Ensemble forecasting with machine learning algorithms for ozone, nitrogen dioxide and PM$_{10}$ on the Prev’Air platform

Debry, E.\textsuperscript{a,*}, Mallet, V.\textsuperscript{b,c,**}

\textsuperscript{a}INERIS BP 2, Parc Technologique Alata, 60550 Verneuil en Halatte, France  
\textsuperscript{b}INRIA, Domaine de Voluceau, Rocquencourt - BP 105, 78153 Le Chesnay Cedex, France  
\textsuperscript{c}CEREA, Joint Laboratory École des Ponts ParisTech/EDF R&D, Université Paris Est, Marne-la-Vallée, France

Abstract

This paper presents the application of an ensemble forecasting approach to the Prev’Air operational platform. This platform aims at forecasting maps, on a daily basis, for ozone, nitrogen dioxide and particulate matter. It relies on several air quality models which differ by their physical parameterizations, their input data and numerical strategies, so that one model may perform better with respect to observations for a given pollutant, at a given time and location. We apply sequential aggregation methods to this ensemble of models, which has already proved good potential in previous research papers. Compared to these studies, the novelties of this paper are the variety of models, the real operational context, which requires robustness assessment, and the application to several pollutants. In this paper, we first introduce the ensemble forecasting methods and the operational platform Prev’Air along with its models. Then, the sequential aggregation performance and robustness are assessed using two different data sets.

The results with the discounted ridge regression method show that the errors of the forecasts are respectively reduced by at least 29\%, 35\% and 19\% for hourly, daily and peak O$_3$ concentrations, by 19\%, 26\% and 20\% for hourly, daily and peak NO$_2$ concentrations, and finally by 17\%, 19\% and 11\% for hourly, daily and peak PM$_{10}$ concentrations. At last, we give a first insight of the ensemble ability to forecast threshold exceedances.

\textsuperscript{*}Tel : +33 3 44 61 82 17, Fax : +33 3 44 55 68 99, Email : Edouard.DEBRY@ineris.fr  
\textsuperscript{**}Tel : +33 1 39 63 55 76, Email : Vivien.Mallet@inria.fr

Preprint submitted to Atmospheric Environment  
April 10, 2014
Keywords: operational forecasting, ensemble forecast, sequential aggregation, Prév’Air platform, chemical transport models, ozone, nitrogen dioxide, particulate matter, threshold exceedance

1. Introduction

Operational air quality models are limited by a number of uncertainty sources. The physical processes are partially understood, and even when they are accurately described, air quality models simplify them either to make the computations tractable or to be consistent with the limited accuracy of the data. The numerous data required by an air quality model are associated with significant uncertainties, e.g., the real emissions may be within a range of ±50%, or even larger, from the model emissions. In addition, the numerical discretization of the models introduces further errors, especially in an operational context where the computational time is constrained.

As a consequence of these uncertainties, different air quality models were developed using different physical formulations, various numerical approximations and different data sources processed in numerous ways. Hence many operational models are nowadays capable of state-of-the-art day-to-day forecasting and are included in operational platforms designed to manage not only a single forecast but an ensemble of simulations. The operational platform Prév’Air, operated by the French National Institute for Industrial Environment and Risks (INERIS), is one of these platforms. It hosts the daily operational forecasts of three base models out of which five to eight simulations are produced. The five- or eight-member ensemble obviously brings more information than a single model, which can be exploited in order to design improved forecasts.

One option is simply to rely on the ensemble mean: every model in the ensemble is given the same weight, and the mean of the model simulations is taken as the new forecast. This approach was applied in air quality and brought some improvement for given targets, especially the correlation (Delle Monache and Stull, 2003; McKeen et al., 2005; van Loon et al., 2007; Potempski and Galmarini, 2009; Solazzo et al., 2012). Similarly, multimodel forecast based on ensemble median was carried out in Riccio et al. (2007). However, with Prév’Air ensemble, the ensemble mean essentially never beats the best model. Consequently distinct weights should be given to the models.
Bias correction techniques were used in Monteiro et al. (2013). The dynamic linear regression (West and Harrison, 1999) was applied to air quality in (Pagowski et al., 2006) to improve the predictions at observation stations. A least-square approach was applied in (Mallet and Sportisse, 2006) with the additional constraint that the same weights were applied at all observed locations. This constraint is retained in this work because it allows to apply the weights at non-observed locations, and therefore to forecast concentration fields.

We applied in an operational mode the machine learning algorithms introduced in (Mallet et al., 2009). Before any forecast, these algorithms attribute for each model a weight based on past observations and past forecasts. The corresponding linear combination produces an aggregated forecast in the form of a concentration field. This is carried out before any forecast and is therefore called sequential aggregation. The learning methods theoretically guarantee that, in the long run, the performance of the aggregated forecast will be at least as good as that of the best linear combination of models. This is satisfied whatever the sequence of observations and model forecasts may be, which makes the methods especially relevant for operational forecasting where robustness is an important feature. The procedure was applied from 2008 and is still applied in an operational setting for daily forecasts. It has been used to forecast ozone, nitrogen dioxide and particulate matter. Compared to (Mallet et al., 2009), the novelties lie in (1) the ensemble which is smaller and composed of other models, (2) the application in a real operational setting, and (3) the application to new pollutants. In addition, the behavior of the aggregated forecasts are further analyzed for the operational application, especially with the reliability of the weights at non-observed locations and the forecast of threshold exceedances.

The methods and their main features are presented in Section 2. The Section 3 explains the operational context: the models involved, the observation network and the forecasting setup. The results are analyzed in Section 4.

2. Ensemble forecasting methods

2.1. Strategy

We consider the observations of O₃, NO₂ and PM₁₀ and the corresponding models’ forecasts at the stations of a monitoring network. For a given pollutant, we denote \( x_{i,m}^{n,h} \) the forecast of model \( m \in \{1, \ldots, M\} \), at time step \( h \in \{1, \ldots, T\} \) and observation station \( i \in \{1, \ldots, N_h\} \). The time step \( h \)
stands either for an hour or a day in this paper. The number of available observation stations vary in time and is denoted \( N_h \). The objective of the ensemble forecast is to predict the observations \( y_{i,h} \) based on all the models’ forecasts. Each model is given a weight \( w_{m,h}^\text{m} \) so that the linear combination \( \hat{x}_{i,h} = \sum_{m=1}^{M} w_{m,h}^\text{m} x_{i,h}^\text{m} \) may be as close as possible to the observations \( y_{i,h} \) at time step \( h \). The weights \( w_{m}^\text{m} \) depend on time since they are computed before every forecast.

In operational context, at time step \( h - 1 \), the weights are computed for time step \( h \) and the system produces the forecast \( \hat{x}_{i,h} \). When the time \( h \) is reached, the observations \( y_{i,h} \) become available and the quality of the forecast \( \hat{x}_{i,h} \) can be evaluated. The weights are then corrected for the next time step, which allows us to produce the next aggregated forecast \( \hat{x}_{i,h+1} \). This procedure is repeated for every time step, hence we refer to it as sequential aggregation.

We decided to rely on weights that do not depend on the observation station. The rationale is that weights independent of the location may be applied at non-observed locations. Previous unpublished studies show that the weights learned at one observation station are irrelevant at other observed locations. On the contrary, when we constrain the weights to be the same at several observed locations, they can be applied at other locations — this is illustrated in Section 4.4. This allows us to combine not only the forecasts at the observation stations, but also the 2D concentration fields, at least in some region covered by the monitoring network.

It is always possible to apply the aggregation algorithm at each station independently of the others. In this case, the weights are only constrained by the observations of the station. It allows the aggregation method to produce weights that are adapted to the target station. It is useful if one wants to improve results at the observation stations only, for instance to forecast threshold exceedances. This is illustrated in Section 4.5.

