Available online at www.sciencedirect.com
ScienceDirect
Atmospheric Environment 41 (2007) 2083–2097
ATMOSPHERIC ENVIRONMENT
www.elsevier.com/locate/atmosenv

Evaluation of long-term ozone simulations from seven regional air quality models and their ensemble

M. van Loon[a], R. Vautard[b,*], M. Schaap[c], R. Bergström[d], B. Bessagnet[e], J. Brandt[f], P.J.H. Builtjes[c], J.H. Christensen[f], C. Cuvelier[g], A. Graff[h], J.E. Jonson[a], M. Krol[i], J. Langner[d], P. Roberts[j], L. Rouil[e], R. Stern[k], L. Tarrasón[a], P. Thunis[g], E. Vignati[g], L. White[l], P. Wind[a]

[a]*EMEP/MSC-W, P.O. Box 43, Blindern 0313, Oslo, Norway*
[b]*LSCE/IPSL Laboratoire CEA/CNRS/UVSQ, 91191 Gif sur Yvette Cedex, France*
[c]*TNO Built Environment and Geosciences, Apeldoorn, The Netherlands*
[d]*SMHI, SE-601 76 Norrköping, Sweden*
[e]*INERIS, Parc Technologique Halata, Verneuil en Halatte, France*
[f]*National Environmental Research Institute, Frederiksborgvej 399, P.O. Box 358, DK-4000 Roskilde, Denmark*
[g]*European Commission DG Joint Research Centre, Institute for Environment and Sustainability, I-21020 Ispra, Italy*
[h]*UBA, Bismarckplatz 1., 14193 Berlin, Germany*
[i]*SRON, Institute of Marine and Atmospheric Research, Utrecht, The Netherlands*
[j]*CONCAWE, Shell Global, HSE Department, P.O. Box 1, Chester CH1 3SH, UK*
[k]*Freie Universität, Berlin, Germany*
[l]*CONCAWE, 42 Blunts Wood Road, Hayward Heath, West Sussex RH16 1NB, UK*

Received 16 July 2006; received in revised form 20 October 2006; accepted 31 October 2006

Abstract

Long-term ozone simulations from seven regional air quality models, the Unified EMEP model, LOTOS-EUROS, CHIMERE, RCG, MATCH, DEHM and TM5, are intercompared and compared to ozone measurements within the framework of the EuroDelta experiment, designed to assess air quality improvement at the European scale in response to emission reduction scenarios for 2020. Modelled ozone concentrations for the year 2001 are evaluated. The models reproduce the main features of the ozone diurnal cycle, but generally overestimate daytime ozone. LOTOS-EUROS and RCG have a more pronounced diurnal cycle variation than observations, while the reverse occurs for TM5. CHIMERE has a large positive bias, which can be explained by a systematic bias in boundary conditions. The other models and the ''ensemble model'', whose concentrations are by definition averaged over all models, represent accurately the diurnal cycle. The ability of the models to simulate day-to-day daily ozone average or maxima variability is examined by means of percentiles, root mean square errors and correlations. In general, daily maxima are better simulated than daily averages, and summertime concentrations are better simulated than wintertime concentrations. Summertime correlations range between 0.5 and 0.7 for daily averages and 0.6 and 0.8 for daily maxima. Two health-related indicators are used, the number of days of exceedance of the $120\,\mu g\,m^{-3}$ threshold for the daily maximal 8-h ozone concentration and the SOMO35. Both are well reproduced in terms of frequency, but the simultaneity of occurrence of exceedance days between observations and simulations is not well captured.

*Corresponding author.
E-mail address: robert.vautard@cea.fr (R. Vautard).

1352-2310/$ - see front matter
doi:10.1016/j.atmosenv.2006.10.073

The advantage of using an *ensemble* of models instead of a single model for the assessment of air quality is demonstrated. The ensemble average concentrations almost always exhibit a closer proximity to observations than any of the models. We also show that the spread of the model ensemble is fairly representative of the uncertainty in the simulations.

Keywords: Air quality; Modelling; Ozone; Model intercomparison; EuroDelta

1. Introduction

The fate of regional air quality in the future depends on many factors. For the near future the most important are likely to be the evolution of the development worldwide, the associated pollutant emission changes and the subsequent evolution of the global atmospheric composition (Akimoto, 2003; Dentener et al., 2005). In the far future, climate change due to radiative forcing modifications may also affect tropospheric chemistry and therefore air quality, via water vapor and temperature and other physical or dynamical changes (Stevenson et al., 2005). In Europe, and for a nearer future (say 2010–2030), air quality is affected by a combination of the evolution of continental emissions and changes in remote emissions (Parrish et al., 1993) which can alter baseline concentrations (Szopa et al., 2006). While the latter factor is impossible to control at the European scale, the former can be modified by concerted efforts in European environmental policies. The likely success of various policies can only be evaluated using numerical models.

The EuroDelta project is designed to evaluate the regional responses to emission reduction scenarios, in support of the European Union Clean Air For Europe (CAFE) Programme and in the framework of the Convention on Long Range Transport of Atmospheric Pollution (CLRTAP, United Nations—Economic Commission for Europe). The originality of the approach in the EuroDelta project is twofold: first, several regional air quality models are used, giving access not only to emission change response but also to an evaluation of the associated uncertainty; second, unlike in previous exercises (Hass et al., 1997), EuroDelta uses long-term simulations, carried out over two meteorological reference years, 1999 and 2001. However, in this article results are only presented for the year 2001. As described in a future article, for each specific scenario a set of "deltas" is available, the differences between the scenario simulation and the base case simulation.

Before one can trust models' response to emission scenarios, a crucial validation stage is required. This paper focuses on models' validation and intercomparison for ozone; aerosols will be investigated in a future paper. Within the evaluation of the Unified EMEP model, a model intercomparison has been conducted (Van Loon et al., 2004), using the years 1999 and 2001, and to a large extent the same models. At city scale, a large model intercomparison has also been carried out (Vautard et al., 2007) within the CityDelta CAFE project (Cuvelier et al., 2007). Intercomparisons have also been performed for regional air quality forecasts (Tilmes et al., 2002). Many of the models have been recently improved by the inclusion of new processes or improvements of process parameterization, thereby warranting ongoing validation and intercomparison. Here, we also consider new indicators that were not taken into account in previous studies such as the "Sum Of Maximum Ozone daily 8-hours means over 35 ppb" (SOMO35: Amann et al., 2005), and the number of exceedance days of health-related concentration thresholds.

Seven European state-of-the-art regional air quality models that have been developed by independent groups participated in this evaluation exercise. The focus of the validation is on ozone over Europe, but additional quantities such as O_x (sum of O_3 and NO_2) are considered as well, since they may provide additional understanding of the performance of the models. The first aim of the article is to compare the simulations issued from these models with a large set of observations, and to intercompare them.

