ECMWF COPERNICUS REPORT

Copernicus Atmosphere Monitoring Service

Annual report on the verification of interim reanalyses

IRA2016

Issued by: METEO-FRANCE/ S. Guidotti

Date: 10/07/2017

Ref:

CAMS50_2015SC2_D50.5.3.1-2016_201707_Annual_verification_report_IRA2016_v1

Contributors

INERIS
F. Meleux
L. Rouïl

METEO-FRANCE
M. Plu
S. Guidotti
N. Assar
V. Arteaga-Pichard

Table of Contents

Executive summary

The present report provides a performance analysis of the regional interim air quality reanalyses throughout Europe, produced by the CAMS for the year 2016.

The CAMS regional services include the provision of ENSEMBLE air quality reanalyses, resulting from the combination of seven well-validated and documented chemistry-transport models' results. So-called "interim" reanalyses 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 they are produced, even if the necessary validation process is not completed. This kind of data is thus flagged as "non-validated" or "non- verified" data. This is the new AQ e-reporting process. 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 reanalyses as first guess of air pollution patterns and levels that developed in Europe in 2016. 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 reanalyses production.

INERIS performed this evaluation process and computed a number of performance indicators and scores for ozone, nitrogen dioxide, PM10 and PM2.5 concentrations. They are presented in this report. Globally the models performed as expected and the ENSEMBLE median model generally gives the good results, but not always the best (see PM simulations). Consistency with previous validated assessment results from MACC projects regarding models' behavior is ensured. However as for the previous exercise in 2015, this evaluation demonstrates the difficulties encountered by the Ensemble approach based on the median to simulate properly exceedances of threshold (limit) values. This should be clearly highlighted as a need for the CAMS Regional service evolution.

The interim reanalyses 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:

- Two few up-to-date observation data was available to perform an extensive evaluation for ozone over the whole of Europe. Eastern and Southern European regions are not correctly covered, what is a pity since they correspond to areas where there are more uncertainties.
- In Western and Central Europe, where there is enough data for the evaluation the models performances are generally satisfactory and compatible with the state of the art with correlation.

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

Best results are obtained for the ensemble: correlation coefficient ranging from 0.8 to 0.9 and root means square error lower than 15-17 μg/m^3.

- For ozone, the performances remain of lower quality than what can be achieved by validated re-analyses and are stable compared to 2015. Decreasing number of up-to-date observation data should be considered carefully in the future.
- For NO_2, performances are quite similar to those obtained in 2015 and consistent with the state of the art. There are few countries for which they are quite good (correlation factor of about 0.7, and RMSE lower than 15 μg/m^3) while other are more difficult to simulate. It is also linked to the "local nature"' of the pollutant largely influenced by local sources that may be not correctly considered by the CAMS models' resolution.
- For PM10, 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 PM10 observation datasets need to be more consolidated in the future. Less observations were available than in 2015 with countries not covered at all: Italy, Slovakia, Romania, Bulgaria, Hungary, Serbia, Croatia, Slovenia …
- Saying, that, it should be noted that were there are enough measurements scores are rather good with RMSE of about 5-7 μg/m^3 and correlation factor reaching 0.8, except in Portugal, Poland and Czech Republic. However, the Ensemble does not always give the best results being hampered by the quite large variability of model responses. This point should be considered carefully in the future.
- 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 19% of the exceedances of the PM10 daily limit values were correctly caught by the Ensemble (against about 70% for some individual models).
- Finally, very few PM2.5 measurement data was available for the evaluation making the conclusions difficult to highlight. Nevertheless, where there are measurements the statistical scores are quite satisfactory (correlation coefficient of about 0.8 and RMSE of about 5 μg/m^3).

Introduction

This report gives an overview of the performances of the European air quality **interim reanalysis** process developed by the CAMS Regional service and implemented to simulate air quality in Europe during the year 2016.

Air quality interim reanalyses result from a combination of chemistry-transport models that simulate the spatio-temporal evolution of air pollutant concentrations, and observations assimilated in every model to correct and improve its results. Data assimilation techniques are used.
The models implemented to perform these interim reanalyses are the set of seven models run in other near-real-time CAMS regional services. The models' set-up is described in a series of reports published in April 2017[2]. 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). A synthesis of the set-up adopted for the 2016 interim reanalyses is appended to this document to support interpretation of the results presented.
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 reanalyses" 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[3]. We estimate that 20 days is appropriate time-lag to get the data and run the reanalyses for a given day.

The set of observation sites reported to the EEA is split into two subsets, one for data assimilation in the interim reanalyses 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. It should be noted that not all the Members States reported UTD data. Consequently, data assimilation and evaluation cannot be performed in some geographical areas, and it will not be possible to draw some clear conclusions about the model capacities in those regions. Southern Europe (included Italy) is more particularly concerned.

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] Reports are referenced as "CAMS50_2015SC2_D50.3.3.2.MODEL-2016_201704_Annual_IRA_Report".
[3] 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 (PM10 and PM2.5). 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 on the basis of classical statistical indicators which 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 behaviour of performance indicators depends on the station typology and 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[4] 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.

[4] 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 various 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 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-effect, a result close to the threshold can fall arbitrary in one or the other category), they can give useful information to compare various models’ behaviour in different geographical regions. GP, FA and ME are expressed in percentage (%).

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 impact human health

In-depth analysis of the interim ENSEMBLE reanalyses 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 the 1st April to 30th September 2016. 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, rather at rural locations. However, it should be noted that evaluation cannot be conducted in several Southern countries (Italy, Greece, Slovenia, Croatia) and Eastern countries (Romania, Bulgaria), because of a lack of reported observation data.
Correlation coefficient is excellent with values higher than 0.9 in most cases. Actually, very few stations in Spain, Portugal, Poland, Czech Republic and Slovakia show poorer performances. The same can be seen for RMSE, although for a major number of stations, RMSE ranges between 10 and 15 µg/m^3, which is very good for interim results, but a bit higher than what is usually achieved with validated assimilated results (rather between 5 and 15 µg/m^3). This can be a consequence of using partial and non-validated observation data, and results should improve when the validated reanalyses are performed. However, results remain very acceptable compared to the state of the art.
Performances decrease for stations around the Mediterranean area (Portugal, Spain) and in Eastern Europe, with values higher than 25 µg/m^3.

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 presents bias, correlation coefficient and RMSE scores for all models at rural stations, while 0 presents the same information at suburban stations (where models best perform). The indicators are sorted per geographical region: Western Europe (EUW), Central Europe (EUC), Southern Europe (EUS), Northern Europe (EUN), Eastern Europe (EUE). First, decreasing number of stations compared to what was available in 2015 should be noted and reported to the EEA. It seems that the up-to-date data reporting process is still unstable and should be improved in the future. Evaluation cannot be reasonably performed for Southern (no stations whatever the typology) and Eastern Europe (only few urban stations).
Where observation data is available, it is interesting to note very few differences in the performances for rural and suburban sites[5], but reasonable consistency between models’ behaviour. Performances are different (and generally better) at urban sites. At rural and suburban sites, all models except LOTOS-EUROS underestimate ozone peak values.
Correlation coefficient is quite high ranging from 0.8 to 0.9 in the best case (the Ensemble model).

[5] The 2015 Interim re-analyses had much more significant differences from a region to another and from a site typology to another

Regarding RMSE, we can note once again good consistency between results even they are slightly better at rural stations. Better results are obtained for the Ensemble and CHIMERE models (from 13 to 17 $\mu g/m^3$) at rural and suburban sites and less than 15 $\mu g/m^3$ at urban sites (0). EURAD and MATCH models gave the worst results, with RMSE reaching 20 $\mu g/m^3$.
Obviously, there are more uncertainties in the models in Southern, Eastern, Northern regions due to uncertainties in emissions, and the complexity of the photochemical processes and meteorology. But 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 perhaps refined when validated reanalyses for 2016 are available.
Nevertheless, overall performances of the models to simulate ozone daily maxima are fairly satisfactory and consistent with previous results obtained in the past and with the state of the art.