### 2.2. Weights update

The weights are computed using the ridge regression (e.g., Cesa-Bianchi and Lugosi, 2006) with a discount in time as proposed in (Mallet et al., 2009). The method is therefore called discounted ridge regression (DRR). The weight vector \( w_{h+1} = (w_{h+1}^{1}, \ldots, w_{h+1}^{M})^T \) is defined as the minimizer of a
quadratic function:

\[ w_{h+1} = \arg\min_{u \in \mathbb{R}^M} \left[ \lambda \|u\|_2^2 + \sum_{h'=1}^{h} \left( 1 + \frac{\gamma}{(h+1-h')^2} \right) \sum_{i=1}^{N_{h'}} \left( \sum_{m=1}^{M} u_m x_{i,h'}^m - y_{i,h'} \right)^2 \right] \tag{1} \]

The quadratic function is the sum of a penalization term \( \lambda \|u\|_2^2 \) and a discrepancy term. The penalization tends to limit the magnitude and the variations of the weights. Its importance in the minimized function depends on the scalar \( \lambda > 0 \) which is a parameter to be determined. The second term measures the distance between all past observations \( y_{i,h'} \) (where \( h' \leq h \)) and the corresponding aggregations \( \sum_{m=1}^{M} u_m x_{i,h'}^m \) with the time-invariant weights \( u_1, \ldots, u_M \). The factor \( \left( 1 + \frac{\gamma}{(h+1-h')^2} \right) \), with \( \gamma > 0 \), is a discount factor that emphasizes the recent past so that the weights should be more adapted to the current atmosphere state. The parameter \( \gamma \) allows to adjust the effect of the discount. The greater \( \gamma \), the higher impact of the discount.

The parameters \( \lambda \) and \( \gamma \) are tuned so that there is the best balance between the penalization term, the discrepancy term and the discount impact. For the most recent observations, at \( h' = h \), the discount factor is \( 1 + \gamma \). For far away past, the discount factor tends to 1. Consequently, \( 1 + \gamma \) directly determines the relative importance given to the recent past. Once \( \gamma \) is set, \( \lambda \) needs to be adjusted so that the penalization contribution is large enough. Note that the contribution of the second term is determined by the number of observation stations, \( N_{h'} \). It also depends on the number of time steps \( h \) under the sum, but \( \lambda \) is anyway kept constant. In practice, it is found that \( \gamma \) is more important than \( \lambda \), and the sensitivity of the results to these parameters remains low (see Section 4.2).

2.3. Algorithm

Following (Mallet et al., 2009), the algorithm reads:

- The first forecast is \( \hat{x}_{i,1} = 0 \), since \( w_1^m = 0 \) for any \( m \).
- For \( h \in \{1, T - 1\} \)
  - Retrieve the new observations \( y_{i,h} \) (\( i \in \{1, \ldots, N_h\} \));
  - Compute \( w_{h+1} \) according to (1);
  - One-step forecast at station \( i \): \( \hat{x}_{i,h+1} = \sum_{m=1}^{M} u_{h+1}^m x_{i,h+1}^m \).
- \( n \)-step forecasts at station \( i \), with \( n > 1 \):

\[
\hat{x}_{i,h+n}^{(n)} = \sum_{m=1}^{M} w_m^{n+1} x_{i,h+n}^m;
\]

The last step is required only in an operational context for this is the only way to compute forecast several days ahead in the future (see Section 3.3). Also note that the weights do not depend on the station \( i \) and may therefore be used to combine the simulated concentrations fields, not just the point values at the observation stations.

### 2.4. Performance guarantee

The discounted ridge regression is a robust method that guarantees good performance in the long run. It can be shown (Cesa-Bianchi and Lugosi, 2006; Mallet et al., 2007a) that, in the long run, the performance of the aggregated forecasts \( \hat{x}_{i,h} \) is at least as good as that of any linear combination of the models with constant (in time) weights. In particular, the aggregated forecast will be at least as good as any model from the ensemble and the ensemble mean. This is guaranteed in the long run, whatever the sequence of observations and the sequence of models forecasts. The result is formulated in terms of mean quadratic error:

\[
\sum_{h=1}^{T} \sum_{i=1}^{N_h} \frac{1}{TN_h} (\hat{x}_{i,h} - y_{i,h})^2 - \min_{u \in \mathbb{R}^M} \left[ \sum_{h=1}^{T} \sum_{i=1}^{N_h} \frac{1}{TN_h} \left( \sum_{m=1}^{M} u_m x_{i,h}^m - y_{i,h} \right)^2 \right] \lesssim O \left( \frac{\ln T}{T} \right). \tag{2}
\]

### 3. Operational platform Prev’Air

Prev’Air is the French national operational system for air pollution forecasting. It has been settled in 2003 from an initiative of the French Ministry for Environment and is currently maintained by INERIS with financial support from the Ministry.

It aims at informing people at large and professionals about the pollutant concentrations behavior and tendencies on the short and long ranges. The French center for weather forecast, Météo-France, and the research center for Dynamic Meteorology (LMD) are closely involved in its development.

On a daily basis, it delivers forecasts up to three days ahead for ozone, nitrogen dioxide and particles (PM\(_{10}\) and PM\(_{2.5}\)) over France and Europe. Forecasts are freely available on the web site [http://www.prevair.org/](http://www.prevair.org/) under the form of peak and daily average concentration maps.
In this section, we aim to give some details about the daily Prev’Air operating and the various models running on it, which are then helpful to understand how ensemble forecast is applied on Prev’Air.

The Prev’Air system consists in several production lines, from raw data retrieving (e.g., meteorological fields and ground observations) to maps delivering. These production lines differ either by the physical model used, the space domain (continental, regional) and its resolution, or by the input data, mainly the meteorological forcing. At the time this article is written, we denote eight of them, listed in Table 1.

<table>
<thead>
<tr>
<th>Production line</th>
<th>Physical model</th>
<th>Domain (WGS84)</th>
<th>Resolution</th>
<th>Meteorological forcing</th>
</tr>
</thead>
<tbody>
<tr>
<td>AWM</td>
<td>CHIMERE</td>
<td>[15°W, 35°E] × [35°N, 70°N]</td>
<td>0.5° × 0.5°</td>
<td>GFS, MM5</td>
</tr>
<tr>
<td>AFM</td>
<td>CHIMERE</td>
<td>[5°W, 10°E] × [41°N, 52°N]</td>
<td>0.15° × 0.1°</td>
<td>GFS, MM5</td>
</tr>
<tr>
<td>AWMA</td>
<td>CHIMERE</td>
<td>[15°W, 35°E] × [35°N, 70°N]</td>
<td>0.5° × 0.5°</td>
<td>Arpege, MM5</td>
</tr>
<tr>
<td>AFMA</td>
<td>CHIMERE</td>
<td>[5°W, 10°E] × [41°N, 52°N]</td>
<td>0.15° × 0.1°</td>
<td>Arpege, MM5</td>
</tr>
<tr>
<td>MOCAG</td>
<td>MOCAGE</td>
<td>global</td>
<td>2° × 2°</td>
<td>ALADIN</td>
</tr>
<tr>
<td>MOCAE</td>
<td>MOCAGE</td>
<td>[11.75°W, 27.75°E] × [34.25°N, 61.75°N]</td>
<td>0.5° × 0.5°</td>
<td>ALADIN</td>
</tr>
<tr>
<td>MOCAF</td>
<td>MOCAGE</td>
<td>[4.95°W, 9.95°E] × [41.05°N, 51.95°N]</td>
<td>0.1° × 0.1°</td>
<td>ALADIN</td>
</tr>
<tr>
<td>Polair3D</td>
<td>Polair3D</td>
<td>[10.5°W, 22.5°E] × [35°N, 57.5°N]</td>
<td>0.5° × 0.5°</td>
<td>GFS, MM5</td>
</tr>
<tr>
<td>ASAFM</td>
<td>AFM + Kriging</td>
<td>[5°W, 10°E] × [41°N, 52°N]</td>
<td>0.15° × 0.1°</td>
<td>GFS, MM5</td>
</tr>
</tbody>
</table>

Table 1: Production lines.

The AWM, AWMA and AFMA production lines which use the CHIMERE physical model (IPSL/LMD) are operated within INERIS. Raw meteorological data come from either the GFS (NOAA) or Arpege (Météo-France) global weather models. The latter benefits from a refined grid over the French domain as it is used for operational forecasts by Météo-France. Raw fields are then refined on each domain with the MM5 meteorological model.

Additionally, two other physical models run on the platform: MOCAGE (Météo-France) and Polair3D, a model from Polyphemus (CEREA\(^1\)). Unlike CHIMERE, they are not operated within INERIS, but respectively at Météo-France and CEREA. Simulation results are then uploaded to INERIS.

\(^1\)Teaching and Research Center for Atmospheric Environment – http://cerea.enpc.fr/
For each production line, at one day D, the simulation period runs from the day before to three days ahead, we usually refer to them as D-1, D+0 (=D), D+1 and D+2 days. Production lines are launched at D-1 in the evening as soon as required data become available; maps are uploaded on the web site early on D+0.