Another important question addressed here deals with the ensemble formulation. As shown by several authors for specific shorter verification periods and for air quality forecasting, the ensemble average of several independent simulations is usually closer to observations than any single simulations (Delle

Monache and Stull, 2003; McKeen et al., 2005, see also Delle Monache et al., 2006a,b; Mallet and Sportisse, 2006). We verify this property with our set of long-term simulations and give mathematical arguments to explain this behavior. An ensemble can only give significant improvements if participating models have complementary strengths and weaknesses, and therefore are representative of the uncertainty in our knowledge. While this specific question is addressed in detail in a separate paper (Vautard et al., 2006), we present here a few arguments in favor of this property for the ensemble presented in this study.

The paper is organized as follows: Section 2 contains a description of the models, and Section 3 contains a description of the observations. Section 4 presents the overall methodology and the simulations. Section 5 is devoted to the validation and the intercomparison of the models. In Section 6 we analyze the ensemble. Section 7 contains the conclusions.

2. Participating models

To represent, as much as possible, the uncertainty in our current knowledge of air quality processes, we allow all models to freely and independently utilize their best estimate for most input data and parameters (horizontal and vertical resolution, meteorological input, boundary conditions, biogenic emissions, etc.). The anthropogenic emissions (see below) are however defined on a common basis, because the final aim of the exercise is the intercomparison of emission control scenarios defined on a common basis.

The participating models in this study are EMEP (http://www.emep.int), CHIMERE (Schmidt et al., 2001; Bessagnet et al., 2004), MATCH (Andersson et al., 2006 and references therein), LOTOS-EUROS (Schaap et al., 2006 and references therein), REM-CALGRID (RCG; Stern et al., 2003), DEHM (Christensen, 1997; Frohn et al., 2002; Geels et al., 2004) and TM5 (Krol et al., 2005). All models, except the global research model TM5, are regional-scale, limited-area models designed for short-term and long-term simulations of oxidant and aerosol formation. The models have different degrees of complexity. The horizontal resolutions of the limited-area models are very similar and range from about 30 to 50 km (see Table 1). TM5 has a coarser horizontal resolution. TM5, EMEP and DEHM describe the whole tropospheric column

Table 1
Participating models and their spatial resolution

Model	Horizontal resolution[a] and number of cells	Vertical resolution	Approx. depth 1st layer (m)
EMEP (EMEP-MSC-W)	50 × 50 km 110 × 100	20 sigma levels up to 100 hPa	90
RCG (UBA)	0.5° × 0.25° 82 × 125	5 layers, surface layer fixed, 4 dynamical layers moving with MH	20
MATCH (SMHI)	0.4° × 0.4° 84 × 106	14 layers (eta coordinates) up to 6 km	60
LOTOS-EUROS (TNO)	0.5° × 0.25° 100 × 140	4 layers, surface layer fixed, 4 dynamical layers moving with MH	25
CHIMERE (INERIS, IPSL)	0.5° × 0.5° 64 × 46	8 layers up to 500 hPa	50
TM5 (JRC)	Eur: 1° × 1° Glob: 6° × 4°	25 levels/hybrid sigma/pressure	50
DEHM (NERI)	Eur: 50 × 50 km Northern hemisph: 150 × 150 km: 96 × 96	20 sigma levels up to 100 hPa	50

[a]The EMEP and DEHM models use a polar stereographic projection. MATCH uses a rotated latitude–longitude grid with shifted pole. All other models use geographical coordinates.

with 20–25 vertical layers, while LOTOS-EUROS, RCG and CHIMERE describe only the lower troposphere, up to above the boundary layer. LOTOS-EUROS and RCG have varying vertical layers, which follow the boundary layer diurnal evolution.

Boundary conditions are either based on observations (EMEP, RCG, LOTOS-EUROS, MATCH (for some components)) or from global (or other large scale) model simulations (CHIMERE, MATCH (for other components), DEHM). Driving meteorology is taken either directly from global analyses (TM5), from a limited-area meteorological model (EMEP, CHIMERE, DEHM, MATCH) or from an optimal interpolation analysis based on observations (RCG, LOTOS-EUROS).

For the base case of EuroDelta, which is discussed in this article, the emissions are in line with the CAFE baseline emissions for the year 2000 (Vestreng, 2003). The choice of year 2001, instead of 2000, as the meteorological year for the simulations is motivated by the fact that Summer 2000 was extremely poor in ozone episodes due to cool and windy conditions. During Summer 2001, classical range of ozone values was found in Europe, with extrema in the rural background exceeding 100 ppb but lower than 150 ppb. The month of July was particularly rich in large-scale ozone episodes. The success of the intensive ESCOMPTE air quality campaign, which was conducted over the Marseille area in June–July 2001, was partly due to weather conditions typical for the development of ozone episodes (Cros et al., 2004; Drobinski et al., 2006). All modelling groups adapted the same annual emission inventory to their model grid and model species. Hourly emissions were obtained using prescribed temporal factors (GENEMIS project and TNO) following the procedure in the CityDelta project (Cuvelier et al., 2007).

In this article the skill of these different models is evaluated and intercompared. We also analyze the skill of the "ensemble model", whose concentrations are the arithmetic average of the concentrations of the seven models.

3. Observations

Given the spatial resolution of the models (30–50 km at finest), it is relevant to validate simulated concentrations against observations at rural locations. Most observations used in this study are gathered from the EMEP (http://www.emep.int) database and some are taken from the AIRBASE (http://dataservice.eea.europa.eu/dataservice) database. From the latter, only stations classified as rural were selected. The EMEP database includes only rural stations. Data for one station in Switzerland were obtained from the Swiss Environmental Agency. In order to prevent overweighting of regions with large data coverage, we selected a reduced set of stations in these areas. This was based on the availability of observations of additional compounds, such as NO_2. Due to the bad representation of coarse-resolution models in mountainous areas, only stations below an altitude of 1000 m were considered. Finally, only stations lying at the intersection of all model domains are considered. For year 2001, this resulted into a total number of 97 stations for O_3, 70 for NO_2, 68 for O_x. Ozone and O_x observations were considered at hourly time steps.

4. Intercomparison methodology

The model validation and intercomparison are carried out using classical statistical indicators: comparison of mean diurnal cycles, biases, root mean square error (RMSE), correlation, relative RMSE (RRMSE), and comparison of simulated versus observed percentiles. Mean diurnal cycles are the average over all available days of concentrations for each of the 24 h. The bias is, by definition, the average difference between simulated and observed values. The RMSE is the square root of the average square of the simulated and observed differences. The RRMSE is the RMSE divided by the observed average. For ozone model skill is evaluated over hourly data, daily averages and daily maximum concentrations. In order to verify the ability of the models to reproduce the day-to-day variability, the ratio between simulated and observed standard deviations is calculated (sigma-ratio statistics).

In addition, two health-related parameters are considered: SOMO35 and the number of exceedance days. The latter is used in the European Union and is defined as the number of days in a calendar year on which the maximum of the 8-hour mean ozone concentrations exceeds $120\,\mu g\,m^{-3}$. SOMO35 (Amann et al., 2005) is an indicator for health impact assessment recommended by the World Health Organization and is defined as the yearly sum of the daily maximum of the 8-hour running average over 35 ppb. All statistics are calculated both over the whole year and over

particular seasons. If not stated otherwise, statistics are calculated over the whole year.

