ECMWF COPERNICUS REPORT

Copernicus Atmosphere Monitoring Service

Annual report on the verification of interim re-analyses

IRA2018

Issued by: METEO-FRANCE/ G. Collin

Date: 19/08/2019

Ref: CAMS50_2018SC1_D5.3.1-2018_201908_Annual_verification_report_IRA2018_v1

IMPLEMENTED BY ECMWF

This document has been produced in the context of the Copernicus Atmosphere Monitoring Service (CAMS).
The activities leading to these results have been contracted by the European Centre for Medium-Range Weather Forecasts, operator of CAMS on behalf of the European Union (Delegation Agreement signed on 11/11/2014). All information in this document is provided "as is" and no guarantee or warranty is given that the information is fit for any particular purpose.
The user thereof uses the information at its sole risk and liability. For the avoidance of all doubts, the European Commission and the European Centre for Medium-Range Weather Forecasts has no liability in respect of this document, which is merely representing the authors view.

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 provides a performance analysis of the Regional interim air quality re-analyses throughout Europe, produced by CAMS for the year 2018.

The CAMS Regional 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 "interim" re-analyses are data assimilated fields of air pollutant concentrations, based on up-to-date observation data. Since October 1st, 2015, according to EU Decision 2011/850/EU *on reciprocal exchange of information and reporting on ambient air quality*, EU Member States must report to the European Environment Agency (EEA) observation data as soon as it is produced, even if the necessary validation process is not completed. Such data is thus flagged as "non-validated" or "non- verified" data. Up-to-Date (UTD) data should be considered as provisional or "interim" data, until they are flagged as "validated" by the Member States, which can formally happen more than one year after their production[1].

Nevertheless, it is interesting to elaborate interim re-analyses as first guess of air pollution patterns and levels that developed in Europe in 2018. Such information can be used to support Member States for the regulatory reporting duty on air quality (according to Directive 2008/50/EC). This is the reason why it is important to carefully evaluate the simulations against observations that are not used for the re-analyses production.

INERIS performed this evaluation process and computed several performance indicators and scores for ozone, nitrogen dioxide, PM_{10} and $PM_{2.5}$ concentrations. They are presented in this report. Globally the models performed as expected and the ENSEMBLE median re-analyses generally gives good results, but not always the best ones especially when analyzing the capability of the re-analyses to detect threshold exceedances for all compounds. Consistency with previous validated interim re-analyses results is ensured.

The interim re-analyses maps can be considered as relevant for policy support, even if some care should be taken, as usual with provisional results.

We can highlight the following points:

- Too little up-to-date observation data was available to perform an extensive evaluation of interim re-analyses over the whole of Europe (except for Central and Western Europe). Very few observations can be available in Eastern Europe and, depending on the pollutant, in Southern and Northern European regions that are not correctly covered. This is frustrating since they correspond to areas where there are more uncertainties (especially because of emissions).
- In Western and Central Europe, where there are more stations for the evaluation of the models' performances, results are generally more representative and correct. The quality of the re-analyses is generally similar to the previous year 2017.

[1] Validated observations related to year Y-1 are reported by September 30th of year Y by the Member States.

- The European Environment Agency is building capacity to strengthen quality assurance procedures in the coming years and more countries are supposed to deliver up-to-date data, which will impact positively the interim re-analyses production process.
- For all pollutants, the performances are always of lower quality than what can be achieved with the validated re-analyses process, for which more stations are available and observation datasets are validated.
- The ENSEMBLE re-analyses give the best results for ozone, when focusing on classical statistical scores. Ozone daily maxima are generally underestimated. Correlation coefficient ranges between 0.8 and 0.9 and RMSE between 15 and 18 μg/m^3 at rural and suburban locations. However, when looking at the threshold exceedances, it is worth noting the low capabilities of the ENSEMBLE re-analyses to detect concentrations above the standards, and its good skills to keep the number of false alarms at a very low level.
- Good model scores for simulating ozone in Western Europe are hampered by inferior performances at few stations in Eastern and Southern Europe.
- The performances of nitrogen dioxide re-analyses are quite stable with previous years and with satisfactory scores. RMSE is around 15 µg/m^3, bias shows an underestimation of 5 µg/m^3 and correlation close to 0.7.
- For PM_{10}, even if the results are quite satisfactory considering the state of the art, the statistical scores remain lower than what is usually achieved with validated re-analyses. Up-to-date PM_{10} observation datasets need to be improved in the future.
- PM_{10} is the pollutant for which model responses range in the largest interval: Correlation coefficient from 0.35 to 0.8 and RMSE from 10 to 25 μg/m^3 depending on the model and the station typology. The results are the best for suburban stations in Western and Central Europe. More frequent overestimations of PM_{10} concentrations occurred over European stations and not only for rural ones.
- Moreover, the evaluation demonstrates how the Ensemble approach, based on a median average of involved models is not appropriate to simulate exceedances of threshold values. Only 35% of good detection of exceedances of the PM_{10} daily limit values were correctly caught by the ENSEMBLE, whereas the best re-analyses got 55 %. As for ozone, the ENSEMBLE re-analyses produce a very low number of false alarms.
- For the first time, we report some overestimation of PM_{10} in 2018 for some models, which is surprising given the usual underestimation reported in the past. This issue may be related to the lack of update in emissions since 2011, and needs to be followed up when assessing the performances of IRA after the update of CAMS-REG-AP_v2.2.1 emissions.
- Finally, despite only little $PM_{2.5}$ measurement data was available for the evaluation, the results obtained for this pollutant are promising. The individual models' responses are quite consistent, and the Ensemble median generally gives the best results. Correlation coefficient ranges from 0.4 to 0.8 according to the location and the station typology and the RMSE from 5 to 17 μg/m^3, which is very reasonable. Once again, the conclusions are limited by the low number of stations available in some geographical areas and should be consolidated and improved in future interim assessments, when the up-to-date data gathering process at the EEA is strengthened.

Introduction

This report gives an overview of the performances of the European air quality **interim re-analysis** process developed by the CAMS Regional services and implemented to simulate air quality in Europe during the year 2018.

Air quality interim re-analyses result from a combination of chemistry-transport models results that simulate the spatio-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 interim 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 (FZJ, Germany), LOTOS-EUROS (KNMI-TNO, The Netherlands), MATCH (SMHI, Sweden), MOCAGE (METEO-FRANCE, France) and SILAM (FMI, Finland).

Observations are issued from the regulatory air quality monitoring networks that report 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. “Interim re-analyses” are so called because the observation data used are not formally validated yet. The 2011 decision stipulates that Member States must report monitoring data as soon as they are produced, in near-real-time, with an appropriate flag indicating that they are not verified or validated yet. This set of data is named “Up-To-Date (UTD) data”. The data is gathered in the commonly named AQ e-reporting database. “Interim data” are UTD data collected on the EEA website within a certain delay, to leave enough time to have a chance to get verified data[2]. We estimate that 20 days is an appropriate time lag to get the data and run the re-analyses for a given day.

The set of observation sites reported to the EEA is split into two subsets, one for data assimilation (with almost 2/3 of the stations) in the interim re-analyses and the other (the remaining 1/3 of the stations) for verification. Those datasets do not overlap, and verification cannot be biased by use of data for both assimilation and verification processes. It should be noted that not all Members States reported UTD data and that other countries just start (like Italy) with partial observed dataset made available. Consequently, data assimilation and evaluation cannot be performed in some geographical areas and the robustness of the results may vary from one area to another, and so it will not be possible to draw some clear conclusions about the model capacities in those regions.

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 and is considered as the best estimate of air pollution patterns and levels, since it combines the strengths of the other models. This is what 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

[2] Member states can check, verify and validate their data when they want and resubmit with the appropriate flag as many times as they wish. Formal validation is expected only in September the year after.

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 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. Comparison of observed and modelled averages is generally considered as well.
Obviously, the behavior of performance indicators depends on the station typology and on the considered pollutant: the models used in the CAMS Regional service run at the European scale and their spatial resolution is about 20 to 10 km in the best case. 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 type 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[3] in evaluation processes:

- Bias indicates, on average, if the simulations under or over-predict 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/m³.

- 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/m³.

- 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.

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

  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.

[3] 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 delimit arcs with bias values and the internal semi-circles correspond to the RMSE values. Therefore, this is a very pedagogic way to present an overview of the relative performances of a set of models, 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 missing 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 arbitrary in one or the other category), they can give a useful information to compare various models’ behavior in different geographical regions. GP, FA and ME are expressed in percentage (%) and referred also 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.

0 shows the Taylor diagram that synthesizes performances of individual CAMS models and the ENSEMBLE to simulate 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, highlighting slightly better scores for the ENSEMBLE with correlation closed to 0.9 and RMSE around 12 µg/m^3.

In-depth analysis of the ENSEMBLE interim 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, 2018. Bias ranges in most parts of Europe between -5 and 5 µg/m^3. Higher bias values (underestimation) can be found in some specific locations in the Southern part of Europe, especially in Italy and in Southeast France. Significant underestimations also occurred in Switzerland and Belgium. However, it should be noted that evaluation cannot be conducted in several Southern countries (Greece, Slovenia, Croatia) and Eastern countries (Romania, Bulgaria), because of a lack of reported interim observation data. It is interesting to note local discrepancies of the quality of the indicator in mountainous areas as well.

Correlation coefficient is excellent with high values, most of them higher than 0.9. The same quality can be seen with RMSE, most of the scores are below 15 µg/m^3 although higher values are found along the Mediterranean area up to 25 µg/m^3 as well as in the Eastern countries.

This can be a consequence of using partial and non-validated observation data, and results should improve when the validated re-analyses are performed. However, results remain acceptable compared to the state of the art.

To help in the interpretation of those maps, one can consider the same performance indicators for each individual model and the ENSEMBLE and various station typologies. 0 and 0 present bias, RMSE, and correlation coefficient scores for all models at rural and suburban stations respectively. The indicators are sorted per geographical region: Western Europe (EUW), Central Europe (EUC), Southern Europe (EUS), Northern Europe (EUN), and Eastern Europe (EUE). The interpretation of the results is hampered by the low (sometimes null) number of stations available for verification in some areas. The number of stations taken in consideration for computing the scores are mentioned on the figures. In Eastern Europe, no station was available for suburban site scores; the verification process has not been performed since very few countries in that area report UTD observation data to the European Environment Agency. The situation is expected to improve in the coming years, as it has improved in this report for Southern countries compared to previous years with the integration of UTD observations from Italy. In Eastern and Northern Europe, the evaluation has been performed against a very low number of stations, which may be a problem regarding the representativeness of the obtained results.

Where observation data is available, the panel of re-analyses shows variabilities in the scores, especially over Western Europe where some slightly underestimate while others slightly overestimate.

Overestimations are more frequent when focusing on suburban sites.

Regarding RMSE, we can note once again good consistency between results even if they are slightly better at rural stations. Better results are obtained for the ENSEMBLE (from 13 to 17 $\mu g/m^3$) at rural and suburban sites.

Correlation coefficient is quite high ranging from 0.8 to 0.9 in best cases among which is the ENSEMBLE. For some re-analyses, significant differences appear with correlations close to 0.5 and far from the best ones.

Obviously, there are more uncertainties in the models in Southern, Eastern and Northern regions due to uncertainties in emissions, and 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 to the weak performance of one or two stations. For this reason, conclusions should be established with care and refined when validated re-analyses for 2018 are available.

Nevertheless, the 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.

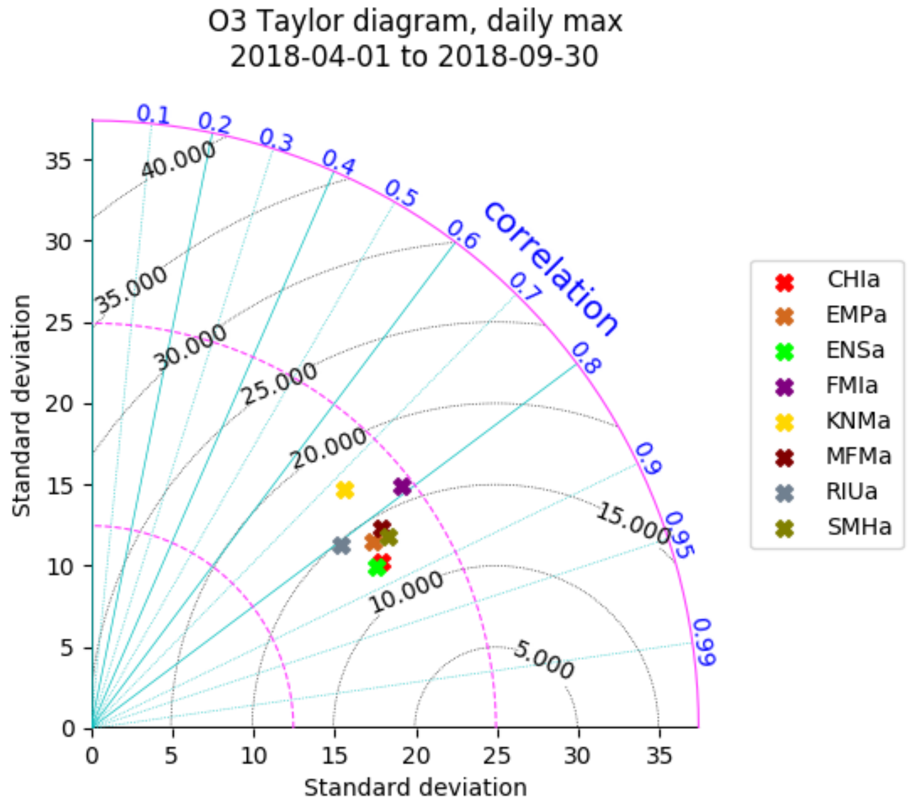
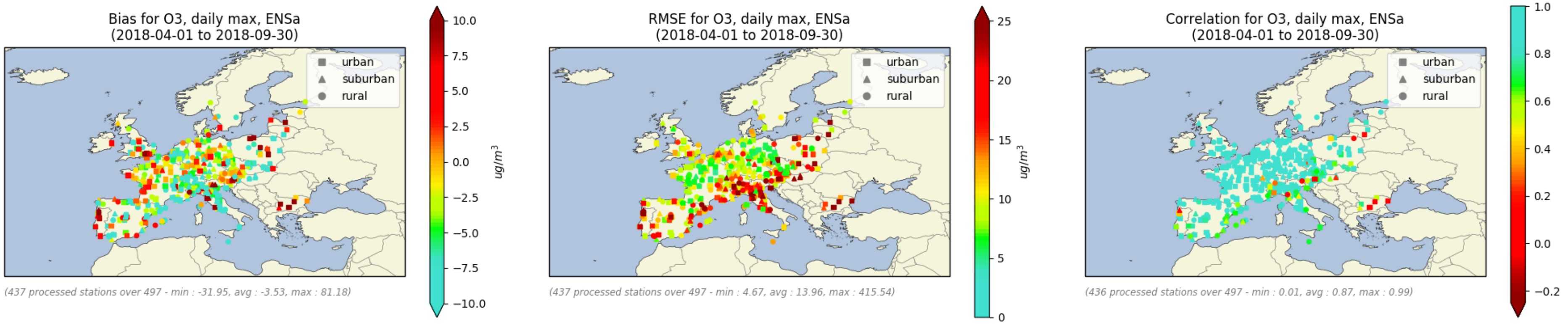


Figure 1 - Taylor diagram presenting performances of all CAMS regional models to simulate summer ozone daily maximum (hourly average).