Furthermore, observational data is retrieved from French and European networks early on D+0, which allows to deliver daily statistics for some CTMs\(^2\) performance evaluation. Among production lines, ASAFM stands as the official production line whose results are produced by correcting concentrations of AFM production line with the kriging of innovations (Honoré et al., 2004) and then used to carry out the maps delivered to the public at large.

In the framework of ensemble modeling, the ASAFM production line may be viewed as an independent model as long as observations used by the kriging of innovations are not the same as those used by the DRR algorithm. This variety in production lines is an opportunity for sequential aggregation. Indeed, the greater variability in models, the higher potential of the ensemble forecasting with sequential aggregation. From now on, the term “model” will refer to one production line: the physical model, its input data and space resolution. Thus, we view the four production lines using CHIMERE as four different models.

3.1. Physical models

The ensemble variety is also that of the three underlying physical models (CTMs). These may use different physical parameterizations for the same process and employ different numerical strategies to solve the coupled advection-diffusion-reaction equations, depending on the particular goal according to which they were developed. In this section, we would like to emphasize these differences and similarities. Table 2 displays the most significant physical parameterizations for each CTM.

\(^2\)Chemistry-Transport Models
<table>
<thead>
<tr>
<th>Parameter</th>
<th>CHIMERE</th>
<th>MOCAGE</th>
<th>Polair3D</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gas chemical mechanism</td>
<td>MELCHIOR (Derognat, 1998)</td>
<td>RACM &amp; REPROBUS (Lefèvre et al., 1999)</td>
<td>RACM (Stockwell et al., 1997)</td>
</tr>
<tr>
<td>Cloud attenuation</td>
<td>(Madronich, 1993)</td>
<td>(Brasseur et al., 1998)</td>
<td>RADM (Madronich, 1987)</td>
</tr>
<tr>
<td>Horizontal transport</td>
<td>Van Leer or PPM (Williamson and Rash, 1989)</td>
<td>(Louis, 1979)</td>
<td>DST (Verwer et al., 2002)</td>
</tr>
<tr>
<td>Gas dry deposition</td>
<td>(Wesely, 1989)</td>
<td>(Wesely, 1989)</td>
<td>(Zhang et al., 2003)</td>
</tr>
<tr>
<td>Wet scavenging</td>
<td>in/below cloud (Bessagnet et al., 2004)</td>
<td>convective/stratiform precipitation (Mari et al., 2000)</td>
<td>in/below cloud (Seinfeld and Pandis, 1998)</td>
</tr>
<tr>
<td>Aqueous mechanism</td>
<td>(Bessagnet et al., 2004)</td>
<td>–</td>
<td>VSRM (Fahey and Pandis, 108)</td>
</tr>
<tr>
<td>Heterogeneous chemistry</td>
<td>(Jacob, 2000)</td>
<td>–</td>
<td>(Jacob, 2000)</td>
</tr>
<tr>
<td>Aerosol dynamics</td>
<td>sectional (8) (Bessagnet et al., 2004)</td>
<td>sectional (20), but dust only</td>
<td>sectional (5) (Debry et al., 2007)</td>
</tr>
<tr>
<td>Inorganic equilibrium</td>
<td>Isorropia (v1.6)</td>
<td>–</td>
<td>Isorropia (v1.7) (Nenes et al., 1998)</td>
</tr>
<tr>
<td>Secondary organic chemistry</td>
<td>(Pun et al., 2002)</td>
<td>–</td>
<td>(Pun and Seigneur, 2007)</td>
</tr>
<tr>
<td>Aerosol dry deposition</td>
<td>(Seinfeld and Pandis, 1998)</td>
<td>(Nho-Kim et al., 2004)</td>
<td>(Zhang et al., 2001)</td>
</tr>
<tr>
<td>Vertical levels</td>
<td>8 from 995 hPa to 500 hPa two step (Verwer, 1994)</td>
<td>47 from 995 hPa to 5 hPa operator splitting</td>
<td>1st order operator splitting $\Delta t = 600$ s</td>
</tr>
<tr>
<td>Numerical time solver</td>
<td>EMEP (2007)</td>
<td>(Dentener et al., 2005)</td>
<td>EMEP</td>
</tr>
<tr>
<td>Anthropogenic emissions</td>
<td>Carbon: (Liouss et al., 1996)</td>
<td>(Guenther et al., 1995) sea salt (Monahan et al., 1986)</td>
<td>(Simpson et al., 1999) sea salt (Monahan et al., 1986)</td>
</tr>
<tr>
<td>Biogenic emissions</td>
<td>(Guenther et al., 1995) sea salt (Monahan et al., 1986)</td>
<td>(Guenther et al., 1995) dust emissions</td>
<td></td>
</tr>
<tr>
<td>Boundary conditions</td>
<td>MOZART &amp; LMDz-INCA or nested simulations</td>
<td>nested simulation</td>
<td>MOZART (Horowitz et al., 2003) &amp; GOCART (Chin et al., 2000)</td>
</tr>
<tr>
<td>Land use</td>
<td>GLCF</td>
<td>GLCF</td>
<td>USGS</td>
</tr>
</tbody>
</table>

Table 2: Physical parameterizations of CHIMERE, MOCAGE and Polair3D as they are used in the Prev’Air production lines.
CHIMERE is a tropospheric photochemical transport model built mainly in order to forecast ozone production over urban areas (Vautard et al., 2001; Schmidt et al., 2001). MOCAGE (Model Of Atmospheric Chemistry At larGE scale) is a multiscale, from regional to global, chemistry and transport model (Josse et al., 2004; Dufour et al., 2005; Bousserez et al., 2007) designed with the aim to provide routine “chemical weather” forecasts. Polair3D is a tropospheric photochemical transport model from regional to continental scale (Tombette and Sportisse, 2007; Sartelet et al., 2007) which is part of the Polyphemus modeling platform (Mallet et al., 2007b).

As the three CTMs are primarily intended to follow gas pollutant concentrations, the gas chemical mechanism is a key component. CHIMERE uses a simplified version of MELCHIOR (Lattuati, 1997) with 44 species and 116 reactions which mainly targets ozone concentrations. MOCAGE and Polair3D use the RACM mechanism which includes 77 species and 237 reactions. It is to note that if both mechanisms share reactions, the rate parameterizations for several of them are not the same, e.g., this the case for $\text{O}_3 + \text{NO} \rightarrow \text{NO}_2$. In addition, MOCAGE uses the REPROBUS mechanism for the stratospheric chemistry.

Horizontal transport or advection is also a critical process in CTMs. CHIMERE uses two numerical schemes, Van Leer or Piecewise Parabolic Method, depending on how slow is the chemical species. Polair3D employs a 3$^{rd}$ order Direct Space Time method with Koren flux limiter for all species, whereas MOCAGE is based on a semi-Lagrangian scheme. Most numerical schemes are designed to be mass conservative, however there may be significant differences between Eulerian (CHIMERE and Polair3D) and semi-Lagrangian approaches in terms of numerical diffusion.

There exist two widely used parameterizations for vertical diffusion in CTMs. CHIMERE and Polair3D share the same. Pollutant concentrations are known to be very sensitive to the vertical diffusion, especially within the atmospheric boundary layer, hence at ground level. Both parameterizations are designed in different ways. Briefly, while Louis (1979) relies on a detailed boundary layer structure, dependent on the CTM vertical discretization, Troen and Mahrt (1986) assumes a particular profile shape for the diffusion coefficient, which leads to distinct computed values.

Removal of pollutants from atmosphere may occur either by dry deposition at the ground surface or by wet scavenging. CHIMERE and MOCAGE share the same parameterization for gas dry deposition, whereas Polair3D uses Zhang et al. (2003). Both parameterizations are based on the multi-
ple resistance analogy in which the canopy resistance is partitioned between
diurnal stomatal and non-stomatal uptakes. There lie differences between
parameterizations, the latter one improving the non-stomatal part.

CHIMERE and Polair3D have a similar wet scavenging approach, including in- and below-cloud removal, which differs from that of MOCAGE. The
latter furthermore distinguishes convective and stratiform precipitations.

Following down Table 2, we come to multiphase chemistry and particle
(cloud droplet and aerosols) modeling, which have often been added later in
CTMs. Aqueous chemistry is mandatory in order to follow sulfate production
in atmosphere, as SO$_2$ dissolves in cloud droplets. CHIMERE and Polair3D
account for sulfate aqueous chemistry and heterogeneous reactions which
produce nitric acid from NO$_2$.