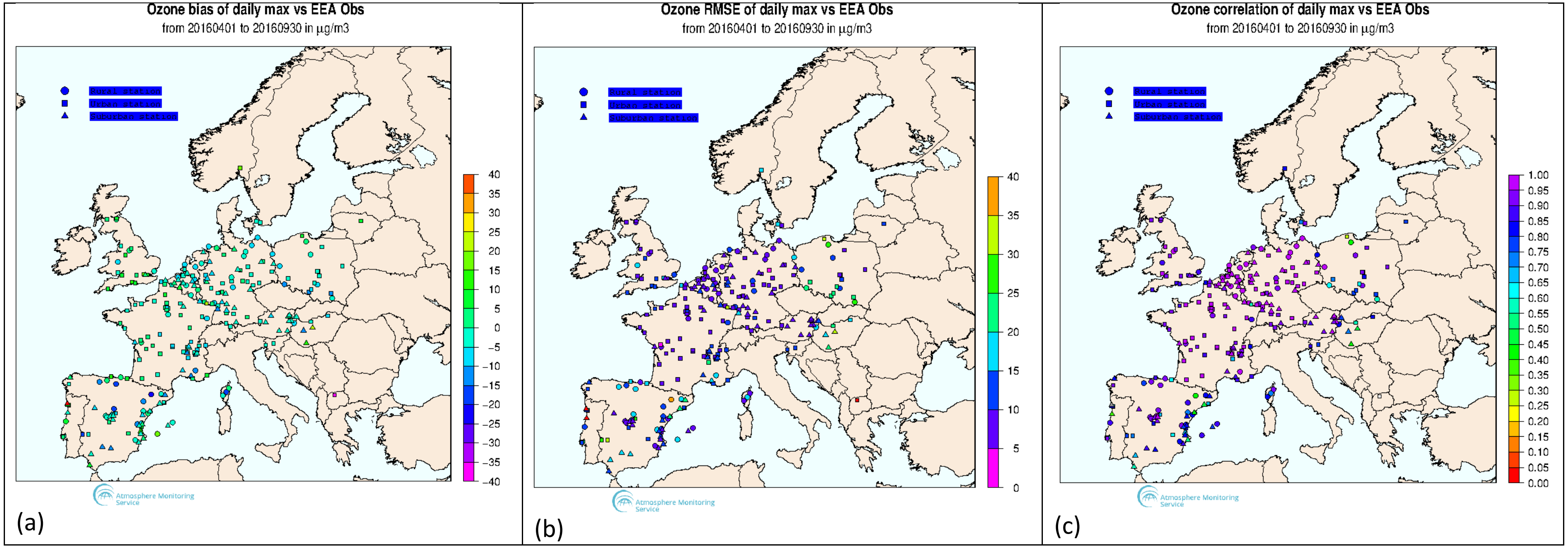
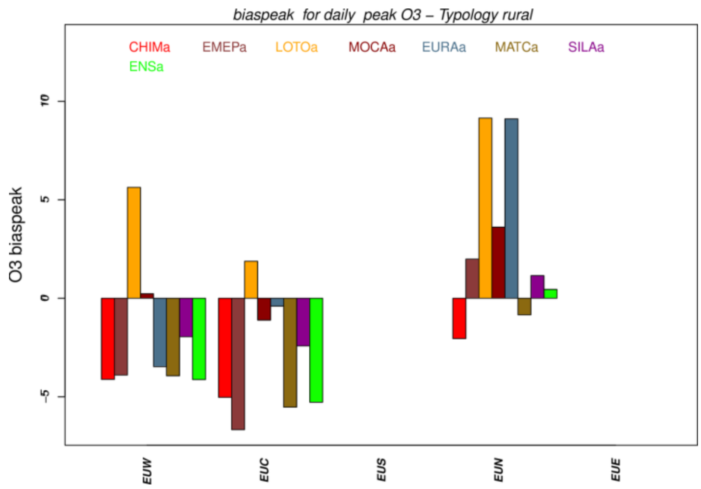
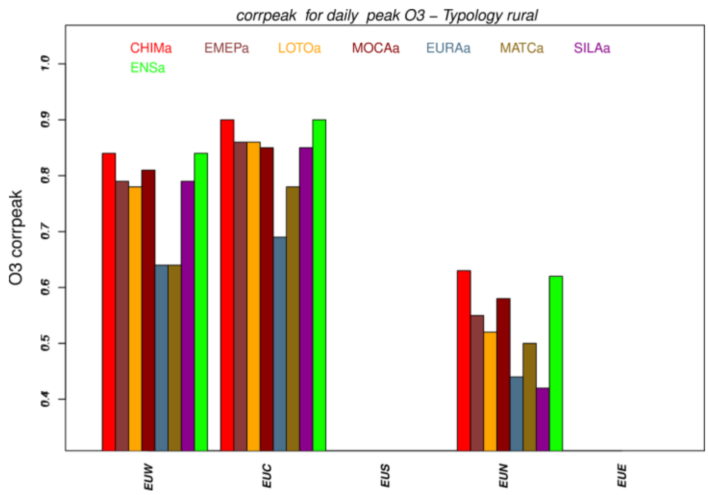


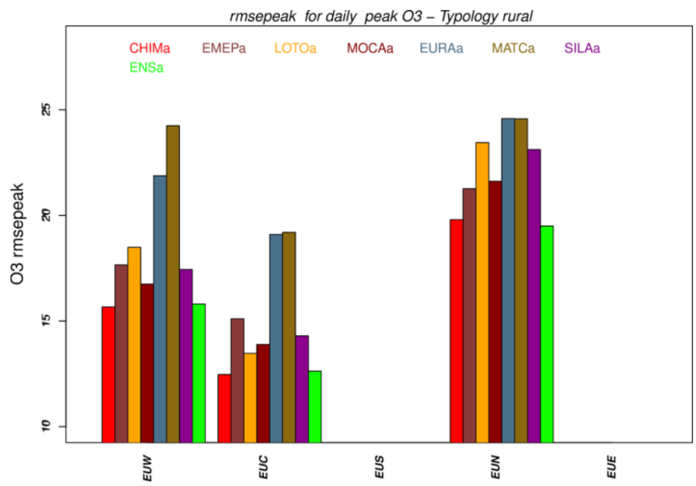
Figure 1 - Maps of Statistical scores of the ENSEMBLE interim reanalyses results against the observation validation dataset from the AQ e-reporting database for the ozone daily maximum from 01/04/2016 to 30/09/2016 (a) bias, (b) root mean square error, (c) correlation coefficient



(a)

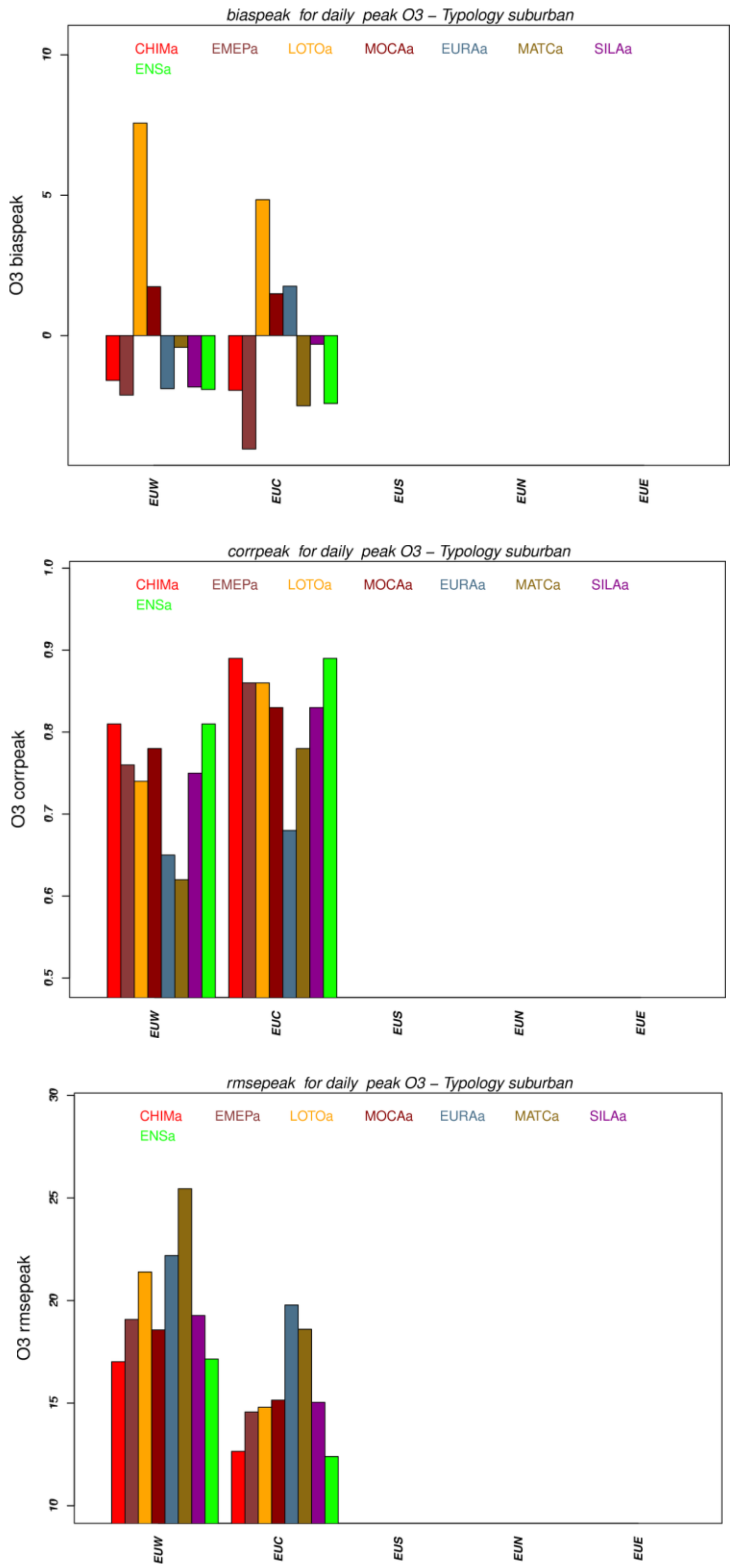


(b)



(c)

Figure 2 - CAMS regional interim reanalyses for predicting daily ozone peak over the summer 2016 throughout European sub-regions (a) Bias (b) Correlation coefficient (c) RMSE at rural stations