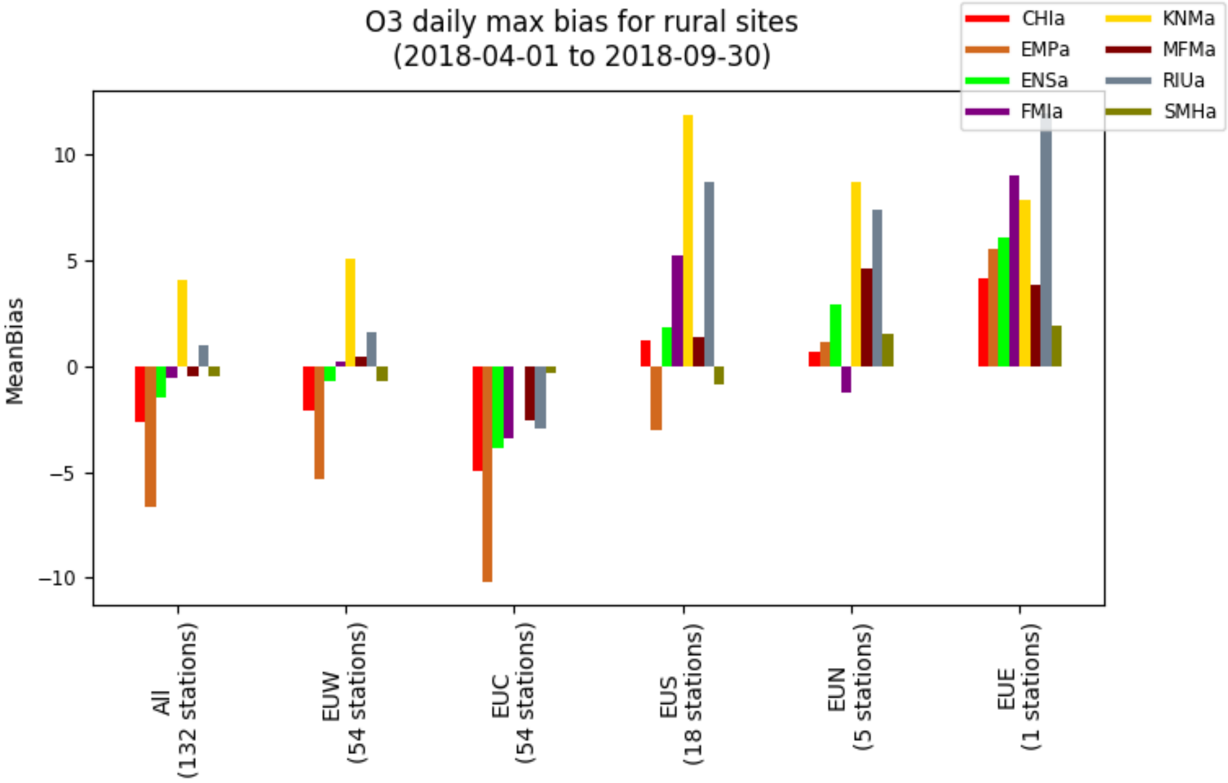

Bias for O3, daily max, ENSa
(2018-04-01 to 2018-09-30)
urban
suburban
rural
(437 processed stations over 497 - min : -31.95, avg : -3.53, max : 81.18)
10.0
7.5
5.0
2.5
0.0
−2.5
−5.0
−7.5
−10.0
ug/m^3
RMSE for O3, daily max, ENSa
(2018-04-01 to 2018-09-30)
urban
suburban
rural
(437 processed stations over 497 - min : 4.67, avg : 13.96, max : 415.54)
25
20
15
10
5
0
ug/m^3
Correlation for O3, daily max, ENSa
(2018-04-01 to 2018-09-30)
urban
suburban
rural
(436 processed stations over 497 - min : 0.01, avg : 0.87, max : 0.99)
1.0
0.8
0.6
0.4
0.2
0.0
−0.2

(a) ozone bias (daily maximum)

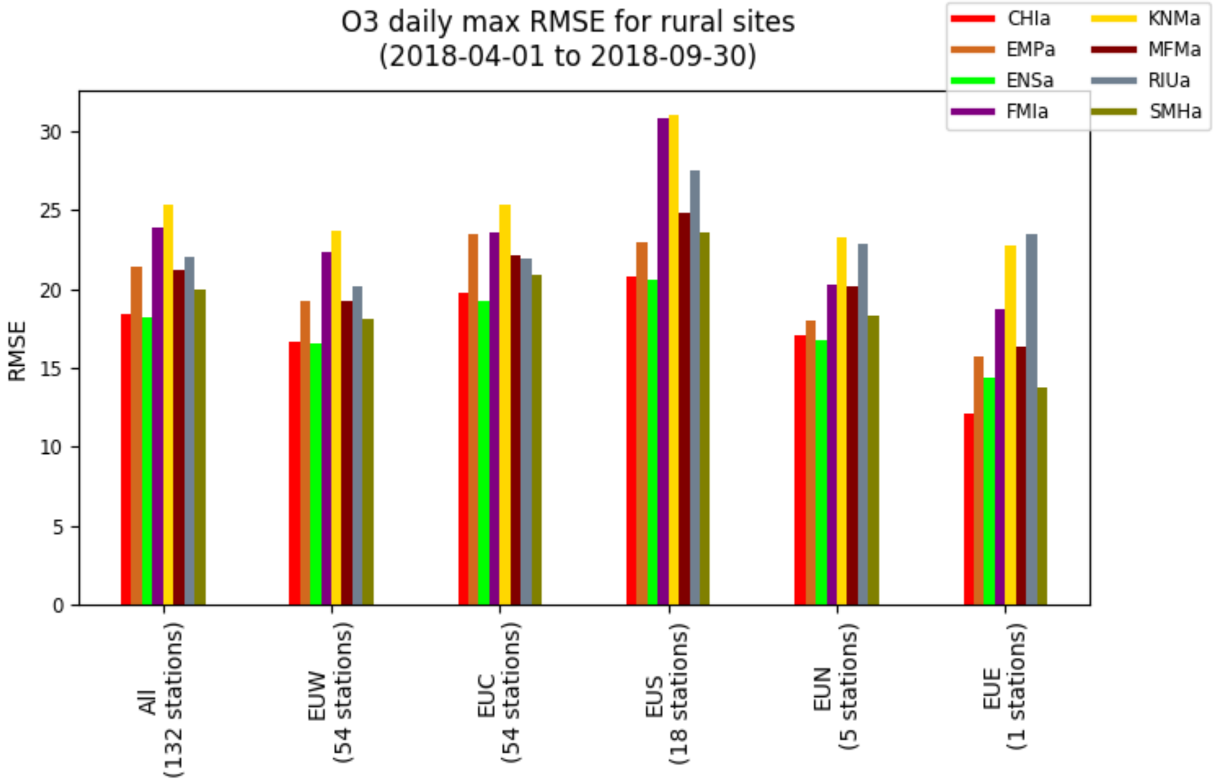
(b) ozone RMSE (daily maximum)

(c) ozone correlation coefficient (daily maximum)

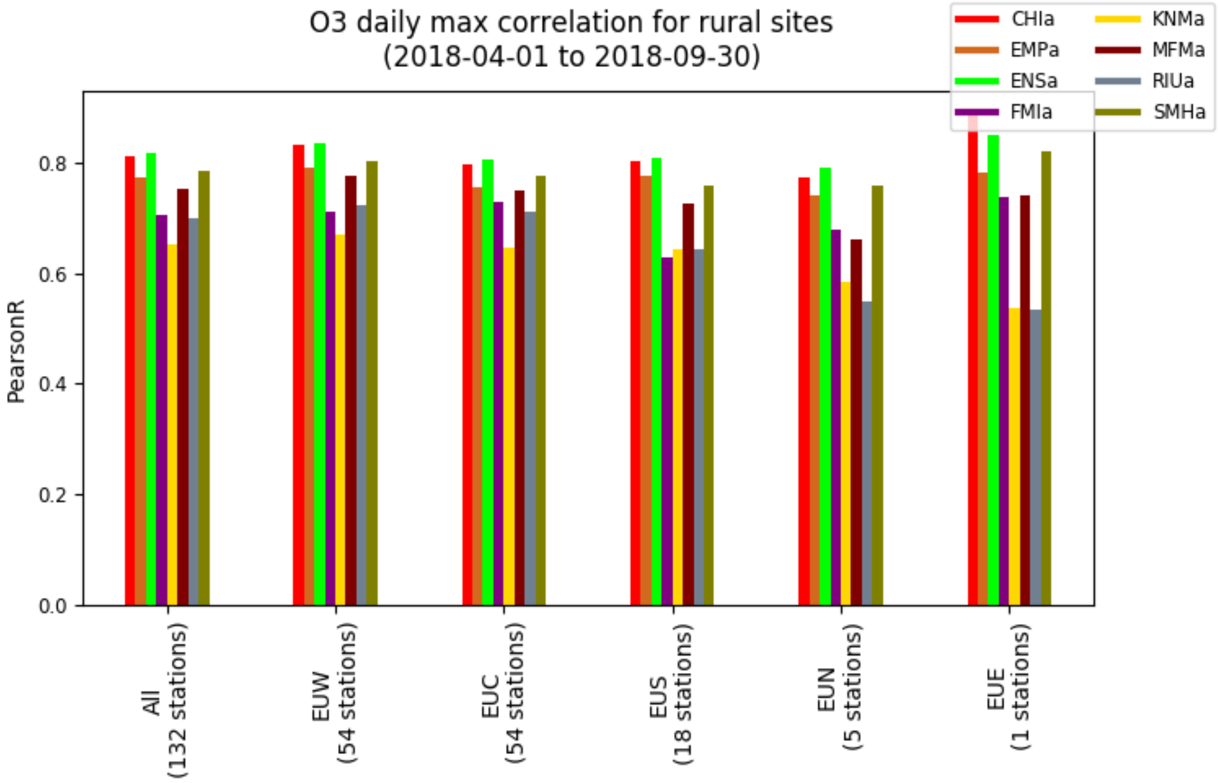
Figure 2 - Maps of Statistical scores of the ENSEMBLE interim re-analyses results against the observation validation dataset from the AQ e-reporting database, for the ozone daily maximum from 01/04/2018 to 30/09/2018: (a) Bias, (b) RMSE, (c) Correlation coefficient.


O3 daily max bias for rural sites
(2018-04-01 to 2018-09-30)
MeanBias
10
5
0
−5
−10
CHIa
EMPa
ENSa
FMIa
KNMa
MFMa
RIUa
SMHa
All (132 stations)
EUW (54 stations)
EUC (54 stations)
EUS (18 stations)
EUN (5 stations)
EUE (1 stations)

(a)

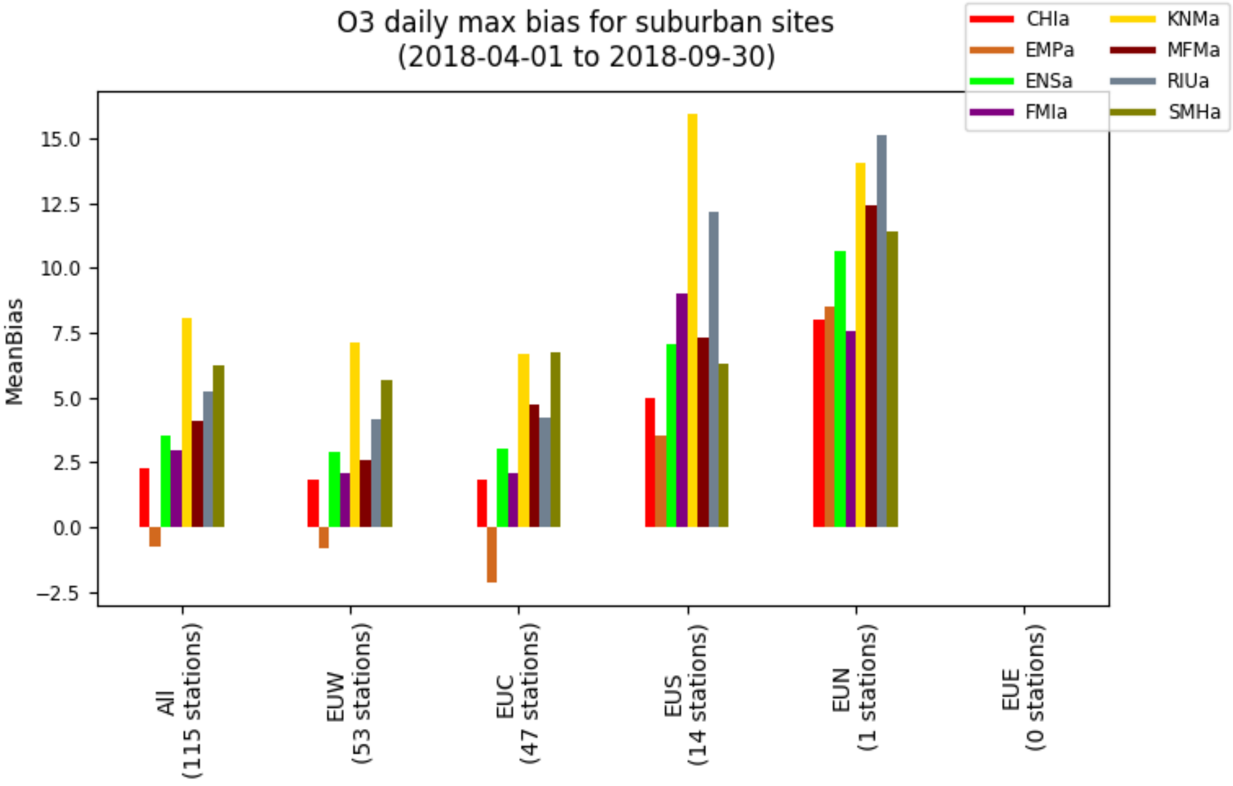

O3 daily max RMSE for rural sites
(2018-04-01 to 2018-09-30)
RMSE
30
25
20
15
10
5
0
CHIa
EMPa
ENSa
FMIa
KNMa
MFMa
RIUa
SMHa
All (132 stations)
EUW (54 stations)
EUC (54 stations)
EUS (18 stations)
EUN (5 stations)
EUE (1 stations)

(b)

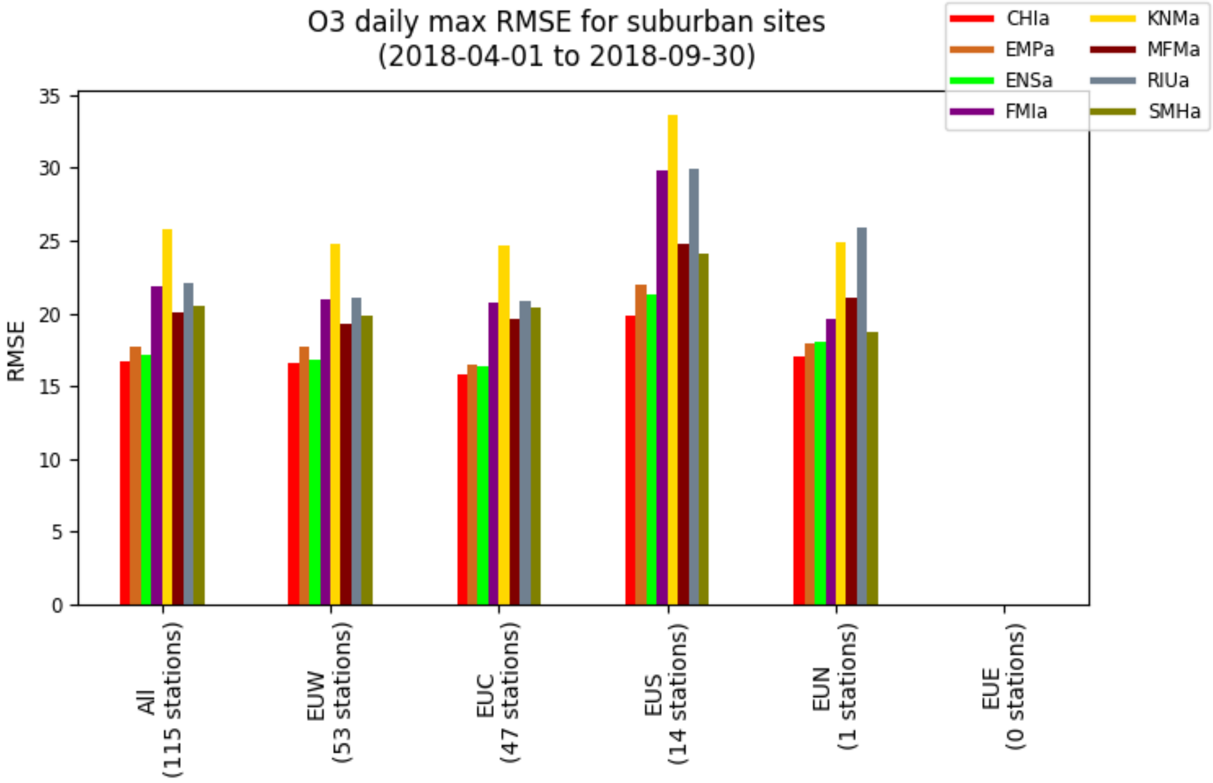


(c)

