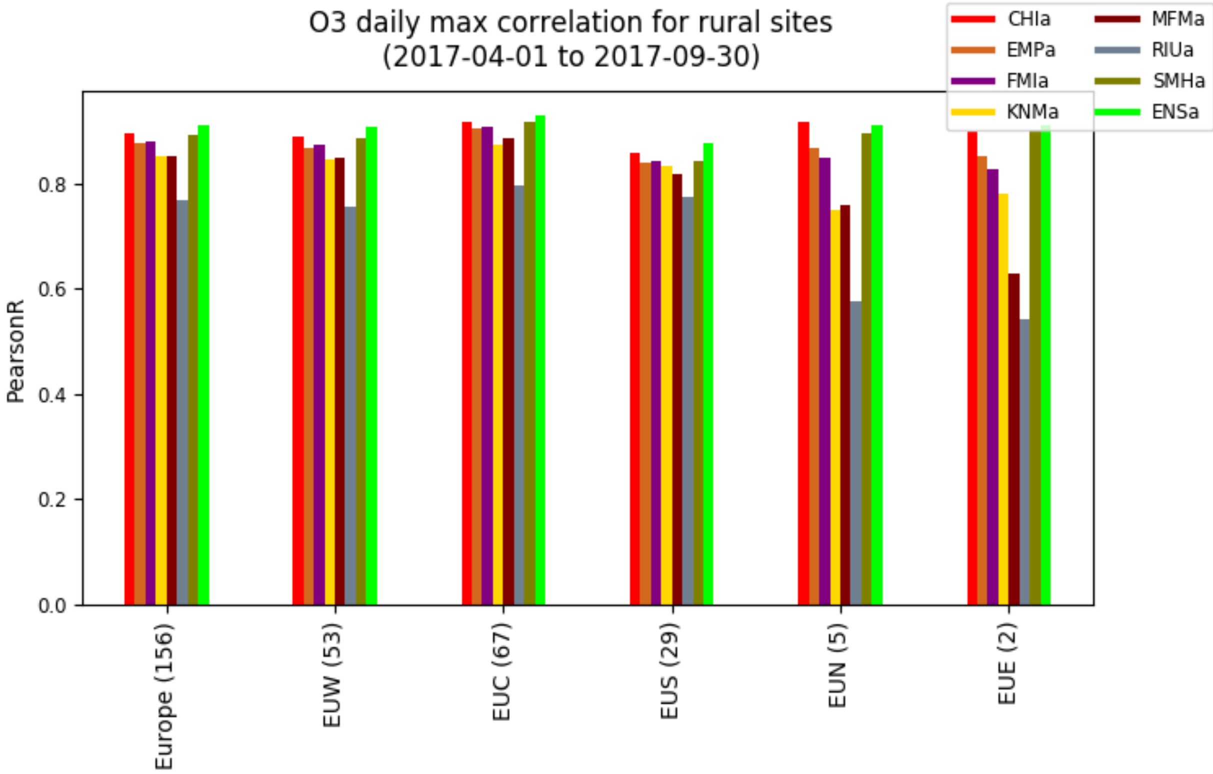
Figure 2 - Maps of Statistical scores of the ENSEMBLE interim re-analyses results against the observation validation dataset from the AQ e-reporting database for the ozone daily maximum, from 01/04/2018 to 30/09/2018 (a) Bias, (b) RMSE, (c) Correlation coefficient


O3 daily max bias for rural sites
(2017-04-01 to 2017-09-30)
CHIa
EMPa
FMIa
KNMa
MFMa
RIUa
SMHa
ENSa
MeanBias
15
10
5
0
−5
Europe (156)
EUW (53)
EUC (67)
EUS (29)
EUN (5)
EUE (2)

(a)

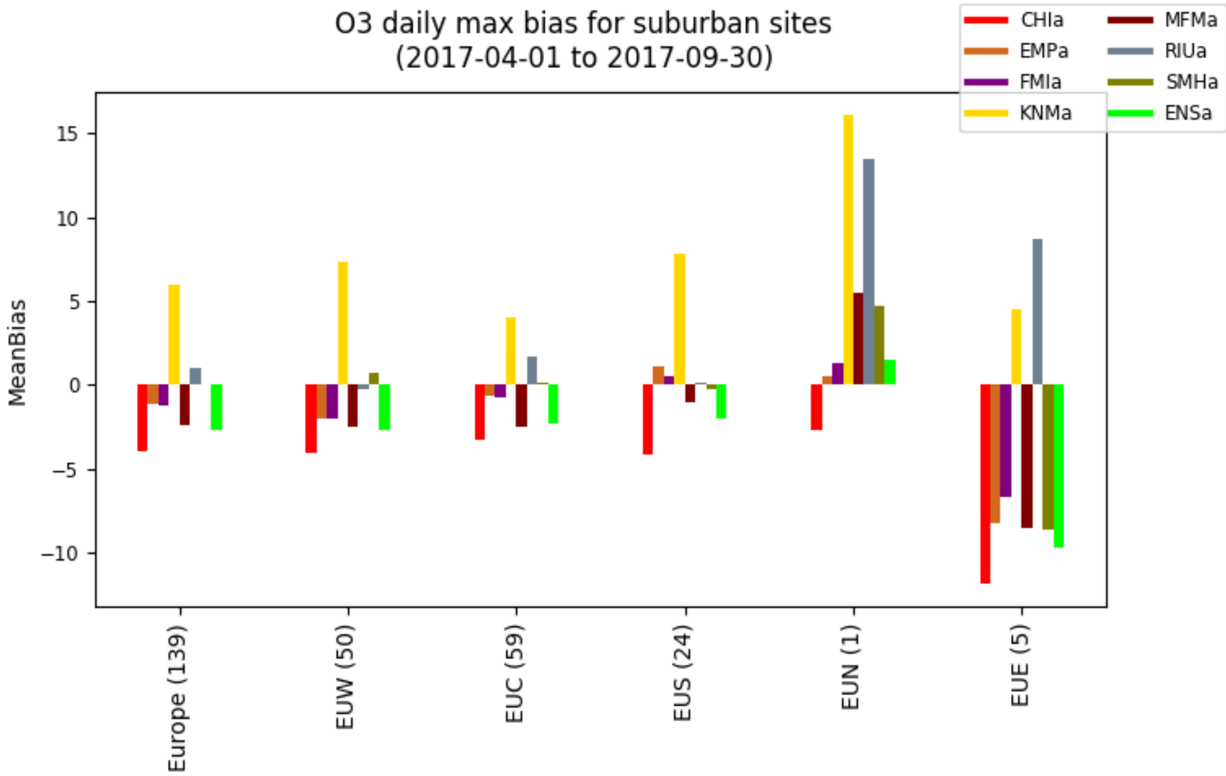

O3 daily max RMSE for rural sites
(2017-04-01 to 2017-09-30)
CHIa
EMPa
FMIa
KNMa
MFMa
RIUa
SMHa
ENSa
RMSE
20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
Europe (156)
EUW (53)
EUC (67)
EUS (29)
EUN (5)
EUE (2)

(b)

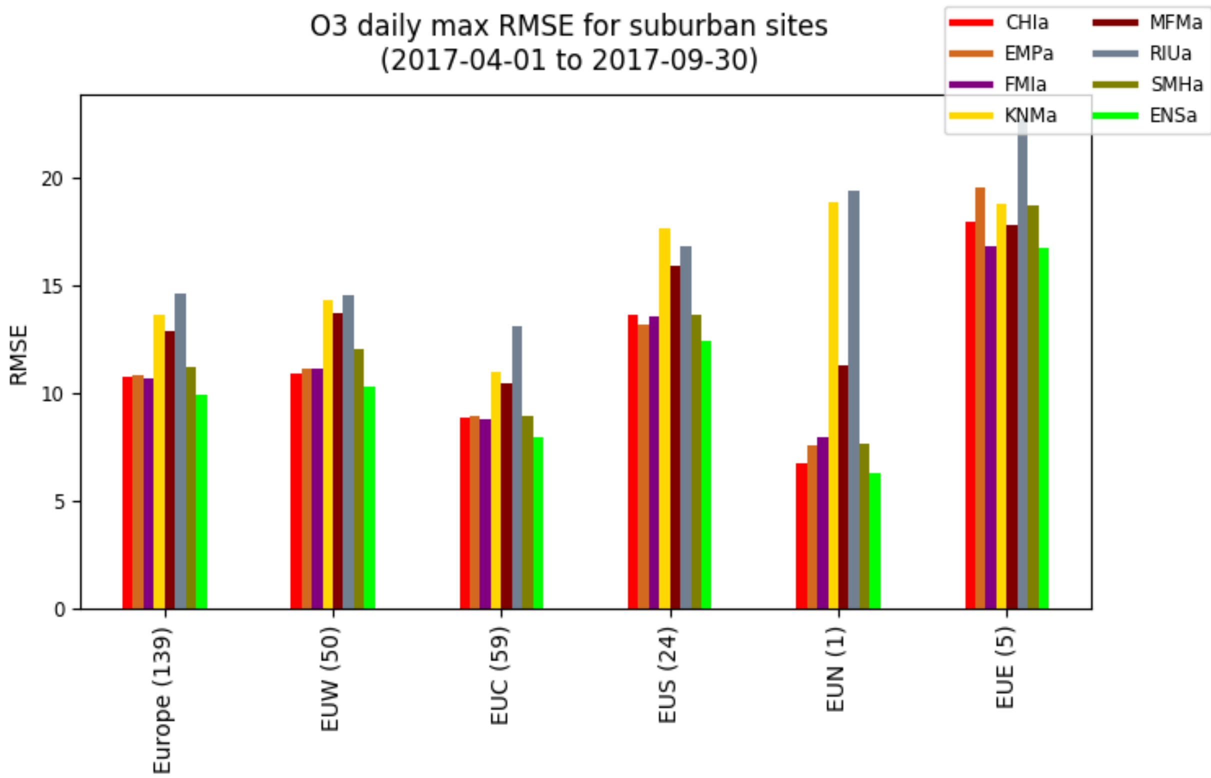


(c)

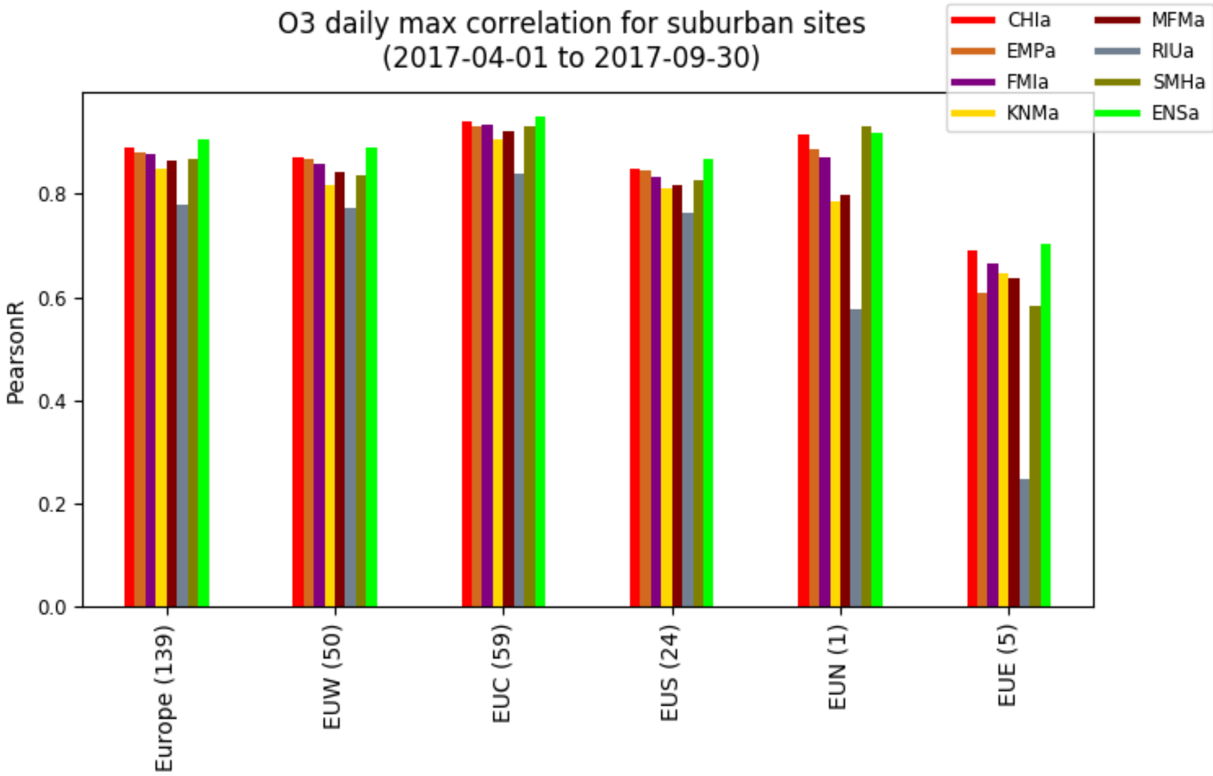
Figure 3 - CAMS regional re-analyses for predicting daily ozone peak over the summer 2017 throughout European sub-regions (a) Bias, (b) RMSE, (c) Correlation coefficient at rural stations


O3 daily max bias for suburban sites
(2017-04-01 to 2017-09-30)
CHIa
EMPa
FMIa
KNMa
MFMa
RIUa
SMHa
ENSa
MeanBias
15
10
5
0
−5
−10
Europe (139)
EUW (50)
EUC (59)
EUS (24)
EUN (1)
EUE (5)

(a)