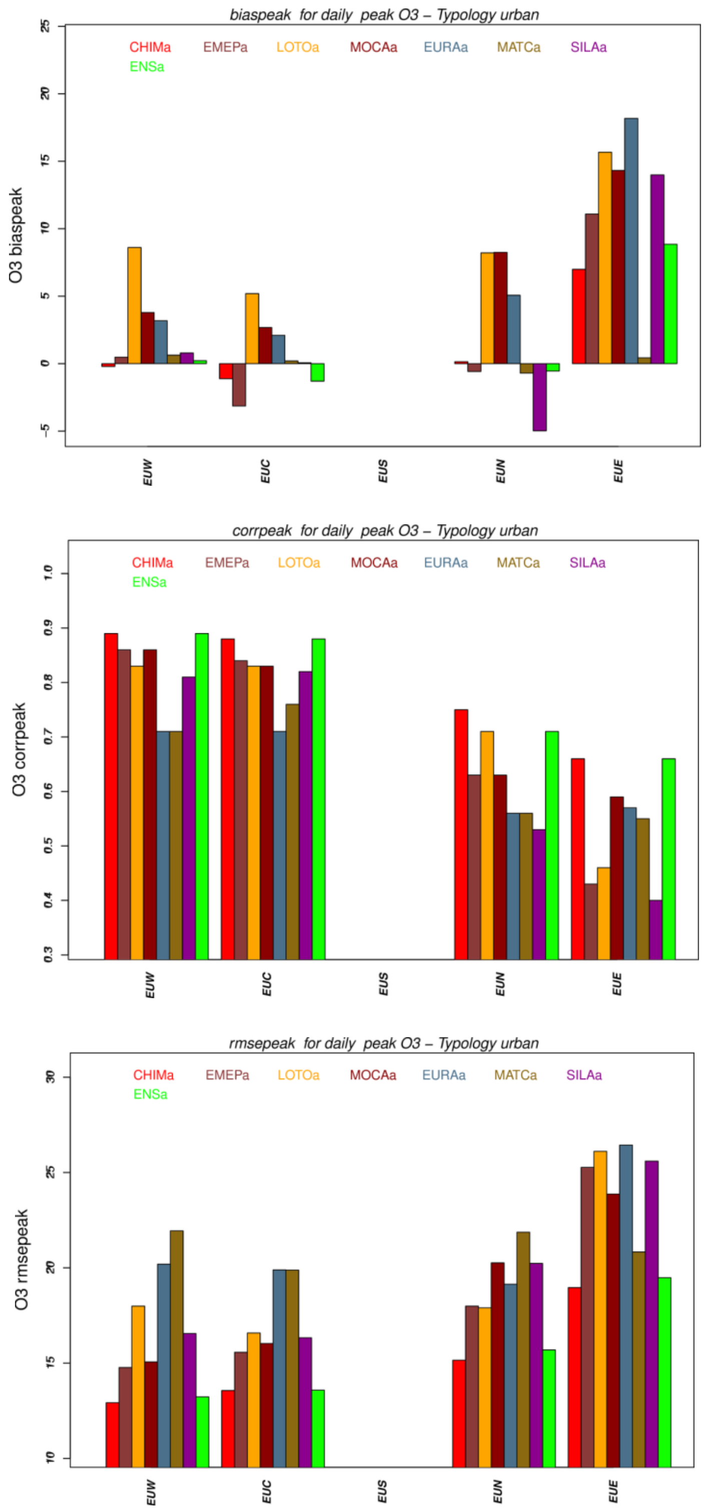


(a)

(b)

(c)

Figure 3 - CAMS regional interim reanalyses for predicting daily ozone peak over summer 2016 throughout European sub-regions (a) Bias (b) Correlation coefficient (c) RMSE at suburban stations



(a)

(b)

(c)

Figure 4 - CAMS regional interim reanalyses for predicting daily ozone peak over summer 2016 throughout European sub-regions (a) Bias (b) Correlation coefficient (c) RMSE at urban 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 the hourly information threshold has been exceeded during the summer time in 2016 (time is presented on the x-axis), sorted per geographical region. Observed and modelled (ENSEMBLE interim reanalyses) are proposed. Clearly, model results are very disappointing with most of the recorded exceedances not correctly caught by the Ensemble model: only 7 situations over 132 were modelled. For this indicator, performance varies largely from a model to another (0)

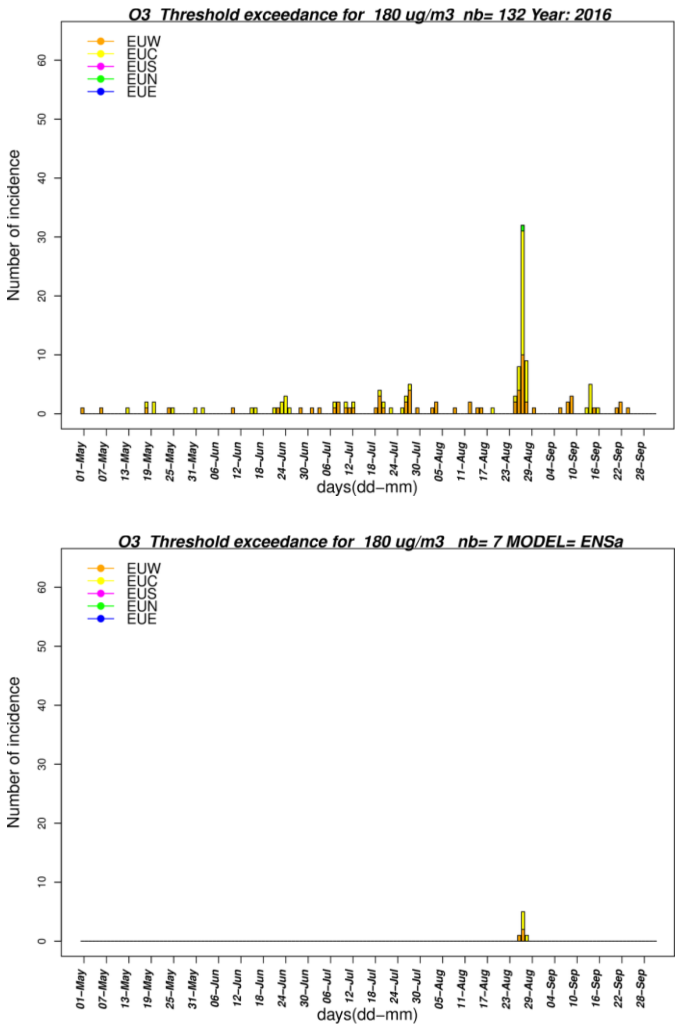


Figure 5 - Number of exceedances of the information threshold value for ozone in summer 2016 – observed (top) and modelled by the ENSEMBLE interim analyses

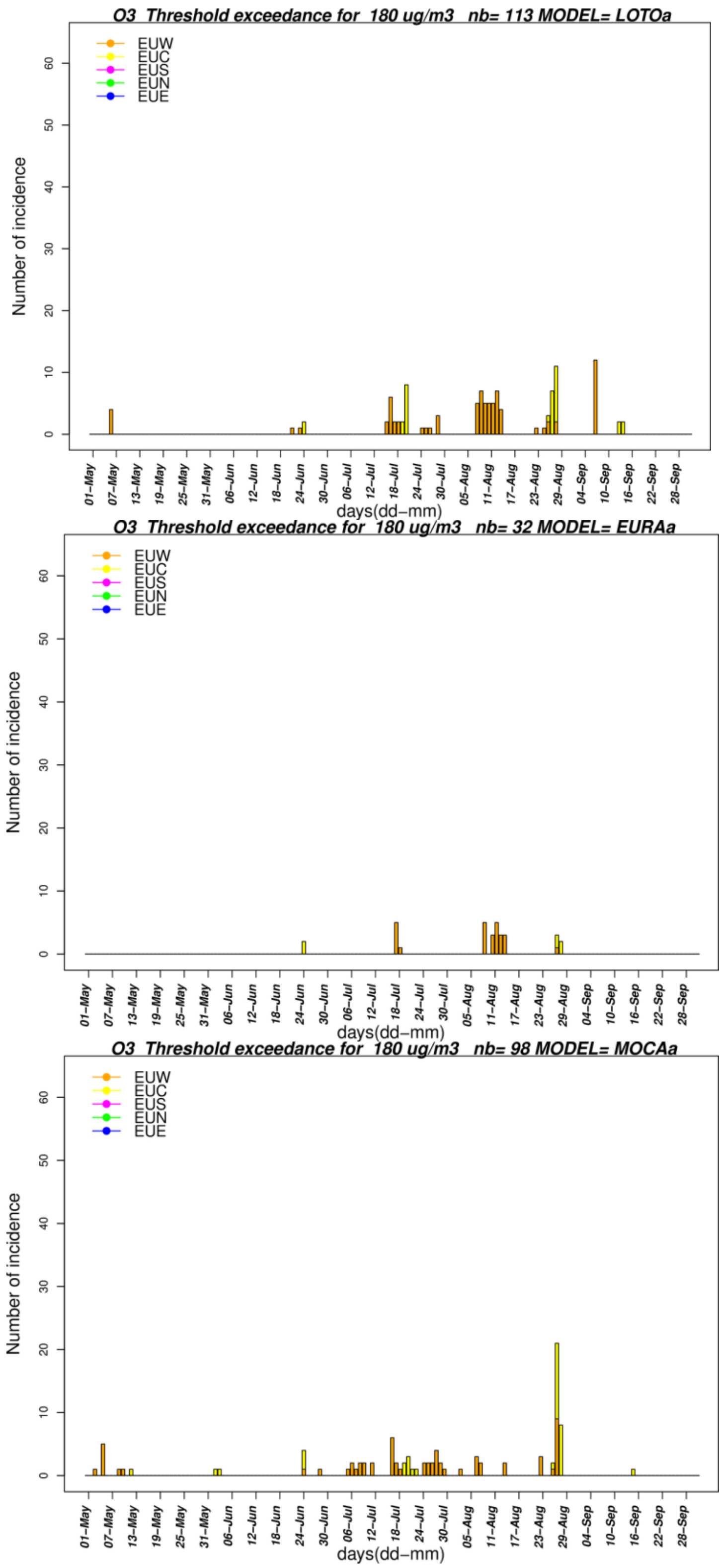


Figure 6 - Number of exceedances of the information threshold value for ozone in summer 2016 – LOTOS-EUROS (top) EURAD (middle), MOCAGE (bottom)

Previous figures show how much it is difficult to simulate exceedance situation with an Ensemble approach based on the median. It 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.

3. Performance indicators for nitrogen dioxide

Warning note: *As already stated from the MACC projects, 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 is about 20 to 10 km, and is not sufficient to catch actual NO_2 concentrations at traffic and industrial sites.*

Figure 7 presents the Taylor diagram for Ensemble CAMS regional interim reanalyses, for the daily maximum (hourly average) of NO_2 concentrations. It shows disperse model performances depending on the station typology. The Correlation coefficient ranges from 0.3 to 0.7 what is slightly lower than what we had for 2015 (0.4 to 0.7-0.75), but still acceptable for NO_2 simulations on coarse modelling grids (NO_2 is considered as a local pollutant). RMSE ranges from 12 to 18 µg/m^3.