5. Evaluation of the model's performances

5.1. Diurnal cycles

Fig. 1 shows the average diurnal cycle of observed and simulated hourly ozone values for the observations and the models, and in Fig. 2 the diurnal cycles relative to each model's daily average are shown. All models are able to catch the mean features of the observed diurnal cycle, with lower values during night-time, higher values during daytime and a daily maximum in the afternoon. However, all models, but DEHM and TM5, overestimate daytime concentrations. The timing of the mean simulated ozone maxima is within an hour of the mean observed one, as shown in Fig. 1.

Significant differences among models are found in absolute values, in particular during night-time, when the spread is largest. During daytime, models apparently have different ozone production rates, as the range of the diurnal cycle is large in some cases,

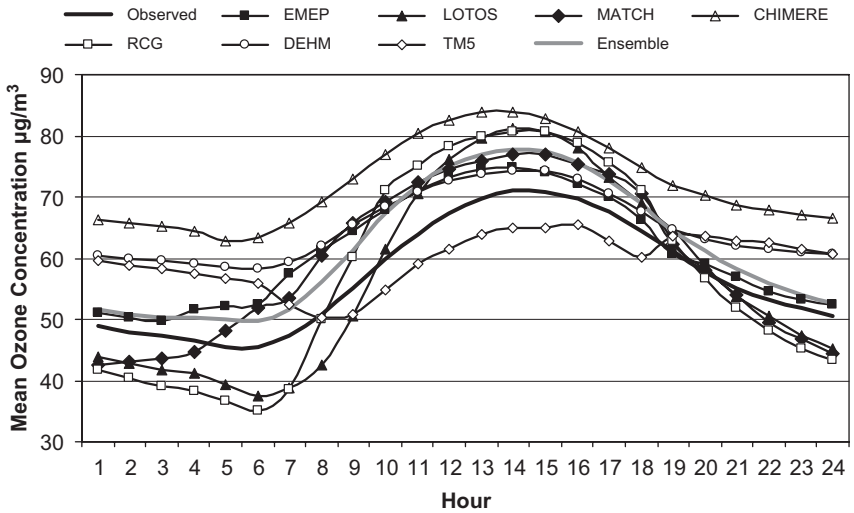

Fig. 1. Yearly mean diurnal cycle of ozone, in $\mu g\,m^{-3}$, as a function of hour, for all models, averaged over all monitoring stations.

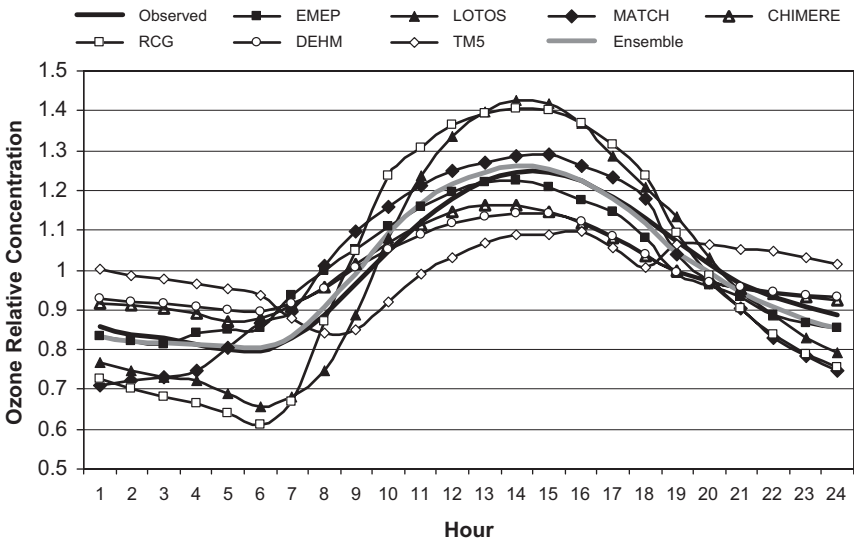

Fig. 2. Yearly average diurnal cycle of ozone, in $\mu g\,m^{-3}$, for all models, averaged over all monitoring stations, relative to the average of each data set. In order to obtain values in this figure, values of Fig. 1 are divided by their average over hours for each model.

e.g. for RCG or LOTOS-EUROS, and small in other cases (TM5, DEHM). Apart from differences in chemical mechanism used by the models, a number of other parameters may influence the diurnal cycle.

First, model ozone concentrations are not taken at the same height. In LOTOS-EUROS and MATCH, a reference height is taken at 3 m for ozone concentrations, assuming a specific profile for concentrations, while in other models the surface layer concentration is used. The assumed height of concentrations is important, in particular in stable surface layers (night-time or wintertime) when deposition can create a large ozone vertical gradient. By contrast, surface resistance can also be underestimated by models due to wet soils (dew, rain, snow) not taken into account. Second, the difference in emission injection height distribution between models can lead to differences in surface-layer ozone titration by fresh nitric oxide emissions.

The model vertical layer structure may also be important: the two models with the thinnest first model layer, LOTOS-EUROS and RCG, show a very similar daily cycle clearly different from the other models: lower concentrations during night-time and a rather steep increase in ozone during the first part of the day. This diurnal variability is overestimated relative to observations (Fig. 2), by dividing the values of Fig. 1 by their average over the diurnal cycle for each model. In addition, both LOTOS-EUROS and RCG use the same meteorology and the same mixing layer concept, with one (LOTOS-EUROS) or two (RCG) dynamic layers above the surface layer and below the mixing height. This may explain the specific behavior of these two models. By contrast, TM5 has a flatter diurnal cycle. Other models have a diurnal variability closer to the observed one. CHIMERE has a large all-day-long bias that may be due to a bias in the ozone boundary condition. Such a bias is also present at the station of Mace Head at the western boundary of the model domain, but this hypothesis is harder to verify on other model boundaries. The ensemble model diurnal cycle follows the observations with a slight positive bias ($5\,\mu g\,m^{-3}$ at maximum) but with very good timing.

In order to examine to what extent the differences in ozone can be attributed to different distributions of total oxidant O_x (the sum of O_3 and NO_2), the diurnal cycle of O_x is shown in Fig. 3. The advantage of using O_x instead of O_3 is that it is insensitive to fast reactions between ozone and NO_x. Thus, incorrect ozone titration intensity should not affect O_x. The daily cycles of models are generally closer to the observed one for O_x than for O_3, showing that wrong ozone titration, possibly due to model vertical resolution and emission injection profiles, explains a significant part of the model-observations differences. CHIMERE still displays a large positive bias, confirming the possible "background" origin of this bias. TM5 diurnal cycle remains too flat, indicating underestimation in photochemical activity. For LOTOS-EUROS and RCG, night-time concentrations are still too low; indicating that overtitration of ozone due to possible meteorological overstability leading