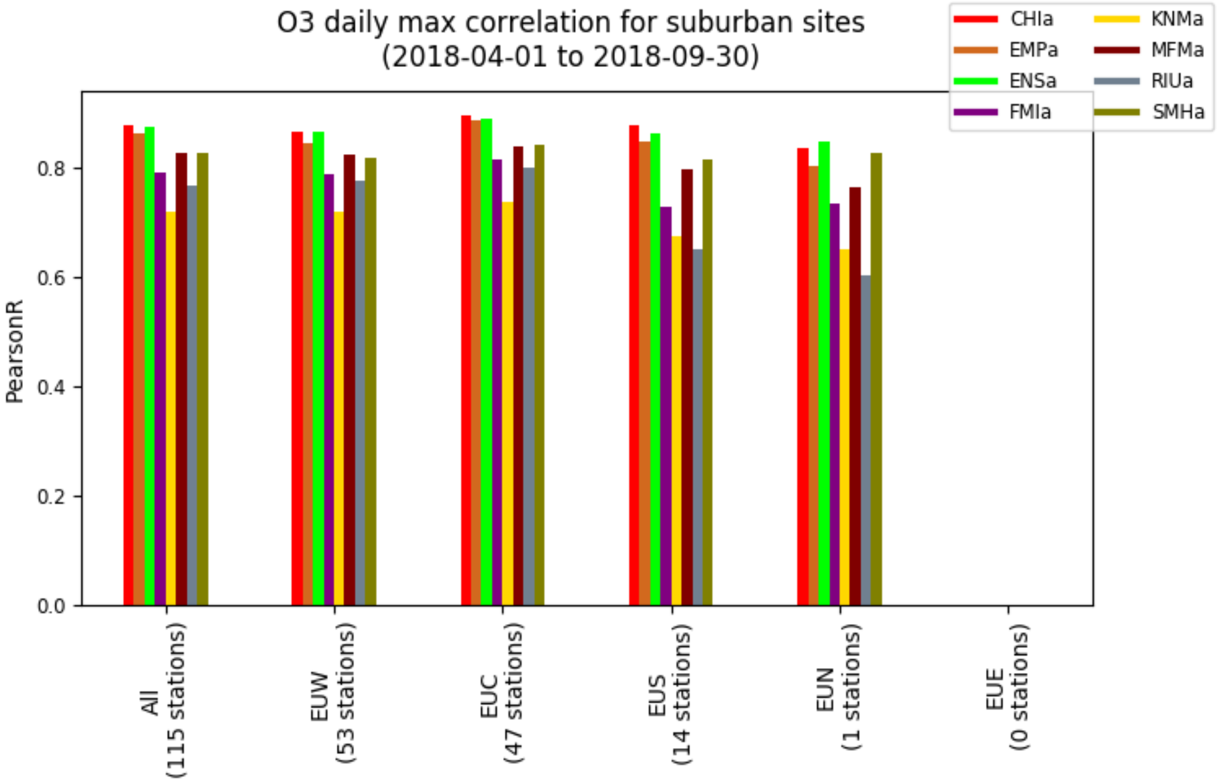
Figure 3 - CAMS Regional interim re-analyses for predicting daily ozone peak over the summer 2018 throughout European sub-regions: (a) Bias, (b) RMSE, (c) Correlation coefficient, at rural stations.


O3 daily max bias for suburban sites
(2018-04-01 to 2018-09-30)
CHIa
EMPa
ENSa
FMIa
KNMa
MFMa
RIUa
SMHa
MeanBias
15.0
12.5
10.0
7.5
5.0
2.5
0.0
−2.5
All
(115 stations)
EUW
(53 stations)
EUC
(47 stations)
EUS
(14 stations)
EUN
(1 station)
EUE
(0 stations)

(a)


O3 daily max RMSE for suburban sites
(2018-04-01 to 2018-09-30)
CHIa
EMPa
ENSa
FMIa
KNMa
MFMa
RIUa
SMHa
RMSE
35
30
25
20
15
10
5
0
All
(115 stations)
EUW
(53 stations)
EUC
(47 stations)
EUS
(14 stations)
EUN
(1 station)
EUE
(0 stations)

(b)



(c)

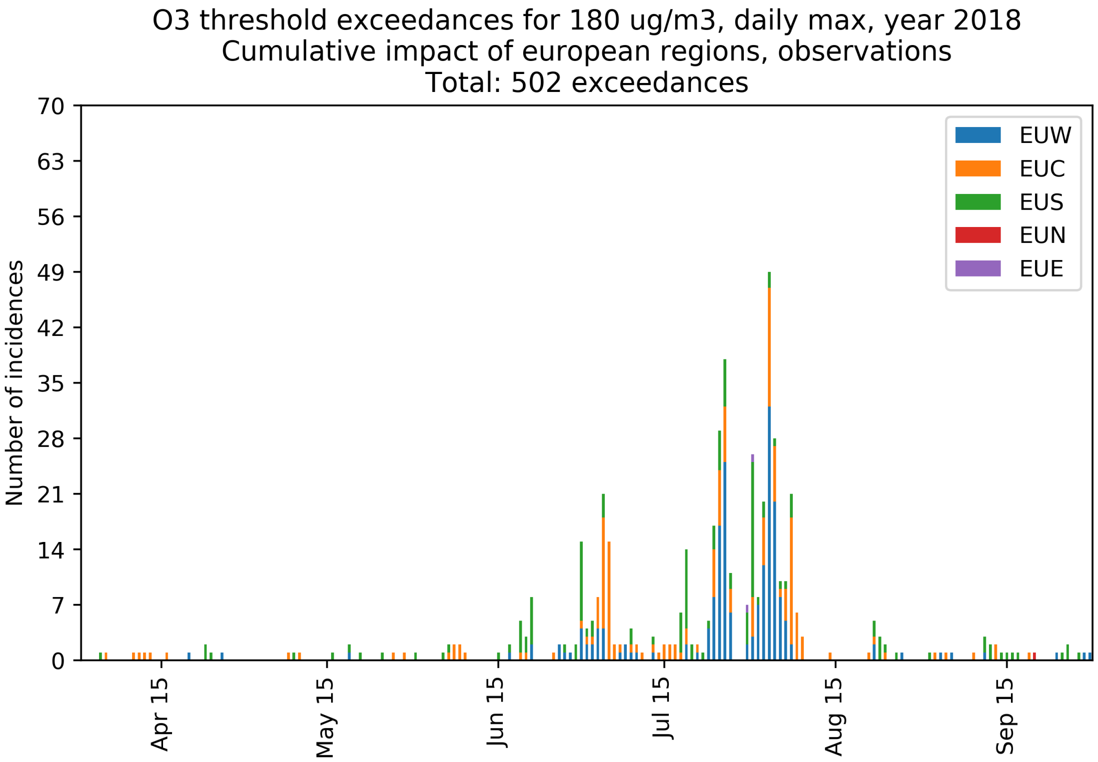
Figure 4 - CAMS Regional interim re-analyses for predicting daily ozone peak over summer 2018 throughout European sub-regions: (a) Bias, (b) RMSE, (c) Correlation coefficient, at suburban stations.

Finally, the models' ability 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 (180µg/m^3) and an alert threshold (240 µg/m^3), over which short-term action plans and communication towards the general public should be implemented by Member States. However, this kind of evaluation against threshold value is very stringent and not always representative of the model quality. Situations above and below the threshold value are counted, but to correctly take into account 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 diagnosis can be seen as a pessimistic analysis of the models' performances.

0 below shows the number of situations when exceedances of the hourly information threshold has been observed during the summer time in 2018 (time is presented on the x-axis), sorted per geographical region (various colors). The first set of histograms shows observed exceedances at ozone stations in Europe. Most of exceedances were located in Central and Western Europe and a few of them in Southern Europe. In total, 502 exceedances were recorded, more than twice the 2017 number. The variability between geographical areas should be interpreted with caution, as it is highly dependent of the number of stations available.

The main episode occurred at the beginning of August 2018 and lasted around one week. Before that, two other episodes were recorded, one at the end of June and another one before the main one at the end of July. The figure in the middle panel displays exceedances modelled by all CAMS models and the ENSEMBLE. If the performances of the ENSEMBLE were good considering statistical indicators, they show very disappointing results for threshold indicators. Less than 20 % of the exceedances were detected by the ENSEMBLE (Figure 6). 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 that detect more exceedances have much more false alarms, therefore also highlighting a positive aspect of the conservative choice of the median.



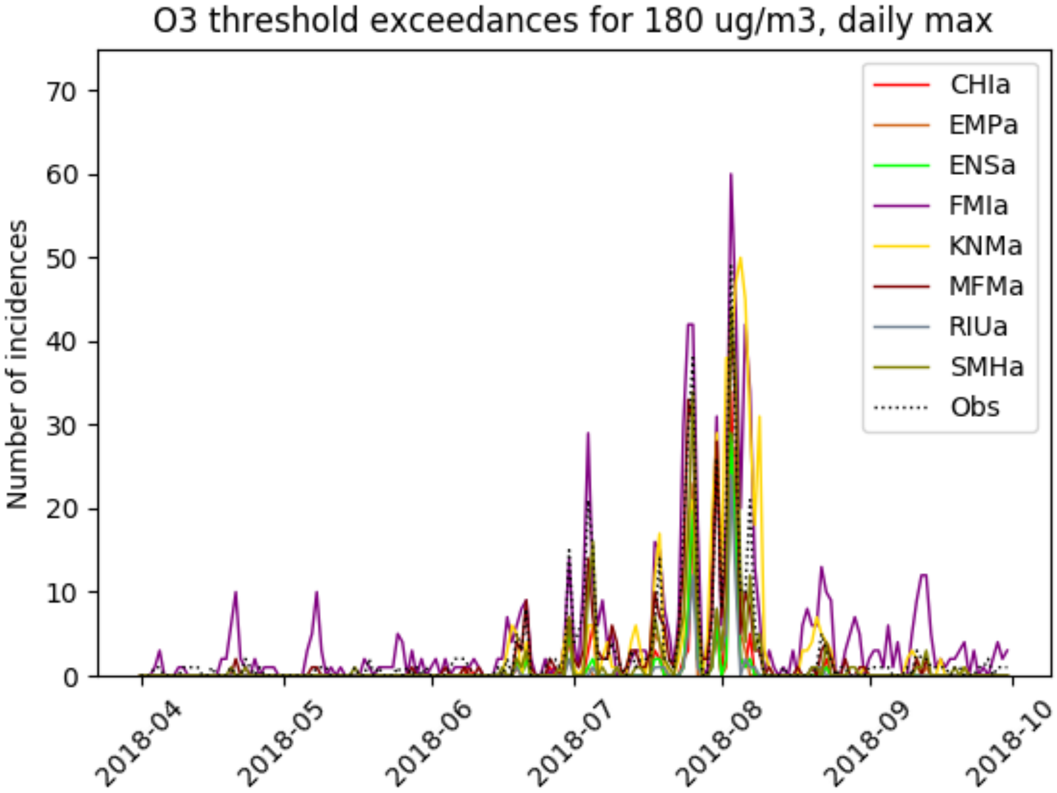


Figure 5 - Number of exceedances of the information threshold value for ozone in summer 2018 - observed (top), modelled by all the interim reanalyses in color lines and observed in black dashed line (bottom).

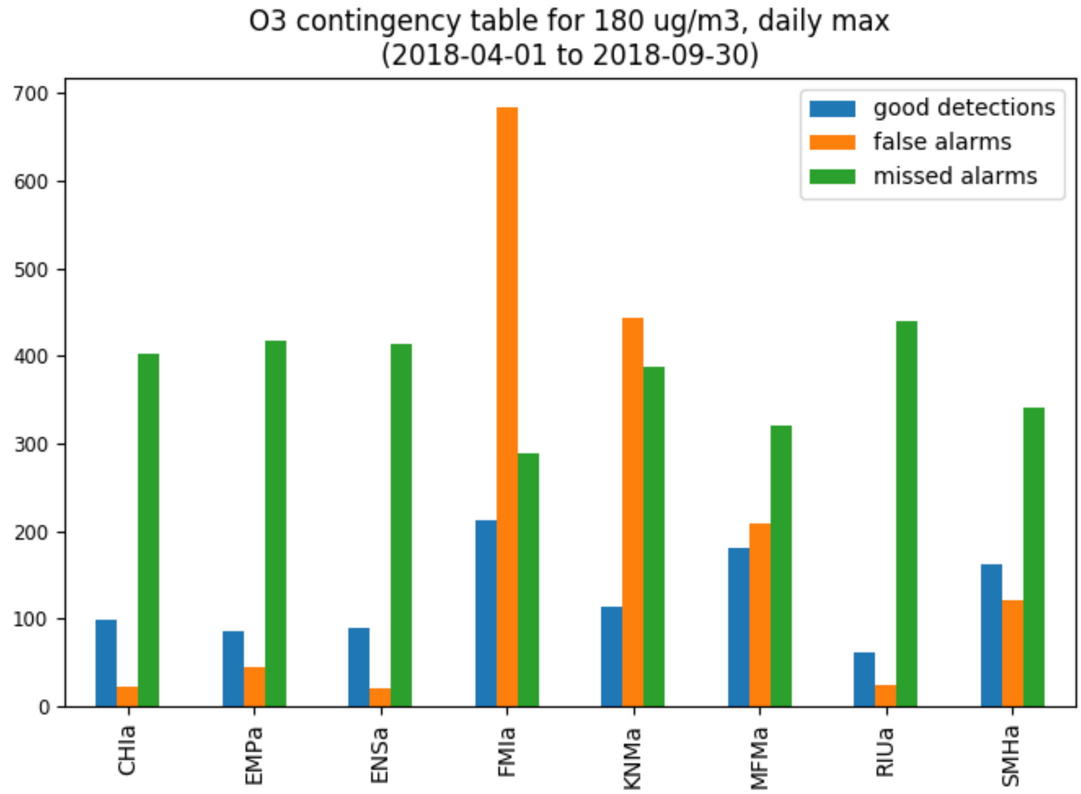


Figure 6 - Histograms describing the models performances regarding the number of exceedances of the ozone thresholds.

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 ENSEMBLE CAMS Regional interim re-analyses and its members, for the daily maximum (hourly average) of NO_2 concentrations. It shows disperse model performances, among them the ENSEMBLE has one of the best ones with correlation close to 0.7 and RMSE around 12 µg/m^3. The worst performances depicted on this diagram are a correlation of 0.4 and RMSE around 15 µg/m^3. It is worth noting that such scores are close to the scores obtained in 2017 (Figure 8).

Maps in Figure 9 allow highlighting a tendency to underestimate NO_2 daily maximum throughout Europe, even if some isolated stations show overestimations. The RMSE in 2018 is slightly better than in 2017 (notice that the scores for 2017 have been reviewed to remove outliers which had affected scores shown in the IRA2017 report), but the improvement is so limited that we can only conclude there is almost no change in terms of NO_2 performances. Opposite interpretation is done for the correlation with several stations indicating a decrease of values in 2018 compared to 2017, but the scores remain at a satisfactory level.

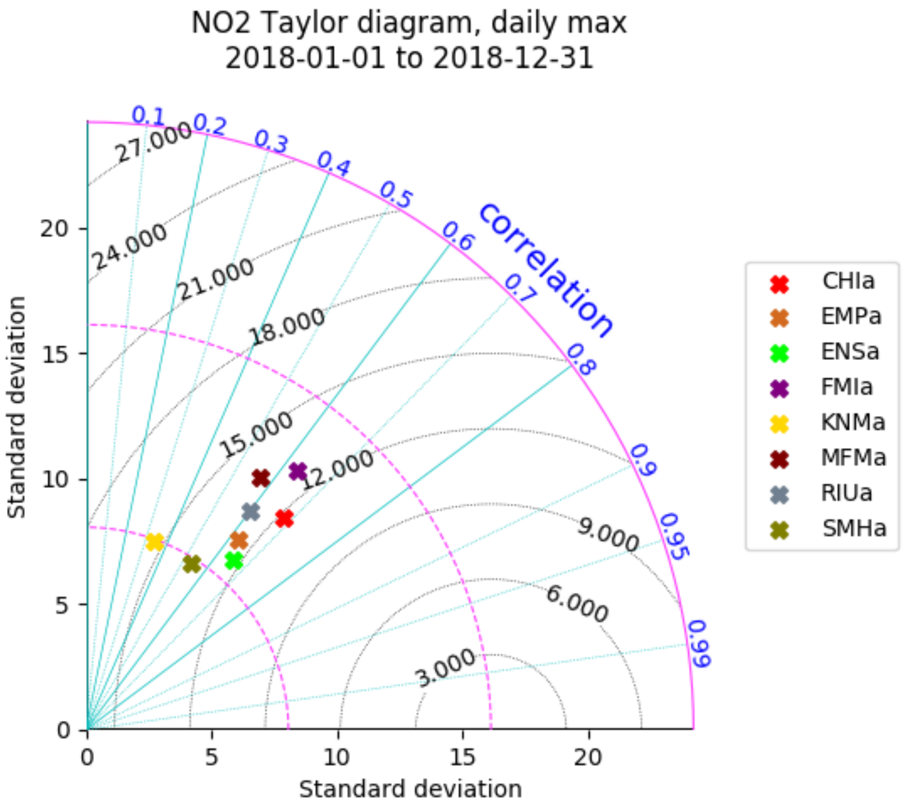


Figure 7 - Taylor diagram presenting the performances of the CAMS Regional interim re-analyses to predict NO_2 daily maxima in 2018.