Maps in 0 allow to better understand these differences. Actually, it seems that the models perform correctly in a limited number of stations located in Benelux, France, Germany, the UK and Austria. Elsewhere, results can be seen as more disappointing, but remain consistent with the state of the art. Results confirm that the models generally underestimate NO_2 concentrations (see bias). For this pollutant, the influence of local sources is very important, but they cannot be accurately taken into account by the adopted model resolution. However, the results obtained are rather promising, keeping in mind that those are only interim results (more stations should be available for the validated reanalyses), and that the indicator considered here is the daily maximum which is much more difficult to catch than the average, in particular because of its dependence to peak.

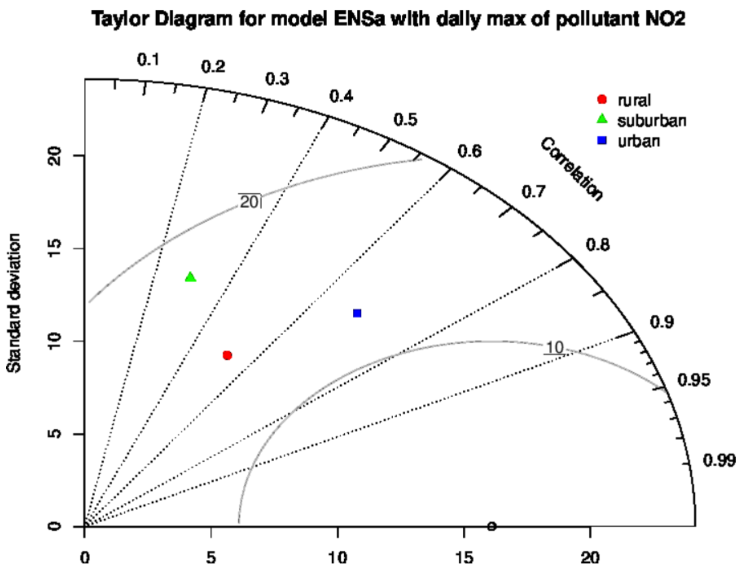


Figure 7 - Taylor diagram presenting the performances of the CAMS regional interim Ensemble reanalyses to predict NO2 daily maxima

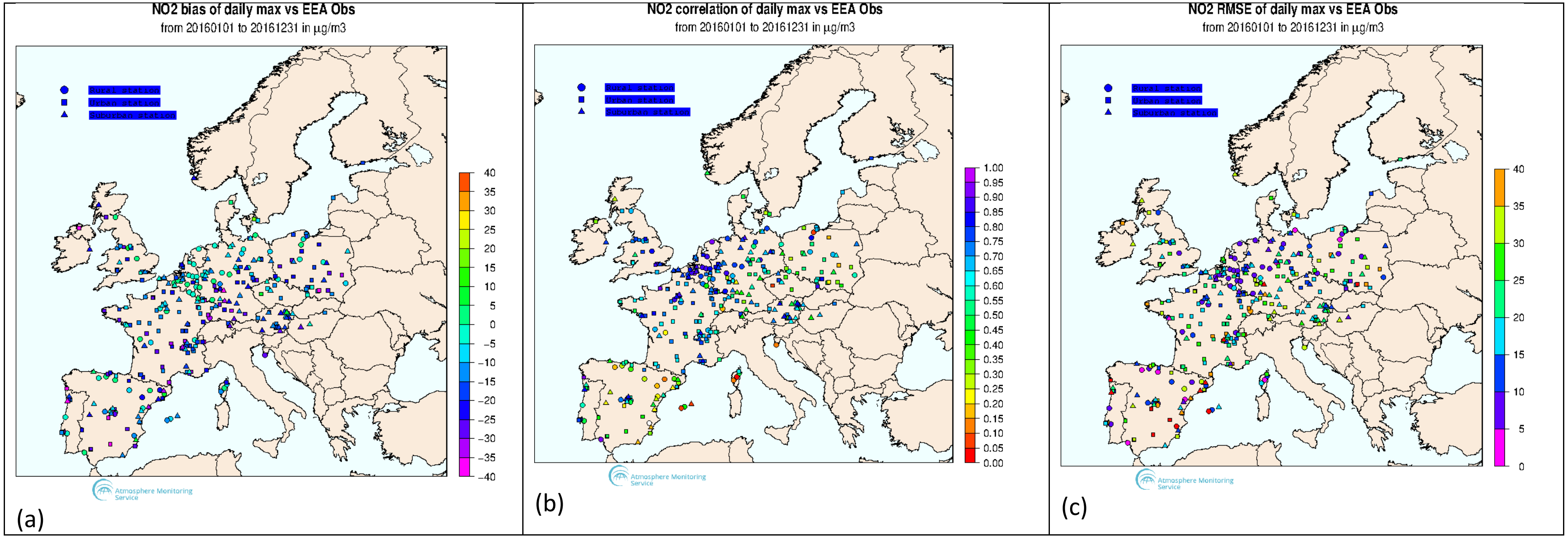


Figure 8 - Maps of Statistical scores of the ENSEMBLE interim reanalyses results against the observation validation dataset from the AQ e-reporting database for the NO2 daily maximum over the year 2016 Bias (a) Correlation coefficient (b) Root mean square error (c)

4. Performance indicators for PM10

0 shows the Taylor diagram obtained for PM10 daily averages over the year 2016, for CAMS regional ENSEMBLE reanalyses. In this case, it is interesting to compare to the results obtained in 2015 which are given as well (0). The dependence of the evaluation to stations typology appears is not so sensitive in 2016 than it was in 2015. In most cases, the best results are achieved for suburban stations and less good ones for the rural stations. This is representative of the impact of the models' resolution regarding the simulated phenomena. However, the rather poor correlation in rural areas should be further investigated, even if it slightly improved in 2016. Quality and quantity of available observations may have an impact. However, it seems that RMSE indicators improved in 2016 as well. But there is still an overall discrepancy, compared to the results we used to have with the CAMS validated re-analyses: 6 to 8 µg/m^3 in the best cases, against 16 to 14 µg/m^3 for the present interim analyses. Thus, we suspect here an impact of the use of up-to-date non validated data that certainly need to be deeply investigated by the data providers.

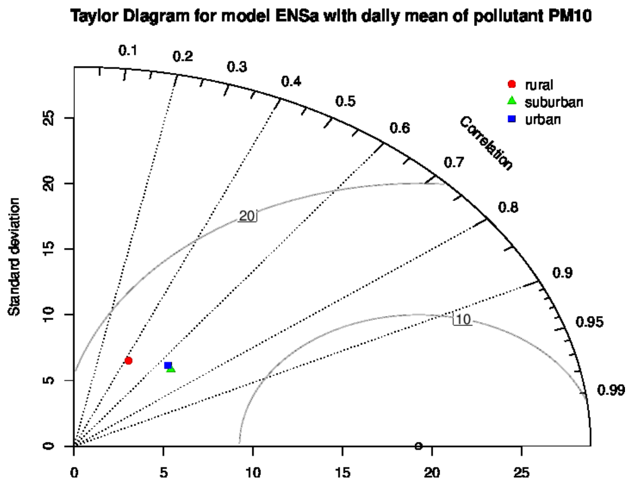


Figure 9 -Taylor diagram presenting the performances of the CAMS regional Ensemble interim reanalyses to predict PM10 daily average in 2016

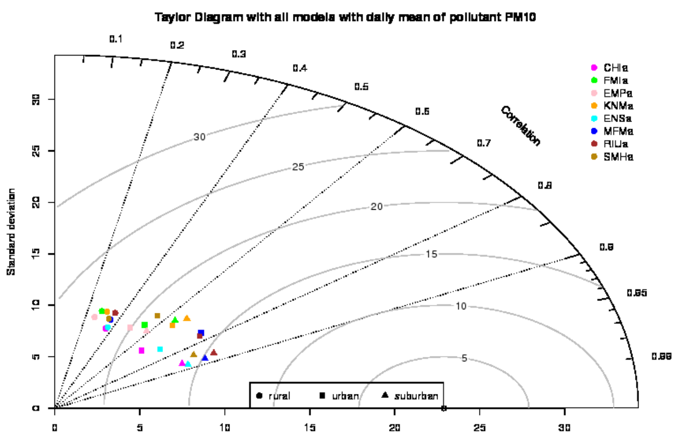


Figure 10 - Taylor diagram presenting the performances of the CAMS regional interim reanalyses to predict PM10 daily average in 2015

0 details the geographical distribution of statistical scores (bias, correlation coefficient and RMSE), for the ENSEMBLE interim reanalyses for the year 2016. Lowest correlation scores are obtained for rural stations located in Portugal, Spain and in the Alps. In several countries (France, Germany, and Benelux), RMSE ranges between 1 and 5 µg/m^3, which is very encouraging, but the overall performance is hampered by high values in Eastern Europe (especially Poland and Czech Republic). PM10 concentrations are always underestimated (see bias), but within an acceptable range (between 5 and 10 µg/m^3).

Local discrepancies can be explained by the complexity (in meteorological terms) of certain areas (mountainous regions), but also by uncertainties in the emission inventories (especially in Eastern Europe), and by the lack of available observation data for data assimilation and evaluation.

Differences between model results can be furthermore 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 interim PM10 reanalyses, with huge gaps in some areas (Southern, Northern and Eastern Europe for rural and suburban stations, and Northern and Southern Europe for urban stations). In several cases (but not systematically) the Ensemble model gives the best scores. Its performance is constrained by the large variability of the models responses. In general, the CHIMERE model gives the best results followed by EURAD and SILAM. The models' performances are not so good at urban stations in Western Europe (correlation coefficient lower than 0.4 and RMSE of about 25 µg/m^3) whatever the model, but improve at rural stations. How this situation evolves with validated data should be carefully investigated in the future.