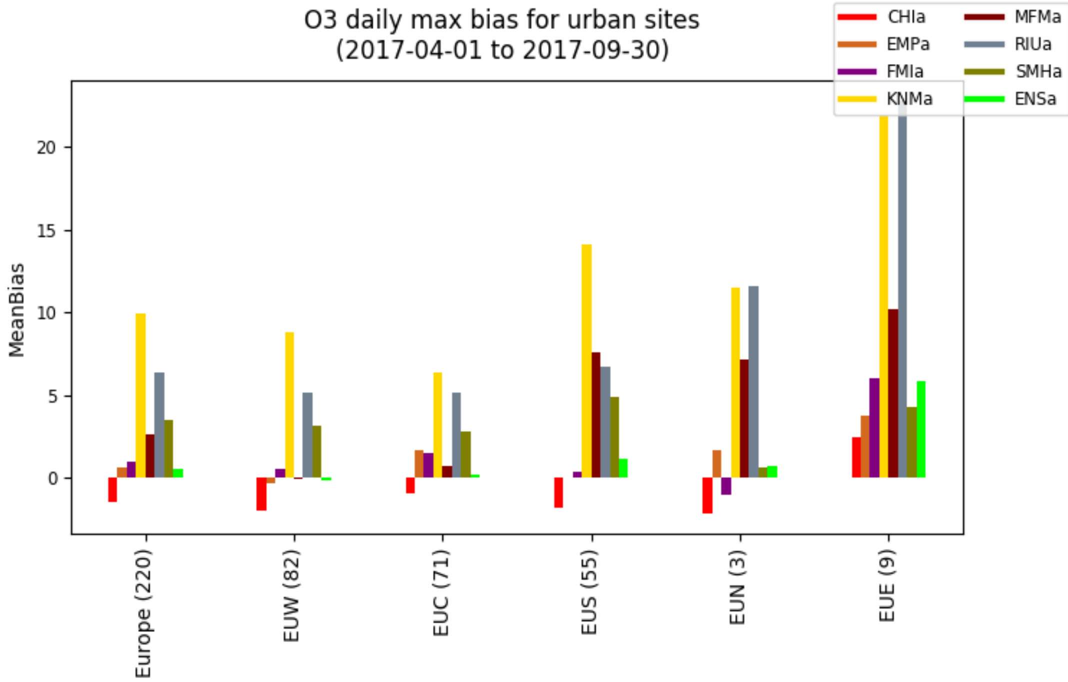

O3 daily max RMSE for suburban sites
(2017-04-01 to 2017-09-30)
CHIa
EMPa
FMIa
KNMa
MFMa
RIUa
SMHa
ENSa
RMSE
20
15
10
5
0
Europe (139)
EUW (50)
EUC (59)
EUS (24)
EUN (1)
EUE (5)

(b)

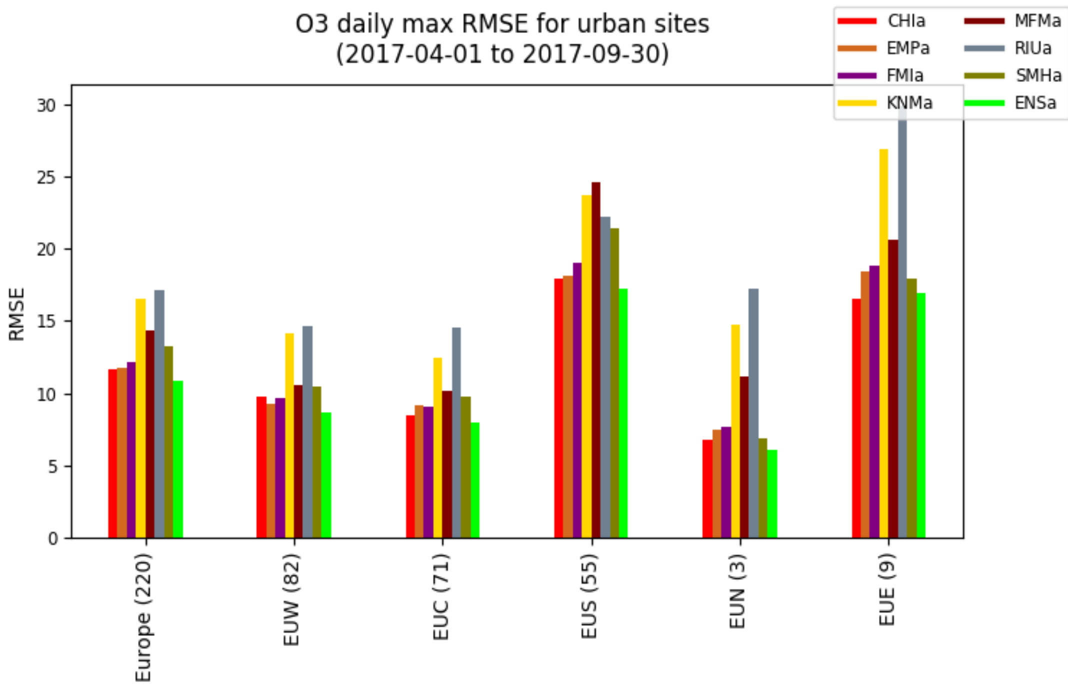


(c)

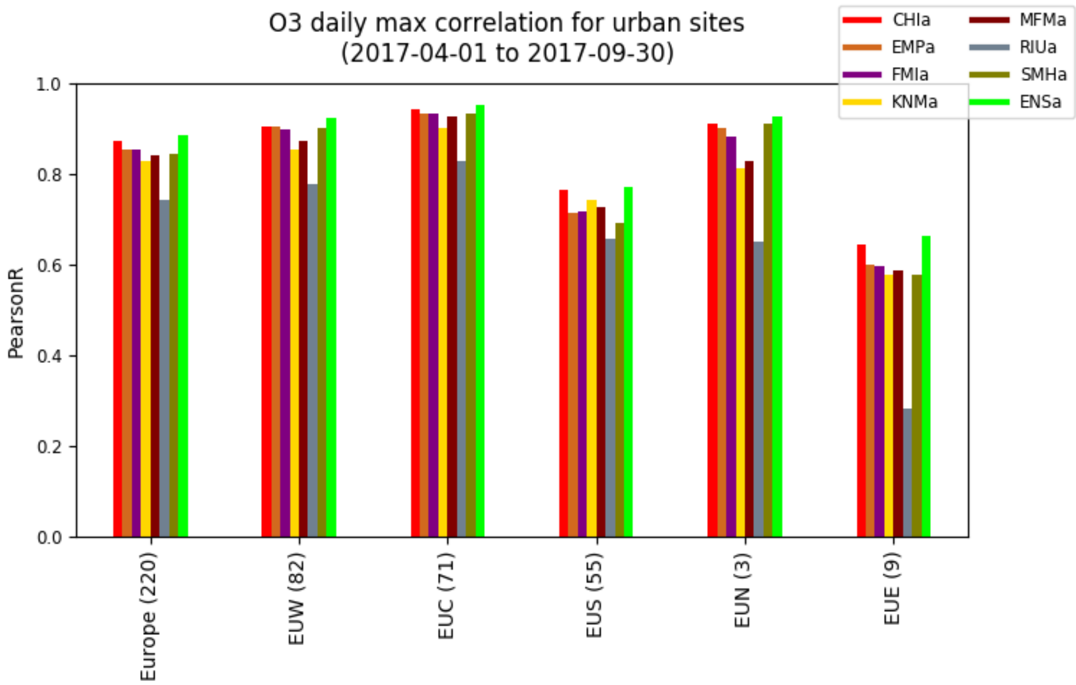
Figure 4 - CAMS regional re-analyses for predicting daily ozone peak over summer 2017 throughout European sub-regions (a) Bias, (b) RMSE, (c) Correlation coefficient at suburban stations


O3 daily max bias for urban sites
(2017-04-01 to 2017-09-30)
CHIa
EMPa
FMIa
KNMa
MFMa
RIUa
SMHa
ENSa
MeanBias
20
15
10
5
0
Europe (220)
EUW (82)
EUC (71)
EUS (55)
EUN (3)
EUE (9)

(a)


O3 daily max RMSE for urban sites
(2017-04-01 to 2017-09-30)
CHIa
EMPa
FMIa
KNMa
MFMa
RIUa
SMHa
ENSa
RMSE
30
25
20
15
10
5
0
Europe (220)
EUW (82)
EUC (71)
EUS (55)
EUN (3)
EUE (9)

(b)



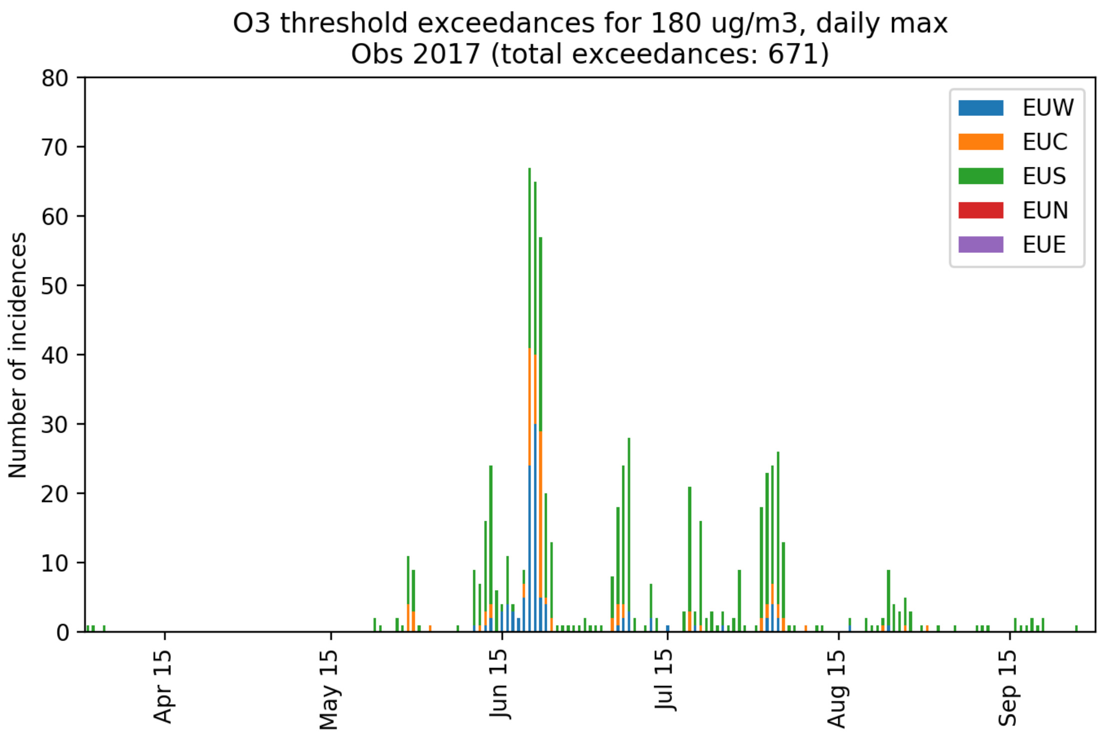
(c)

Figure 5 - CAMS regional re-analyses for predicting daily ozone peak over summer 2017 throughout European sub-regions (a) Bias, (b) RMSE, (c) Correlation coefficient at urban stations

Finally, the ability of the models to simulate the number of exceedances of a given threshold value has also been assessed. This is important for ozone, since the EU legislation (Directive 2008/50/EC) sets quality objectives with an information threshold (180 $\mu g/m^3$) and an alert threshold (240 $\mu g/m^3$) over which short-term action plans and communication towards the general public should be activated by the Member States. However, this kind of evaluation against threshold value is very stringent and not always representative of the model quality. The number of situations above and below the threshold value is calculated, but to correctly account for model uncertainty it would be necessary to take a range of acceptable values around the threshold. This is not done in the present study. Therefore, the diagnostic can be seen as a pessimistic analysis of the models' performances.

Error! Reference source not found. 6 on the next page shows the number of situations where the hourly observations exceed the information threshold during the year 2017 (time is presented on the x-axis), sorted by geographical regions and computed only with the set of stations dedicated to evaluation.
The Information ozone threshold has been exceeded 671 times in 2017 (almost twice more than in 2016), mostly located in Southern Europe.
The main episode started on June 20th and lasted during almost one week. Beside this, two other less intensive episodes occurred in July and August.
Figure 7 displays exceedances modelled by all CAMS models and the ENSEMBLE. If the performances of the ENSEMBLE were good considering statistical indicators, its capacity to detect threshold exceedances is not among the best re-analyses even if its performances are correct.
The re-analyses and observations are quite consistent regarding the exceedances, with a very good temporal agreement. Only a few exceedances are produced by re-analyses for days when it was not observed (temporal false alarm).

The ENSEMBLE detected around 30 % of the exceedances (Figure 7). This can be explained by the nature of the indicator (no range of uncertainty is taken into account), but also by the way the ENSEMBLE is built up. It is based on the median of individual model results with performance varying largely from one model to another. The median smooths the indicator (evaluation against threshold values) and the obtained results cannot be considered as representative of the actual quality and accuracy of the models. Overall performances of the ENSEMBLE also show negative bias, which usually does not help to detect the exceedances.
However, the contingency plot highlights the low number of false alarms made by the ENSEMBLE while other models which detect more exceedances have much more false alarms, therefore also highlighting a positive aspect of the conservative choice of the median.
The re-analysis with the best detection ability (46 %) is also the one with the highest number of false alarms (mostly spatial false alarms).