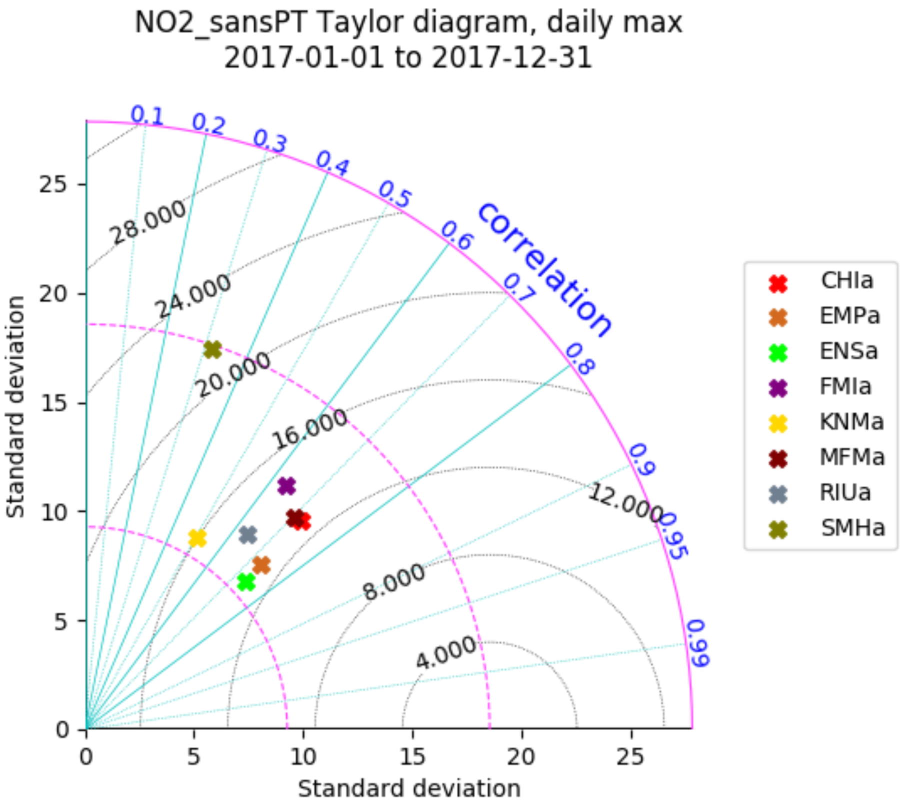


Figure 8 - Taylor diagram presenting the performances of the CAMS Regional interim re-analyses to predict NO_2 daily maxima in 2017.

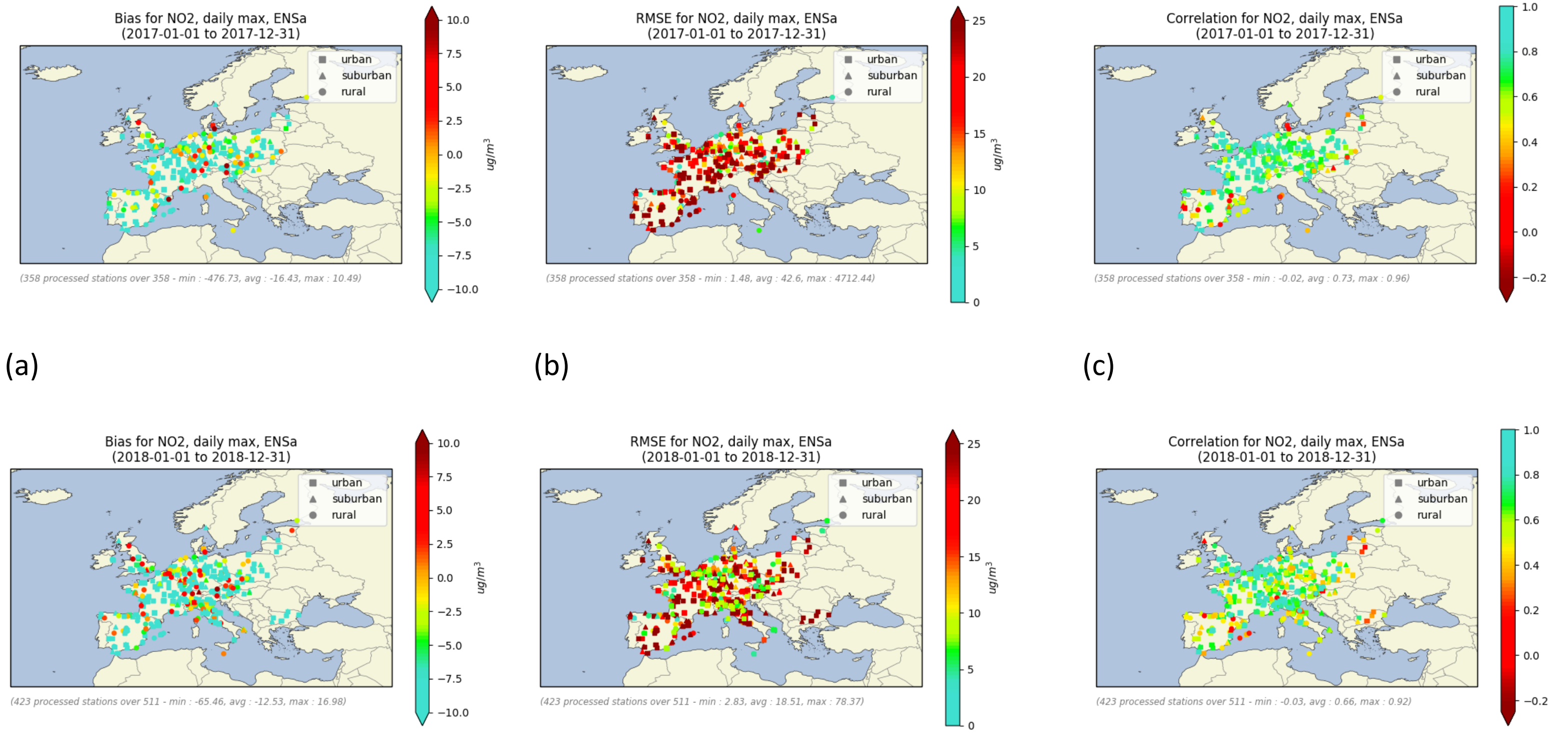

Bias for NO2, daily max, ENSa
(2017-01-01 to 2017-12-31)
urban
suburban
rural
(358 processed stations over 358 - min : -476.73, avg : -16.43, max : 10.49)
10.0
7.5
5.0
2.5
0.0
−2.5
−5.0
−7.5
−10.0
ug/m3
RMSE for NO2, daily max, ENSa
(2017-01-01 to 2017-12-31)
urban
suburban
rural
(358 processed stations over 358 - min : 1.48, avg : 42.6, max : 4712.44)
25
20
15
10
5
0
ug/m3
Correlation for NO2, daily max, ENSa
(2017-01-01 to 2017-12-31)
urban
suburban
rural
(358 processed stations over 358 - min : -0.02, avg : 0.73, max : 0.96)
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
(a)
(b)
(c)
Bias for NO2, daily max, ENSa
(2018-01-01 to 2018-12-31)
urban
suburban
rural
(423 processed stations over 511 - min : -65.46, avg : -12.53, max : 16.98)
10.0
7.5
5.0
2.5
0.0
−2.5
−5.0
−7.5
−10.0
ug/m3
RMSE for NO2, daily max, ENSa
(2018-01-01 to 2018-12-31)
urban
suburban
rural
(423 processed stations over 511 - min : 2.83, avg : 18.51, max : 78.37)
25
20
15
10
5
0
ug/m3
Correlation for NO2, daily max, ENSa
(2018-01-01 to 2018-12-31)
urban
suburban
rural
(423 processed stations over 511 - min : -0.03, avg : 0.66, max : 0.92)
1.0
0.8
0.6
0.4
0.2
0.0
−0.2

Figure 9 - Maps of Statistical scores of the ENSEMBLE interim re-analyses results against the observation validation dataset from the AQ e-reporting database for the NO_2 daily maximum, over the year 2017 (top) and the year 2018 (bottom): (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 2018, for CAMS Regional individual and ENSEMBLE re-analyses. The results are very encouraging with the ENSEMBLE correlation coefficient ranging from 0.6 to 0.7 for the best models, among which there is the ENSEMBLE. Two re-analyses performances lie out of the main group, with correlation around 0.4 and RMSE close to 15. The ENSEMBLE RMSE is around 12 µg/m^3, which is better than what was obtained for the 2017 re-analyses (15 µg/m^3), but less good than the validated re-analyses usual performances (RMSE lower than 10 µg/m^3).

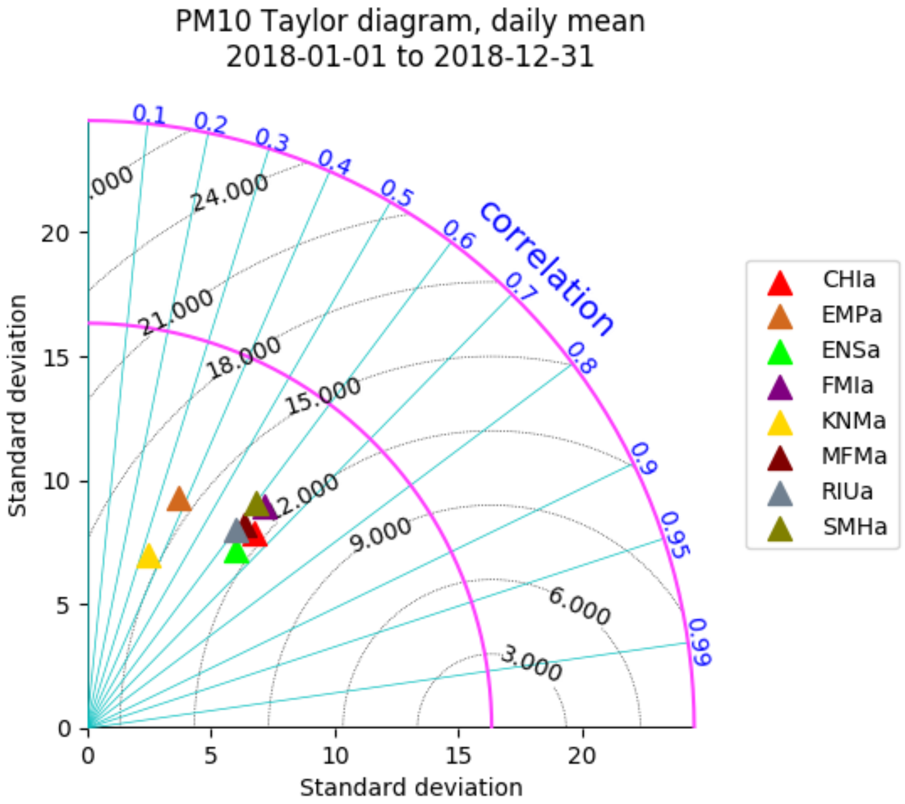


Figure 10 - Taylor diagram presenting the performances of the CAMS Regional ENSEMBLE interim re-analyses to predict PM_{10} daily average in 2018.

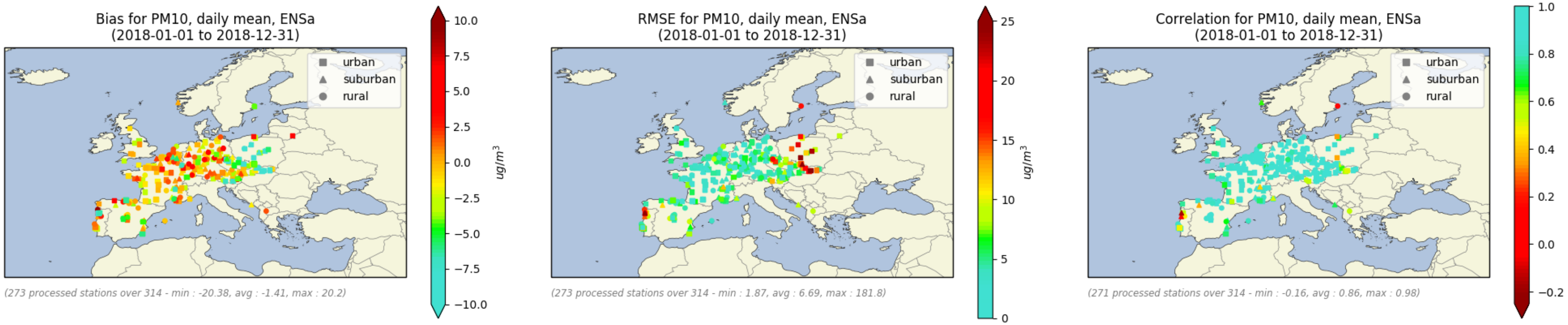
0 details the geographical distribution of statistical scores (bias, correlation coefficient and RMSE), for the ENSEMBLE interim re-analyses for the year 2018. The lowest scores (bias, correlation and RMSE) are obtained for stations located in Portugal and in the Central-eastern parts of Europe. In several countries (France, Germany, Benelux and the UK), RMSE ranges between 1 and 5 µg/m^3, which is very encouraging. A positive bias is frequent over European stations in 2018 for the ENSEMBLE, which is different compared to the score of the previous years and the general behavior of CTMs that usually shows an underestimation. This issue may be related to the lack of update in emissions since 2011, and needs to be followed up when assessing the performances of IRA after the update of CAMS-REG-

AP_v2.2.1 emissions. The bias geographical variabilities is significant whereas RMSE and correlation have quite good values in most of the countries, with low geographical variabilities.

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} interim re-analyses, with huge gaps in some areas (Southern, Northern and Eastern Europe for all station typologies). Once again, results are only robust over Western and Central Europe. For these two areas, the ENSEMBLE performances are quite similar, overestimating the PM_{10} concentrations over rural stations and slightly underestimating them over suburban and urban stations.

RMSEs are better for rural sites and less good over suburban sites.

Despite the variabilities of the model responses, especially with two re-analyses well out of the group in terms of performances even if their scores remain acceptable, the ENSEMBLE is most of the time the re-analysis providing the best description of the PM_{10} distribution over Europe.