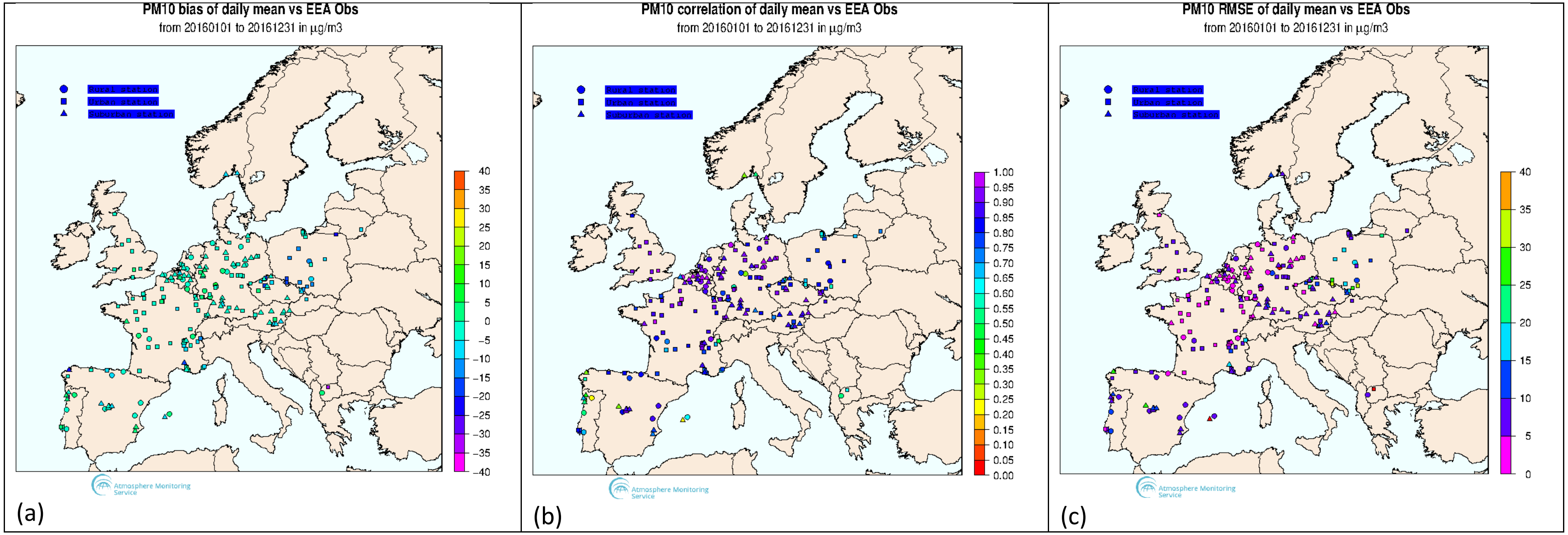
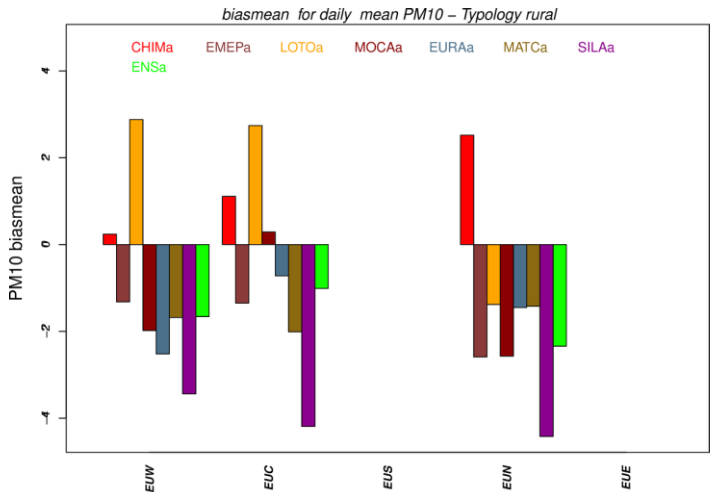
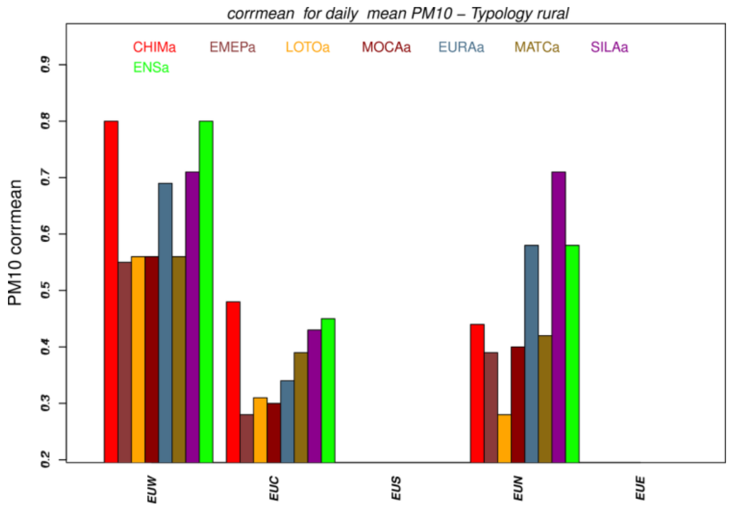


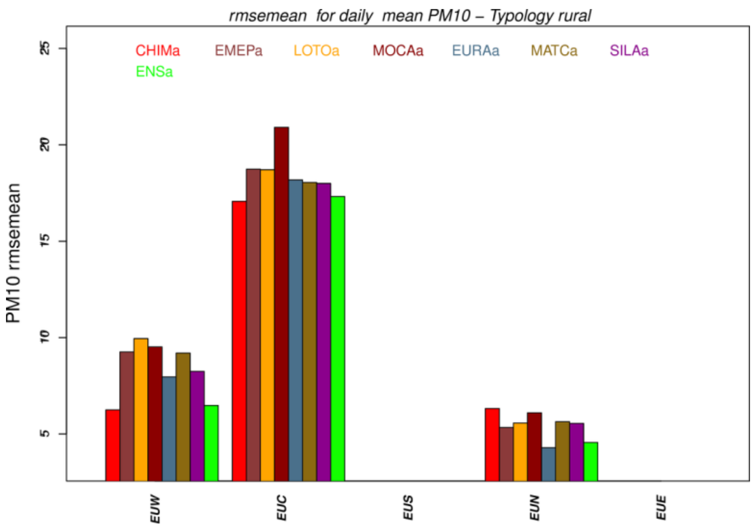
Figure 11 - Maps of Statistical scores of the ENSEMBLE interim reanalyses results against the observation validation dataset from the AQ e-reporting database for the PM10 daily average over the year 2016 Bias (a) Correlation coefficient (b) Root mean square error (c)



(a)

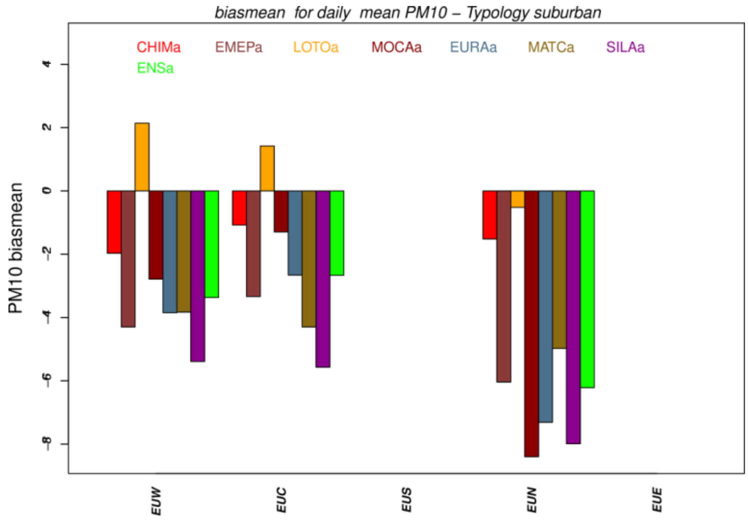


(b)

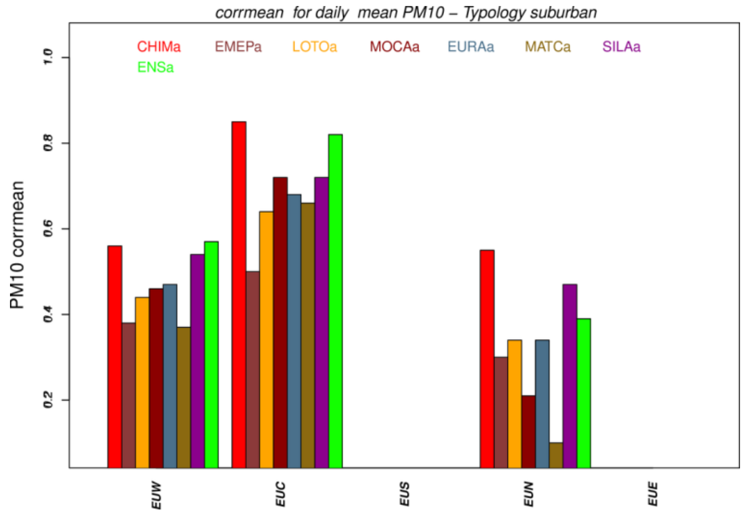


(c)