Both CTMs have a sectional aerosol module with respectively 8 sections
from 40 nm to 10 $\mu$m for CHIMERE and 5 from 10 nm to 10 $\mu$m for Polair3D.
These modules include coagulation, condensation/evaporation of semivolatile
species and nucleation processes. Coagulation is usually based on Brownian
motion (Seinfeld and Pandis, 1998) and has a weak effect on mass concentra-
tions relatively to other processes. There are numerous nucleation parameter-
izations, known to differ by as much as 18 orders of magnitude (Zhang et al.,
2010a). CHIMERE employs the binary H$_2$O−H$_2$SO$_4$ parameterizations Kul-
mala et al. (1998); Polair3D gives the choice between such a parameterization
(Korhonen et al., 2003) or a ternary H$_2$O−H$_2$SO$_4$−NH$_3$ one (Napari et al.,
2002). Condensation/evaporation is driven by the thermodynamic equilib-
rium between gas and particulate phases. For CHIMERE and Polair3D, this
one is respectively solved by Isorropia and AEC for inorganic and organic
species. There are small differences depending on the version and options
used for each module. A more consistent difference is that while this process
is time-resolved for CHIMERE, assuming equilibrium is not reached within
one time step, Polair3D usually assumes this equilibrium is achieved at least
for smallest sections (Edouard Debry and Sportisse, 2006). Additionally,
Polair3D takes into account the activation process which transfers instantly
the content of so-called activated aerosols (CCN) to cloud droplets when the
liquid water content reaches a critical value. Today, the CTM MOCAGE in-
cludes an aerosol module. Nevertheless, at the time these tests were carried
out, only dust was available in MOCAGE outputs (Martet et al., 2011).

Apart from physical parameterizations, CTMs also strongly differ in the
numerical strategy used to solve advection-diffusion-reaction equations. There
are two main approaches. On the one hand, one may compute evolution
rates for all concentrations and hand them over to an independent numerical
solver which takes care of the time integration over one time step\(^3\), possibly sub-cycling and automatically adapting the inner time step. This is the
CHIMERE choice. On the other hand, one may choose an operating splitting
technique, integrating successively each process with a specialized numerical
solver, as Rosenbrock implicit method for Polair3D gas chemistry.

In the lowest part of Table 2, we detail some input data which, although
not being part of the CTM itself, are usually strongly associated with it. That
is to note CHIMERE uses an improved emission inventory for carbon par-
ticles. Together with MOCAGE, they use their own continental simulations
to provide initial and boundary conditions for their regional domain. The
land use cover used by Polair3D is USGS (http://landcover.usgs.gov/).

Finally, any CTM may have a physical parameterization or a numerical
strategy which will perform better, with respect to observations, for a given
pollutant, and at a given time and location. This is what we illustrate in
Figures 1. Each model (as denoted in Table 1) is associated with one color.
At any time, each space location of the domain is given the color of the model
which competes best with respect to nearest observations. That is to note
how rapidly the best model changes with space and time. This point was
already highlighted by Garaud and Mallet (2010), but with a great number
of models. We show here it also remains true with a limited number of
models. In the sequel, the various models will be made anonymous as we
are not interested in each model performance, but rather in the ensemble
performance with respect to the best model.

\(^3\)For CHIMERE, the main step is 1 hour.
3.2. Observation/monitoring network

As for data assimilation, the observation network is a key point for ensemble modeling in an operational context: the denser the network, the more reliable the field aggregation.

Prev’Air benefits from the BASTER\textsuperscript{4} network which gathers stations positioned by several associations for air quality monitoring (AASQA) in their respective administrative region. Data from these stations are collected in near real-time by the French environment and energy management agency (ADEME\textsuperscript{5}) and put at the disposal of several public organizations. Since March 2011, this network is directly managed by INERIS.

Figure 2 displays the 729 available stations from the BASTER network over the French metropolitan territory in 2008/2009: 58% of the stations are classified as urban (inside the city itself or its suburbs), 16% as industrial, 12% are devoted to road traffic and 10% to rural areas. The station network appears then to be mostly concentrated in urban and industrial areas and to be relatively sparse between them. This can be highlighted by computing the average station-to-network distance and its standard deviation. The distance of a station to the rest of the network is defined as the distance to the closest station. The average of this distance on all stations is $0.09^\circ \pm 0.12^\circ$.

\textsuperscript{4}http://www.buldair.org/
\textsuperscript{5}http://www.ademe.fr/
Figure 2: BASTER monitoring network.
Provided data consists in hourly concentrations for $O_3$, $NO_2$ and $PM_{10}$. These hourly concentrations result from the average of 4 consecutive measurements, taken every quarter of hour. Whether to compare these concentrations to the computed concentrations at the beginning or the end of the given hour is the choice of the modeler and has never seemed to affect significantly the model/measure comparisons.

In this operational framework, the reliability of the observations has to be addressed, since they are directly targeted by the ensemble forecasting method, which covers various aspects: the instrumental errors (errors of the measuring device) at first, then the representativeness of the monitoring station (e.g., when it is influenced by very local sources) and at last, how fast the observational data is faithfully retrieved.

Measuring devices always come with an error or uncertainty range, even when one-hour averages are used. We point out that, in the weights computation (Eq. 1), the observations from all available stations are included. We can speculate that the various observational errors tend to cancel out in the global minimization, thus reducing the impact of non-representative or corrupted values.

In practice, we first filter out negative or uncommonly high values. In addition, only stations which provided observations for at least 30% of the given period are considered reliable and included in the weights computations and performance evaluation. There usually remain 431 stations for $O_3$, 469 for $NO_2$ and 358 for $PM_{10}$. The differences are due to the fact that not all pollutants are monitored at each station.

### 3.3. Operational set up

Ensemble forecasts have been running on a daily basis on the operational platform Prev’Air since 2008 for pollutants $O_3$, $NO_2$ and $PM_{10}$. In the sequel, we detail the set-up applied for any species and at a given date, which is illustrated by Figure 3. Hours are specified in local time.

The first step consists in gathering all data required for the learning period. As explained in Section 3, each model of the Prev’Air platform is run on a daily basis at D-1, thus giving a forecast for D+0 (today), D+1 (tomorrow) and D+2 (the day after tomorrow). This then produces several interlaced predictions, each day having its D+0, D+1 and D+2 forecasts from previous simulations. For each model, the D+0 forecasts are concatenated from the beginning of learning period up to yesterday. If it is not available, the D+1 is used instead, or the D+2 if the D+1 is also missing. Thus,
according to our experience, most of unplanned breaks in the computing chain can be overcome, avoiding “holes” in the learning period. At the same time, observational data is retrieved for the same period. The observations are retrieved until 5h00–6h00 a.m. of D+0, in near real time, so that all D-1 observations until 23h00 at best are available and assimilated. The length of the learning period is at least 30 days and grows by one day each day.

Then, the computation of the aggregation weights begins with the discounted ridge regression algorithm (see Section 2). Weights may be computed on a daily or hourly basis. When on an hourly basis, the computation is done independently for each hour in the day, retrieving only the given hour in each day of the learning period; when on a daily basis, that is to say when weights are computed for the daily peak or mean, the simulation and observation data are filtered accordingly. The values of the algorithm’s parameters are those recommended in Mallet et al. (2007a): $\lambda = 1000$ and $\gamma = 100$. Ensemble methods are implemented in C++ and are not time consuming, compared to model simulations. As an example, the CPU time for ozone peak and 108 days is 15–20 seconds\(^6\) on a Dual-Core AMD Opteron (2218) @2.6GHz. The CPU time does not grow significantly with the number of observations.

The result of the computation is a sequence of weights (one per model) for today (W+0), with which the D+0 of each model’s latest simulation are aggregated, thus producing ensemble forecasts at the stations and an aggregated map (using the same weights), for today and the given species. These are ready before 7h00 a.m. The same weights are again used to aggregate the D+1 and D+2 forecasts of each simulation.

\(^6\) Elapsed real time without parallel computing.
Figure 3: Operational set-up for a given day.
This set-up implies two major assumptions: first, the weights are applied outside of the (observed) locations for which they were computed and second, they are applied on further days ahead of the one for which they were computed.

These assumptions have been introduced in Section 2 and already addressed in previous publications (Mallet et al., 2009), nevertheless they appear even more crucial in this operational context where unplanned breaks in the computing chain or loss of internet connection are not rare and all required data may not always come in time. As an example, if the latest observations could not be retrieved, still weights can be computed with older observations, but one may wonder whether it is still valid to apply them to the current day. In the same manner, if a given area has very few observational data, one may ask to what extent weights may be applied within this area.