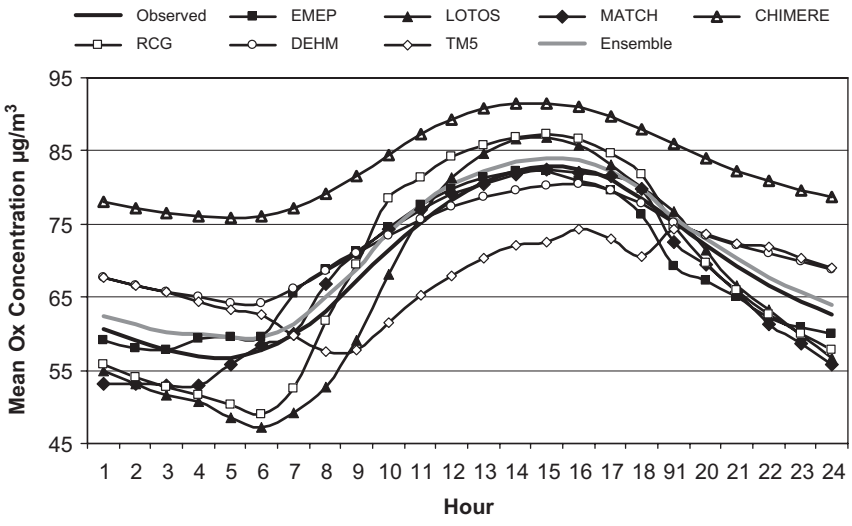

Fig. 3. Yearly average diurnal cycle of $O_x = O_3 + NO_2$, in $\mu g\,m^{-3}$, for all models, averaged over all monitoring stations.

to too high NO_x concentrations cannot be the leading factor for this discrepancy. The other models (MATCH, DEHM, EMEP) faithfully follow the observed O_x diurnal cycle. The ensemble model exhibits an almost perfect O_x diurnal cycle.

5.2. Percentiles

Fig. 4 shows a number of percentiles, averaged over all stations. Each percentile is computed on a stationwise basis and then an average value is computed over the stations percentiles in Figs. 4(a) and (b). Percentiles are correctly reproduced by most models, except for CHIMERE, which overestimates low percentiles. This can be a consequence of a bias in boundary conditions, as suggested before, since low ozone percentiles occur in windy conditions rapidly transporting ozone within the center of the model domain. It can also be due to the other problems mentioned above: difference between the model surface layer middle height and the monitoring station height, or too high mixing during night-time or in daytime cloudy conditions smoothing near-ground ozone vertical gradients. High percentiles, which correspond to sunny summertime daytime conditions, are fairly well simulated by all models including CHIMERE. LOTOS-EUROS and RCG slightly overestimate

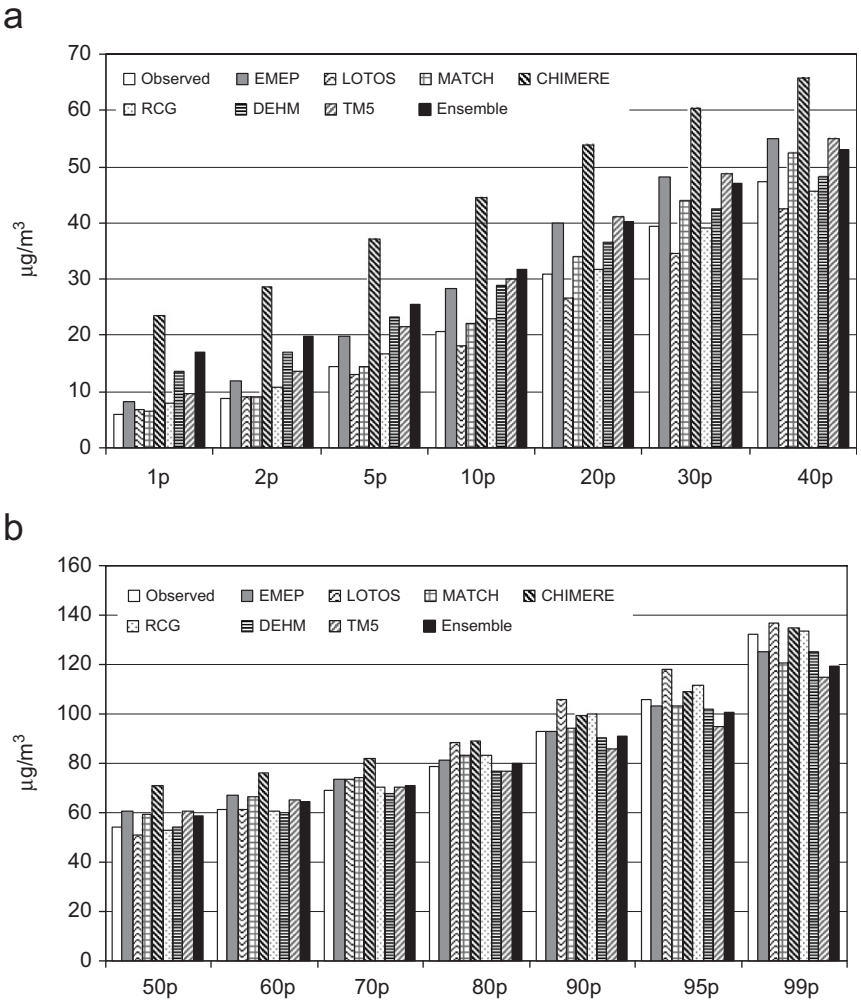

Fig. 4. Ozone concentration percentiles, calculated over all stations, for each model, in $\mu g\,m^{-3}$. All ozone concentrations are put together, for a given model, and sorted. Then for each percentile p, the ozone concentration found at the $p*N/100$ rank is taken as the pth percentile for this model.

high percentiles, which is consistent with the diurnal cycles and an overestimation of ozone production in the boundary layer.

5.3. Error statistics

In this section we present seasonal statistics of the models skill. The seasons considered are winter (December, January and February), spring (March, April, May), summer (June, July, August), and autumn (September, October, November). The biases and relative RMSEs for daily average and daily ozone maximum are given in Tables 2 and 3, respectively.

On a yearly basis all models, except CHIMERE and DEHM, show relative low positive biases. The positive bias in 2001 is mainly caused by a large overestimation during the autumn season, shown by all models. A closer inspection of time series reveals that ozone is severely overestimated in September (by all models except RCG) and October (by all models). The overestimation is observed over all stations. One possible explanation is that boundary conditions, during autumn 2001, are lower than those taken from the climatological data. At Mace Head, at the western coast of Ireland, ozone concentrations from the so-called clean sector are lower than their 20-year average during these months, particularly during September 2001 (clean sector concentrations about 8–9 $\mu g\,m^{-3}$ lower than average). This partly agrees with the overestimations by the models which are 13 $\mu g\,m^{-3}$ on average (range 8–25 $\mu g\,m^{-3}$), except for RCG which did not show an overestimation in that month. For October this bias is 13 $\mu g\,m^{-3}$ (range 8–26 $\mu g\,m^{-3}$), while the clean sector at Mace Head is only about 4 $\mu g\,m^{-3}$ lower than average. Some models (MATCH, EMEP) use observed ozone for boundary conditions but still exhibit overestimation. TM5, which is a global model, also exhibits overestimation. Fall ozone overestimation cannot be fully explained by anomalous boundary conditions that could be due to the combination of weather and normal emissions. In simulations of autumn 1999 overestima-

Table 2
Relative bias of the simulated daily average and daily maximal ozone concentrations in 2001 as percent of the observed average

	Daily average					Daily maximum				
	Year	DJF	MAM	JJA	SON	Year	DJF	MAM	JJA	SON
EMEP	10	**1**	**2**	**7**	33	**1**	**−5**	**−4**	**−3**	18
LOTOS	**2**	−25	**4**	11	10	7	−14	11	8	18
MATCH	6	6	**4**	**3**	14	**2**	6	**1**	**−5**	12
CHIMERE	29	35	18	20	56	10	15	**4**	**3**	24
RCG	**3**	−9	7	**−1**	16	7	**−2**	12	**0**	17
DEHM	18	−5	12	21	42	**−1**	−20	**−3**	**3**	11
TM5	6	7	**−5**	**3**	29	−7	−6	−13	−11	6
Ensemble	11	**1**	6	9	28	**−1**	−8	**−2**	**−4**	11

Boldface numbers stand for biases less than 5 $\mu g\,m^{-3}$ in absolute values.