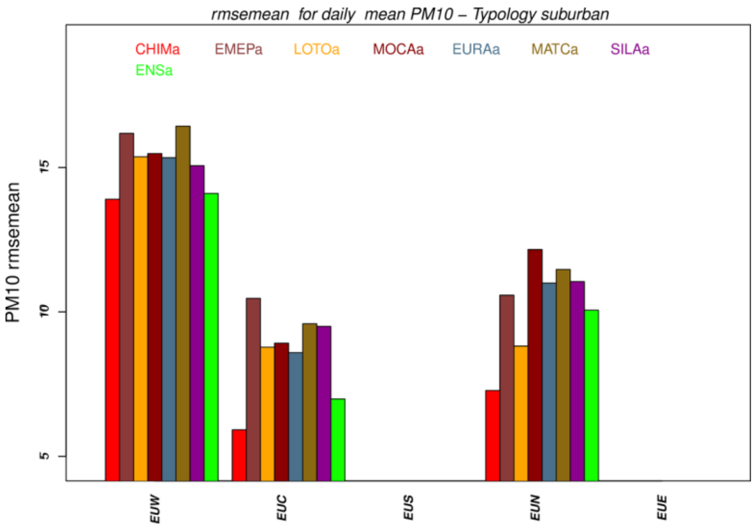
Figure 12 - CAMS regional interim reanalyses for predicting PM10 daily average over the year 2016 throughout European sub-regions (a) Bias (b) Correlation coefficient (c) RMSE at rural stations



(a)

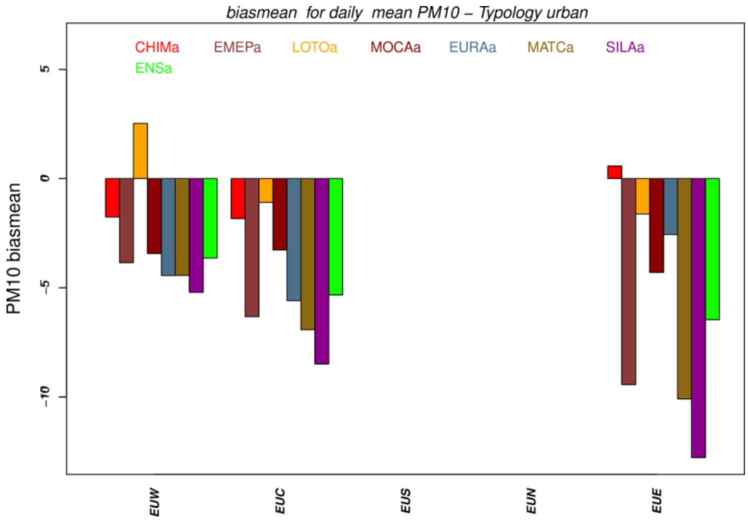


(b)

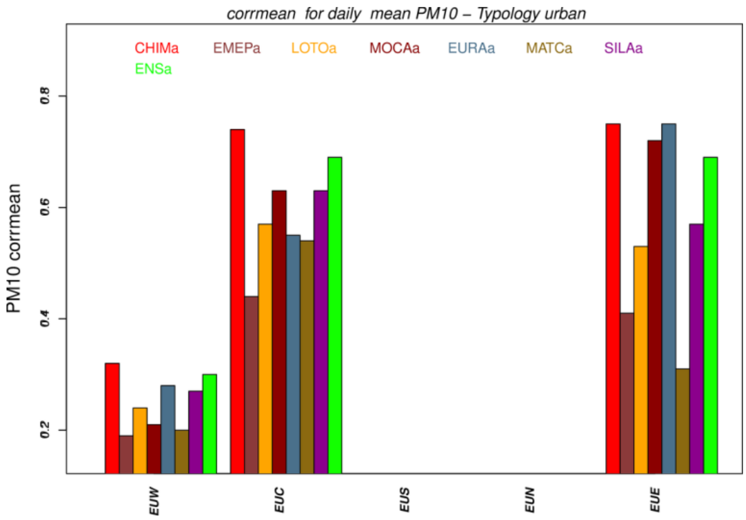


(c)

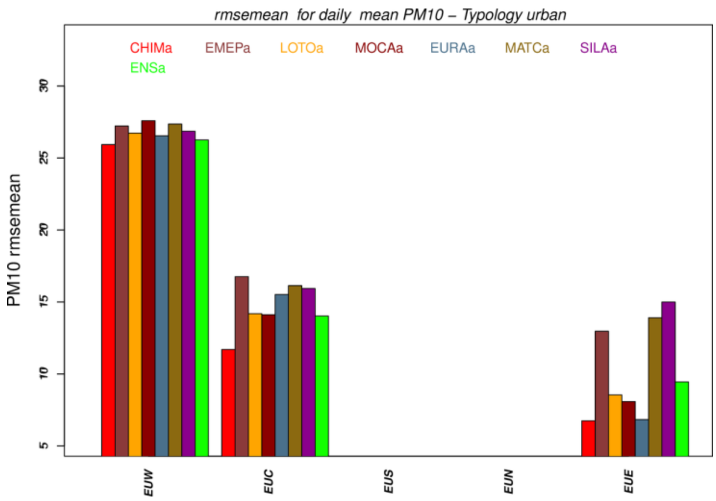
Figure 13 - CAMS regional interim reanalyses for predicting PM10 daily average over the year 2016 throughout European sub-regions (a) Bias (b) Correlation coefficient (c) RMSE at suburban stations



(a)

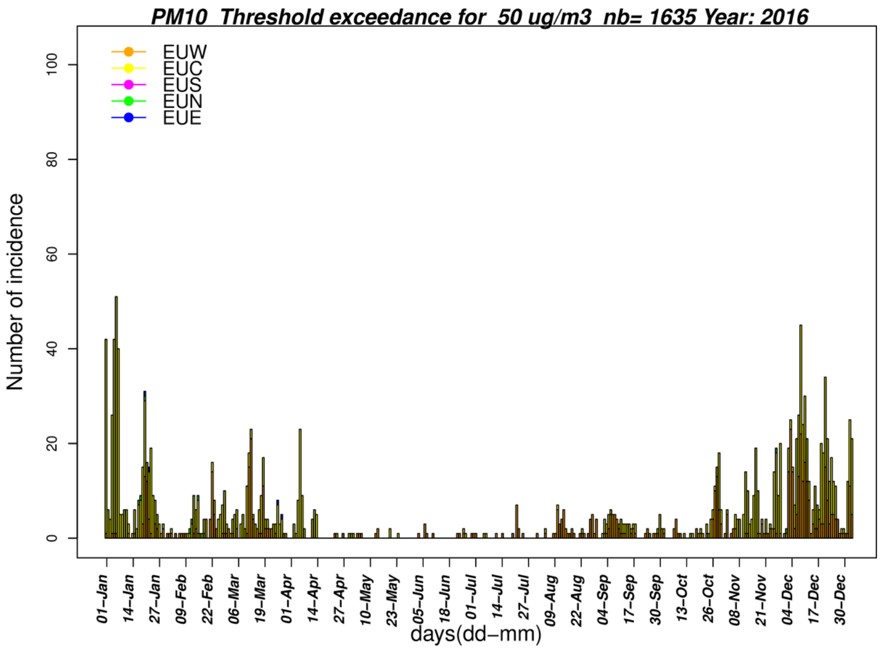


(b)



(c)

Figure 14 - CAMS regional interim reanalyses for predicting PM10 daily average over the year 2016 throughout European sub-regions (a) Bias (b) Correlation coefficient (c) RMSE at urban stations



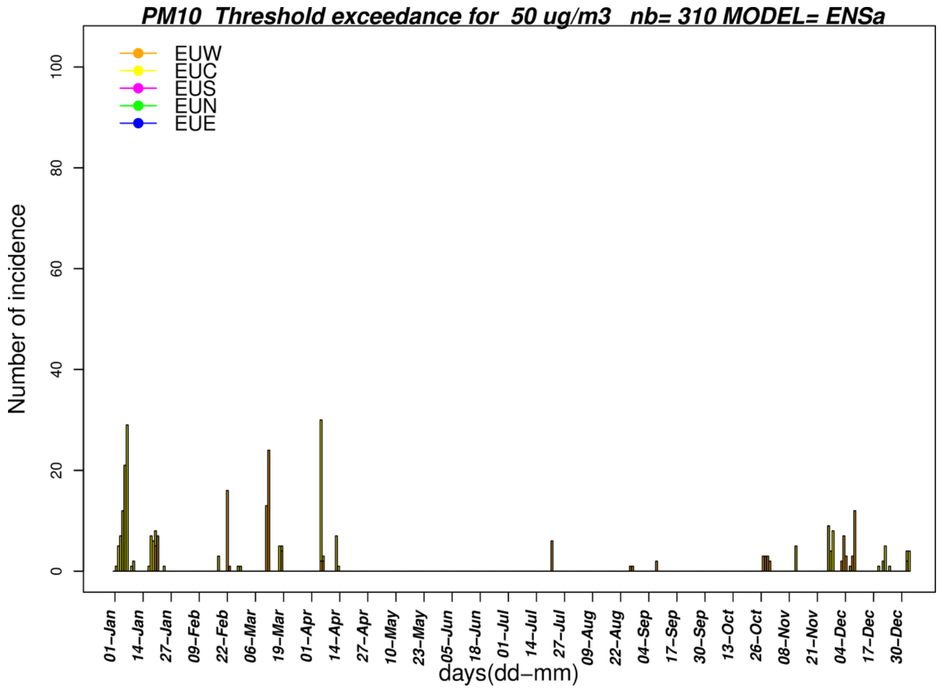


Figure 15 - Number of exceedance of daily limit value for PM10 in 2016 – observed (top) and modelled by the ENSEMBLE interim reanalyses

We can have a look at the indicators related to threshold exceedances. 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 (from the ENSEMBLE) are presented and compared. Logically, a number of exceedances are missed by the model, but as for ozone it does not give information about the model' skills or accuracy. The indicator does not allow to account for intrinsic uncertainty around the threshold value and is very severe for model evaluation. 19 % of the total number of exceedances have been correctly predicted by the interim reanalyses (it was a quarter in 2015) Nevertheless, we can see that a quarter of the total number of exceedances has been correctly predicted by the interim reanalyses. In particular, spring and autumn episodes in Western and Central Europe are correctly simulated. CHIMERE model managed to catch 70% of the exceedances (0), which demonstrates, as for ozone, that the adopted Ensemble approach in CAMS services is not really appropriate to monitor episode high concentration values.