In the framework of this article, we placed ourselves in the same forecast conditions, except that observations and simulations are supposed to be always on time. We then addressed the previous questions with detailed performance analysis, and we considered new pollutants, compared to previous studies.

### 3.4. Pollutants and periods of interest

The major concern of the Prev’Air platform is the ozone concentration level. Indeed, ozone is an atmospheric oxidant well-known for its health adverse effects. It is produced in the troposphere through complex photochemical reactions between nitrogen oxides and volatile organic compounds (VOC). These chemical components, known as ozone precursors, are emitted in large amount above cities, thus producing ozone during summer due to solar radiation. Above cities, ozone reacts fastly with emitted nitrogen monoxide, nevertheless when pollutants are transported away from cities, where nitrogen monoxide is less present, ozone concentration begins to grow up to form large plumes which regularly occur in the south-east of France.

The ozone concentration level may then exceed levels defined by the European Community (EC, 2002): the *information threshold* and the *alert threshold*, respectively set to 180 µg m$^{-3}$ and 240 µg m$^{-3}$ on an hourly basis. As soon as these thresholds are exceeded, specific actions must be taken by public authorities.

Nitrogen dioxide and particulate matter are also monitored by the Prev’Air platform. NO$_2$ and particles are usually present in large amount within cities
mainly due to road traffic emissions and wood heating, leading to visible particle pollution fogs above cities. Especially, particles may be responsible for multiple adverse effects on health for the ultrafine ones (Oberdörster et al., 2005), on visibility and building surface for coarse ones. The NO$_2$ and particulate matter concentrations are usually higher during winter due to higher emissions and a lesser solar radiation. The European Community has also determined thresholds for these pollutants (EC, 1999; Air Quality Directive, 2008): NO$_2$ and PM$_{10}$ daily concentrations must not exceed respectively 200 $\mu$g m$^{-3}$ more than 18 times per year and 50 $\mu$g m$^{-3}$ more than 35 days per year.

As a consequence, we focused on summer for ozone, and winter for NO$_2$ and PM$_{10}$. For each season, two periods have been studied in order to ensure the results not being dependent on the period. Table 3 summarizes the chosen periods. The learning period lasts at least 30 days for all periods, the beginning date depends on the data availability. The evaluation period starts just after the end of learning period, it lasts 107/108 days for summer and 121 days for winter.

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Beginning learning period</th>
<th>Beginning evaluation</th>
<th>End (incl.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>2008-05-17</td>
<td>2008-06-16</td>
<td>2008-09-30</td>
</tr>
<tr>
<td>S2</td>
<td>2009-05-01</td>
<td>2009-05-31</td>
<td>2009-09-15</td>
</tr>
<tr>
<td>W1</td>
<td>2008-11-02</td>
<td>2008-12-02</td>
<td>2009-04-01</td>
</tr>
<tr>
<td>W2</td>
<td>2009-11-01</td>
<td>2009-12-01</td>
<td>2010-03-31</td>
</tr>
</tbody>
</table>

Table 3: Summer and winter time periods.

4. Results

In this section, we detail the results obtained during years 2008/2010 on the Prev’Air platform for the pollutants and periods of interest. First, forecasting scores are computed. Then, we assess to which extent one may rely on ensemble forecasts for operational purposes. Finally, we investigate the threshold exceedance detection with the ensemble.

4.1. Forecasting scores

The performance indicators used in the sequel include the main ones with which models running on Prev’Air are daily evaluated (normalized mean
Their exact definitions are given in Appendix A. The forecasting scores are computed for the whole evaluation period (see Table 3). When computing bias factor and NMSE, ozone concentrations lower than 40 µg m$^{-3}$ and those of NO$_2$ and PM$_{10}$ lower than 1 µg m$^{-3}$ are excluded. The filtering is based on observation concentrations.

Tables 4, 5 and 6 respectively display the performance for pollutants O$_3$, NO$_2$ and PM$_{10}$. The performance of the discounted ridge regression (DRR) can be compared with the best model (BM) of the ensemble according to the root mean square error (RMSE) for the given pollutant and the best constant linear combination (BCLC) which minimizes the least-square error with time-invariant weights.

Firstly, as a general remark, one can note that the BCLC always performs better than the BM, which is essentially a consequence of the BCLC definition: this one is the best of all possible combinations (in the least square sense), while the best model is simply one of the possible combinations. Secondly, looking at the RMSE, the DRR always succeeds to compete with the BCLC, hence fulfilling the asymptotic theoretical guarantee (2).

For ozone, the greatest improvement are observed for the daily average concentrations whose RMSE is reduced by at least 35%. Hourly concentrations are strongly improved as well, with a 29% decrease of the RMSE. The RMSE on the peak concentrations decreases by at least 19% over both periods.

Other indicators are improved, even though they are not directly targeted by the DRR algorithm (Eq. 1): the simulated mean becomes almost equal to the observed one, whereas it is shifted by at least +7 µg m$^{-3}$ for the best model (except fo S2 peak), and the correlation lies in the range 70–85%, whereas it stays below 70% for the best model, except for peak concentrations in period S1.

For NO$_2$, RMSE is improved by at least 19% on both periods for hourly concentrations. This improvement is even greater for peak (20%) and daily (26%) concentrations. The simulated mean becomes closer to the observed mean, and is even equal for daily concentrations in period W2. Bias factors get superior and closer to unity for all frequencies and periods. Nevertheless, NMSE is worsened in most cases, except for W1 peak concentrations.

A similar analysis can be carried out for PM$_{10}$. RMSE is improved by at least 17%, 19% and 11% on both periods for respectively hourly, daily and peak concentrations. NMSE is reduced in most cases except for W1 peak
and W2 hourly concentrations. Correlation is also significantly improved, acquiring at least 17 points for hourly concentrations, 13 for daily concentrations and 5 for peak ones. It even exceeds 50% for daily concentrations. The bias factor displays the same behavior as for NO₂; it is generally inferior to unity for best model and becomes greater for DRR.

<table>
<thead>
<tr>
<th>Period</th>
<th>Frequency</th>
<th>Model</th>
<th>Mean</th>
<th>RMSE</th>
<th>NMSE</th>
<th>Correlation (%)</th>
<th>Bias Factor</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>hourly</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>BM</td>
<td>67.6 (+11.1)</td>
<td>27.8</td>
<td>0.35</td>
<td>56.8</td>
<td>1.07</td>
</tr>
<tr>
<td></td>
<td></td>
<td>BCLC</td>
<td>57.2 (+0.7)</td>
<td>18.8</td>
<td>0.25</td>
<td>76.9</td>
<td>0.93</td>
</tr>
<tr>
<td></td>
<td></td>
<td>DRR</td>
<td>58.2 (+1.7)</td>
<td>18.4</td>
<td>0.24</td>
<td>78.4</td>
<td>0.95</td>
</tr>
<tr>
<td>S1</td>
<td>daily</td>
<td>BM</td>
<td>67.6 (+11.1)</td>
<td>21.2</td>
<td>0.34</td>
<td>46.9</td>
<td>1.16</td>
</tr>
<tr>
<td></td>
<td></td>
<td>BCLC</td>
<td>57.0 (+0.5)</td>
<td>13.6</td>
<td>0.19</td>
<td>64.9</td>
<td>0.98</td>
</tr>
<tr>
<td></td>
<td></td>
<td>DRR</td>
<td>58.1 (+1.5)</td>
<td>12.8</td>
<td>0.19</td>
<td>70.5</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td>peak</td>
<td>BM</td>
<td>94.8 (+7.7)</td>
<td>18.8</td>
<td>0.25</td>
<td>77.1</td>
<td>1.11</td>
</tr>
<tr>
<td></td>
<td></td>
<td>BCLC</td>
<td>88.2 (+1.2)</td>
<td>15.4</td>
<td>0.20</td>
<td>80.6</td>
<td>1.04</td>
</tr>
<tr>
<td></td>
<td></td>
<td>DRR</td>
<td>88.8 (+1.8)</td>
<td>15.1</td>
<td>0.20</td>
<td>81.5</td>
<td>1.04</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>hourly</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>BM</td>
<td>76.8 (+12.8)</td>
<td>26.6</td>
<td>0.32</td>
<td>65.5</td>
<td>1.13</td>
</tr>
<tr>
<td></td>
<td></td>
<td>BCLC</td>
<td>64.4 (+0.4)</td>
<td>18.9</td>
<td>0.23</td>
<td>79.1</td>
<td>0.96</td>
</tr>
<tr>
<td></td>
<td></td>
<td>DRR</td>
<td>65.0 (+1.0)</td>
<td>18.7</td>
<td>0.23</td>
<td>79.6</td>
<td>0.97</td>
</tr>
<tr>
<td>S2</td>
<td>daily</td>
<td>BM</td>
<td>77.1 (+13.1)</td>
<td>19.8</td>
<td>0.32</td>
<td>61.0</td>
<td>1.22</td>
</tr>
<tr>
<td></td>
<td></td>
<td>BCLC</td>
<td>64.1 (+0.2)</td>
<td>13.2</td>
<td>0.19</td>
<td>69.4</td>
<td>1.01</td>
</tr>
<tr>
<td></td>
<td></td>
<td>DRR</td>
<td>64.9 (+1.0)</td>
<td>12.8</td>
<td>0.18</td>
<td>72.1</td>
<td>1.02</td>
</tr>
<tr>
<td></td>
<td>peak</td>
<td>BM</td>
<td>93.4 (-2.3)</td>
<td>20.0</td>
<td>0.22</td>
<td>69.6</td>
<td>1.01</td>
</tr>
<tr>
<td></td>
<td></td>
<td>BCLC</td>
<td>96.1 (+0.4)</td>
<td>15.3</td>
<td>0.18</td>
<td>83.3</td>
<td>1.03</td>
</tr>
<tr>
<td></td>
<td></td>
<td>DRR</td>
<td>96.7 (+1.0)</td>
<td>15.1</td>
<td>0.18</td>
<td>83.9</td>
<td>1.03</td>
</tr>
</tbody>
</table>