Table 3
RRMSE of the modelled daily average O_3 and modelled daily maximum concentrations in 2001

	Daily average					Daily maximum				
	Year	DJF	MAM	JJA	SON	Year	DJF	MAM	JJA	SON
EMEP	35	**40**	26	28	59	27	**29**	21	22	38
LOTOS	38	53	28	29	57	31	38	23	24	46
MATCH	**31**	**40**	**25**	**25**	**42**	**24**	30	**19**	21	**30**
CHIMERE	45	57	33	33	74	25	32	20	**20**	36
RCG	36	44	28	28	53	29	31	24	25	39
DEHM	46	46	36	37	78	29	36	25	22	40
TM5	38	43	30	31	56	29	**29**	27	28	**30**
Ensemble	**32**	**37**	**25**	**25**	**53**	**22**	**27**	**17**	**18**	31

Boldface numbers stand for the lower two values of each column.

tions of ozone of that amplitude are not found (not shown), indicating that these are not due to systematic model biases. During summer months LOTOS-EUROS, MATCH, TM5 and RCG exhibit averages close to the observed values. Only RCG has a slight negative bias. CHIMERE has a large positive bias (20%). A slight overestimation is also found in spring, while the bias sign in winter is less systematic.

A low bias does not necessarily imply low RRMSE values, as can be concluded from a comparison of the values in Tables 2 and 3, although CHIMERE and DEHM have both biases and RRMSE larger than the other models for daily averages. In winter and fall, model skill is poorer than in spring and summer, even for models having weak biases. Since winter ozone is mostly controlled by other processes than photochemistry (deposition, titration, boundary conditions) it is expected that ozone be sensitive to model formulation (vertical resolution, emission injection vertical profile, deposition velocities). RRMSE for daily maximum concentrations are lower than for daily average concentrations, without exception. For daily averages the models that perform best both in terms of bias and RRMSE are MATCH and EMEP.

A bias in daily averages does not necessarily lead to a bias in daily maxima, since, for instance, daytime ozone values can be less biased than nighttime ones, such as for CHIMERE (Fig. 1). The results for the biases and diurnal cycles are clear indications for this (Table 2). LOTOS-EUROS and to some extent also RCG seem to have a too flat distribution with the peak in the distribution function clearly shifted to the left side of the corresponding peak value in the observations, whereas the other models tend to shift to the right (Fig. 4). For values higher than about $95\,\mu g\,m^{-3}$ both models show higher frequencies than the observations. As could be expected from its large positive bias for mean ozone and much smaller bias for daily maximum ozone, CHIMERE and DEHM show a shift in the distribution towards the right (Fig. 4). From about $120\,\mu g\,m^{-3}$ onwards, these models show a good agreement with the observed distribution, in line with its low biases for daily maximum ozone values in the period March–August.

Daily maxima are best reproduced by CHIMERE, MATCH, DEHM and EMEP, especially in summer (RRMSE between 20% and 22%), the season when acute ozone episodes occur. Daily maxima variability is mostly driven by photochemistry and the variability of radiation (cloud representation), wind and boundary layer height. RRMSE of the ensemble simulation is almost always smaller than that of any individual model, a fact that will be explained in Section 6.

It is important to examine whether a model is able to simulate a realistic variability in ozone concentrations. A measure of this ability is the sigma ratio, i.e. the standard deviation of the modelled time series divided by the standard deviation of the observed time series. Values for the sigma ratio are given in Table 4. On a seasonal basis, models generally underestimate the observed variability both of daily averages and daily maxima, except in the fall season. This underestimation can be due to a lack of variability in models' forcings (boundary conditions, emissions) or to a lack of horizontal or vertical resolution. The ensemble simulation exhibits even lower variability, due to variance reduction by averaging.

Table 4
Sigma ratio of the modelled and observed daily average and daily maximum O_3 concentrations in 2001

	Daily average					Daily maximum				
	Year	DJF	MAM	JJA	SON	Year	DJF	MAM	JJA	SON
EMEP	**0.92**	**0.93**	0.74	0.72	1.15	**0.88**	0.88	0.76	0.75	1.22
LOTOS	1.15	0.77	**0.96**	0.66	1.45	1.15	**0.95**	**1.08**	0.64	1.65
MATCH	0.89	0.83	0.87	0.72	**1.00**	0.78	0.76	0.81	0.62	**0.98**
CHIMERE	0.84	0.78	0.80	0.79	0.89	0.82	0.71	0.83	**0.82**	0.79
RCG	**0.95**	0.73	**0.93**	0.74	1.03	**1.00**	0.84	**1.13**	0.73	1.18
DEHM	1.17	0.53	1.18	**1.15**	1.43	**1.12**	0.52	1.22	**0.95**	1.43
TM5	0.89	**1.00**	0.80	**0.82**	1.21	0.8	**0.95**	0.73	0.73	1.07
Ensemble	0.88	0.66	0.73	0.62	**0.99**	0.85	0.65	0.78	0.61	**1.03**

Boldface numbers stand for the two sigma ratios closest to 1.

The temporal correlation coefficients for hourly O_3 values are listed in Table 5. For all models, higher correlation coefficients are seen for the whole year than for any of the individual seasons. The reason for this is presumably that the yearly pattern of ozone with high values in summer and lower values in winter is reproduced by the models and therefore contributes positively to the correlation coefficients.

The models exhibit higher correlation coefficients in the summer half year than in the winter half year, especially when looking at the daily maximum. The exception is TM5, which has best correlations in winter. Interestingly models which have large biases or RRMSE do not have a low correlation. Such is the case of CHIMERE, which exhibits the best correlation in summer for both daily averages and daily maxima. For this model, the bias is due to overestimation of boundary conditions, which contributes to degrade model skill for mean values but not for daily variability as these boundary conditions are kept constant for each month. In winter and fall, MATCH, EMEP, TM5 and CHIMERE have higher and comparable correlations.

Again, the ensemble simulation correlation is higher than that of any individual model, both on a yearly and on monthly basis. This average simulation therefore behaves in a more "robust" way, with relatively small fluctuations in the monthly biases (except for September and October due to background values). This property will be commented in Section 6.