Bias for PM10, daily mean, ENSa (2018-01-01 to 2018-12-31)

(273 processed stations over 314 - min : -20.38, avg : -1.41, max : 20.2)

RMSE for PM10, daily mean, ENSa (2018-01-01 to 2018-12-31)

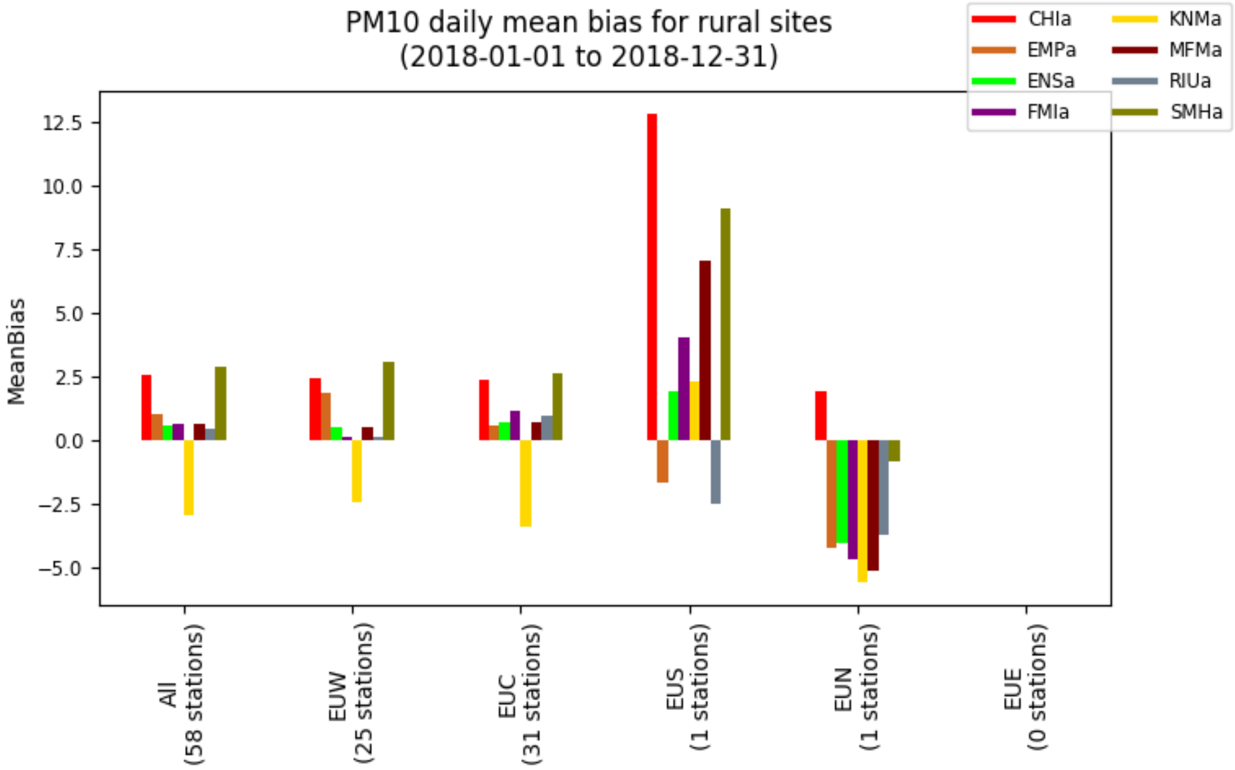
(273 processed stations over 314 - min : 1.87, avg : 6.69, max : 181.8)

Correlation for PM10, daily mean, ENSa (2018-01-01 to 2018-12-31)

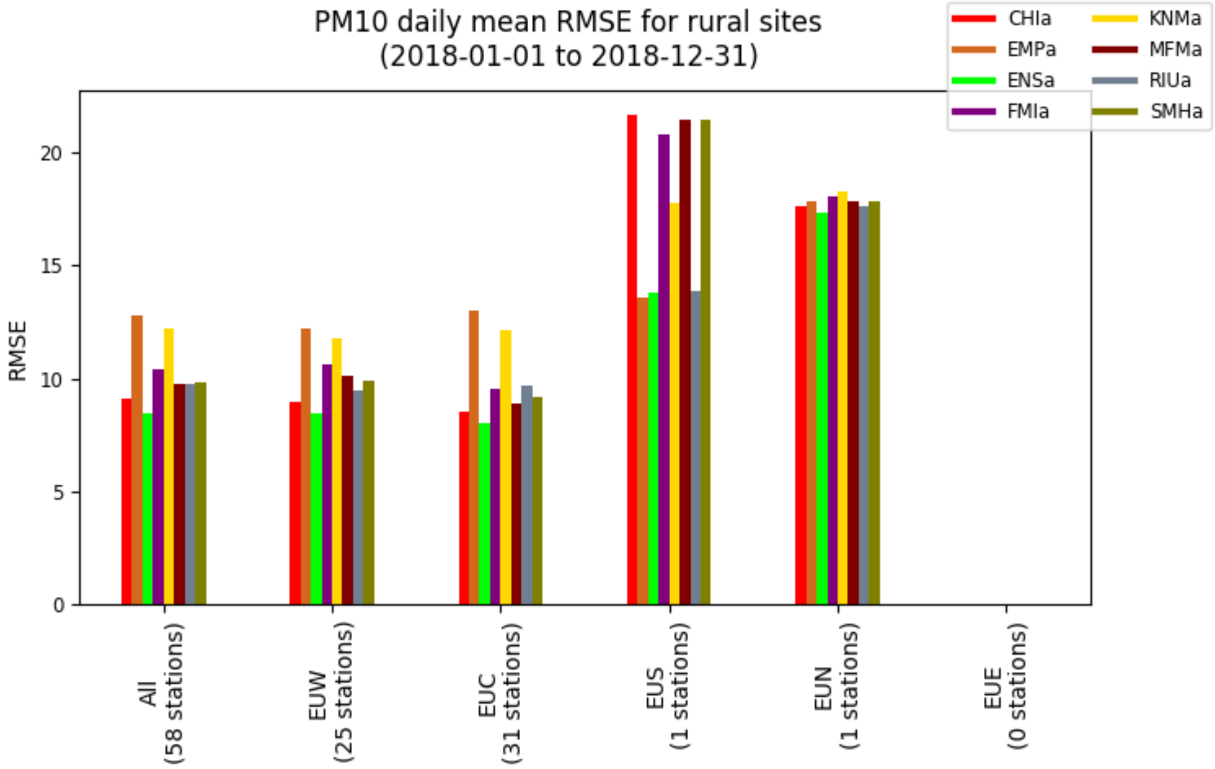
(271 processed stations over 314 - min : -0.16, avg : 0.86, max : 0.98)

(a) (b) (c)

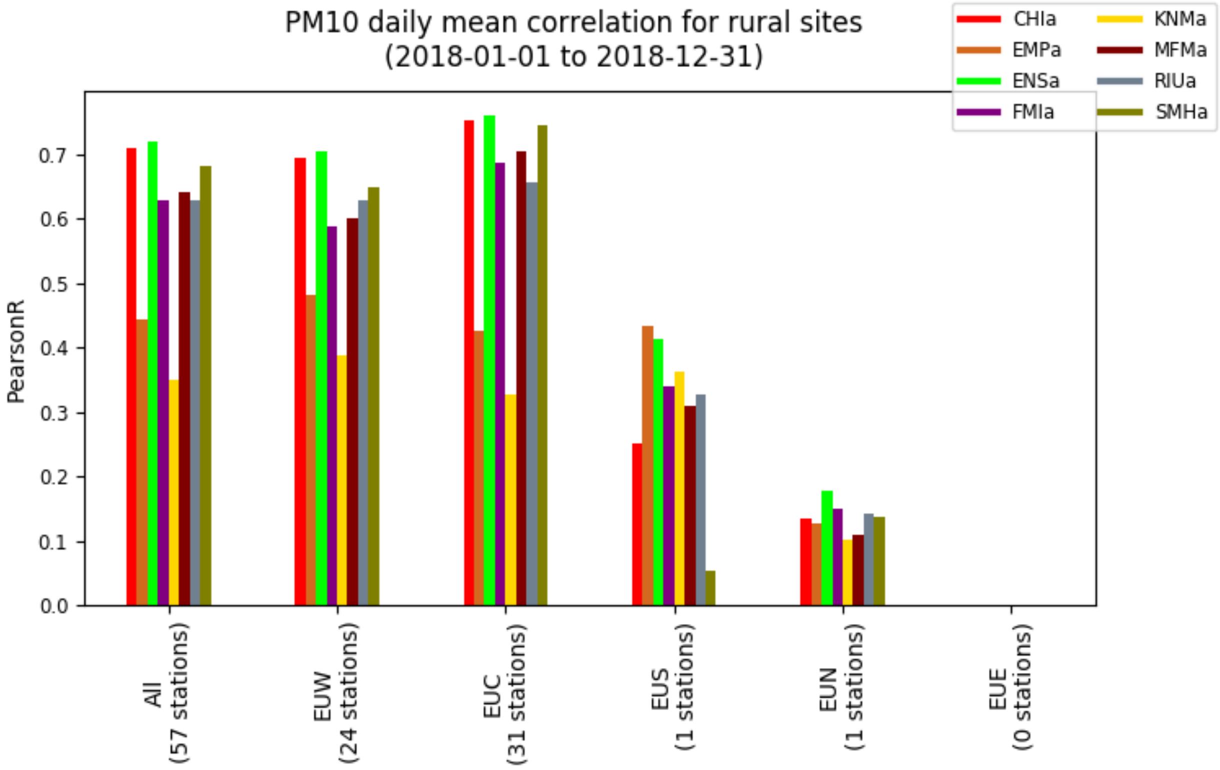
Figure 11 - Maps of Statistical scores of the ENSEMBLE interim re-analyses results against the observation validation dataset from the AQ e-reporting database for the PM_{10} daily average, over the year 2018: (a) Bias, (b) RMSE, (c) Correlation coefficient.


PM10 daily mean bias for rural sites
(2018-01-01 to 2018-12-31)
CHIa
EMPa
ENSa
FMIa
KNMa
MFMa
RIUa
SMHa
MeanBias
12.5
10.0
7.5
5.0
2.5
0.0
−2.5
−5.0
All
(58 stations)
EUW
(25 stations)
EUC
(31 stations)
EUS
(1 stations)
EUN
(1 stations)
EUE
(0 stations)

(a)

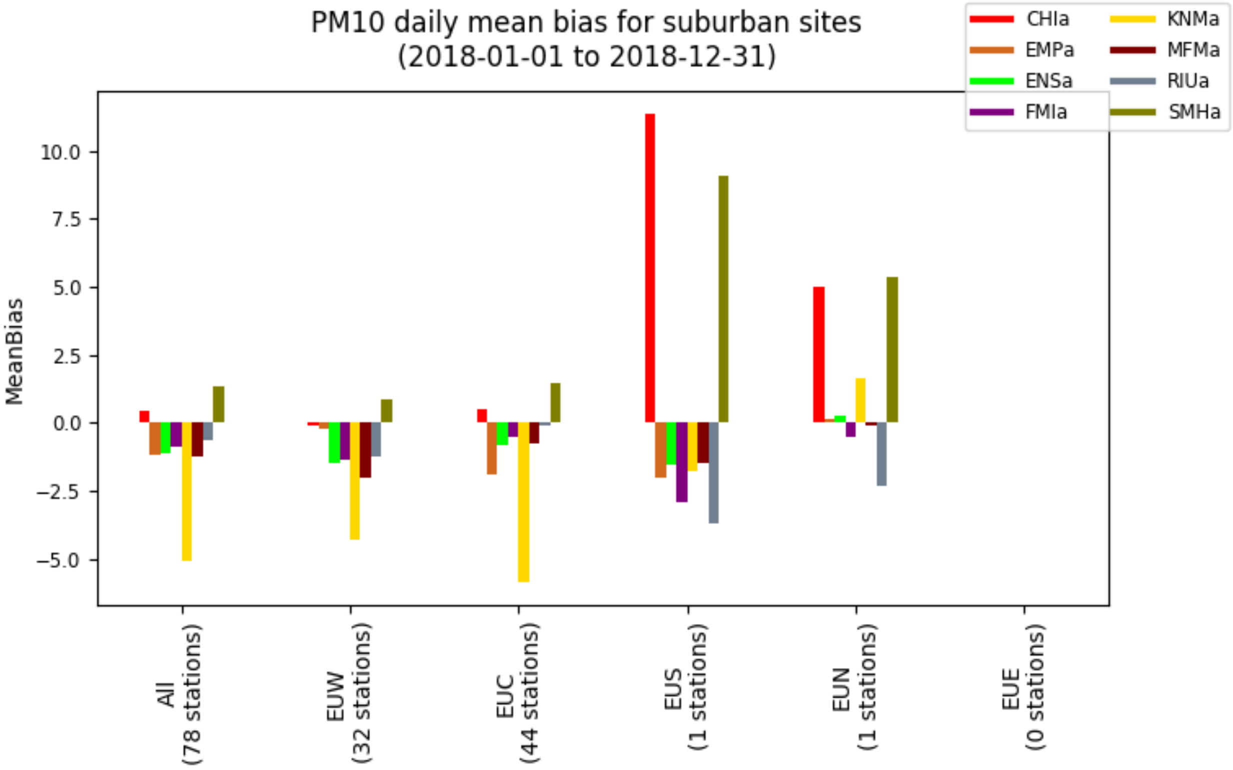

PM10 daily mean RMSE for rural sites
(2018-01-01 to 2018-12-31)
CHIa
EMPa
ENSa
FMIa
KNMa
MFMa
RIUa
SMHa
RMSE
20
15
10
5
0
All
(58 stations)
EUW
(25 stations)
EUC
(31 stations)
EUS
(1 stations)
EUN
(1 stations)
EUE
(0 stations)

(b)

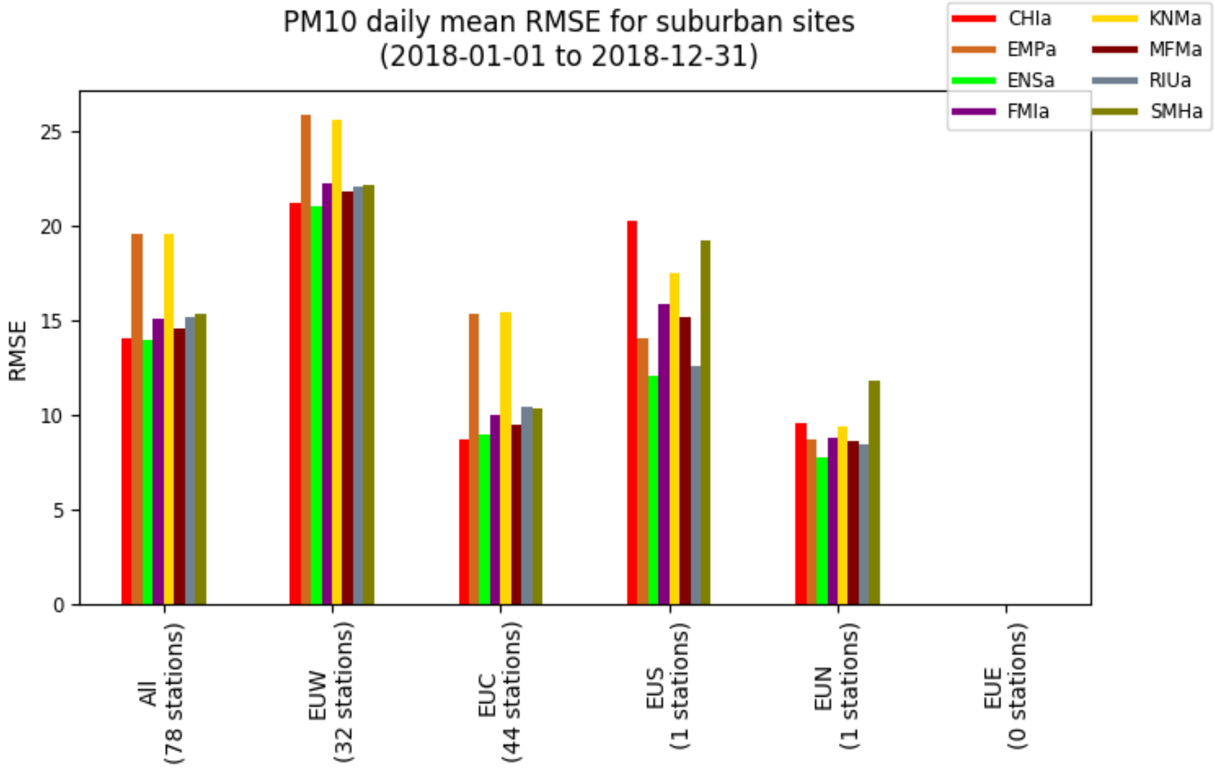


(c)

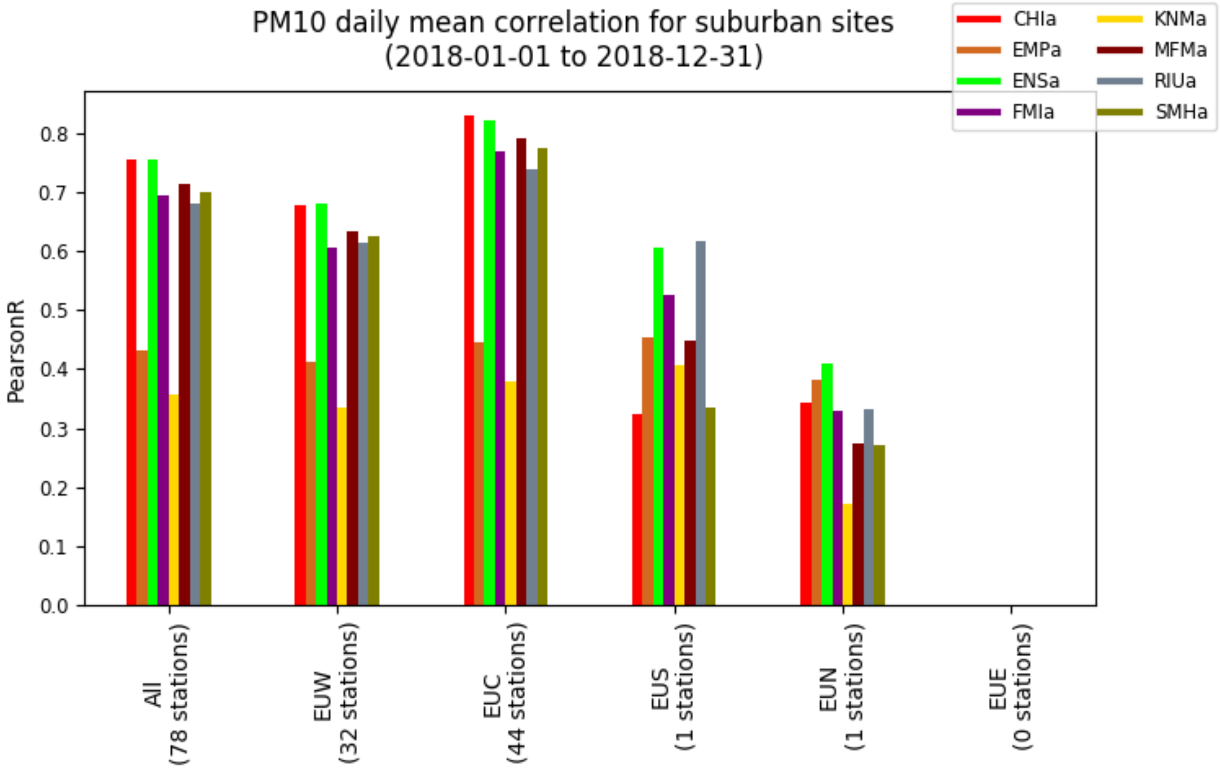
Figure 12 - CAMS Regional interim re-analyses for predicting PM_{10} daily average over the year 2018 throughout European sub-regions: (a) Bias, (b) RMSE, (c) Correlation coefficient, at rural stations.