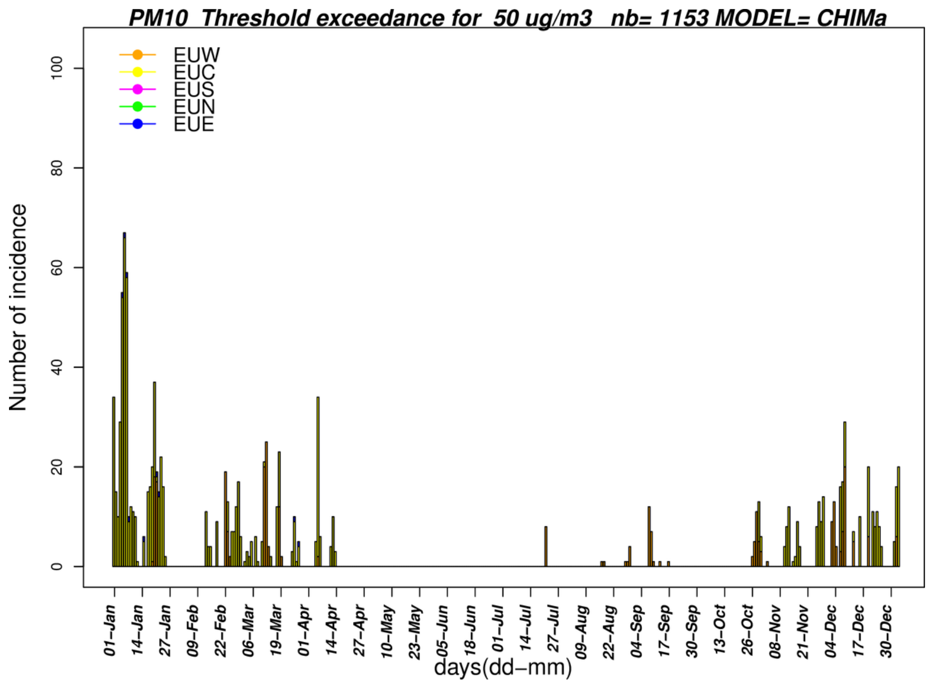


Figure 16 - Number of exceedance of daily limit value for PM10 in 2016 – observed (top) and modelled by the CHIMERE interim reanalyses

5. Performance Indicators for PM2.5

The evaluation of models' performances for PM2.5 was constrained by the low number of stations available, which is even lower than the number we had for the 2015 IRA verification. This limit is clearly highlighted considering the maps on 0. However, where there are some measurements, the results are rather good: bias ranges from -5 to 10 μg/m^3, correlation coefficient can exceed 0.8, except in some specific locations, and RMSE stays generally below 10 μg/m^3 except in Poland and specific location in Southern and Eastern Europe. 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. Evaluation results will be reconsidered carefully in the future evaluation of validated reanalyses.

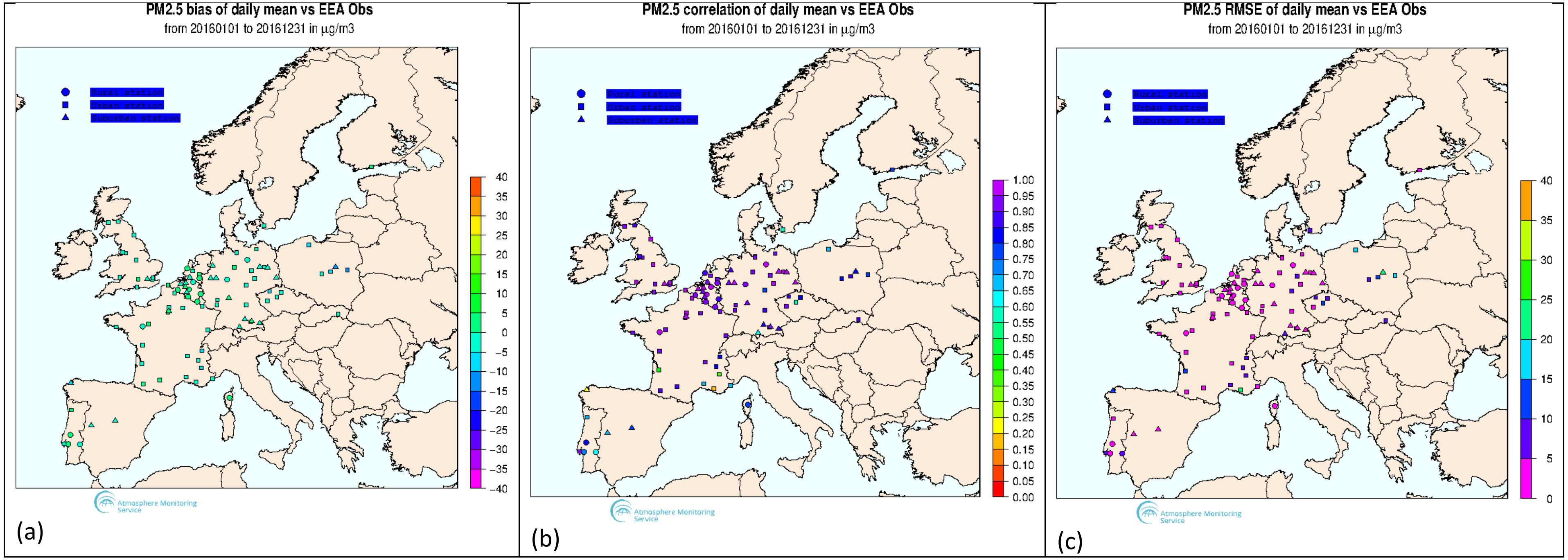


Figure 17 - Maps of Statistical scores of the ENSEMBLE interim reanalyses results against the observation validation dataset from the AQ e-reporting database for the PM2.5 daily average over the year 2016 Bias (a) Correlation coefficient (b) Root mean square error (c)

Conclusion

The present report presents an analysis of the performances of the interim air quality reanalyses throughout Europe, produced by the CAMS Regional service for the year 2016. It focuses on ENSEMBLE air quality reanalyses resulting from the combination of seven well-validated and documented chemistry-transport models results. We call here “interim” reanalyses 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 reanalyses as first guess of air pollution patterns and levels that developed in Europe in 2016. 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 reanalyses production.
INERIS run this process and computed a number of performance indicators and scores for ozone, nitrogen dioxide, PM10 and PM2.5 concentrations. They are presented in this report using maps, Taylor diagrams and histograms. The main conclusions arising from this analysis are the following:

- Two few up-to-date observation data was available to perform an extensive evaluation for ozone over the whole of Europe. Eastern and Southern European regions are not correctly covered, what is a pity since they correspond to areas where there are more uncertainties.
- In Western and Central Europe, where there is enough data for the evaluation the models performances are generally satisfactory and compatible with the state of the art with correlation. Best results are obtained for the ensemble: correlation coefficient ranging from 0.8 to 0.9 and root means square error lower than 15-17 $\mu g/m^3$.
- For ozone, the performances remain of lower quality than what can be achieved by validated re-analyses and are stable compared to 2015. Decreasing number of up-to-date observation data should be considered carefully in the future.
- For NO_2, performances are quite similar to those obtained in 2015 and consistent with the state of the art. There are few countries for which they are quite good (correlation factor of about 0.7, and RMSE lower than 15 $\mu g/m^3$) while other are more difficult to simulate. It is also linked to the “local nature”’ of the pollutant largely influenced by local sources that may be not correctly considered by the CAMS models’ resolution.
- 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 more consolidated in the future. Less observations were available than in 2015 with countries not covered at all: Italy, Slovakia, Romania, Bulgaria, Hungary, Serbia, Croatia, Slovenia …
- Saying, that, it should be noted that were there are enough measurements scores are rather good with RMSE of about 5-7 $\mu g/m^3$ and correlation factor reaching 0.8, except in Portugal, Poland and Czech Republic. However, the Ensemble does not always give the best results being hampered by the quite large variability of model responses. This point should be considered carefully in the future.

- 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 19 % of the exceedances of the PM10 daily limit values were correctly caught by the Ensemble (against about 70% for some individual models).
- Finally, very few PM2.5 measurement data was available for the evaluation making the conclusions difficult to highlight. Nevertheless, where there are measurements the statistical scores are quite satisfactory (correlation coefficient of about 0.8 and RMSE of about 5 $\mu g/m^3$).