5.4. Ozone exceedances

An exceedance day is defined as a day on which the maximum 8-hour average ozone concentration during a day exceeds $120\,\mu g\,m^{-3}$. The number of such days within a calendar year shall not exceed 25 according to the new EU target for 2010. Fig. 5 shows the performance of the models with respect to exceedance days. All models have difficulties in reproducing the correct number of exceedance days. LOTOS-EUROS has a large number of successful exceedance predictions, but also a large number of false alarms. CHIMERE has more successes than false alarms or missed events. However, none of the models has more correct events than wrongly classified events.

5.5. SOMO35

For all models seasonal contributions to SOMO35 are computed and reported in Table 6. Models generally are able to catch the observed SOMO35 levels and their seasonal and spatial variation. The overestimation in the autumn by all models is also found in SOMO35. Other years will show higher observed values of SOMO35 in autumn.

When averaged over the stations considered in this study about 90% of the SOMO35 value is produced during spring and summer. There are, however, large differences among models in both absolute levels (0–7365 ppb day) and the percentage of SOMO35 that is produced in the winter half (0–35%, average 9%). At locations where freshly emitted ozone precursors are available, local ozone is produced in the summer half of the year, while in the winter half NO titration lowers the ozone levels. Hence, at such locations a significant amount of the yearly SOMO35 is produced in the summer half, with only a very small contribution from the winter half. At less polluted locations ozone production in

Table 5
Correlation coefficients for daily average and daily maximum O_3

	Daily average					Daily maximum				
	Year	DJF	MAM	JJA	SON	Year	DJF	MAM	JJA	SON
EMEP	0.72	0.67	0.55	0.50	0.55	0.75	0.60	0.59	0.61	0.53
LOTOS	0.70	0.49	0.54	0.49	0.43	0.76	0.47	0.70	0.66	0.48
MATCH	**0.80**	0.68	**0.66**	0.60	**0.67**	0.81	0.58	0.68	0.7	**0.61**
CHIMERE	0.76	0.62	0.58	**0.64**	0.60	**0.84**	0.62	**0.71**	**0.77**	**0.62**
RCG	0.71	0.58	0.59	0.52	0.36	0.76	0.56	0.70	0.61	0.44
DEHM	0.64	0.45	0.41	0.56	0.31	0.75	0.45	0.60	0.68	0.45
TM5	0.67	**0.69**	0.44	0.35	**0.62**	0.72	**0.63**	0.47	0.51	0.58
Ensemble	**0.79**	**0.74**	**0.66**	**0.68**	0.58	**0.84**	**0.69**	**0.76**	**0.78**	0.59

Boldface numbers stand for the highest two correlations in each column.

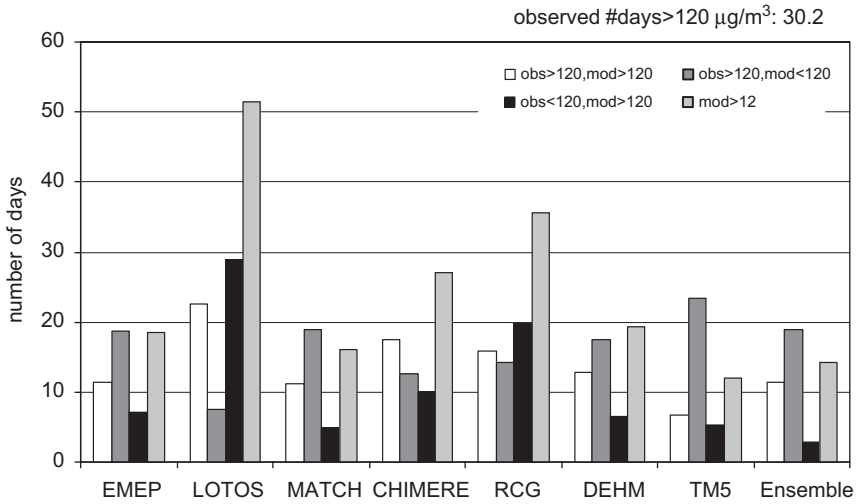

Fig. 5. Simultaneous exceedances or non-exceedances of the 120 µg m^{-3} threshold, for the daily maximum 8-h average ozone concentration, for each model and the ensemble. From left to right: successes of the exceedances, missed exceedances, false alarms, and predicted exceedances. The numbers are calculated over all stations considered together.

Table 6
Observed and modelled SOMO35 values (ppb day) and spatial correlation coefficients for the models

	SOMO35 levels				
	Year	DJF	MAM	JJA	SON
Observed	**2563**	**83**	**940**	**1361**	**174**
EMEP	**2648**	43	833	**1327**	439
LOTOS	3935	50	1426	1870	578
MATCH	2667	**79**	1031	1233	**317**
CHIMERE	3380	150	1147	1613	460
RCG	3480	**76**	1452	1489	453
DEHM	3031	6	**938**	1649	432
TM5	1615	39	451	893	**227**
Ensemble	**2603**	24	**923**	**1323**	327

Boldface numbers stand for the two models closest to the observations.

the summer is weaker, while in the winter ozone concentrations often lie near background values due to the absence of the titration effect. Since background values outside the summer are generally over 35 ppb and drop to around 30 ppb during the summer months, at unpolluted sites the highest contributions to SOMO35 can be expected outside the summer. This effect is seen in the observed SOMO35 values, which show contributions up to 35% in the winter half at the western borders of Europe. Because of summer-condition persistence at the southernmost locations, a significant contribution to SOMO35 is produced in the autumn, mainly in September, while contribution at latitudes higher than 50°N is always very modest.

6. Ensemble performance and uncertainty

In most of the above model skill results, the "ensemble model", whose ozone concentration is the average of the concentrations of all models, exhibits a superior performance to that of individual models. This fact has also been shown in recent studies using several air quality-forecasting models over shorter periods (Delle Monache and Stull, 2003; McKeen et al., 2005). In this section we give a tentative explanation for this fact. Prior to this we discuss the relation between the range of simulated ozone concentrations and their uncertainty. For the simulation of ozone (or other pollutant) concentrations, models not only use the equations of the physics, but also a number of parameterizations, with parameters determined from reduced sets of observations or empirically. The uncertainty of all these values translates into a global uncertainty on the simulated ozone concentrations themselves. In the best case, modellers have selected, independently from one to another, their model options or parameter values. The range of available choices reflects the uncertainty. In the worst case, all modellers have selected the same options, or missed key processes, in which case simulation errors are identical from one model to another.

One way to test whether the model ensemble is representative of the uncertainty consists in checking that the probability distribution of the simulated ozone concentrations is consistent with the distribution of the observations. The rank histogram (Talagrand et al., 1998) is a measure of such a property. Using n models, one obtains a population of n estimates of the actual concentration at a given time. If simulated values are statistically representative of the range of uncertainty, the rank of the observation among the estimates (an integer value between 0 and n) must have an equiprobable distribution.

As an example Fig. 6 shows the rank histogram of summertime (April to September) daily ozone maxima, all stations and days being put together. The first two bins (0 and 1) have a number of counts larger than the other bins, which reflects a general difficulty of models to simulate low daily maxima, thus a small bias of the ensemble. This "ensemble bias" can be calculated for each station by subtracting the mean observed concentration from the mean ensemble average simulated concentration. When removing this bias at each station, the rank histogram becomes flatter. In this bias-free case the ensemble gives a better, but not perfect, account of the uncertainty. However, it should be noted that all models use the same basic emission inventory. As emissions are not perfect, the whole range of uncertainty is not spanned by the model ensemble, which is a limitation of this study.