PM10 daily mean bias for suburban sites
(2018-01-01 to 2018-12-31)
MeanBias
10.0
7.5
5.0
2.5
0.0
−2.5
−5.0
All (78 stations)
EUW (32 stations)
EUC (44 stations)
EUS (1 stations)
EUN (1 stations)
EUE (0 stations)
CHIa
EMPa
ENSa
FMIa
KNMa
MFMa
RIUa
SMHa

(a)

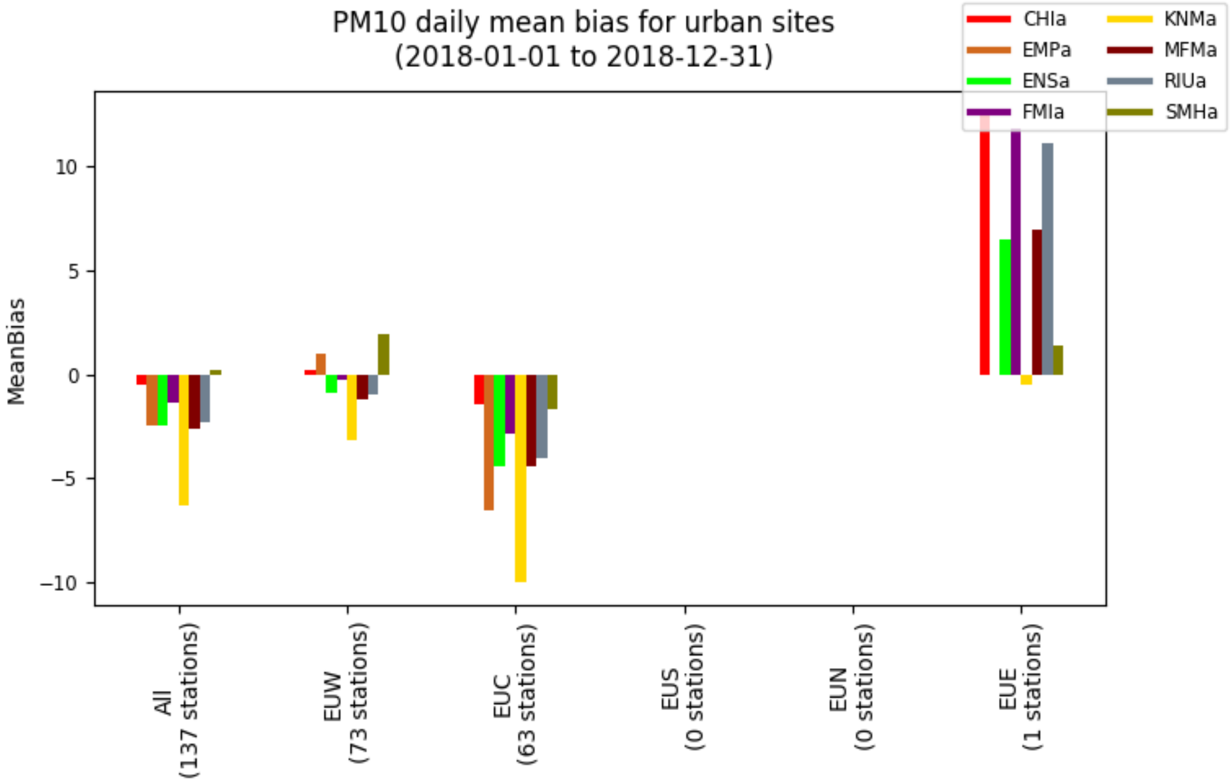

PM10 daily mean RMSE for suburban sites
(2018-01-01 to 2018-12-31)
RMSE
25
20
15
10
5
0
All (78 stations)
EUW (32 stations)
EUC (44 stations)
EUS (1 stations)
EUN (1 stations)
EUE (0 stations)
CHIa
EMPa
ENSa
FMIa
KNMa
MFMa
RIUa
SMHa

(b)

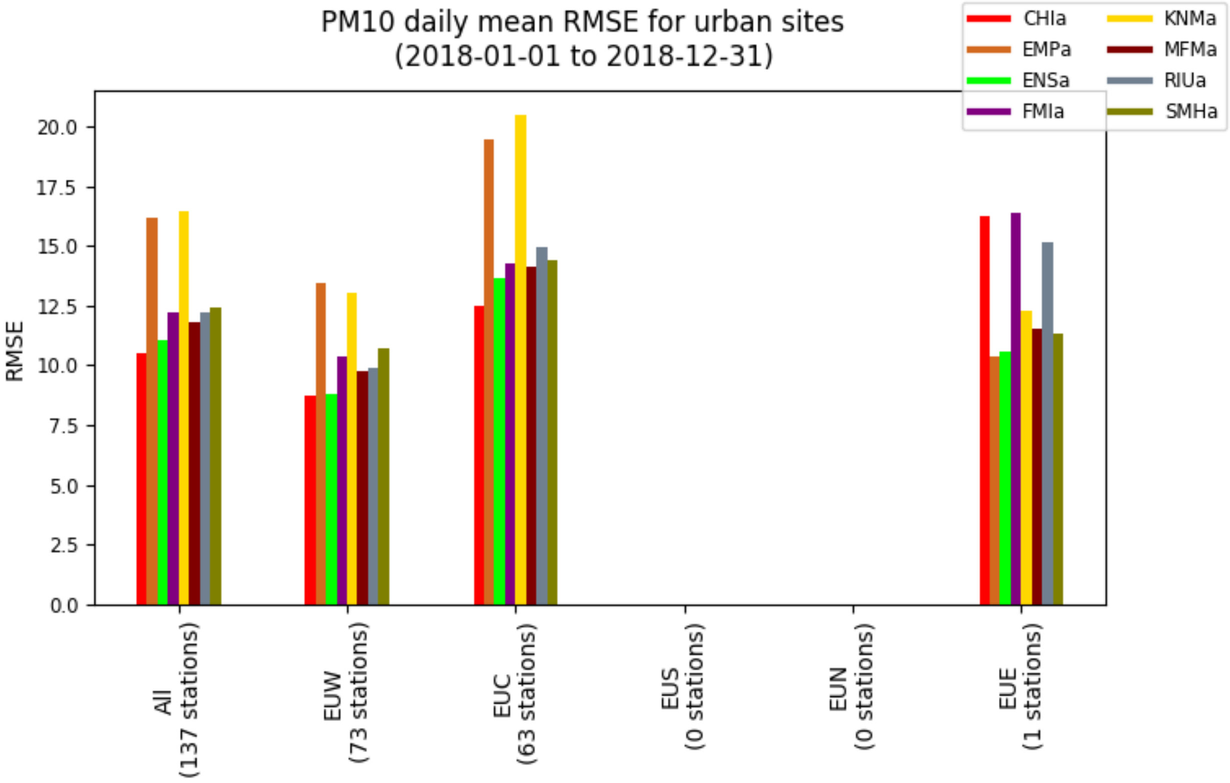


(c)

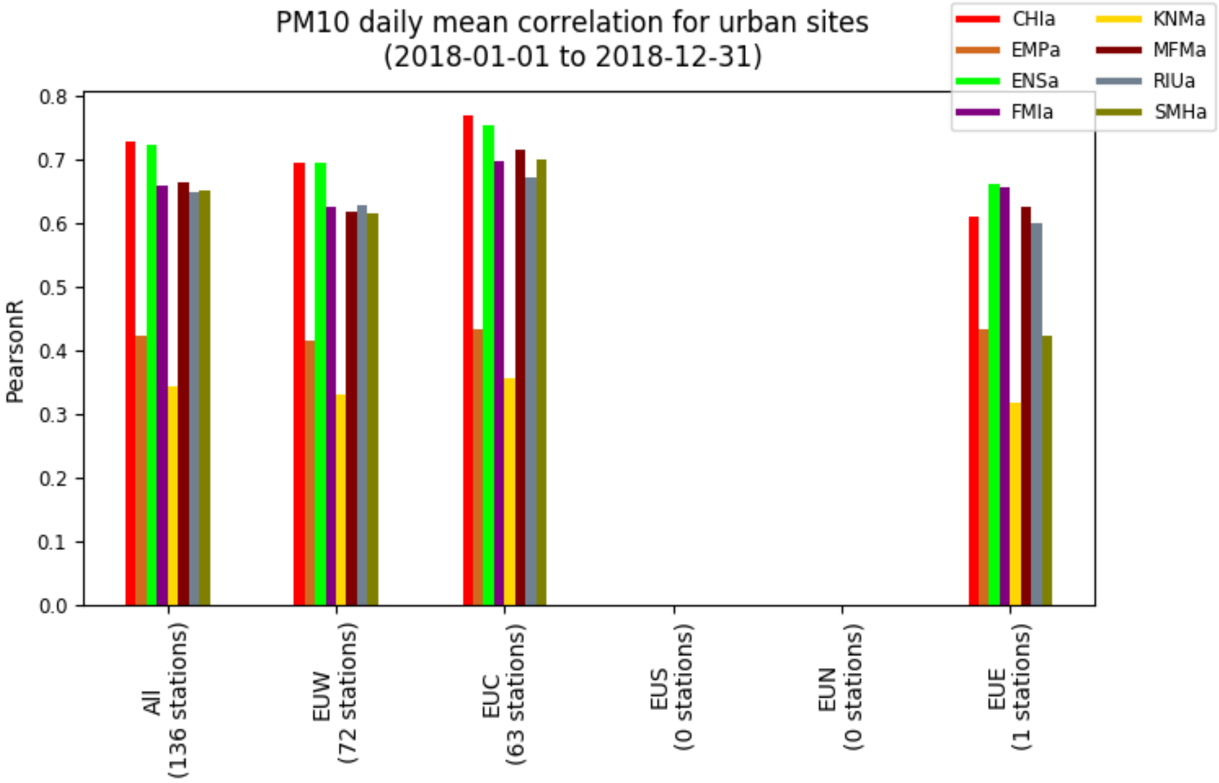
Figure 13 - CAMS Regional interim re-analyses for predicting PM_{10} daily average over the year 2018 throughout European sub-regions: (a) Bias, (b) RMSE, (c) Correlation coefficient, at suburban stations.


PM10 daily mean bias for urban sites
(2018-01-01 to 2018-12-31)
CHIa
EMPa
ENSa
FMIa
KNMa
MFMa
RIUa
SMHa
MeanBias
10
5
0
−5
−10
All
(137 stations)
EUW
(73 stations)
EUC
(63 stations)
EUS
(0 stations)
EUN
(0 stations)
EUE
(1 stations)

(a)

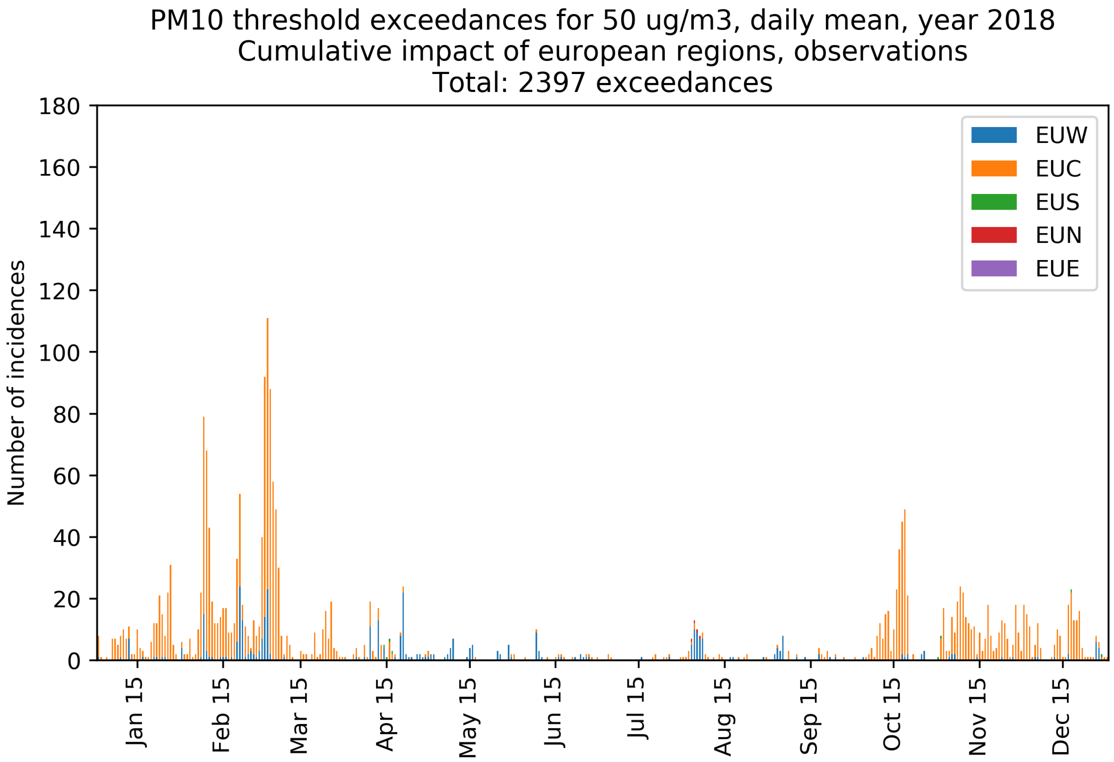

PM10 daily mean RMSE for urban sites
(2018-01-01 to 2018-12-31)
CHIa
EMPa
ENSa
FMIa
KNMa
MFMa
RIUa
SMHa
RMSE
20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
All
(137 stations)
EUW
(73 stations)
EUC
(63 stations)
EUS
(0 stations)
EUN
(0 stations)
EUE
(1 stations)

(b)



(c)

Figure 14 - CAMS Regional interim re-analyses for predicting PM_{10} daily average over the year 2018 throughout European sub-regions: (a) Bias, (b) RMSE, (c) Correlation coefficient, at urban stations.