Table 4: Forecasting scores for ozone. Mean and RMSE are in µg m⁻³. In the “Mean” column, the value in brackets is the variation with respect to the observed mean, a positive value means the model overestimates.
<table>
<thead>
<tr>
<th>Period</th>
<th>Frequency</th>
<th>Model</th>
<th>Mean</th>
<th>RMSE</th>
<th>NMSE</th>
<th>Correlation (%)</th>
<th>Bias Factor</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>hourly</td>
<td>BM</td>
<td>14.5 (-19.1)</td>
<td>31.0</td>
<td>0.81</td>
<td>29.4</td>
<td>0.62</td>
</tr>
<tr>
<td></td>
<td></td>
<td>BCLC</td>
<td>27.2 (-6.4)</td>
<td>24.9</td>
<td>1.13</td>
<td>44.7</td>
<td>1.13</td>
</tr>
<tr>
<td></td>
<td></td>
<td>DRR</td>
<td>27.4 (-6.2)</td>
<td>25.0</td>
<td>1.20</td>
<td>44.4</td>
<td>1.15</td>
</tr>
<tr>
<td>W1</td>
<td>daily</td>
<td>BM</td>
<td>14.5 (-19.1)</td>
<td>27.1</td>
<td>0.70</td>
<td>34.4</td>
<td>0.54</td>
</tr>
<tr>
<td></td>
<td></td>
<td>BCLC</td>
<td>27.6 (-6.0)</td>
<td>20.7</td>
<td>0.86</td>
<td>45.8</td>
<td>1.02</td>
</tr>
<tr>
<td></td>
<td></td>
<td>DRR</td>
<td>27.9 (-5.7)</td>
<td>20.0</td>
<td>0.88</td>
<td>47.9</td>
<td>1.03</td>
</tr>
<tr>
<td></td>
<td>peak</td>
<td>BM</td>
<td>21.7 (-22.3)</td>
<td>39.9</td>
<td>0.98</td>
<td>28.4</td>
<td>0.61</td>
</tr>
<tr>
<td></td>
<td></td>
<td>BCLC</td>
<td>35.1 (-9.0)</td>
<td>31.9</td>
<td>0.96</td>
<td>37.6</td>
<td>1.04</td>
</tr>
<tr>
<td></td>
<td></td>
<td>DRR</td>
<td>35.6 (-8.5)</td>
<td>31.3</td>
<td>0.98</td>
<td>39.6</td>
<td>1.06</td>
</tr>
<tr>
<td></td>
<td>hourly</td>
<td>BM</td>
<td>11.3 (-19.9)</td>
<td>29.5</td>
<td>0.83</td>
<td>41.9</td>
<td>0.52</td>
</tr>
<tr>
<td></td>
<td></td>
<td>BCLC</td>
<td>26.6 (-4.7)</td>
<td>23.3</td>
<td>1.66</td>
<td>43.0</td>
<td>1.31</td>
</tr>
<tr>
<td></td>
<td></td>
<td>DRR</td>
<td>27.5 (-3.8)</td>
<td>23.3</td>
<td>1.79</td>
<td>43.3</td>
<td>1.36</td>
</tr>
<tr>
<td>W2</td>
<td>daily</td>
<td>BM</td>
<td>11.3 (-20.0)</td>
<td>26.1</td>
<td>0.76</td>
<td>43.3</td>
<td>0.47</td>
</tr>
<tr>
<td></td>
<td></td>
<td>BCLC</td>
<td>29.6 (-1.7)</td>
<td>17.8</td>
<td>1.51</td>
<td>40.3</td>
<td>1.35</td>
</tr>
<tr>
<td></td>
<td></td>
<td>DRR</td>
<td>31.4 (+0.0)</td>
<td>17.2</td>
<td>1.62</td>
<td>44.8</td>
<td>1.42</td>
</tr>
<tr>
<td></td>
<td>peak</td>
<td>BM</td>
<td>17.5 (-25.1)</td>
<td>36.8</td>
<td>0.82</td>
<td>35.2</td>
<td>0.56</td>
</tr>
<tr>
<td></td>
<td></td>
<td>BCLC</td>
<td>38.4 (-4.2)</td>
<td>29.2</td>
<td>1.69</td>
<td>31.0</td>
<td>1.37</td>
</tr>
<tr>
<td></td>
<td></td>
<td>DRR</td>
<td>40.3 (-2.3)</td>
<td>29.1</td>
<td>1.86</td>
<td>33.0</td>
<td>1.44</td>
</tr>
</tbody>
</table>

Table 5: Forecasting scores for nitrogen dioxide. Mean and RMSE are in µg m$^{-3}$. 
<table>
<thead>
<tr>
<th>Period</th>
<th>Frequency</th>
<th>Model</th>
<th>Mean</th>
<th>RMSE</th>
<th>NMSE</th>
<th>Correlation (%)</th>
<th>Bias Factor</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>hourly</td>
<td>BM</td>
<td>23.7 (-6.8)</td>
<td>23.7</td>
<td>0.92</td>
<td>32.1</td>
<td>0.99</td>
</tr>
<tr>
<td></td>
<td></td>
<td>BCLC</td>
<td>26.1 (-4.4)</td>
<td>20.3</td>
<td>0.84</td>
<td>44.7</td>
<td>1.08</td>
</tr>
<tr>
<td></td>
<td></td>
<td>DRR</td>
<td>27.0 (-3.5)</td>
<td>19.4</td>
<td>0.86</td>
<td>49.8</td>
<td>1.11</td>
</tr>
<tr>
<td>W1</td>
<td>daily</td>
<td>BM</td>
<td>23.7 (-6.8)</td>
<td>19.5</td>
<td>0.60</td>
<td>40.5</td>
<td>0.89</td>
</tr>
<tr>
<td></td>
<td></td>
<td>BCLC</td>
<td>27.3 (-3.2)</td>
<td>16.7</td>
<td>0.55</td>
<td>50.8</td>
<td>1.03</td>
</tr>
<tr>
<td></td>
<td></td>
<td>DRR</td>
<td>28.0 (-2.5)</td>
<td>15.7</td>
<td>0.55</td>
<td>57.2</td>
<td>1.05</td>
</tr>
<tr>
<td></td>
<td>peak</td>
<td>BM</td>
<td>27.4 (-10.6)</td>
<td>26.8</td>
<td>0.64</td>
<td>38.0</td>
<td>0.87</td>
</tr>
<tr>
<td></td>
<td></td>
<td>BCLC</td>
<td>33.3 (-4.7)</td>
<td>24.2</td>
<td>0.67</td>
<td>42.0</td>
<td>1.06</td>
</tr>
<tr>
<td></td>
<td></td>
<td>DRR</td>
<td>33.7 (-4.4)</td>
<td>23.8</td>
<td>0.67</td>
<td>43.9</td>
<td>1.08</td>
</tr>
<tr>
<td></td>
<td>hourly</td>
<td>BM</td>
<td>16.9 (-10.4)</td>
<td>21.4</td>
<td>0.85</td>
<td>25.8</td>
<td>0.83</td>
</tr>
<tr>
<td></td>
<td></td>
<td>BCLC</td>
<td>23.1 (-4.2)</td>
<td>18.0</td>
<td>0.86</td>
<td>43.0</td>
<td>1.07</td>
</tr>
<tr>
<td></td>
<td></td>
<td>DRR</td>
<td>23.2 (-4.0)</td>
<td>17.7</td>
<td>0.86</td>
<td>44.9</td>
<td>1.08</td>
</tr>
<tr>
<td>W2</td>
<td>daily</td>
<td>BM</td>
<td>23.8 (-3.5)</td>
<td>17.8</td>
<td>0.75</td>
<td>41.5</td>
<td>0.99</td>
</tr>
<tr>
<td></td>
<td></td>
<td>BCLC</td>
<td>24.1 (-3.1)</td>
<td>14.3</td>
<td>0.60</td>
<td>50.7</td>
<td>1.02</td>
</tr>
<tr>
<td></td>
<td></td>
<td>DRR</td>
<td>24.3 (-3.0)</td>
<td>13.7</td>
<td>0.59</td>
<td>54.2</td>
<td>1.03</td>
</tr>
<tr>
<td></td>
<td>peak</td>
<td>BM</td>
<td>28.1 (-6.5)</td>
<td>25.2</td>
<td>0.78</td>
<td>33.8</td>
<td>0.97</td>
</tr>
<tr>
<td></td>
<td></td>
<td>BCLC</td>
<td>30.0 (-4.5)</td>
<td>22.1</td>
<td>0.67</td>
<td>39.8</td>
<td>1.06</td>
</tr>
<tr>
<td></td>
<td></td>
<td>DRR</td>
<td>29.8 (-4.7)</td>
<td>22.0</td>
<td>0.67</td>
<td>40.1</td>
<td>1.05</td>
</tr>
</tbody>
</table>