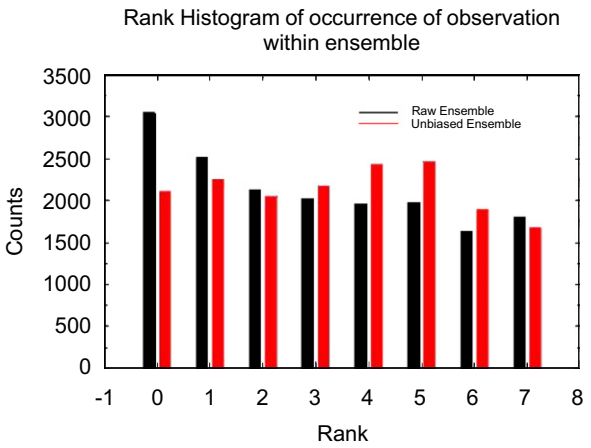

Fig. 6. Rank histogram of summertime ozone daily maxima, combining data from all stations and days, from April to September. Black bars show the rank histogram from the raw ensemble. Light-shaded bars stand for the histogram of unbiased ensemble concentrations, the bias being removed separately for each model and each station.

From these results we conclude that, apart from the bias, the actual ozone concentration has similar statistical properties as any of the simulated ensemble members. This point is discussed in more details in Vautard et al. (2006). For a given day, let us call x_k $(k = 1, \ldots, K = 7)$ the kth model concentration, x_a the actual value and x_m the ensemble average, and b the ensemble bias, which depends on the station.

We assume that the ensemble has a fixed global bias,

$$b = \overline{\left(\frac{1}{K}\sum_{k=1}^{K} x_k - x_a\right)}, \tag{1}$$

where the overbar represents the expectation operator, which is the theoretical average over possible realizations of the process. We also assume that, for a given day, the simulated and observed concentrations are samples randomly drawn from identical distributions (apart from the bias shift) of standard deviation σ, which is a measure of the ensemble spread. The above results obtained with Rank histograms suggest that this condition is not far from being satisfied on average over all days and stations. Under these assumptions, the expected RMSE of the ensemble for that given day,

$$S_{ens} = \overline{\left(\frac{1}{K}\sum_{k=1}^{K} x_k - x_a\right)^2}, \tag{2}$$

can be written, after developing the r.h.s. of (2):

$$S_{ens} = \left(1 + \frac{1}{K}\right)\sigma^2 + b^2, \tag{3}$$

while the variance error of a given model ($K = 1$) is larger,

$$S_{mod} = 2\sigma^2 + b^2, \tag{4}$$

corresponding to $K = 1$. Therefore, for a given day, we expect the simulated ensemble average concentration to be statistically closer to the observation than any individual model. Note that the variance σ depends on the day, some days (windy) being less uncertain than other days (stagnant episodes).

From (4) the RMS error of the ensemble average concentration can be expressed as

$$\mathrm{RMS}_{ens} = \sqrt{\left(1 + \frac{1}{K}\right)\hat{\sigma}^2 + b^2}, \tag{5}$$

where $\hat{\sigma}^2$ is the mean variance of the daily distributions of the ensemble. These statistical

arguments show that the RMS error is therefore a decreasing function of the number of the ensemble members, for the idealized scenario assumed for this derivation. We have verified (not shown) that formula (5) is approximately verified for summertime ozone daily maxima. Ensemble averaging also increases correlation, but analytical calculations are not detailed here. Finally, note that Eq. (5) also shows that the spread of the ensemble, σ, is related to the model skill, a fact that is discussed in more detail in Vautard et al. (2006).

7. Summary and conclusions

This article was designed to evaluate and intercompare seven atmospheric chemistry-transport models, six of which being of regional extent, over Europe, using long-term simulations for the year 2001. We used several statistical and health indicators in order to perform this evaluation. This evaluation is a prerequisite to the use of models for emission reduction scenarios. In this article only the simulation of ozone and O_x is examined.

Most of the models reproduce the observed ozone diurnal cycle, the daily average and daily maxima variabilities with significant realism. These results suggest that the models should be able to simulate the response to emission reduction scenarios. Except for TM5 and DEHM, daytime ozone concentrations are overestimated. LOTOS-EUROS and RCG have a more pronounced diurnal cycle variation than observed, while the reverse occurs for TM5. CHIMERE has a large positive bias, which probably results from a bias in boundary conditions. The other models and the ''ensemble model'', whose concentrations are by definition the average concentration, represent accurately the diurnal cycle. The diurnal cycle of $O_x(= NO_2 + O_3)$ is better simulated than the ozone diurnal cycle, showing that titration of ozone by NO is not well simulated. We could not elucidate the main factors leading to this difference, which results from a combination of several processes: vertical resolution, mixing, dry deposition, emission injection height, representation of ozone profile in the lower boundary layer.

In general, daily maxima are better simulated than daily averages, and summertime concentrations are better simulated than wintertime concentrations. Concentrations in a stable atmosphere, such as often found during night-time and wintertime, are sensitive to low-level stability, boundary conditions, diurnal emission profiles and deposition velocities while daytime and summertime maxima are sensitive to photochemistry. The simulation of the ozone day-to-day variability is assessed by means of percentiles of the ozone distributions, RMSE and correlation. Two health-related indicators are used, the number of days of exceedance and the SOMO35. Both are well reproduced, but the simultaneity of occurrence of exceedance days is not well captured.

For most indicators the ensemble average of the seven simulated values, denoted here as the ''ensemble model'', almost always exhibits a superior skill compared to any individual model, even though it has a too weak variability. We show that the spread of ensemble-model values is fairly representative of the uncertainty of summertime ozone daily maxima, as the occurrence of the observation within the model values range has rather flat distribution, when the bias is removed. One concludes that, for a given day, the probability distribution of occurrence of the observation is well represented by the distribution of the simulated values. This property allowed us to show that the ''ensemble model'' has a theoretical skill, superior to that of any individual model using statistical arguments.

Note, however, that one dimension of uncertainty is not accounted for in the ensemble we study here. All models use the same annual emission inventory. The overall spread of simulated concentrations should increase when taking into account emissions uncertainty. This study could not allow to quantify this effect.

The model intercomparison we have presented here is the first stage of a larger project (EuroDelta) whose aim is to evaluate, with the ensemble of models, the impact of reduction of emission of primary pollutants on air quality at the scale of the European continent. The results presented here show that the ensemble of models used here is fairly representative of the uncertainty in our knowledge of regional air quality, at least for ozone. The key question whether the spread of responses of this ensemble is representative of uncertainty in a scenario mode, for the evaluation of emission control policies, remains, however, open.

Acknowledgments

The manuscript has been improved by the careful reading of two anonymous reviewers. One

of the reviewers made thorough style and English corrections.

References

Akimoto, H., 2003. Global air quality and pollution. Science 302, 1702–1719.