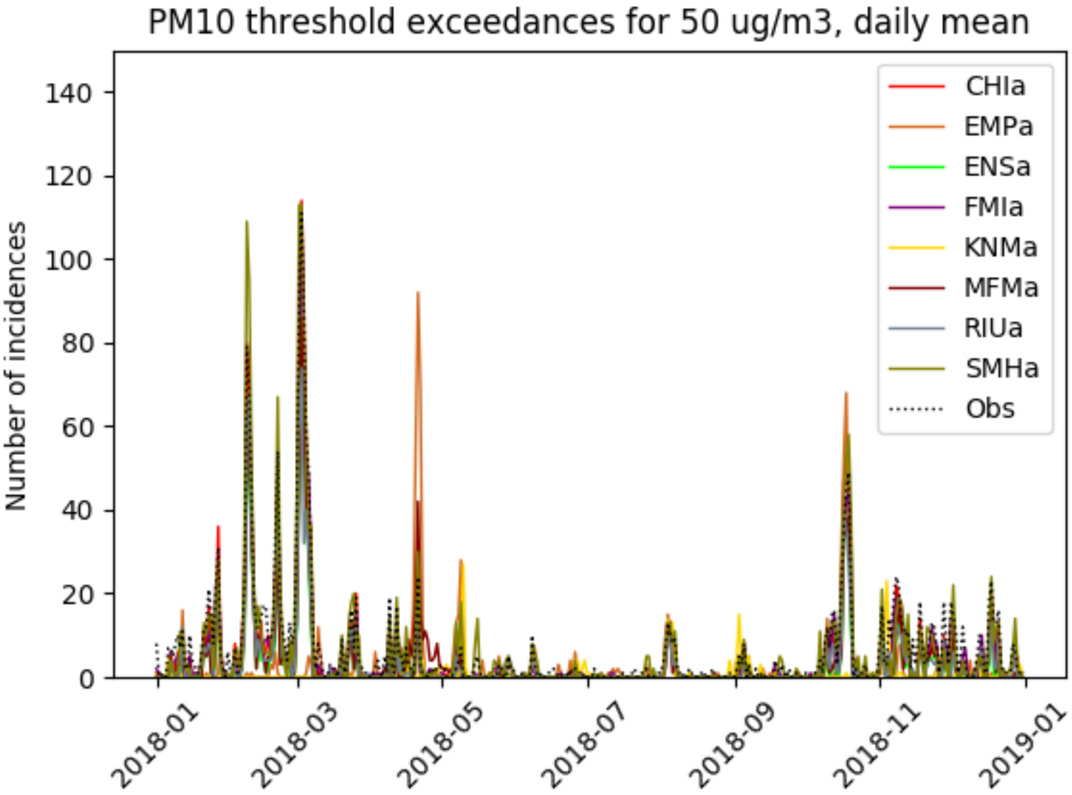


Figure 15 - Number of exceedance of daily limit value for PM_{10} in 2018 - observed (top) and modelled by interim re-analyses (bottom).

0 shows the number of exceedances of the PM_{10} daily limit value (50 µg/m^3) sorted per region. Both observed and re-analyzed data are presented and compared.

Re-analyses are able to capture PM_{10} pollution episode at the right time and with the correct duration. Nonetheless, ENSEMBLE skills can be improved: indeed, it is not the best re-analyses to describe exceedances, with only ~35 % of good detections (Figure 16) while the best individual re-analyses manage to capture 55 % of the exceedances. However, it is worth noting that the ENSEMBLE makes a low number of false alarms, meaning that 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 PM_{10} concentrations. Other manage to capture well part of the exceedances, but with also a large number of false alarms. This is problematic as it leads to describe as polluted, areas which are not so.

As Figure 15 indicates that temporal false alarms are low, it means that re-analyses tend to overestimate geographical areas with concentrations above the standards.

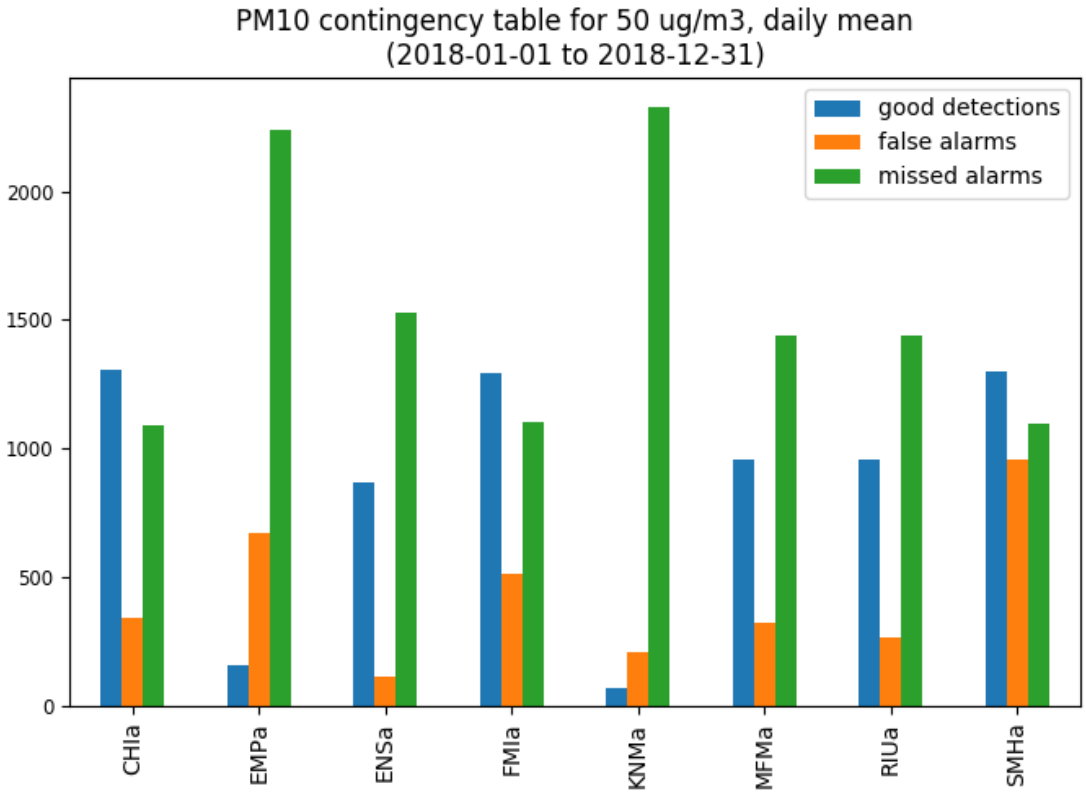
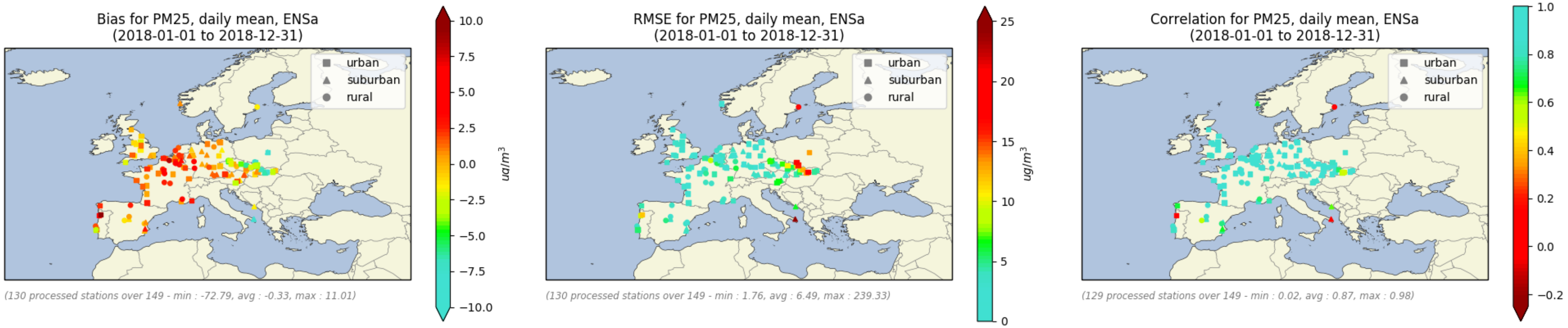


Figure 16 - Number of exceedances of daily limit value for PM_{10} in 2018 modelled by interim re-analyses.

5. Performance Indicators for $PM_{2.5}$

The evaluation of models' performances for $PM_{2.5}$ was constrained by the low number of available stations. This limit is clearly highlighted considering the maps on 0. However, where some measurements are available, the results are rather good: bias is around 0 for most of the European countries, except those of Eastern Europe where scores show a significant underestimation (and a large RMSE). Correlation coefficient can exceed 0.8, except in some specific locations, and RMSE generally stays below 10 $\mu g/m^3$. 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 07 and 08). Model responses are generally similar to PM_{10} responses and the ENSEMBLE re-analyses are amongst the best. The statistical scores are quite satisfactory with correlation coefficient generally higher than 0.5 and RMSE generally lower than 10 $\mu g/m^3$, except for the urban sites located in central Europe. The tendency is to have a more pronounced underestimation in Central Europe than in Western Europe and RMSE also higher in Central Europe than in Western Europe. Yet, those results are very difficult to interpret considering the low number of stations available for the evaluation.



Bias for PM25, daily mean, ENSa (2018-01-01 to 2018-12-31)

(130 processed stations over 149 - min : -72.79, avg : -0.33, max : 11.01)

RMSE for PM25, daily mean, ENSa (2018-01-01 to 2018-12-31)

(130 processed stations over 149 - min : 1.76, avg : 6.49, max : 239.33)

Correlation for PM25, daily mean, ENSa (2018-01-01 to 2018-12-31)

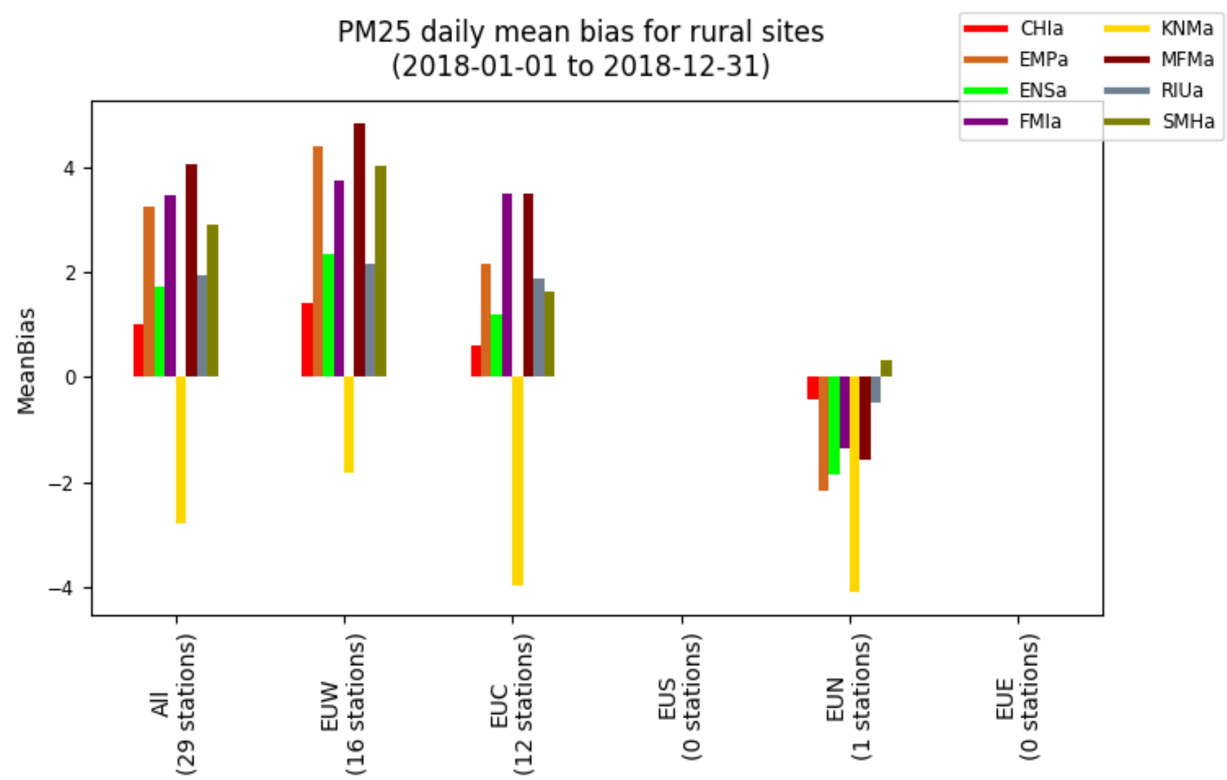
(129 processed stations over 149 - min : 0.02, avg : 0.87, max : 0.98)

(a)

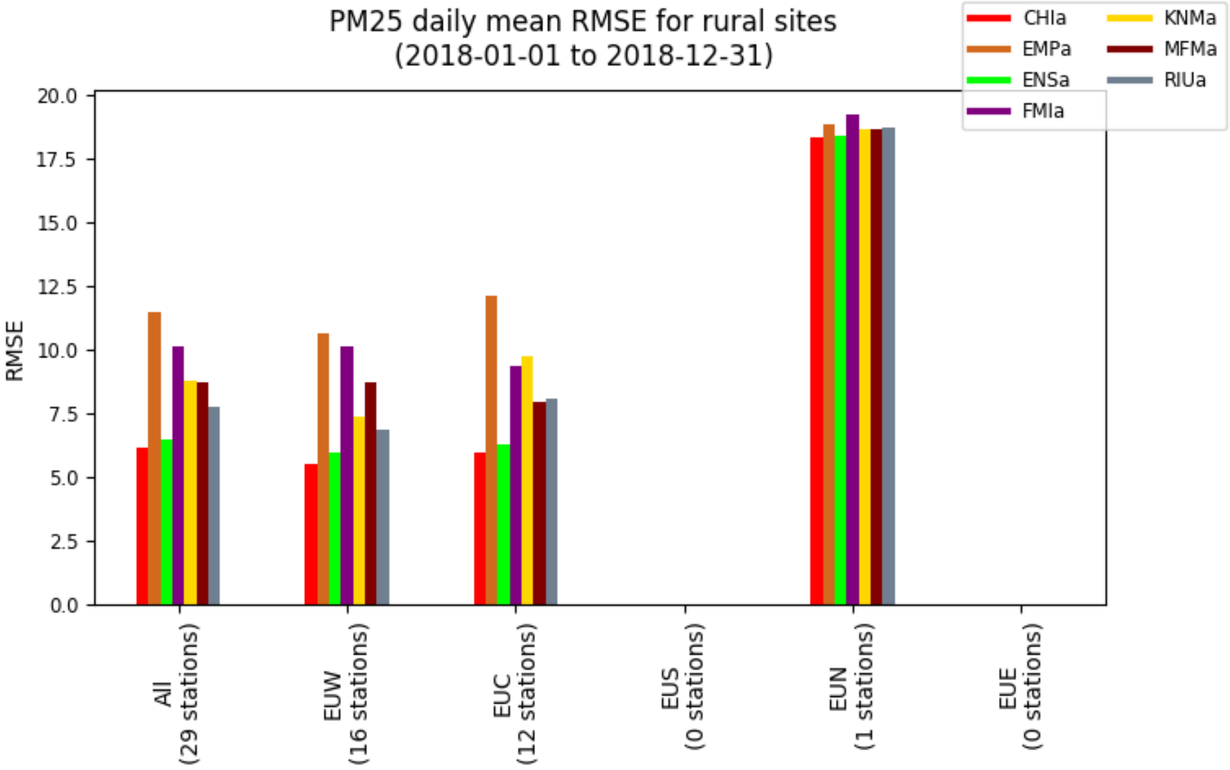
(b)

(c)

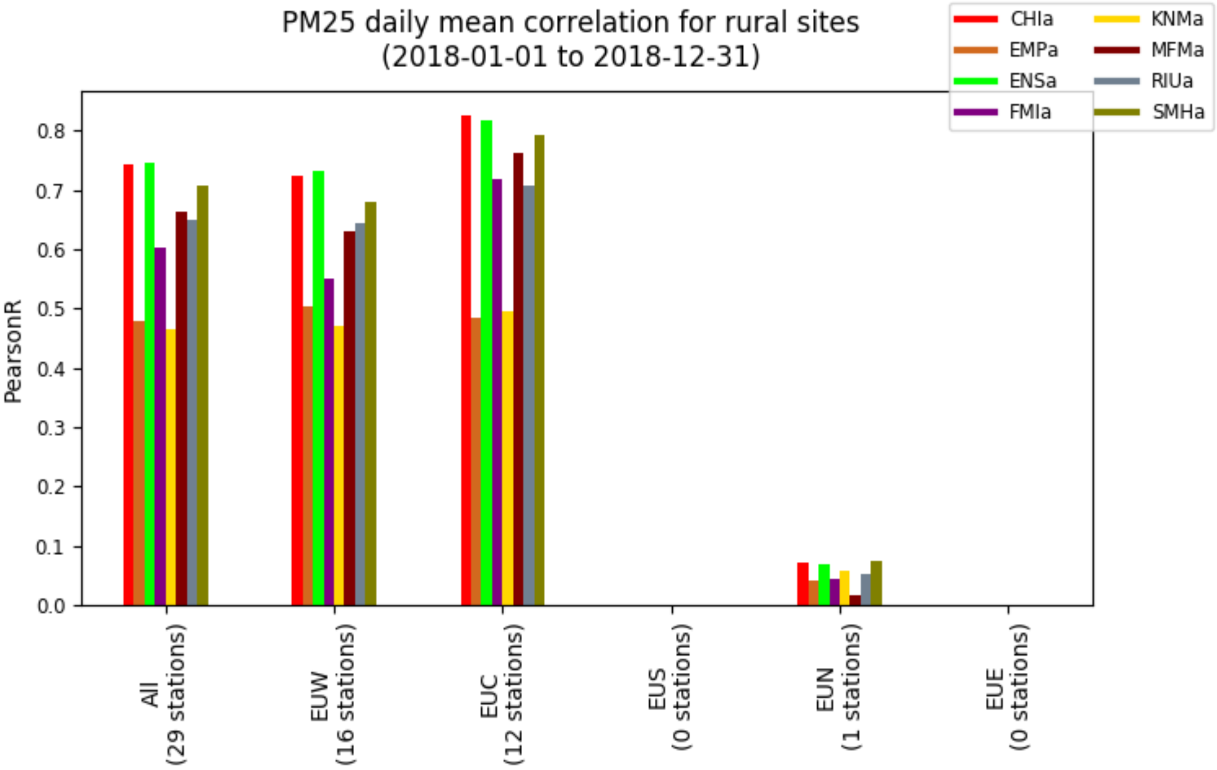
Figure 17 - Maps of Statistical scores of the ENSEMBLE interim re-analyses results against the observation validation dataset from the AQ e-reporting database for the $PM_{2.5}$ daily average over the year 2018: (a) Bias, (b) RMSE, (c) Correlation coefficient.