Table 6: Forecasting scores for PM$_{10}$. Mean and RMSE are in µg m$^{-3}$. 
To sum up for all pollutants, the DRR gives consistent results for both periods: the RMSE is significantly improved, along with the simulated mean and, most of the time, correlation. The greatest improvements are usually observed for hourly and daily concentrations.

Bias factor is an indication of the model tendency to underestimate or overestimate concentrations. Further results (not shown here) reveal indeed that most models tend to underestimate NO\textsubscript{2} and PM\textsubscript{10} concentrations and to overestimate those of ozone, especially low night-time ones. It happens to be the opposite for DRR concentrations. For NO\textsubscript{2} and PM\textsubscript{10}, the DRR bias factor may significantly increase above unity. This overestimation is mainly due to small concentrations as increasing the cutoff value gradually lowers both bias factor and NMSE. As an example, for a 10 \(\mu\text{g m}^{-3}\) cutoff value, the DRR NMSE begins to compete against the best model NMSE in all cases. For ozone, the DRR bias factor is slightly reduced, becoming sometimes inferior to unity. Further investigations show that, when diminishing the cutoff value for NMSE and bias factor, the best model NMSE increases more rapidly than that of the DRR for daily and hourly frequencies, which means the best model has greater difficulties to represent low concentrations than DRR.

In order to illustrate the ensemble performance, Figure 4 shows the daily RMSE (left hand side) and mean (right hand side) for ozone peak concentrations during August 2009. Figure 5 shows the averaged ozone daily profile for summer 2009. Colored lines are the ensemble members, the thick black line stands for DRR, and the diamonds are for the observations. The figures clearly illustrate that the aggregated forecast beats any model. They show that on average, the aggregation gets really close to the observations. The daily peaks averaged over all stations often coincide with the average observations. On the daily profiles, the aggregated one, which is a forecast, essentially coincides with the observations, while the models are missing both the average level and the amplitude of the profile. We may conclude that the aggregated forecast shows higher improvement as the forecast target gets averaged over a larger set of values.
4.2. Influence of parameters

In order to understand the influence of each parameter, let us rewrite the inner part of Eq. (1) in the following way:

\[ w_{h+1} = \lambda \arg \min_{u \in \mathbb{R}^M} \left[ \|u\|_2^2 + \sum_{h'=1}^{h} \left( \frac{\gamma}{\lambda} \frac{1}{(h + 1 - h')^2} \right) \sum_{i=1}^{N_{h'}} \left( \sum_{m=1}^{M} u_m x_{i,h'}^m - y_{i,h'} \right)^2 \right] \]

where we have just factorized the penalization parameter (\( \lambda \)).
The minimization process is essentially sensitive to the ratio $\gamma/\lambda$. Indeed, the most recent observations ($h' = h - 1$) are weighted by $1/\lambda + \gamma/\lambda$, thus the greater the ratio, the greater the influence of latest observations. On the contrary, oldest observations (at time $h' = 1$) are weighted by $1/\lambda$. Then, the greater $\lambda$, the less influential observations.

In practice, the ensemble performance shows only little sensitivity to the parameter $\lambda$ which should, as discussed in Section 2.2, depend mainly on the number of stations included, so that the penalization and the discrepancy terms remain balanced. It is more sensitive to the parameter $\gamma$, which determines how strong the discount in time is.

Figure 6 displays the ratio between the DRR RMSE and the BCLC RMSE against the parameter $\gamma$. It exhibits several values for $\gamma$ that allow for the DRR to beat the BCLC.

In this case, the results remain reasonably sensitive to the parameter $\gamma$. This indicates that the weights are already well constrained by the number of observations. It becomes more sensitive when the DRR is applied on a per station basis.

4.3. Impact of the number of models

The number of models is another key parameter in ensemble modeling. Indeed, as the CPU burden may linearly increase with the number of models,
one may wonder how many models are necessary to largely improve performance.

Figure 7 addresses this question. The performance is plotted with respect to the number of models for each pollutant. For the sake of comparison, the performance is measured by the RMSE divided by the RMSE of the forecasts aggregating all available models. The RMSE not only depends on the number of models, but also on the combination of selected models. Thus, for each number of models, the average RMSE on all possible combinations is computed, together with its standard deviation (error bars).

Figure 7: Performance against the number of models. The performance is measured by the RMSE divided by the RMSE of the forecasts aggregating all available models. The RMSE clearly decreases with the number of models, for all pollutants. The standard deviation also decreases, starting with a peak for one model, which reflects the dispersion among the models’ performance. The RMSE for ozone seems to reach a plateau after 6 models, which indicates that there is redundancy in the ensemble. This is however not the case for NO\textsubscript{2} and PM\textsubscript{10}, for which adding more models could be a way of improvement.

4.4. Weight relevance in space and time

As ensemble modeling is developed mainly for forecasting purposes, its robustness is to be assessed. First, weights are computed with a discrete
station network. They can be safely applied at station locations, but one may wonder to which extent can they be used outside of the station network.

To address this question, we carried out a sort of cross validation. When the aggregation is performed at one given station S, the weights are computed without the stations in the vicinity. More precisely, we exclude the stations whose geographical distance to the station S is lower than a given value \( d \). This allows to check whether the weights can be applied at distance \( d \) from the network. Since the results may strongly depend on the chosen station, it has been carried out for at least 20 distinct stations chosen randomly.

Figure 8 displays the average RMSE at the chosen stations against the distance \( d \), in degree. For the sake of comparison, the RMSE is divided by that obtained with the shortest distance, which only excludes the selected station, as if there were no station at this point. The standard deviation of the RMSE (error bars) reflects the test variability due to the station selection.

![Figure 8: Normalized RMSE against the distance (in degrees) to the network.](image)

As expected, the RMSE increases with the distance, since the weights become less and less relevant as they are applied far from the network. However, it is striking to see that even up to 2° far from the network, the RMSE only worsens by at most 5%, even considering the standard deviation. In view that any location in the French metropolitan domain is at shorter distance from the network than 2°, we conclude that the weight relevance in
space is acceptable to aggregate 2D fields over the full territory in operational context.

Second, in an operational context, several failures may arise, involving simulation and/or observation availability. Then, if today’s weights cannot be computed due to the lack of data, to which extent older weights may be applied to the present situation?