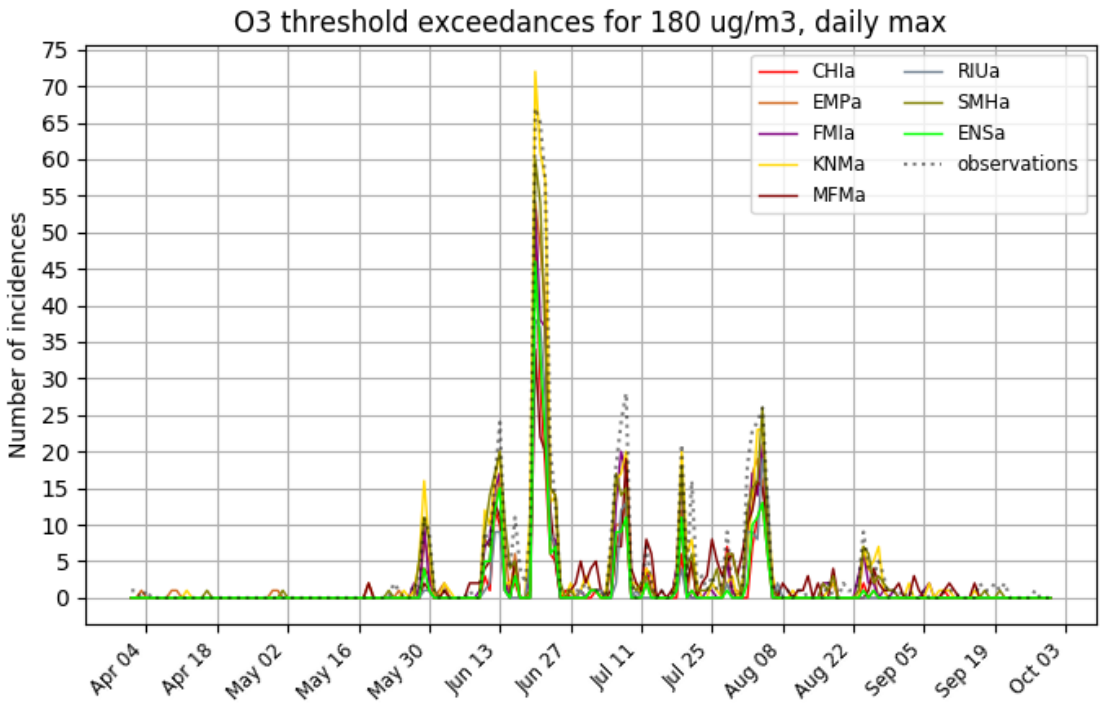


Figure 6 - Number of exceedances of the information threshold value for ozone in summer 2017 - observed (top), modelled by all the re-analyses in colour lines and observed in black dashed line (bottom)

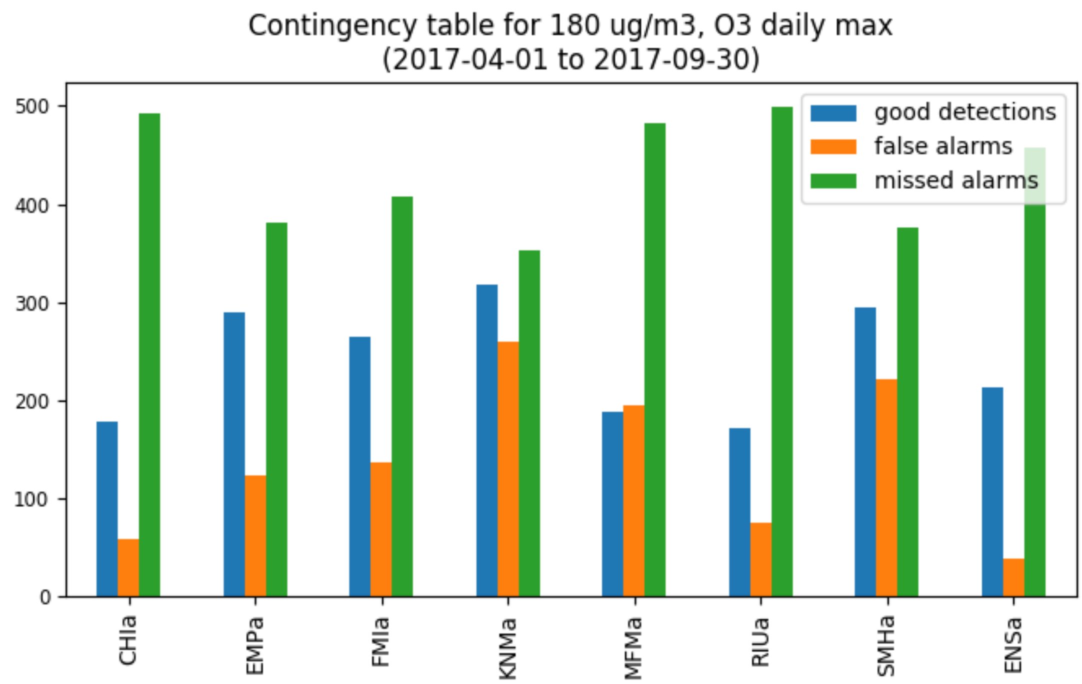


Figure 7 - Histograms describing the models' performances regarding the number of exceedances of the ozone thresholds

3. Performance indicators for nitrogen dioxide

Warning note: *It should be reminded that the CAMS regional mapping system is not fitted to deal with local hot spot situations, such as those that develop near busy roads or on industrial sites. Actually, the model resolution of 10 km is not sufficient to catch actual NO_2 concentrations at traffic and industrial sites.*

0 presents the Taylor diagram for the CAMS regional re-analyses of the ENSEMBLE and its members, for the daily maximum (hourly average) of NO_2 concentrations. It shows stable model performances compared to past results. Among them, the ENSEMBLE has one of the best ones with correlation close to 0.7 and RMSE around 13 µg/m^3. The worst performances depicted on this diagram are a correlation of 0.6 and RMSE around 15 µg/m^3.

Maps in Figure 8 allow highlighting a tendency to underestimate NO_2 daily maximum throughout Europe for the ENSEMBLE, even if some geographical areas and isolated stations show overestimations. Nevertheless, the overall bias is negative (-12 µg/m^3). The RMSE in 2017 is higher in the South than in the North, with stations depicting more than 30 µg/m^3.
Correlations are for most of the stations above 0.8 except for some geographical areas in Spain, Italy and Eastern Europe.

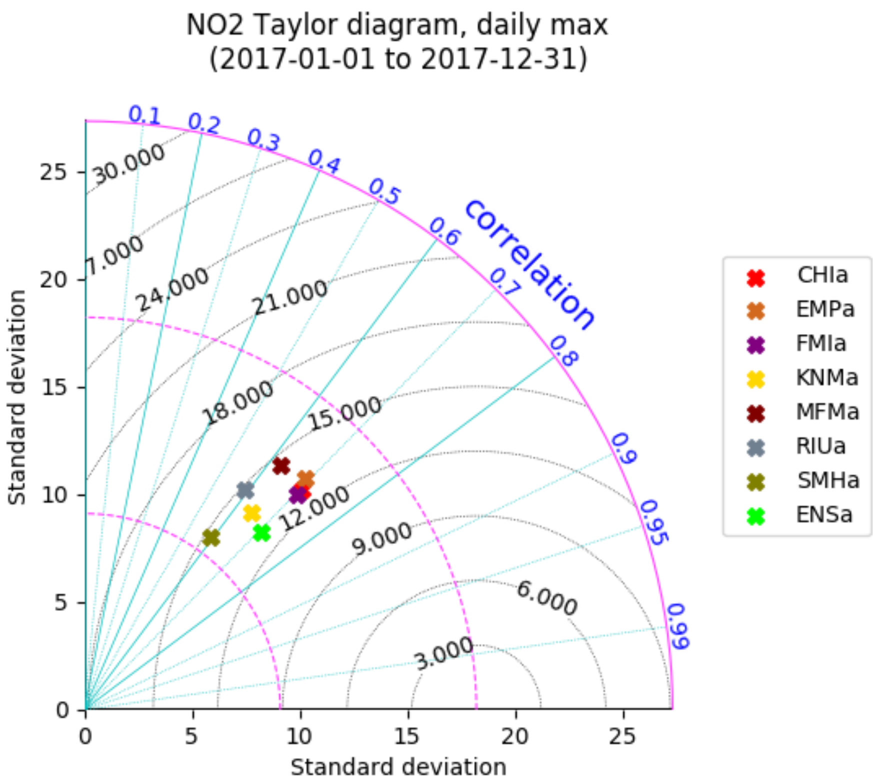
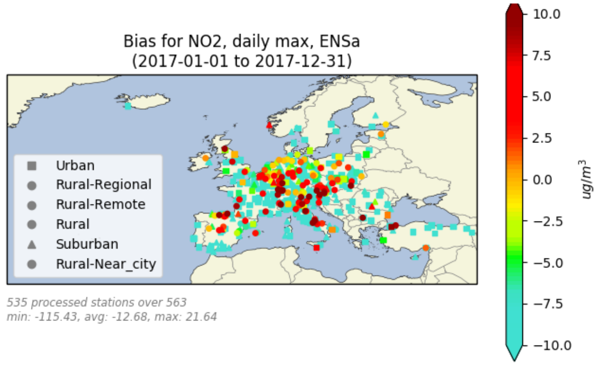
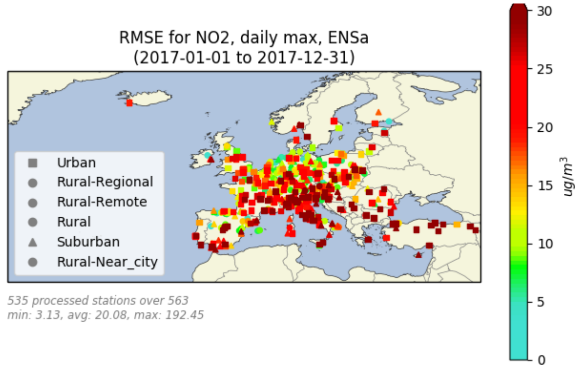


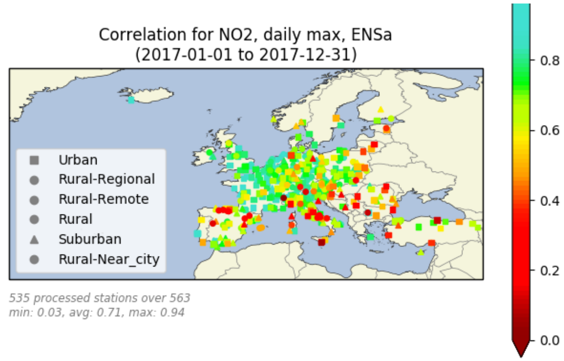
Figure 8 - Taylor diagram presenting the performances of the CAMS regional re-analyses to predict NO_2 daily maxima in 2017 at all background stations



(a) NO_2 bias (daily maximum)



(b) NO_2 RMSE (daily maximum)



(c) NO_2 correlation coefficient (daily maximum)

Figure 9 - Maps of Statistical scores of the ENSEMBLE re-analyses results against the observation validation dataset from the AQ e-reporting database for the NO_2 daily maximum over the year 2017 (a) Bias, (b) RMSE, (c) Correlation coefficient

4. Performance indicators for PM_{10}

0 shows the Taylor diagram obtained for PM_{10} daily averages over the year 2017, for CAMS regional individual and ENSEMBLE re-analyses. The diagram shows disperse scores, with correlation coefficient ranging from 0.6 to 0.9 and RMSE between 7 and 12 µg/m^3. The ENSEMBLE is among the group of re-analyses with the best performances.
Even if the ENSEMBLE scores are similar to those of the previous year (2016), the group of re-analyses is more disperse meaning that for some productions the quality has decreased.

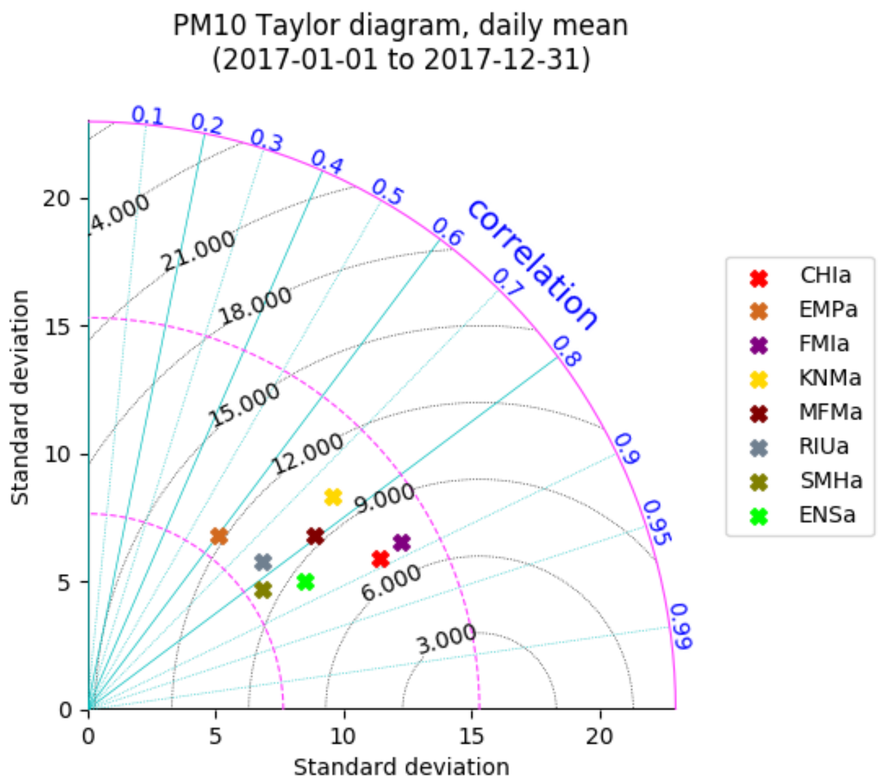


Figure 10 - Taylor diagram presenting the performances of the CAMS regional ENSEMBLE re-analyses to predict PM_{10} daily average in 2017 for all background stations

0 details the geographical distribution of statistical scores (bias, correlation coefficient and RMSE), for the ENSEMBLE re-analyses for the year 2017. The lowest scores (Bias, correlation and RMSE) are obtained for stations located in Western and Central parts of Europe. In several countries (France, Germany, Benelux, the UK), RMSE ranges between 1 and 5 µg/m^3, which is very encouraging. In contrast, Eastern countries and more particularly Poland and Southern countries like Turkey (which was not included in the previous assessment) exhibit high RMSE above 25 µg/m^3, highlighting significant variabilities of scores throughout the domain. The overall bias of the ENSEMBLE shows a tendency to underestimate the PM_{10} concentrations, partly driven by the performances of some

countries (Poland and Turkey). The correlations are pretty high everywhere in Europe except for some isolated stations and those from Turkey which range from 0.3 to 0.7.

