Evaluation of air quality forecasts
CCA-WG1

Frédérik Meleux

Institut National de l’Environnement Industriel et des Risques
CCA WG1: Forecasting

Steering Group [JRC, VITO, NILU, U. Strasbourg, DG ENV, EEA]

<table>
<thead>
<tr>
<th>Time</th>
<th>Session</th>
<th>Speaker</th>
</tr>
</thead>
<tbody>
<tr>
<td>14:00</td>
<td>Introduction &amp; Evaluation of MACC forecast</td>
<td>F. Meleux</td>
</tr>
<tr>
<td>14:15</td>
<td>CALIOPE forecasts evaluated by DELTA</td>
<td>M.T. Pay</td>
</tr>
<tr>
<td>14:30</td>
<td>Evaluation of DELTA forecast functionality</td>
<td>J. Stocker</td>
</tr>
<tr>
<td>14:45</td>
<td>Discussion</td>
<td>All</td>
</tr>
</tbody>
</table>

14:00-15:00

WG1 – CCA: Forecasting
• Workplan 2014-2016
  – Q1: Can we use DELTA and its reporting template adapted for forecasting as in the current version? Or do we need to add other indicators/diagrams? If so which ones?
  – Q2: Do we need to add new MQO addressing the detection of threshold exceedances?
AQ forecasting system

• European and national air quality platforms produce daily forecasts for the D+0, D+1 and D+2.
  • provide every day information related to the air quality levels
  • targetted pollutants: O3, NO2, PM10, PM2.5
• In case of pollution episode:
  • Support to policy users
    – Provide recommendations to the public
    – to identify the likely causes
    – to assess population exposure
  – to set-up the efficient measures (short term action plans)
MACCII regional forecasts over Europe

- Based on an ensemble of 7 European models using the same input data (met, emissions, boundary conditions)

<table>
<thead>
<tr>
<th>Model</th>
<th>Current geometry</th>
<th>Assimilation method</th>
</tr>
</thead>
<tbody>
<tr>
<td>CHIMERE</td>
<td>0.1°, L8, top : 500hpa</td>
<td>Optimal Interpolation</td>
</tr>
<tr>
<td>INERIS, CNRS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>EMEP</td>
<td>0.25°x0.125°, L20, top : 100hpa</td>
<td>Variational 3d-var</td>
</tr>
<tr>
<td>met.no</td>
<td></td>
<td></td>
</tr>
<tr>
<td>EURAD</td>
<td>15km, L23, top : 100hpa</td>
<td>Variational, 3d-var</td>
</tr>
<tr>
<td>FRIUUK</td>
<td></td>
<td></td>
</tr>
<tr>
<td>L-EUROS</td>
<td>0.25°x0.125°, L4, top : 3.5km</td>
<td>Ensemble Kalman Filter</td>
</tr>
<tr>
<td>TNO, KNMI</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MATCH</td>
<td>0.2°, L40, top : 100hpa</td>
<td>Variational, 3d-var</td>
</tr>
<tr>
<td>SMHI</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MOCAGE</td>
<td>0.2°, L47, top : 5hpa</td>
<td>Variational, 3d-var</td>
</tr>
<tr>
<td>MF, CERFACS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SILAM</td>
<td>0.2°, L46/8, top : 100hpa</td>
<td>Variational, 4d-var</td>
</tr>
<tr>
<td>FMI</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
MACC AQ forecast evaluations on a daily basis

- MACCII operational evaluation relies on NRT obs data for computing:
  - Bias, RMSE and correlation
  - Calculated for the 96 hours
- Timeseries of scores from D+0 to D+3 averaged over the last week and the last three months
- Taylor diagrams
Additional statistical verification of model forecasts and ensemble done a posteriori: 6-monthly reports

- Evaluations focus on RMSE, bias, correlation averaged over quarters
  - Analyses of the model scores compared to the ones computed one year before and compared to the ensemble performances
  - Teams provide explanations regarding the changes on their model behaviours
MACCII policy

• There are no scores referring to daily values and to regulatory threshold
  – To provide useful products for national or local applications dealing with air quality management
  – scientific evaluation of the model forecasts and of the ensemble multi-model.

• This approach might change with the operational set-up of the Copernicus atmospheric service (in 2015)
  – ENSEMBLE can test the FAIRMODE procedures for forecasting evaluation
A target has been designed in the previous FAIRMODE phase for forecasting applications:

\[
\text{target} = \sqrt{\frac{1}{N} \sum_{1}^{N} (M_i - O_i)^2} - \sqrt{\frac{1}{N} \sum_{1}^{N} (O_{i-1} - O_i)^2}
\]

Stating that the worst acceptable model is the persistent model, so at a given station the forecast (D+0) provides the observation of the eve (D-1).
What are the policy objectives for using AQ forecasts:

- Predict the development of a pollution episode
  - How able are the model to reproduce the transition between non polluted regime to polluted regime?
  - How able the model are to detect / anticipate threshold exceedances?
    - The triggering of measures is based on threshold values
    - Use of the group function in Delta tool to select an adapted range of station for such evaluation (including observation uncertainty)
  - How stable are the forecast scores from D+0 to D+n (usually n = 2 or 3)?
    - Policy measures are more efficient when they are taken earliest – so the goal is really to provide confident forecasts at least at D+1 (D+2 would be even better)
Focus on daily mean and daily maximum

Various assessments should be considered:

- A global approach (whole domain and whole year)
- Sub-geographical groups (in linkages to the heterogeneity of the geographical distribution of the emission sources)
- Sub-temporal groups (seasonal evaluations...)

For instance, the model ability to reproduce PM10 levels may vary from one season to another.
French evaluation for episodes

- The evaluation of the French PREV’AIR system relies on the classical skill scores (RMSE...)
- Additional assessment are produced focusing on episodes

PM10 daily mean concentrations (µg/m³); March 2014
French evaluation for episodes

- Timeseries of the number of threshold exceedances
- Contingency tables
  - D+0 to D+2
  - Regional representation
Conclusions

• Forecast model should be assessed using classical scores and target defined in the delta tool -> but the best models doesn’t mean ability to detect threshold exceedances

• Using forecast for AQ managements, additional evaluations should be requested
  – Is there possible adaptation of the Delta target for forecasting threshold exceedances ?
  – How can we connect this forecasting target to indicators dedicated to threshold detection ability (contingency table, odds ratio skill scores ...) ?

• Do we need strict evaluation of the co-located threshold exceedances in obs & mod ?