Appendix: modelling set-up for the 2016 interim reanalyses

Modelling system: CHIMERE	
Horizontal resolution	0.15°x0.1°
Vertical resolution	Variable, 8 levels from the surface up to 500 hPa
Gas phase chemistry	MELCHIOR2, comprising 44 species and 120 reactions (Derognat, 2003)
Heterogeneous chemistry	NO_2, HNO_3, N_2O_5
Aerosol size distribution	9 bins from 10 nm to 40 µm
Inorganic aerosols	Primary particle material, nitrate, sulphate, ammonium
Secondary organic aerosols	Biogenic, anthropogenic
Aqueous phase chemistry	Sulphate
Dry deposition/sedimentation	Classical resistance approach
Mineral dust	Dusts are considered
Sea Salt	Inert sea salt
Boundary values	Values provided by CAMS global
Initial values	24h forecast from the day before
Anthropogenic emissions	MACC-TNO inventory 2011
Biogenic emissions	MEGAN
Assimilation system	
Assimilation method	Kriging-based analysis
Observations	Surface ozone and PM10
Frequency of assimilation	Every hour over the day before
Meteorological driver	00:00 UTC operational IFS forecast for the day before

Modelling system: EMEP	
Horizontal resolution	0.25° x 0.125° lon-lat (native model grid; downscaling to 0.1° x 0.1° is done in post-processing)
Vertical resolution	20 layers (sigma) up to 100 hPa, with approximately 10 in the Planetary Boundary layer
Gas phase chemistry	Evolution of the 'EMEP scheme', comprising 70 species and 140 reactions (Andersson-Sköld and Simpson, 1999; Simpson et al. 2012)
Heterogeneous chemistry	MARS (Binkowski and Shankar, 1995), oxidation of NO_2 by ozone on aerosols (night and winter)
Aerosol size distribution	2 size fractions PM2.5 and PM10-2.5
Inorganic aerosols	Thermodynamic equilibrium for the H^+-NH_4^+-SO_4^{2-}-NO_3^--H_2O system
Secondary organic aerosols	EmChem09soa (Simpson et al., 2012, Bergström et al, 2012)
Aqueous phase chemistry	SO_2 oxidation by ozone and N_2O_2
Dry deposition/sedimentation	Resistance approach for gases and for aerosol, including non-stomatal deposition of NH_3
Mineral dust	Boundary conditions from global C-IFS are used
Sea Salt	Boundary conditions from global C-IFS are used
Boundary values	Boundary conditions from global C-IFS are used
Initial values	24h forecast from the day before
Anthropogenic emissions	TNO-MACC emission data for 2011
Biogenic emissions	Included
Assimilation system	
Assimilation method	Intermittent 3d-var
Observations	NO_2 columns from OMI, NO_2 and O_3 surface concentrations from in situ data distributed by Meteo-France (with option to assimilate SO_2 surface concentrations)
Frequency of assimilation	6-hourly
Meteorological driver	12:00 UTC operational IFS forecast (yesterday's)

Modelling system: EURAD-IM	
Horizontal resolution	15 km on a Lambert conformal projection
Vertical resolution	23 layers up to 100 hPa Lowest layer thickness about 35 m About 15 layers below 2 km
Gas phase chemistry	RACM-MIM
Heterogeneous chemistry	N_2O_5 hydrolysis: RH dependent parameterization
Aerosol size distribution	Three log-normal modes: two fine + one coarse, fixed standard deviation
Inorganic aerosols	Thermodynamic equilibrium for the $H+-NH_4+-SO_4^{2}--NO_3--H_2O$ system
Secondary organic aerosols	Updated SORGAM module
Aqueous phase chemistry	10 gas/aqueous phase equilibria 5 irreversible S(IV) -> S(VI) transformations
Dry deposition/sedimentation	Resistance approach/size dependent sedimentation velocity
Mineral dust	DREAM model
Sea Salt	Included
Boundary values	C-IFS forecast
Initial values	3d-var analysis for the previous day
Anthropogenic emissions	TNO MACC-III (2011) inventory with 0.125° x 0.0625° resolution
Biogenic emissions	MEGAN V2.10 (Guenther et. al, 2012) GFAS wild fire emission data
Assimilation system	
Assimilation method	Intermittent 3d-var
Observations	NRT surface in situ data distributed by Meteo-France, NO_2 column retrievals from AURA/OMI and METOP/GOME-2, MOPITT CO profiles, IASI CO data
Frequency of assimilation	Hourly
Meteorological driver	WRF forced by the operational IFS analysis for the previous day

Modelling system: LOTOS-EUROS	
Horizontal resolution	0.25° (longitude) x 0.125° (latitude)
Vertical resolution	4 layers, top at 3.5 km above sea level
Gas phase chemistry	Modified version of the original CBM-IV
Heterogeneous chemistry	N_2O_5 hydrolysis
Aerosol size distribution	Bulk approach: PM2.5 and PM2.5-10
Inorganic aerosols	ISORROPIA-2
Secondary organic aerosols	Not included in this version
Aqueous phase chemistry	Linearized
Dry deposition/sedimentation	Resistance approach, following Erisman et al. (1994). Zhang (2001) deposition scheme is used for particles, explicitly including particle size and sedimentation
Mineral dust	Emissions after Marticorena & Bergametti (1995) with soil moisture inhibition as described by Fécan et al (1999).
Sea Salt	Parameterized emissions based on wind speed at 10m following (Monahan et al., 1986) and sea-surface temperature (Martensson et al., 2003).
Boundary values	C-IFS forecast (lateral and top).
Initial values	24h forecast from the day before
Anthropogenic emissions	TNO-MACC-III (2011) inventory
Biogenic emissions	Following Guenther et al. (1993) using 115 tree types over Europe
Assimilation system	
Assimilation method	Ensemble Kalman filter
Observations	Ensemble Kalman filter
Frequency of assimilation	In-situ surface observations (O_3, NO_2, PM10, PM2.5) distributed by Meteo-France and OMI NO_2 columns
Meteorological driver	Hourly, performed once a day for the previous day

Modelling system: MATCH	
Horizontal resolution	0.2° (since 10 January 2011)
Vertical resolution	26 levels (using reduction of IFS levels)
Gas phase chemistry	Based on EMEP (Simpson et al., 2012), with modified isoprene chemistry (Carter, 1996; Langner et al., 1998)
Heterogeneous chemistry	HNO_3-formation from N_2O_5; equilibrium reactions for NH_3-HNO_3
Aerosol size distribution	2bins: 0.01–2.5, 2.5–10 µm
Inorganic aerosols	Sulphate, Nitrate, Ammonium
Secondary organic aerosols	Not in operational version yet (VBS-scheme based on Bergström et al., 2012, is implemented)
Aqueous phase chemistry	SO_2 oxidation by H_2O_2 and O_3
Dry deposition/sedimentation	Resistance approach/size dependent sedimentation velocity
Mineral dust	Yes
Sea Salt	Yes
Boundary values	C-IFS forecast for the day before (zero boundaries for sea-salt)
Initial values	MATCH 24h forecasts from the day before
Anthropogenic emissions	TNO MACC-III emission inventory for the year 2011 (Kuenen et al., 2014; Denier van der Gon et al., 2015a), with 1/8° × 1/16° resolution
Biogenic emissions	Isoprene (Simpson, 1995; updated biogenic emissions of isoprene and monoterpenes, based on Simpson et al., 2012, implemented but not yet in operational version)
Assimilation system	
Assimilation method	3Dvar
Observations	NRT in-situ observations (O_3, NO_2, CO, SO_2, PM10, PM2.5) distributed by Meteo-France
Frequency of assimilation	Hourly, performed once a day for the previous day
Meteorological driver	IFS forecast and analyses 00Z for the same day (0.2°, 78 levels)

Modelling system: MOCAGE	
Horizontal resolution	0.2° regular lat-lon grid
Vertical resolution	47 layers up to 5 hPa Lowest layer thickness about 40 m About 8 layers below 2 km
Gas phase chemistry	RACM (tropospheric) and REPROBUS (stratospheric)
Heterogeneous chemistry	Only reactions on Polar Stratospheric Clouds (stratosphere) yet
Aerosol size distribution	Bins
Inorganic aerosols	Included: ISORROPIA module (Guth et al, 2016)
Secondary organic aerosols	Not implemented in current CAMS version
Aqueous phase chemistry	Aqueous reactions for sulphate production
Dry deposition/sedimentation	Resistance approach (Michou et al., 2004) for gases, (Nho-kim et al., 2005) for aerosol
Mineral dust	Included: see evaluation by Sic et al. (2014)
Sea Salt	Included: see evaluation by Sic et al. (2014)
Boundary values	Global CAMS chemical and aerosol fields
Initial values	24h forecast from the day before
Anthropogenic emissions	TNO (2011) inventory binned at 0.2° resolution
Biogenic emissions	Fixed monthly biogenic emission, based upon Simpson approach.
Assimilation system	
Assimilation method	3d-var
Observations	O_3, NO_2 and PM10 in situ data provided by INERIS are assimilated
Frequency of assimilation	Hourly
Meteorological driver	00:00 UTC operational IFS forecast

Modelling system: SILAM	
Horizontal resolution	0.1° regular lat-lon grid
Vertical resolution	69 layers for meteorological pre-processor (IFS hybrid levels 69 to 137, covering the troposphere), 9 layers for chemistry and vertical sub-grid-scale mixing calculations
Gas phase chemistry	CBM-4 gas-phase transformation, inorganic chemistry scheme with input to heterogeneous transformations (Sofiev, 2000)
Heterogeneous chemistry	Sofiev (2000)
Aerosol size distribution	Bins. Varies: for anthropogenic source, follows the emission of PM 2.5-10 split, for sea salt uses 5 bins from 10nm up to 30µm, dust is split into 4 bins from 10nm up to 30 µm
Inorganic aerosols	SO_4, NO_3, NH_4, Primary BC, OC, sea salt, desert dust
Secondary organic aerosols	Volatility Basis-Set. Being validated, not used in operational runs
Aqueous phase chemistry	SO_2 oxidation, nitrate formation (Sofiev, 2000)
Dry deposition/sedimentation	Resistance approach (Wesely et al., 1989) for gases, (Kouznetsov & Sofiev, 2012) for aerosol
Mineral dust	Taken from the C-IFS boundary conditions
Sea Salt	Updated source term Sofiev et al (2011)
Boundary values	C-IFS values for all available species, except for sea salt, which is taken from SILAM global forecasts
Initial values	24h forecast from the day before
Anthropogenic emissions	MACC-2011 inventory binned at 0.1° resolution
Biogenic emissions	Dynamic biogenic emissions, based upon Poupkou et al. (2010)
Assimilation system	
Assimilation method	Operational intermittent 3d-var for analysis; 4dvar for pollen reanalysis
Observations	In-situ surface data O_3, NO_2, PM2.5 operational; and vertically integrated columns in research mode (NO_2, AOD)
Frequency of assimilation	Hourly
Meteorological driver	00:00 UTC operational IFS forecasts up to +24h

ECMWF - Shinfield Park, Reading RG2 9AX, UK

Contact: info@copernicus-atmosphere.eu