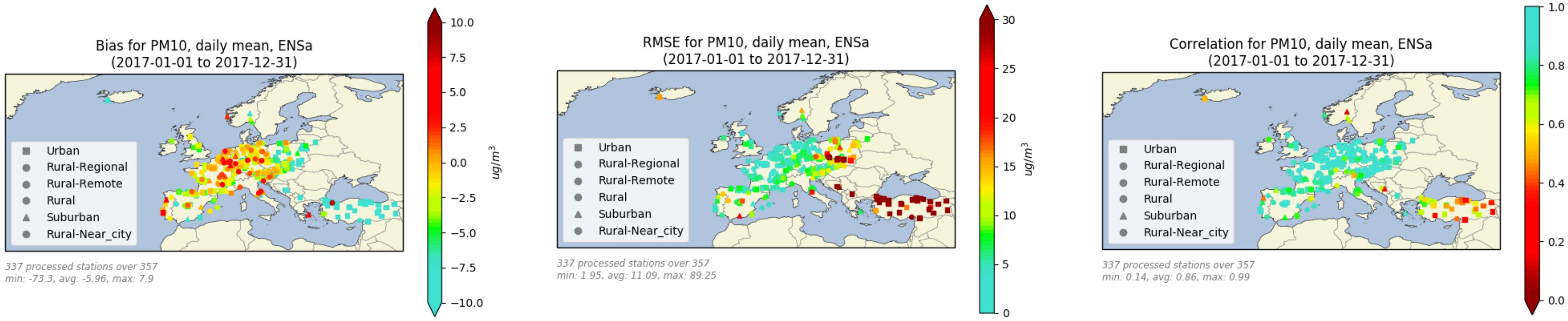
Differences between model results can be further investigated considering histograms of scores per region and for each model. 0, 0, and 0 show these results for rural, suburban and urban stations respectively. They confirm the low number of stations available for the verification of PM_{10} re-analyses, with huge gaps in some areas (scores are plotted for Northern and Eastern Europe for all station typologies but will not be discussed here, due to their low representativity).

Compared to previous years, results are robust not only over Western and over Central Europe, but also for Southern Europe especially over urban stations. For these three areas the ENSEMBLE performances are quite similar, with a slight underestimation (less than 1 µg/m^3) of PM_{10} concentrations over rural stations, despite a large range in the model responses.

Moving from suburban to urban sites, the bias becomes more negative between -2 µg/m^3 and -5 µg/m^3, even higher for Southern Europe where the negative bias reaches -30 µg/m^3 over urban sites. This certainly spots an anomaly in treatment for this part of Europe. A large range of re-analyses bias is shown on figure 14; two individual results looks good with low bias and RMSE when the 4 others are very high, leading to bad scores for the ENSEMBLE, especially over Turkey as shown on Figure 11. This result will be further analysed to understand why model responses are so different for this area.

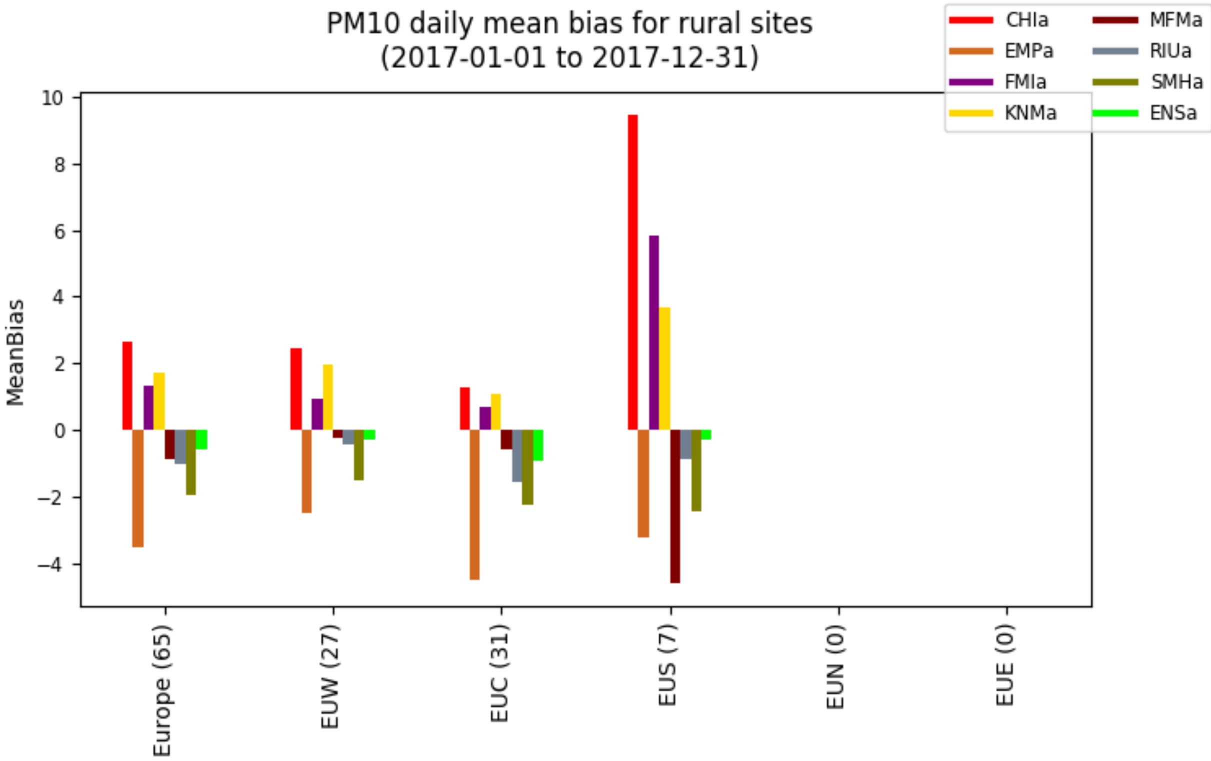
RMSE exhibits same values for rural and suburban sites while it increases over urban sites, with the outstanding contribution of Southern Europe.

Despite the variabilities of the model responses, their scores remain satisfactory; the ENSEMBLE is most of the time among the best productions, even if CHIMERE and SILAM usually performed better. Regarding the correlations, the ENSEMBLE is again among the best three scores, despite a significant variability in the model responses. Whatever the typology of the station considered, it remains above 0.8 except for Southern Europe where correlations are around 0.5.

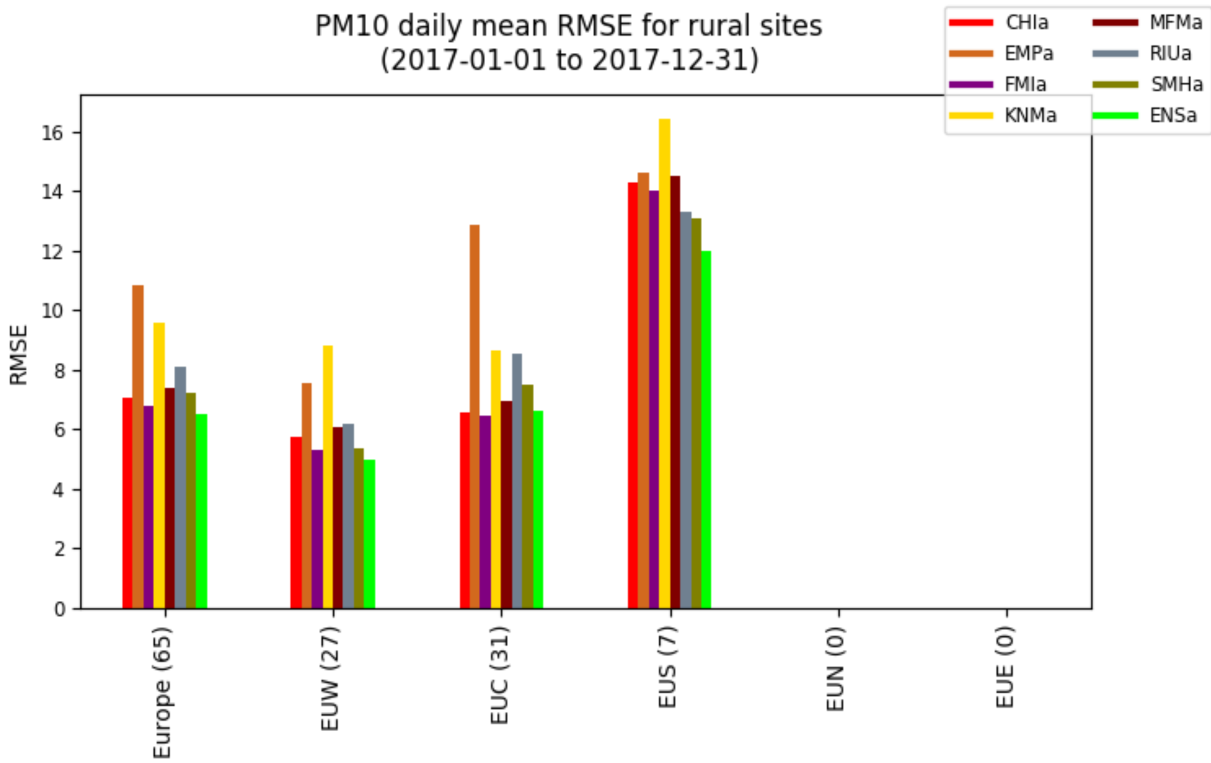


(a) (b) (c)

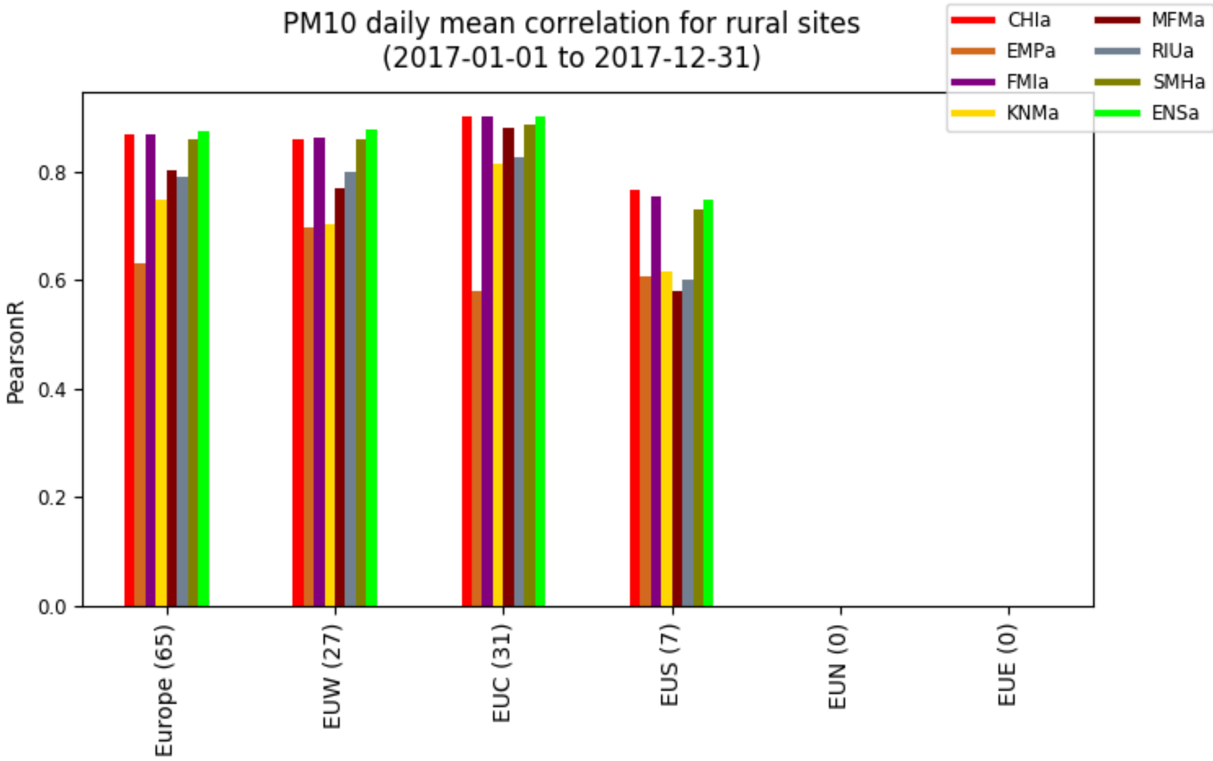
Figure 11 - Maps of Statistical scores of the ENSEMBLE re-analyses results against the observation validation dataset from the AQ e-reporting database for the PM_{10} daily average over the year 2017 (a) Bias, (b) RMSE, (c) Correlation coefficient


PM10 daily mean bias for rural sites
(2017-01-01 to 2017-12-31)
CHIa
EMPa
FMIa
KNMa
MFMa
RIUa
SMHa
ENSa
MeanBias
10
8
6
4
2
0
−2
−4
Europe (65)
EUW (27)
EUC (31)
EUS (7)
EUN (0)
EUE (0)

(a)