PM25 daily mean bias for rural sites
(2018-01-01 to 2018-12-31)
CHIa
EMPa
ENSa
FMIa
KNMa
MFMa
RIUa
SMHa
MeanBias
4
2
0
−2
−4
All
(29 stations)
EUW
(16 stations)
EUC
(12 stations)
EUS
(0 stations)
EUN
(1 stations)
EUE
(0 stations)

(a)

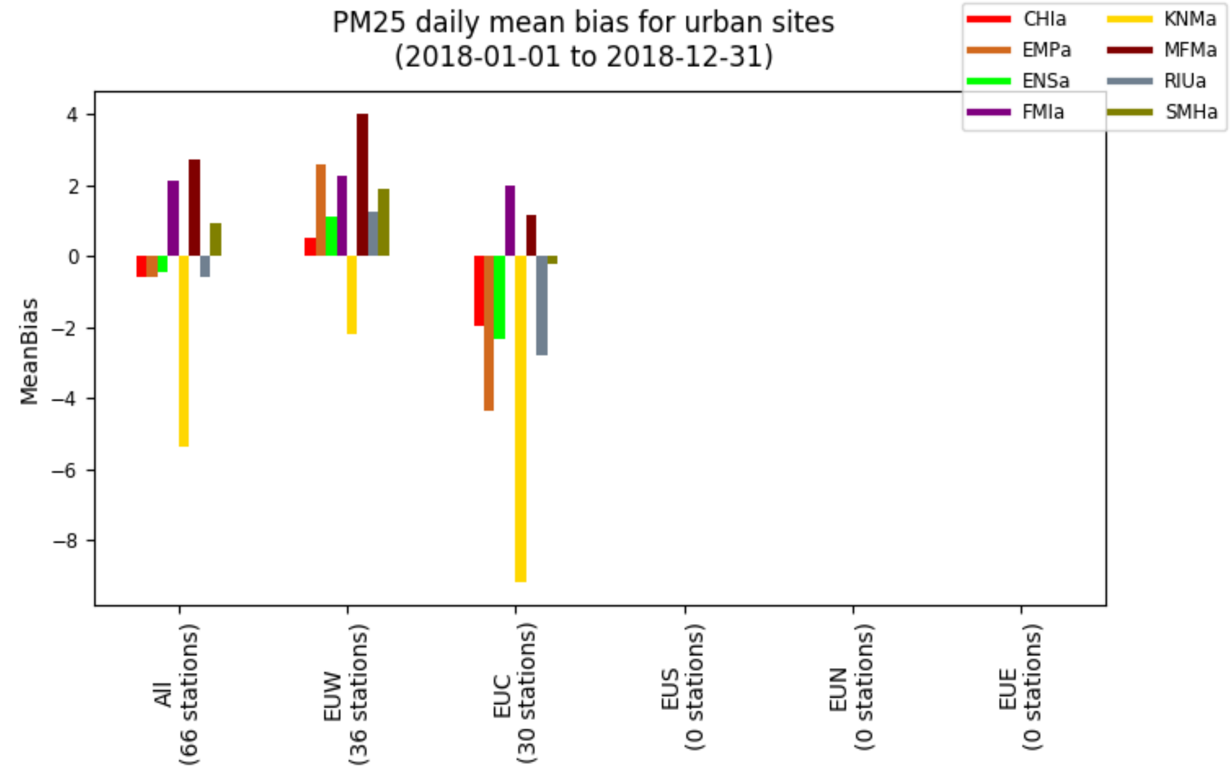

PM25 daily mean RMSE for rural sites
(2018-01-01 to 2018-12-31)
CHIa
EMPa
ENSa
FMIa
KNMa
MFMa
RIUa
RMSE
20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
All
(29 stations)
EUW
(16 stations)
EUC
(12 stations)
EUS
(0 stations)
EUN
(1 stations)
EUE
(0 stations)

(b)

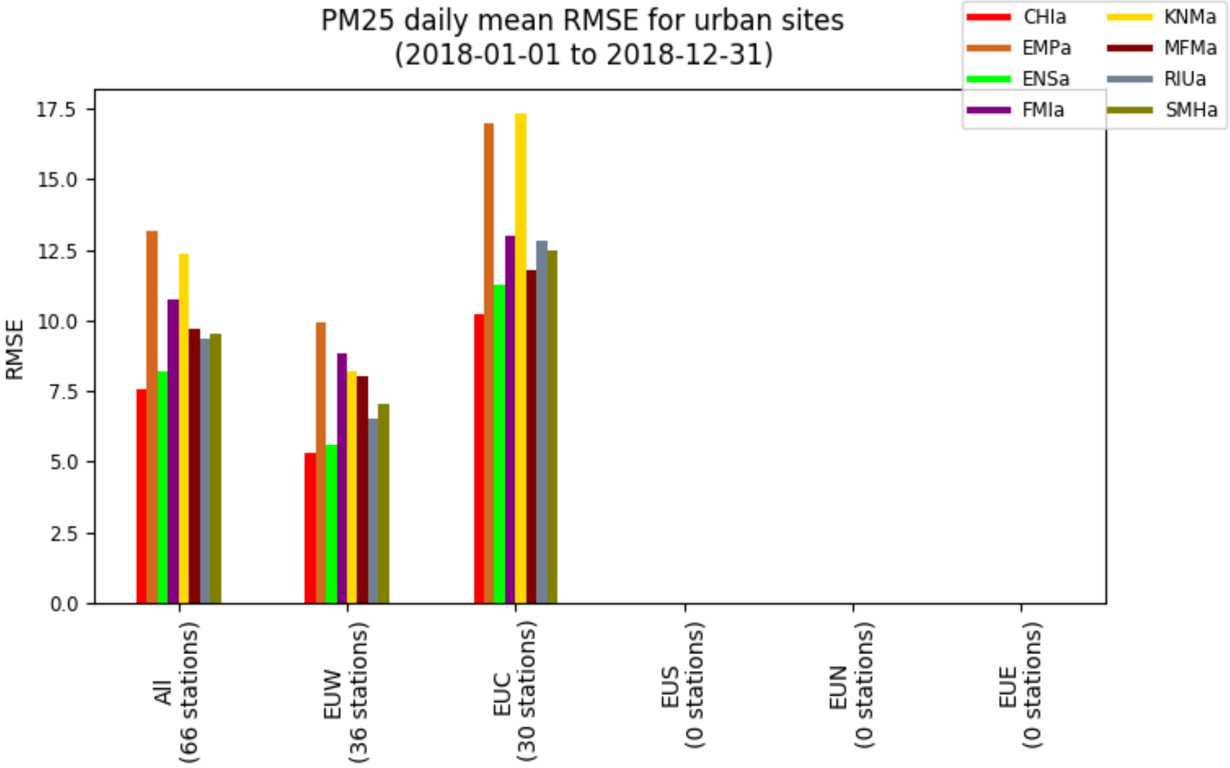


(c)

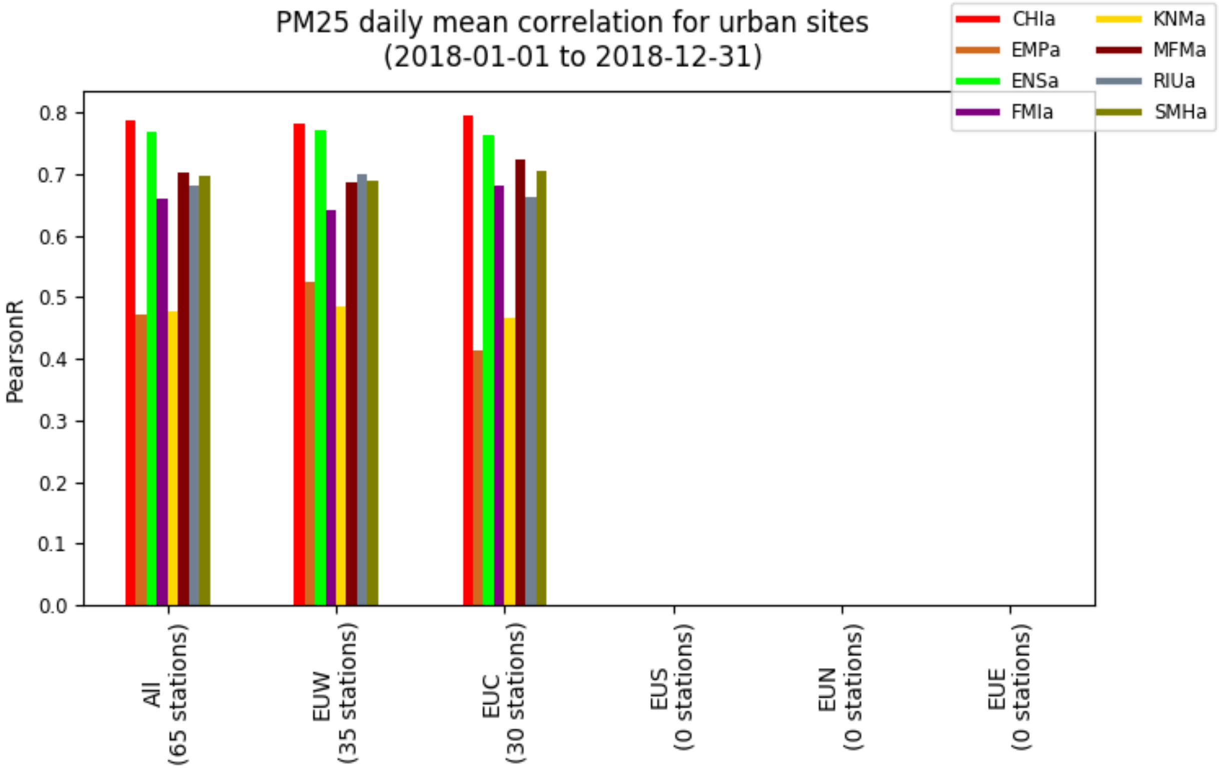
Figure 18 - CAMS Regional interim re-analyses for predicting $PM_{2.5}$ daily average over the year 2018 throughout European sub-regions: (a) Bias, (b) RMSE, (c) Correlation coefficient, at rural stations.


PM25 daily mean bias for urban sites
(2018-01-01 to 2018-12-31)
CHIa
EMPa
ENSa
FMIa
KNMa
MFMa
RIUa
SMHa
MeanBias
4
2
0
−2
−4
−6
−8
All (66 stations)
EUW (36 stations)
EUC (30 stations)
EUS (0 stations)
EUN (0 stations)
EUE (0 stations)

(a)


PM25 daily mean RMSE for urban sites
(2018-01-01 to 2018-12-31)
CHIa
EMPa
ENSa
FMIa
KNMa
MFMa
RIUa
SMHa
RMSE
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
All (66 stations)
EUW (36 stations)
EUC (30 stations)
EUS (0 stations)
EUN (0 stations)
EUE (0 stations)

(b)



(c)

Figure 19 - CAMS Regional interim re-analyses for predicting $PM_{2.5}$ daily average over the year 2018 throughout European sub-regions: (a) Bias, (b) RMSE, (c) Correlation coefficient, at urban stations.

Conclusion

The present report presents an analysis of the performances of the interim air quality re-analyses throughout Europe, produced by the CAMS Regional service for the year 2018. It focuses on ENSEMBLE air quality re-analyses resulting from the combination of seven well-validated and documented chemistry-transport models results. We call here “interim” re-analyses data assimilated fields of air pollutant concentrations based on up-to-date observation data. Because such data is quickly available after their production, the validation process it is submitted to is not necessarily achieved and the data should be considered as “interim” data. Nevertheless, we found interesting to elaborate interim re-analyses as first guess of air pollution patterns and levels that developed in Europe in 2018. Such information can be used to support Member States for the regulatory reporting duty on air quality (according to Directive 2008/50/EC). This is the reason why it is important to evaluate carefully the simulations against observations that are not used for the re-analyses production.

INERIS ran this process and computed a number of performance indicators and scores for ozone, nitrogen dioxide, PM_{10} and $PM_{2.5}$ concentrations. They are presented in this report using maps, Taylor diagrams and histograms. The main conclusions arising from this analysis are the following:

- Too little up-to-date observation data was available to perform an extensive evaluation of interim re-analyses over the whole of Europe (except for Central and Western Europe). Very few observations can be available in Eastern Europe and, depending on the pollutant, in Southern and Northern European regions that are not correctly covered. This is frustrating since they correspond to areas where there are more uncertainties (especially because of emissions).
- In Western and Central Europe, where there are more stations for the evaluation the models’ performances, results are generally more representative and correct. The quality of the re-analyses is generally similar to the previous year 2017.
- The European Environment Agency is building capacity to strengthen quality assurance procedures in the coming years and more countries are supposed to deliver up-to-date data, which will impact positively the interim re-analyses production process.
- For all pollutants, the performances are always of lower quality than what can be achieved with the validated re-analyses process, for which more stations are available and observation datasets are validated.
- The ENSEMBLE re-analyses give the best results for ozone when focusing on classical statistical scores. Ozone daily maxima are generally underestimated. Correlation coefficient ranges between 0.8 and 0.9 and RMSE between 15 and 18 $\mu g/m^3$ at rural and suburban locations. However, when looking at the threshold exceedances, it is worth noting the low capabilities of the ENSEMBLE re-analyses to detect concentrations above the standards, and its good skills to keep the number of false alarms at a very low level.
- Good model scores for simulating ozone in Western Europe are hampered by inferior performances at few stations in Eastern and Southern Europe.

- The performances of nitrogen dioxide re-analyses are quite stable with previous years, and with satisfactory scores. RMSE is around 15 µg/m^3, bias shows an underestimation of 5 µg/m^3 and correlation close to 0.7.
- For PM_{10}, even if the results are quite satisfactory considering the state of the art, the statistical scores remain lower than what is usually achieved with validated re-analyses. Up-to-date PM_{10} observation datasets need to be improved in the future.
- PM_{10} is the pollutant for which model responses range in the largest interval: correlation coefficient from 0.35 to 0.8 and RMSE from 10 to 25 µg/m^3, depending on the model and the station typology. The results are the best for suburban stations in Western and central Europe. More frequent overestimations of PM_{10} concentrations occurred over European stations and not only for rural ones.
- Moreover, the evaluation demonstrates how the Ensemble approach, based on a median average of involved models is not appropriate to simulate exceedances of threshold values. Only 35% of good detection of exceedances of the PM_{10} daily limit values were correctly caught by the ENSEMBLE, whereas the best re-analyses got 55 %. As for ozone, the ENSEMBLE re-analyses produce a very low number of false alarms.
- Finally, despite only few $PM_{2.5}$ measurement data was available for the evaluation, the results obtained for this pollutant are promising. The individual models' responses are quite consistent, and the Ensemble median generally gives the best results. Correlation coefficient ranges from 0.4 to 0.8 according to the location and the station typology and the RMSE from 5 to 17 µg/m^3, which is very reasonable. Once again, the conclusions are limited by the low number of stations available in some geographical areas and should be consolidated and improved in future interim assessments, when the up-to-date data gathering process at the EEA is strengthened.

Copernicus Atmosphere Monitoring Service

ECMWF - Shinfield Park, Reading RG2 9AX, UK

Contact: info@copernicus-atmosphere.eu

atmosphere.copernicus.eu copernicus.eu ecmwf.int