In order to assess the persistence in time of the weights performance, we introduced an offset of several days between the weights computation and the actual aggregation. Figure 9 displays the RMSE, divided by that obtained without offset, against of the offset in days, for all pollutants.

![Figure 9: Normalized RMSE against the number of days offset.](image)

The RMSE is not monotonic. We have no clear explanation of this behavior. We can just hint that some physical situations may reproduce themselves along time, so that weights computed for one given time may also fit to another episode or similar event. This holds especially for NO\(_2\), whose concentrations are strongly influenced by anthropogenic emissions which roughly follow a weekly pattern, so that physical situations may reproduce themselves from week to week. In all cases, the RMSE is worsened, and nothing can replace the most recent weights. However, the RMSE is only a little worsened, at most by 3% for ozone, with 10 days offset, and of at most 6% for NO\(_2\) and 8% for PM\(_{10}\), after 20 days. This means that the system will be robust.
in operational forecasting.

4.5. Extreme events

As already stated, the Prev’Air platform is particularly concerned with the accurate detection of ozone threshold exceedances, two of which are being monitored on hourly averages: the information threshold (180 µg m$^{-3}$) and the alert threshold (240 µg m$^{-3}$).

Each time one of these thresholds is exceeded, warning or alert messages are sent to local authorities so that they can warn public at large and take appropriate measures. Nevertheless, whether or not to trigger an alert under simulation response raises questions, and depends strongly on how confident the modeler is in the forecast(s). Indeed, one isolated forecast exceedance might be considered as a model error (a false alarm), whereas several exceedances, coinciding with large plumes and stable meteorological conditions, should be seriously taken into account.

Ensemble modeling has been developed on Prev’Air with the aim to supply the modeler with more reliable exceedance detections, which is what one would expect when using several models. The purpose of this section is not to fully address this issue, which would involve much deeper work, but to give first insights of the aggregation detection performance. In addition to ozone, we display results for NO$_2$ and PM$_{10}$ threshold exceedances in winter.

Given one threshold value, Table 7 summarizes the various events which may occur.

<table>
<thead>
<tr>
<th>Simulation</th>
<th>Observation</th>
</tr>
</thead>
<tbody>
<tr>
<td>exceed? yes</td>
<td>yes</td>
</tr>
<tr>
<td>exceed? no</td>
<td>no</td>
</tr>
<tr>
<td>miss</td>
<td>no</td>
</tr>
</tbody>
</table>

Table 7: Threshold exceedance events.

Let us denote respectively $N_{\text{obs}}^1$ and $N_{\text{obs}}^0$ the number of observations on a given period which exceeded the threshold and those which did not. Simulation values are partitioned between those which hit the threshold exceedance event ($N_{\text{hit}}$), those which missed it ($N_{\text{miss}}$), those which wrongly predicted a threshold exceedance event ($N_{\text{fa}}$), and finally those which correctly predicted no exceedance ($N_{\text{cn}}$). These numbers are linked together by the following
relationships:

\[ N_{\text{obs}}^1 = N_{\text{hit}} + N_{\text{miss}}, \quad N_{\text{obs}}^0 = N_{\text{fa}} + N_{\text{cn}}. \]  

(4)

Naturally, what the modeler desires is to maximize the number of hits, while minimizing the number of false alarms. Following (Drifi et al., 2009), model and ensemble performances may be assessed with the threat score:

\[ R = \frac{N_{\text{hit}}}{N_{\text{hit}} + N_{\text{miss}} + N_{\text{fa}}}. \]  

(5)

The greater the threat score, the better.

Thereafter, we present the detection performance of the ensemble compared with that of the best model according to the threat score. The ensemble may be used either directly, assuming that an exceedance is detected when half of its members detected it (method labelled ENS), or indirectly by using the DRR forecast. In the latter case, the aggregated forecast is computed on a per station basis, i.e., the weights are learned independently at each station.

Figures 10 display \( R \) for various thresholds and each pollutant, using peak concentrations.
Figure 10: Threat score $R$ against the threshold for various pollutants.
First, the ENS gets very poor results for NO$_2$ and PM$_{10}$, compared to those of the BM. This is mainly due to the fact that all members severely underestimate concentrations for these pollutants. Therefore, the best model performance is worsened by the other models, and the ensemble as a whole cannot bring directly any improvement. Only for O$_3$ and low thresholds, ENS performs slightly better than the BM.

Second, DRR significantly improves exceedance detection for NO$_2$ and PM$_{10}$, for all threshold levels, but is disappointing for O$_3$, especially for the aforesaid regulation thresholds. This is explained by the fact that the DRR tends to underestimate ozone concentrations and overestimate NO$_2$ and PM$_{10}$ ones.

Several aggregation algorithms for event prediction (as opposed to concentration prediction) were tested in (Drifi et al., 2009), but only little improvements were found for ozone. The difficulty in improving threshold exceedance detection partly lies in the fact that there are only few events from which to learn. This is also the reason why we cannot fully conclude on ensemble abilities.

5. Conclusion

In this paper, we presented the application of an ensemble forecasting approach (Mallet et al., 2009) to the Prev’Air operational platform. The ensemble method takes advantage of the multiple models available on the platform (CHIMERE, MOCAGE, Polyphemus/Polair3D) with various resolutions and forcing, in order to improve predictions with sequential aggregation. The discounted ridge regression computes new weights for the models before every new prediction. The same weights are computed at once for all stations, so that the forecast fields of the ensemble can be aggregated, not just the forecasts at the stations. In the long run, the approach is theoretically guaranteed to be at least as good as any linear combination with constant weights (in particular, any model in the ensemble or the ensemble mean).

Using the discounted ridge regression, the RMSE of the forecasts respectively drops by at least 29%, 35% and 19% for hourly, daily and peak O$_3$ concentrations, by 19%, 26% and 20% for hourly, daily and peak NO$_2$ concentrations, and finally by 17%, 19% and 11% for hourly, daily and peak PM$_{10}$ concentrations. Having used several distinct periods for each pollutant is the guarantee that these results are not dependent on the period. These
improvements always come along with better correlations and bias factors. This study also shows that DRR is able to correct the models tendencies to underestimate or overestimate observations, which may sometimes worsen the NMSE indicator depending on the occurrence frequency of low concentrations.

Several tests were carried out to determine whether the aggregated forecasts are reliable enough for an operational context. The results show that weights can be safely applied up to 2° away from the observation network, hence they can be used to combine the forecast fields. The weights are also robust in time since they remain relevant for forecasts several days ahead. Considering that the BASTER observation network is enough to cover the whole French metropolitan territory, and that the number of forecast days ahead does not exceed 3–5 days, we can conclude that ensemble approach is suitable for the operational context. This justifies why this approach has been applied since 2008 on the operational Prev’Air platform.

Future work should address the issue of threshold exceedance forecast. There is a need for aggregation methods that would estimate the uncertainty of the aggregated forecasts. Also, the spatial distribution of the aggregated forecasts should also be improved. Indeed, constraining weights with data from a large and widely spread monitoring network is likely to reduce spatial variability. This drawback can be overcome by the spatialization of the weights. This can be achieved in two different ways: either by partitioning the monitoring network or by forecasting analyzed fields (Mallet, 2010; Thiberville, 2010) instead of observations, which allows to compute weights in every grid cells.
Appendix A. Performance indicators

Let \((x_i)_{i=1}^n\) and \((y_i)_{i=1}^n\) be respectively simulation and observation data. We define the following indicators:

- **Mean**
  \[
  \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i , \quad \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i \tag{A.1}
  \]

- **Root mean square error**
  \[
  \text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (x_i - y_i)^2} \tag{A.2}
  \]

- **Correlation**
  \[
  \text{Correlation} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \times \sum_{i=1}^{n} (y_i - \bar{y})^2}} \tag{A.3}
  \]

- **Normalized mean square error**
  \[
  \text{NMSE} = \sqrt{\sum_{i=1}^{n} \left( \frac{x_i - y_i}{y_i} \right)^2} \tag{A.4}
  \]

- **Bias factor**
  \[
  BF = \frac{1}{n} \sum_{i=1}^{n} \frac{x_i}{y_i} \tag{A.5}
  \]
References


Dentener, F., Stevenson, D., Cofala, J., Meckler, R., Amann, M., Bergamaschi, P., Raes, F., Derwent, R., 2005. The impact of air pollutant and


37


Madronich, S., 1993. UV radiation in the natural and perturbed atmosphere. Lewis Publisher, Boca Raton.