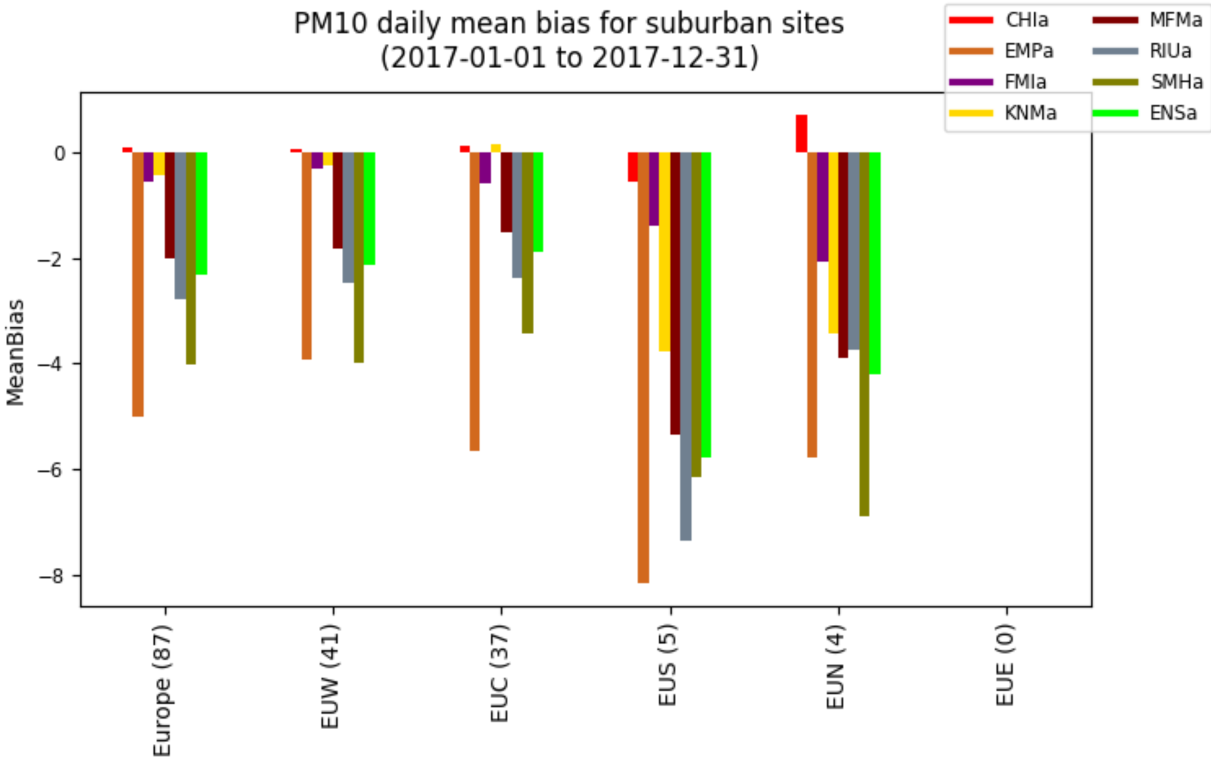

PM10 daily mean RMSE for rural sites
(2017-01-01 to 2017-12-31)
CHIa
EMPa
FMIa
KNMa
MFMa
RIUa
SMHa
ENSa
RMSE
16
14
12
10
8
6
4
2
0
Europe (65)
EUW (27)
EUC (31)
EUS (7)
EUN (0)
EUE (0)

(b)

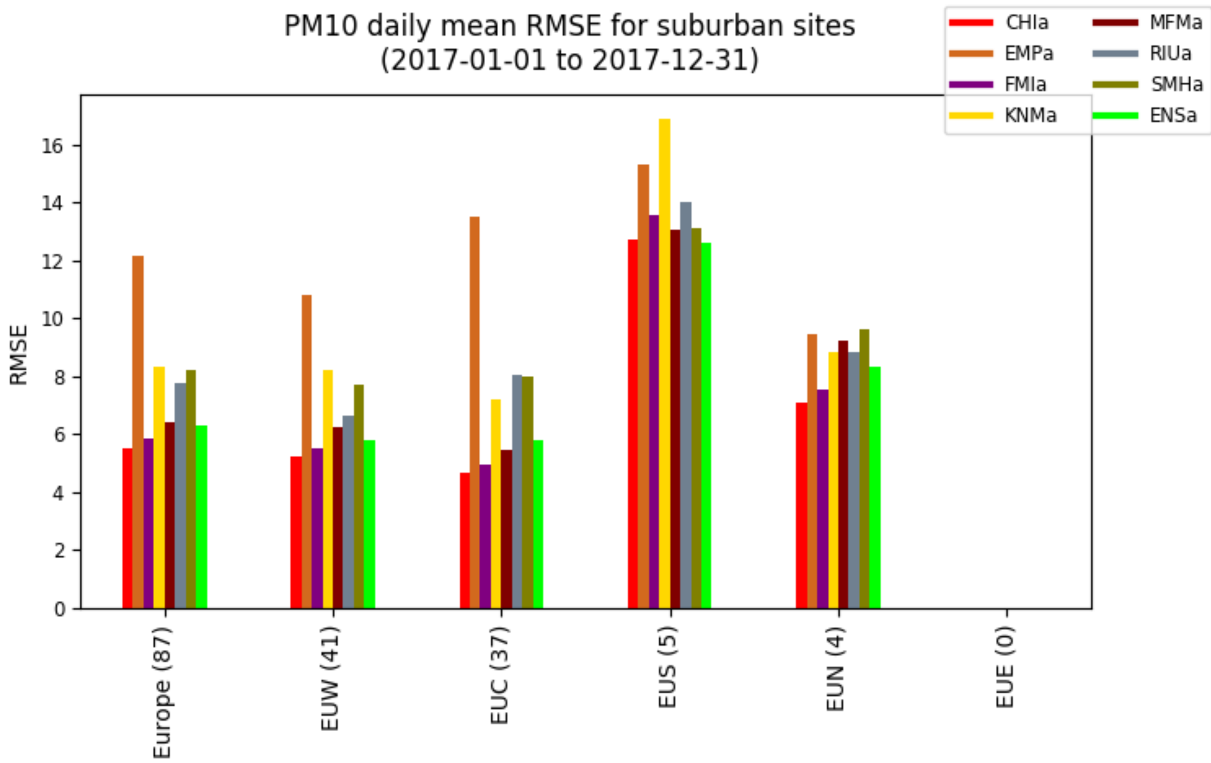


(c)

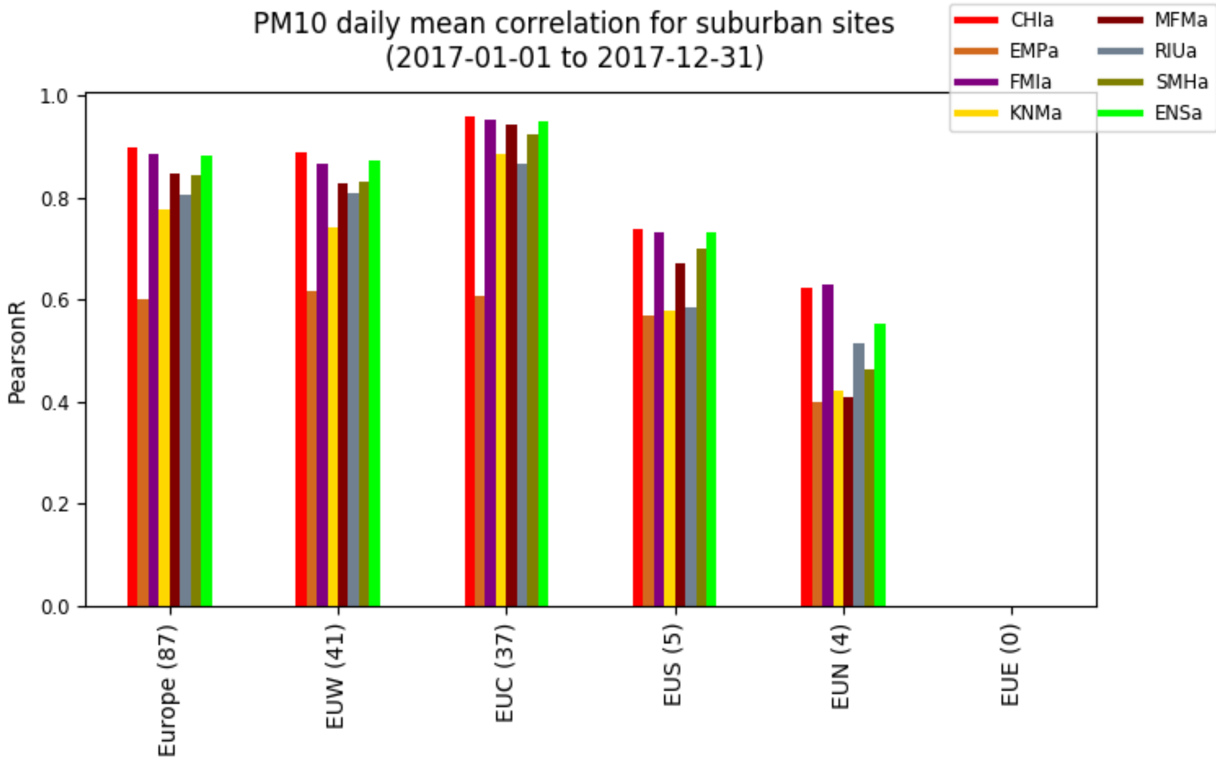
Figure 12 - CAMS regional re-analyses for predicting PM_{10} daily average over the year 2017 throughout European sub-regions (a) Bias, (b) RMSE, (c) Correlation coefficient at rural stations


PM10 daily mean bias for suburban sites
(2017-01-01 to 2017-12-31)
CHIa
EMPa
FMIa
KNMa
MFMa
RIUa
SMHa
ENSa
MeanBias
0
−2
−4
−6
−8
Europe (87)
EUW (41)
EUC (37)
EUS (5)
EUN (4)
EUE (0)

(a)

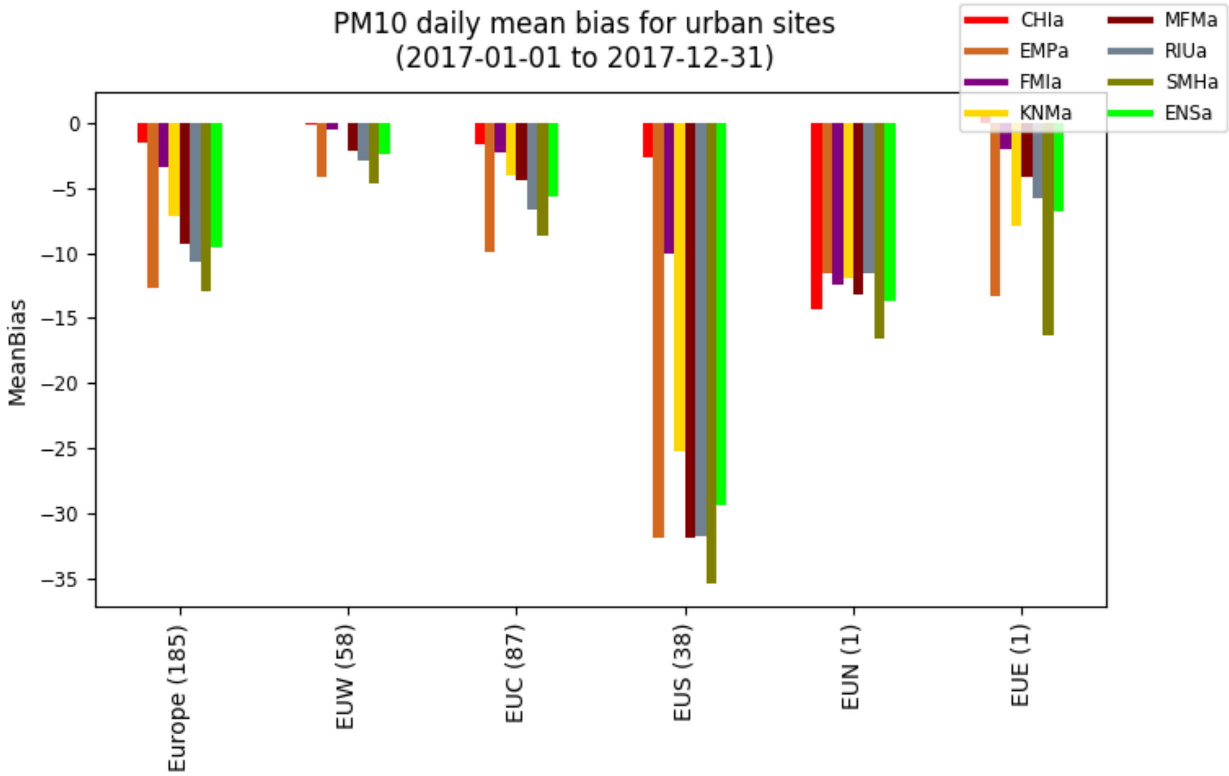

PM10 daily mean RMSE for suburban sites
(2017-01-01 to 2017-12-31)
CHIa
EMPa
FMIa
KNMa
MFMa
RIUa
SMHa
ENSa
RMSE
16
14
12
10
8
6
4
2
0
Europe (87)
EUW (41)
EUC (37)
EUS (5)
EUN (4)
EUE (0)

(b)

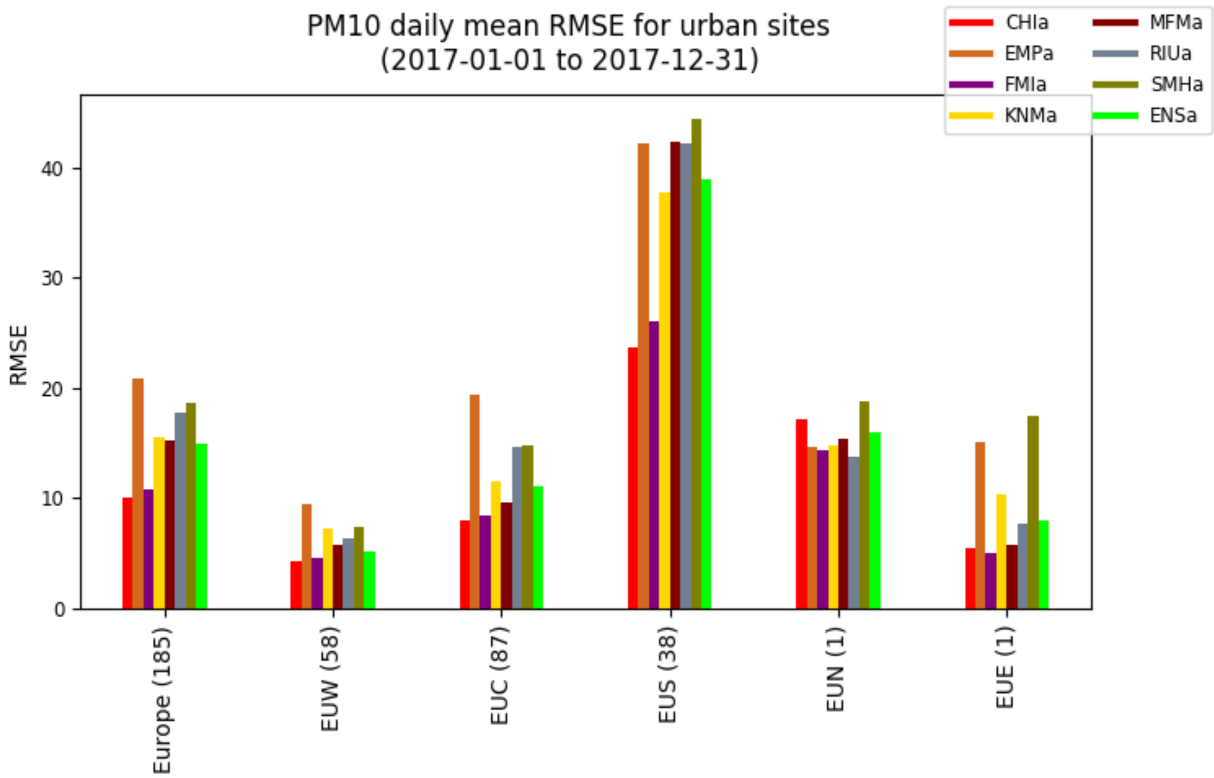


(c)

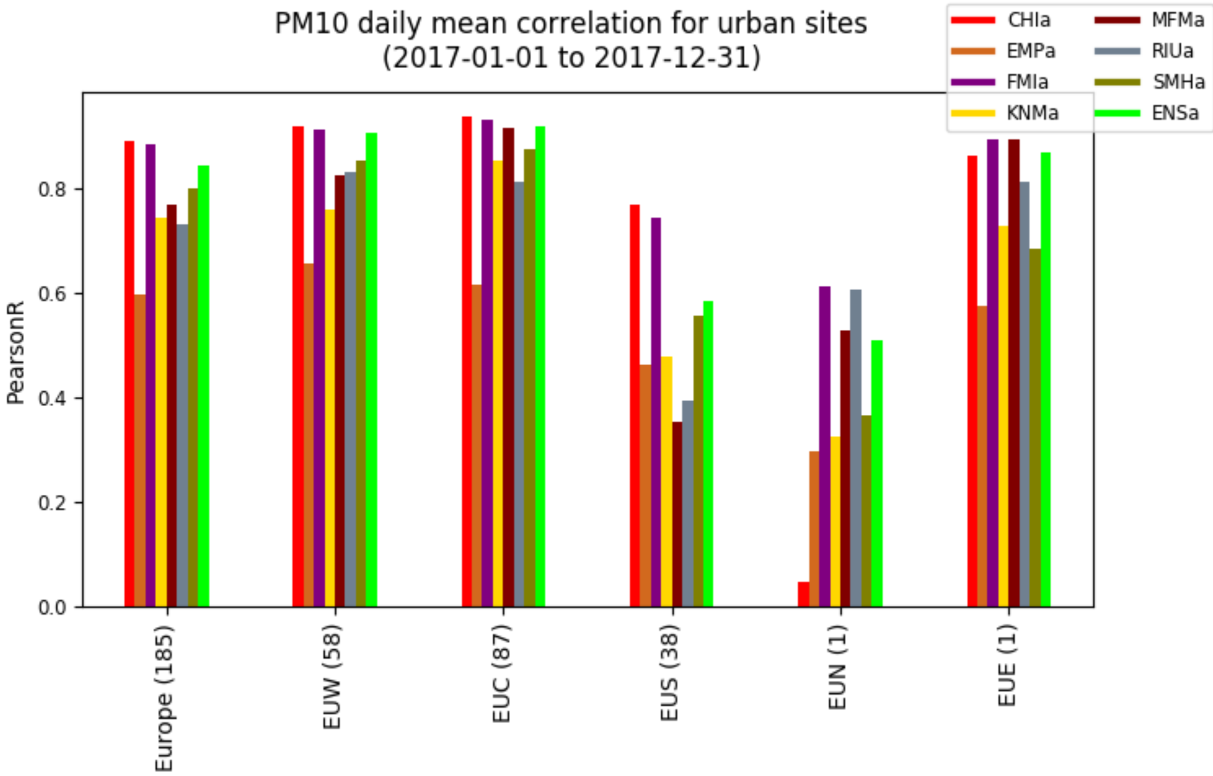
Figure 13 - CAMS regional re-analyses for predicting PM_{10} daily average over the year 2017 throughout European sub-regions (a) Bias, (b) RMSE, (c) Correlation coefficient at suburban stations


PM10 daily mean bias for urban sites
(2017-01-01 to 2017-12-31)
CHIa
EMPa
FMIa
KNMa
MFMa
RIUa
SMHa
ENSa
MeanBias
0
−5
−10
−15
−20
−25
−30
−35
Europe (185)
EUW (58)
EUC (87)
EUS (38)
EUN (1)
EUE (1)

(a)

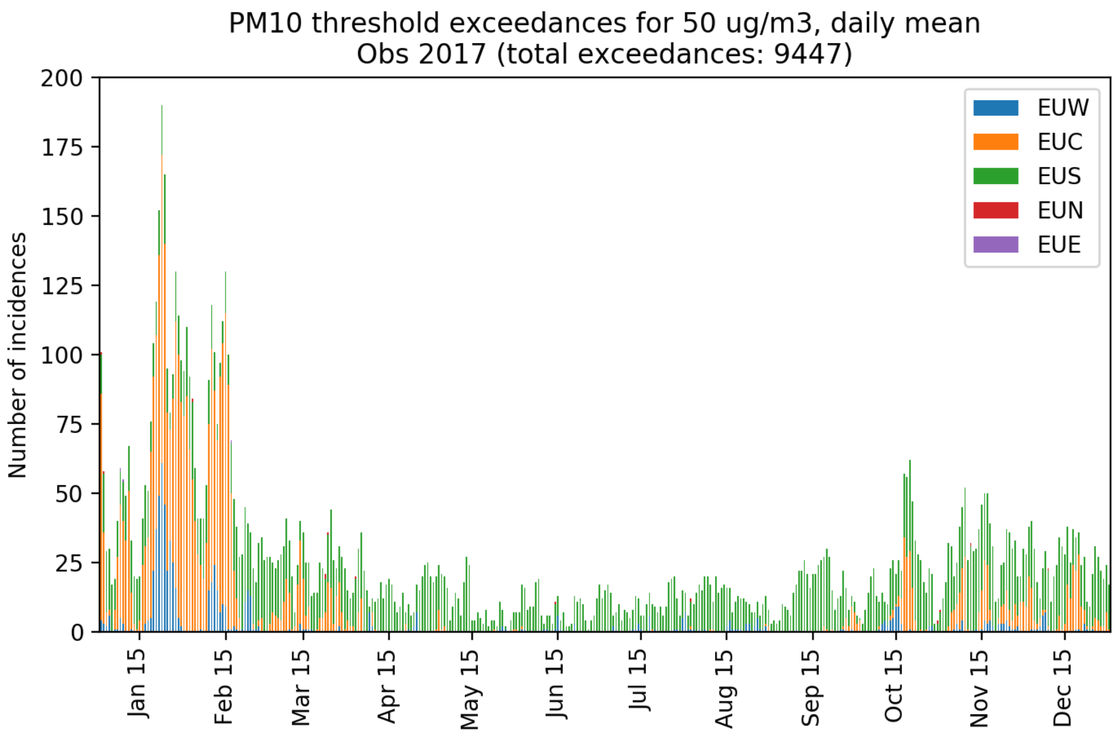

PM10 daily mean RMSE for urban sites
(2017-01-01 to 2017-12-31)
CHIa
EMPa
FMIa
KNMa
MFMa
RIUa
SMHa
ENSa
RMSE
0
10
20
30
40
Europe (185)
EUW (58)
EUC (87)
EUS (38)
EUN (1)
EUE (1)

(b)



(c)

Figure 14 - CAMS regional re-analyses for predicting PM_{10} daily average over the year 2017 throughout European sub-regions (a) Bias, (b) RMSE, (c) Correlation coefficient at urban stations



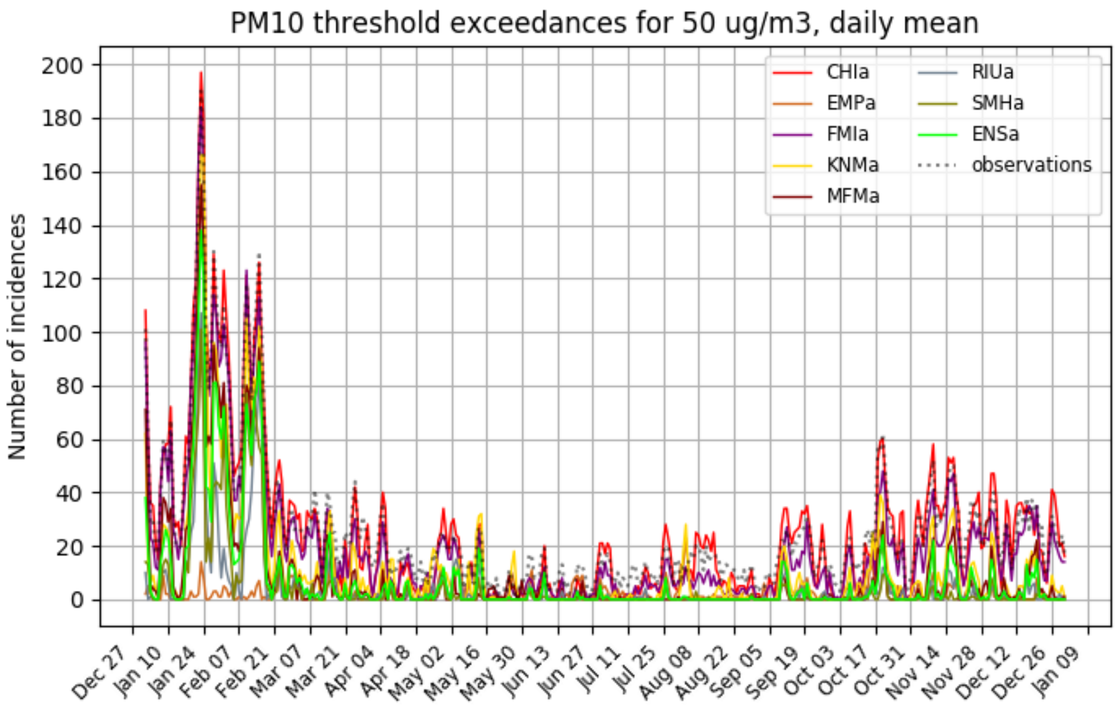


Figure 15 - Number of exceedance of daily limit value for PM_{10} in 2017 – observed (top) and modelled by re-analyses (bottom)

0 shows the number of exceedances of the PM_{10} daily limit value (50 µg/m^3), sorted per region. Both observed and re-analysed data are presented and compared.
9447 daily exceedances were recorded, based on the CAMS_50 verification stations in 2017 (Figure 15). This is three times more than in 2016, with a large contribution of PM_{10} episodes that occurred in January and February over Central and Western Europe. For the rest of the year, the frequent episodes occurring in Southern Europe are the main reason for the increase in the number of exceedances. This is explains the gap in the exceedances between 2016 and 2017 since a higher number of stations were considered in 2017, especially due to reporting made by the Turkish stations. The plot also highlights the correct capacity of the re-analyses to capture time variability in the episode severity.
The timeseries show that the ENSEMBLE tends to underestimate the exceedances. This is confirmed on Figure 16 with ~2800 detections, about only 30 % of the observed number of exceedances. Again, CHIMERE and SILAM are the two models showing a good ability of detection, respectively around 70% and 60%.
However, it is worth noting that the ENSEMBLE makes a low number of false alarms, meaning that the ENSEMBLE re-analyses indicate exceedances with a good level of confidence.
Other re-analyses show a large panel of responses in terms of ability to detect threshold exceedances. Some have very poor scores due to their chronic underestimation of the PM_{10} concentrations. Other manage to capture well part of the exceedances, but also with many false alarms. This is problematic as it leads to describing as polluted, areas or days that are not.
Part of the skills of the re-analyses are explained by their behaviours over Southern Europe.

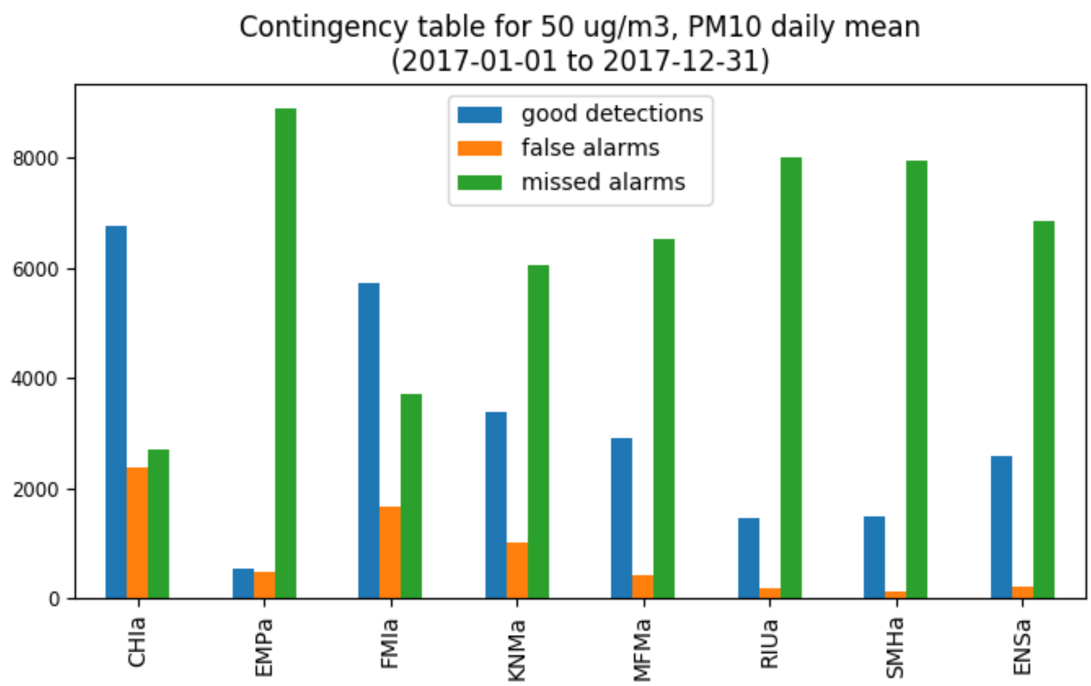


Figure 16 - Number of exceedances of daily limit value for PM_{10} in 2017 modelled by re-analyses.

5. Performance Indicators for $PM_{2.5}$

The evaluation of models' performances for $PM_{2.5}$ was constrained by the low number of stations available. This limit is clearly highlighted considering the maps on 0. However, where some measurements are available, the results for the ENSEMBLE are rather good: bias is within a couple of µg/m^3 for most of the European countries except Poland and Turkey, where scores show a significant underestimation (and a large RMSE). Correlation coefficient exceeds 0.8 everywhere (0.9 in average), except in some isolated locations, and RMSE stays generally below 5 µg/m^3. Even if some concerns about the representativeness of these scores can be raised considering the low number of stations, we can consider those figures as encouraging. The values are remarkably homogeneous regarding the geographical location of the stations.

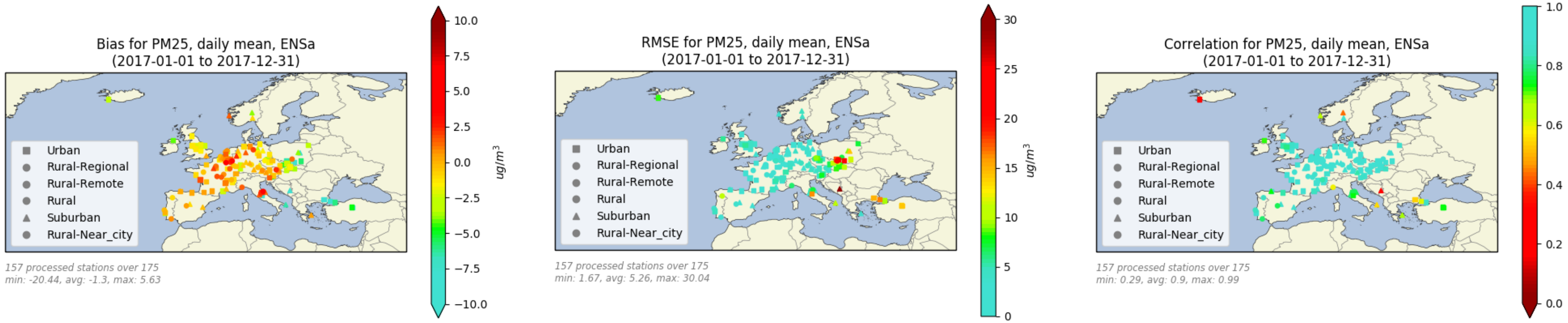
Those conclusions are confirmed by the analyses of the histograms by sub regions, showing correlation coefficient and RMSE estimated for each model and for the various station typologies (rural, suburban and urban respectively on 0 and 0).

The re-analyses bias are predominantly positive (up to 6 µg/m^3 in EUS) over rural stations moving to a negative bias (up to -12 µg/m^3 in EUS) over urban stations. The variability of the model response is important and dependent of the region considered, generally higher in EUC than EUW and higher in EUS than in EUC.

Same conclusions are drawn for RMSE, with lowest values for EUW. The RMSE does not change that much when going from rural stations to urban stations.

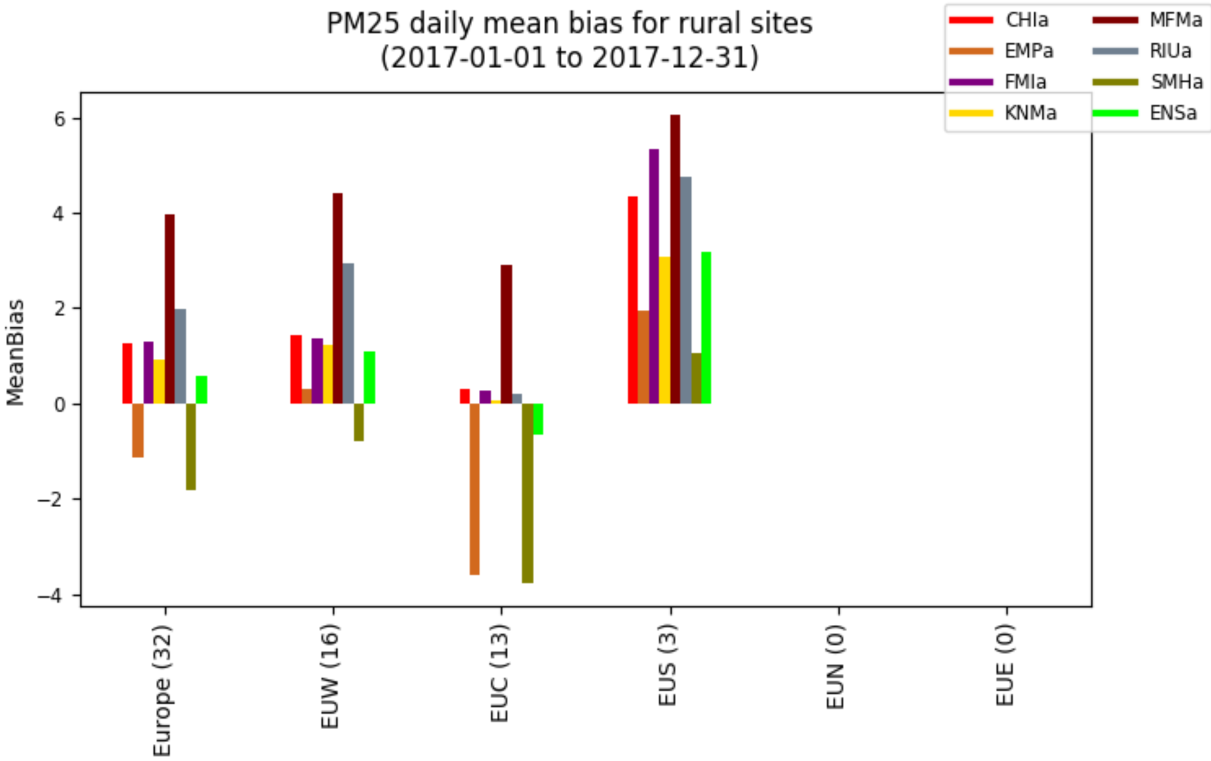
Correlations are above 0.8 for most of the models and most of the regions over rural areas, and slightly decrease for urban stations especially in EUS.

The ENSEMBLE results are quite robust and remain among the best ones whatever the region and the station typology. Results are considered very encouraging and can be considered for policy use of the CAMS_50 validated re-analyses.

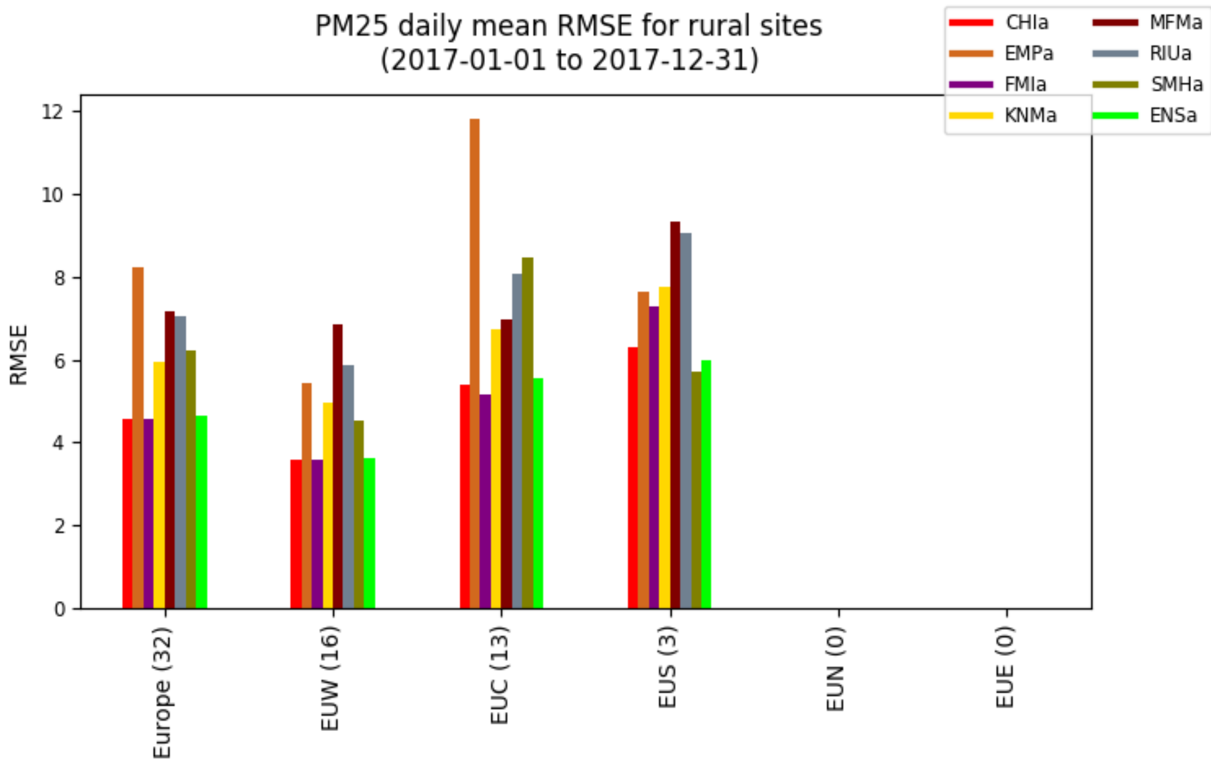


(a) (b) (c)

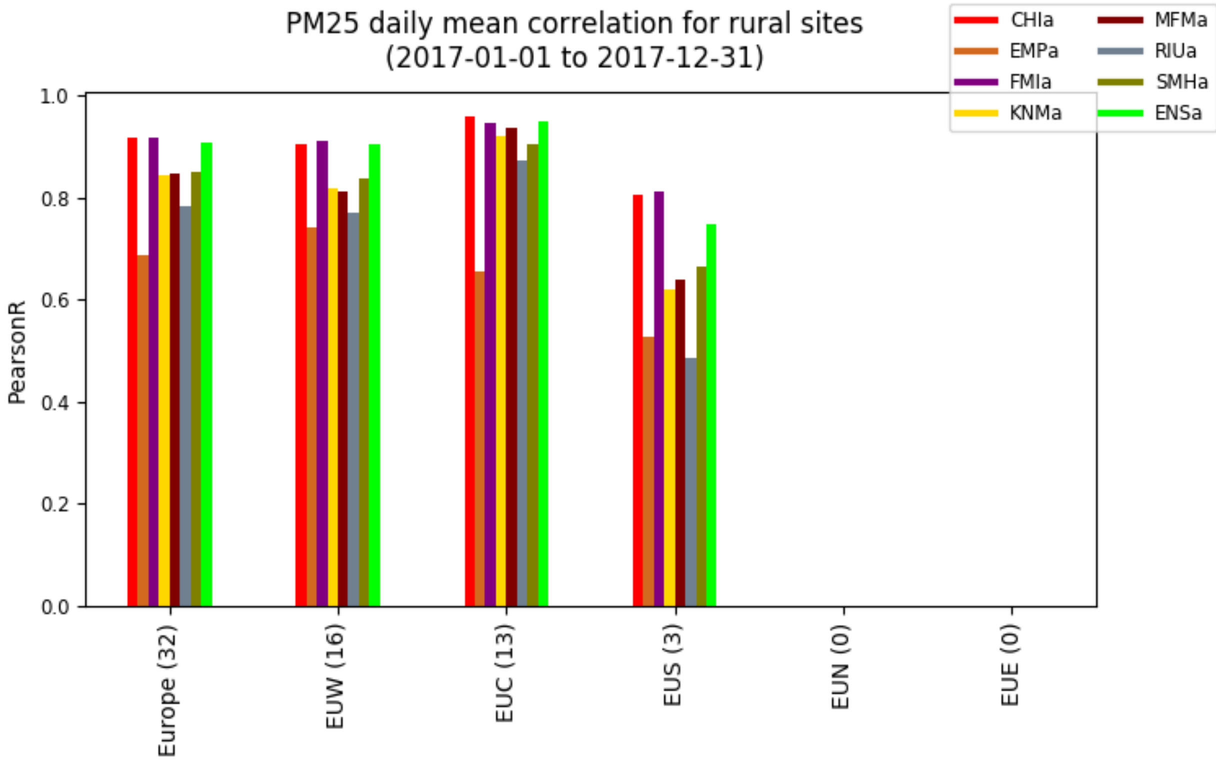
Figure 17 - Maps of Statistical scores of the ENSEMBLE re-analyses results against the observation validation dataset from the AQ e-reporting database for the $PM_{2.5}$ daily average over the year 2017 (a) Bias, (b) RMSE, (c) Correlation coefficient


PM25 daily mean bias for rural sites
(2017-01-01 to 2017-12-31)
CHIa
EMPa
FMIa
KNMa
MFMa
RIUa
SMHa
ENSa
MeanBias
6
4
2
0
−2
−4
Europe (32)
EUW (16)
EUC (13)
EUS (3)
EUN (0)
EUE (0)

(a)

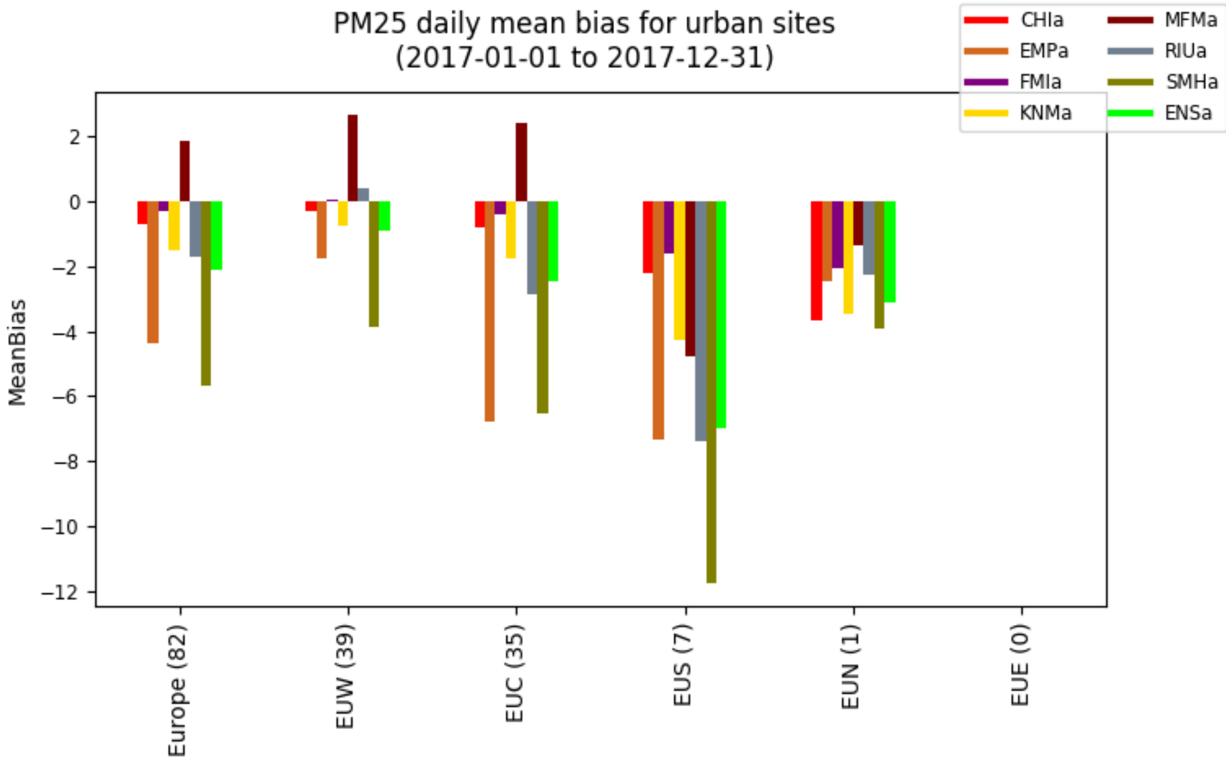

PM25 daily mean RMSE for rural sites
(2017-01-01 to 2017-12-31)
CHIa
EMPa
FMIa
KNMa
MFMa
RIUa
SMHa
ENSa
RMSE
12
10
8
6
4
2
0
Europe (32)
EUW (16)
EUC (13)
EUS (3)
EUN (0)
EUE (0)

(b)

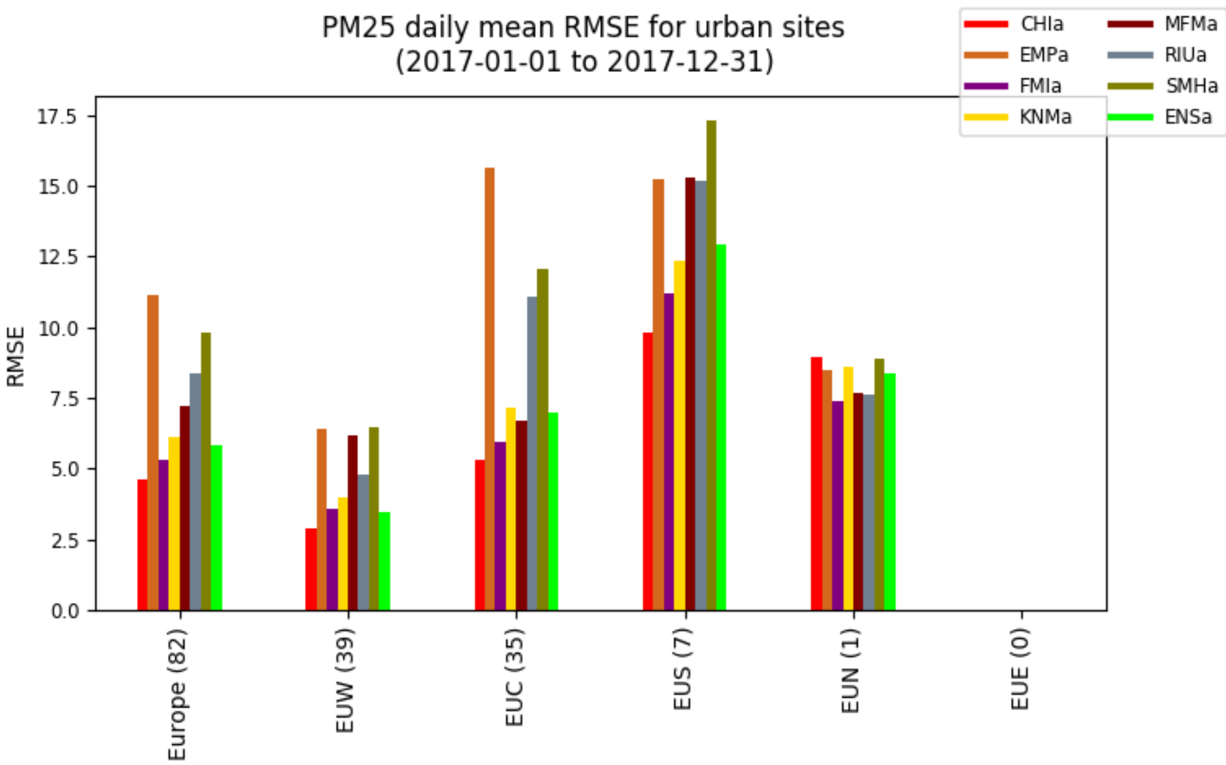


(c)

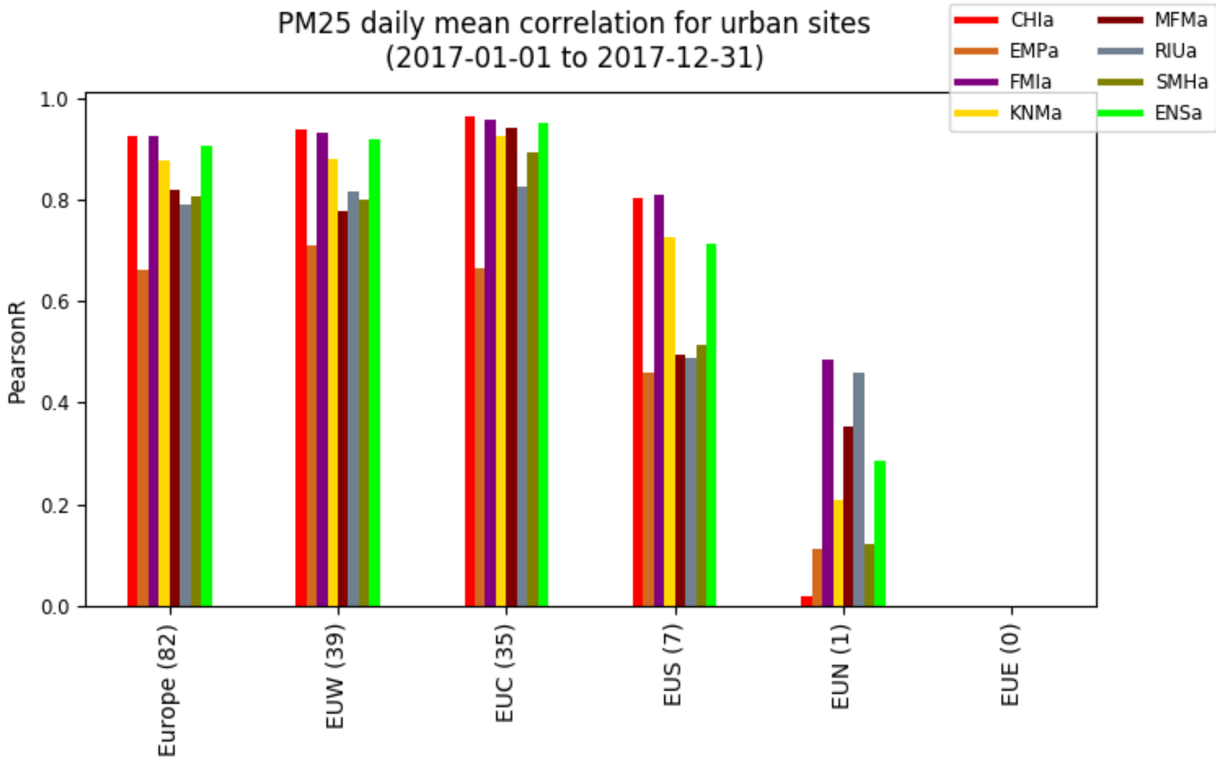
Figure 18 - CAMS regional re-analyses for predicting $PM_{2.5}$ daily average over the year 2017 throughout European sub-regions (a) Bias, (b) RMSE, (c) Correlation coefficient at rural stations


PM25 daily mean bias for urban sites
(2017-01-01 to 2017-12-31)
CHIa
EMPa
FMIa
KNMa
MFMa
RIUa
SMHa
ENSa
MeanBias
2
0
−2
−4
−6
−8
−10
−12
Europe (82)
EUW (39)
EUC (35)
EUS (7)
EUN (1)
EUE (0)

(a)


PM25 daily mean RMSE for urban sites
(2017-01-01 to 2017-12-31)
CHIa
EMPa
FMIa
KNMa
MFMa
RIUa
SMHa
ENSa
RMSE
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
Europe (82)
EUW (39)
EUC (35)
EUS (7)
EUN (1)
EUE (0)

(b)



(c)

Figure 19 - CAMS regional re-analyses for predicting $PM_{2.5}$ daily average over the year 2017 throughout European sub-regions (a) Bias, (b) RMSE, (c) Correlation coefficient at urban stations

Conclusion

This report gives an overview of the performances of the European air quality **validated re-analysis** (VRA) process developed by the CAMS_50 services and implemented in 2019 to simulate air quality in Europe during the year 2017.

Air quality validated re-analyses result from a combination of chemistry-transport models that simulate the spatial and temporal evolution of air pollutant concentrations, and observations assimilated in the model to correct and improve its results. Data assimilation techniques are used.
The models implemented to perform the service are the set of seven models run in other near-real-time CAMS_50 services. The models are CHIMERE (INERIS, France), EMEP (MET Norway, Norway), EURAD-IM (Univ. Köln, Germany), LOTOS-EUROS (TNO/KNMI, The Netherlands), MATCH (SMHI, Sweden), MOCAGE (METEO-FRANCE, France) and SILAM (FMI, Finland).

Observations that are assimilated in the chemistry-transport models are issued from the regulatory air quality monitoring networks that report air quality data to the European Environment Agency (EEA) according to air quality Directive 2008/50/EC and Decision 2011/850/EU *on reciprocal exchange and reporting on ambient air quality*. Observation datasets have followed a stringent validation process by the national experts on the one side and by EEA on the other side.
Some models also use assimilation of additional datasets such as satellite observations.
To run this evaluation process, INERIS computed a number of performance indicators and scores for ozone peaks, nitrogen dioxide, PM_{10} and $PM_{2.5}$. They are presented in this report. Globally, the models performed as expected and the ENSEMBLE median very often gives the best results. Consistency with previous validated assessment results is ensured.

The following points can be highlighted:

- Even if model scores for ozone are good in a large part of Europe and consistent with the previous year, performances are better in Western and Central Europe than in Southern Europe (Mediterranean countries). This can be explained by uncertainties in meteorological fields and by the complexity of the photochemistry and deposition processes in those regions that may not be correctly approximated in the models. This is a well-known issue, but it should be carefully considered for future improvements of the CAMS_50 modelling system. Some difficulties are also noted in some part of Eastern Europe that should be further investigated, but the limited coverage of such region by stations makes it difficult to interpret the results.

- For ozone, a stability of the ENSEMBLE performances should be noted compared to last year. For the detection of the exceedances of the air quality objectives (information threshold in particular), a slight increase compared to last year was seen (30% of good detection of 180 µg/m^3 threshold), partly thanks to a more intensive summer in terms of ozone episodes than last year. The ENSEMBLE is among the best but not the best one for threshold detection. However, its behaviour produces a very limited number of false alarms.
- For nitrogen dioxide, the results are better than in 2016, with a general improvement for all re-analyses including the ENSEMBLE, which has very good scores. Nonetheless, observations are

lacking in Eastern and Southern parts of Europe to draw robust conclusions for these parts of Europe.

- The same concern holds for PM_{10} and $PM_{2.5}$, too few observation data are available for evaluation in Eastern countries. Nevertheless, it may change in the coming years as noticed in 2017 in Southern countries: indeed, many data for these countries were reported in 2017 for PM_{10}, allowing for a robust evaluation for this part of Europe. The results are very good for Western and Central Europe, with correlation coefficients higher than 0.85 and RMSE lower than 5 $\mu g/m^3$; for Southern Europe however, there is still some progress to be made to correct the regular underestimation of the PM_{10} concentrations, particularly over urban stations. The heterogeneity of the model responses for PM_{10} in Southern countries and especially Turkey should be investigated.
- For $PM_{2.5}$, the results are similar to PM_{10} in terms of re-analyses responses. The scores are good with low bias and RMSE values for most of the stations and correlation around 0.9. The number of stations is growing year after year, which improves the representativeness of this evaluation.
- Concerning the capability of the ENSEMBLE to detect threshold exceedances for O_3 and PM_{10}, it is worth noting that rates of false alarm are very low. This means that the representations of areas with exceedance are reliable and certainly underestimated, due to the tendency of the ENSEMBLE to have a significant number of missing exceedances.
- The overall performances of the ENSEMBLE re-analyses are quite satisfactory (and in line with the state of the art) for the set of pollutants analysed in this report (O_3, NO_2, PM_{10} and $PM_{2.5}$).
- This report provides insights of CAMS_50 re-analysis skills that may help users to rely on these data.

ECMWF - Shinfield Park, Reading RG2 9AX, UK

Contact: info@copernicus-atmosphere.eu