Amann, M., Bertok, I., Cofala, J., Gyarfas, F., Heyes, C., Klimont, Z., Schöpp, W., Winiwater, W., 2005. Clean Air For Europe (CAFE) Programme Final Report, Laxenburg, Austria.

Andersson, C., Langner, J., Bergström, R., 2006. Inter-annual variation and trends in air pollution over Europe due to climate variability during 1958–2001 simulated with a regional CTM coupled to the ERA40 reanalysis. Tellus B. doi:10.1111/j.1600-0889.2006.00196.x.

Bessagnet, B., Hodzic, A., Vautard, R., Beekmann, M., Cheinet, S., Honoré, C., Liousse, C., Rouil, L., 2004. Aerosol modeling with CHIMERE: preliminary evaluation at the continental scale. Atmospheric Environment 38, 2803–2817.

Christensen, J., 1997. The Danish Eulerian hemispheric model—a three dimensional air pollution model used for the Arctic. Atmospheric Environment 31, 4169–4191.

Cros, B., Durand, P., Cachier, H., Drobinski, P., Fréjafon, E., Kottmeier, C., Perros, P.-E., Peuch, V.-H., Ponche, J.-L., Robin, D., Saïd, F., Toupance, G., Wortham, H., 2004. The ESCOMPTE program: an overview. Atmospheric Research 69, 241–279.

Cuvelier, C., Thunis, P., Vautard, R., Amann, M., Bessagnet, B., Bedogni, M., Berkowicz, R., Brandt, J., Brocheton, F., Builtjes, P., Coppalle, A., Denby, B., Douros G., Graf, A., Hellmuth, O., Honoré, C., Hodzic, A., Jonson, J., Kerschbaumer, A., de Leeuw, F., Minguzzi, E., Moussiopoulos, N., Pertot, C., Pirovano, G., Rouil, L., Schaap, M., Stern, R., Tarrason, L., Vignati, E., Volta, M., White, L., Wind, P., Zuber, A., 2007. CityDelta: a model intercomparison to explore the impact of emission reductions in 2010 in European cities in 2010, Atmospheric Environment 41, 189–207.

Delle Monache, L., Stull, R.B., 2003. An ensemble air quality forecast over western Europe during an ozone episode. Atmospheric Environment 37, 3469–3474.

Delle Monache, L., Deng, X., Zhou, Y., Stull, R., 2006a. Ozone ensemble forecasts: 1. A new ensemble design. Journal of Geophysics Research 111 doi:10.1029/2005JD006310.

Delle Monache, L., Deng, X., Zhou, Y., Stull, R., 2006b. Ozone ensemble forecasts: 2. A Kalman filter predictor bias correction. Journal of Geophysics Research 111 doi:10.1029/2005JD006311.

Dentener, F., Stevenson, D., Cofala, J., Mechler, R., Amann, M., Bergamaschi, P., Raes, F., Derwent, R., 2005. The impact of air pollutant and methane emission controls on tropospheric ozone and radiative forcing: CTM calculations for the period 1990–2030. Atmospheric Chemistry and Physics 5, 1731–1755.

Drobinski, P., Saïd, F., Ancellet, G., Arteta, J., Augustin, P., Bastin, S., Brut, A., Caccia, J.L., Campistron, B., Cautenet, S., Colette, A., Cros, B., Corsmeier, U., Coll, I., Dabas, A., Delbarre, H., Dufour, A., Durand, P., Guénard, V., Hasel, M., Kalthoff, N., Kottmeier, C., Lemonsu, A., Lohou, F., Masson, V., Menut, L., Moppert, C., Peuch, V.H., Puygrenier, V., Reitebuch, O., Vautard, R., 2006. Regional transport and dilution during high pollution episodes in southeastern France: summary of findings from the ESCOMPTE experiment. Journal of Geophysics Research, in press.

Frohn, L.M.J., Christensen, H., Brandt, J., 2002. Development of a high resolution nested air pollution model—the numerical approach. Journal of Computational Physics 179, 68–94.

Geels, C., Doney, S., Dargaville, R., Brandt, J., Christensen, J.H., 2004. Investigating the sources of synoptic variability in atmospheric CO_2 measurements over the Northern Hemisphere continents—a regional model study. Tellus 56B, 35–50.

Hass, H., Builtjes, P.J.H., Simpson, D., Stern, R., 1997. Comparison of model results obtained with several European regional air quality models. Atmospheric Environment 31, 3259–3279.

Krol, M., Houweling, S., Bregman, B., van den Broek, M., Segers, A., van Velthoven, P., Peters, W., Dentener, F., Bergamaschi, P., 2005. The two-way nested global chemistry-transport zoom model TM5: algorithm and applications. Atmospheric Chemistry and Physics 5, 417–432.

Mallet, V., Sportisse, B., 2006. Uncertainty in a chemistry-transport model due to physical parameterizations and numerical approximations: an ensemble approach applied to ozone modelling. Journal of Geophysics Research 111 doi:10.1029/2005JD006149.

McKeen, S., Wilczak, J., Grell, G., Djalalova, I., Peckham, S., Hsie, E.-Y., Gong, W., Bouchet, V., Menard, S., Moffet, R., McHenry, J., McQueen, J., Tang, Y., Carmichael, G.R., Pagowski, M., Chan, A., Dye, T., Frost, G., Lee, P., Mathur, R., 2005. Assessment of an ensemble of seven real-time ozone forecasts over eastern North America during the summer of 2004. Journal of Geophysics Research 110 doi:10.1029/2005JD005858.

Parrish, D.D., Holloway, J.S., Trainer, M., Murphy, P.C., Forbes, G.L., Fehsenfeld, F.C., 1993. Export of North American ozone pollution to the North Atlantic Ocean. Science 259, 1436–1439.

Schaap, M., Timmermans, R.M.A., Roemer, M., Boersen, G.A.C., Builtjes, P.J.H., 2006. The LOTOS-EUROS model: description, validation and latest developments. International Journal of Environment and Pollution, in press.

Schmidt, H., Derognat, C., Vautard, R., Beekmann, M., 2001. A comparison of simulated and observed ozone mixing ratios for the summer of 1998 in Western Europe. Atmospheric Environment 36, 6277–6297.

Stern, R., Yamartino, R., Graff, A., 2003. Dispersion modelling within the European community's air quality directives: long term modelling of O_3 PM10 and NO_2. 26th ITM on Air Pollution Modelling and its Application. 26–30 May 2003, Istanbul, Turkey.

Stevenson, D., Doherty, R., Sanderson, M., Johnson, C., Collins, B., Derwent, D., 2005. Impacts of climate change and variability on tropospheric chemistry. Faraday Discussions 130, 41–57 doi:10.1039/b417412g.

Szopa, S., Hauglustaine, D.A., Vautard, R., Menut, L., 2006. Future global tropospheric ozone changes and impact on European air quality. Geophysical Research Letters 33, L14805 doi:10.1029/2006GL025860.

Talagrand, O., Vautard, R., Strauss, B., 1998. Evaluation of probabilistic prediction systems. Proceedings of the Seminar on Predictability, Reading, UK, ECMWF, pp. 1–26.

Tilmes, S., Brandt, J., Flatoy, F., Bergrstrom, R., Flemming, J., Langner, J., Christensen, J.H., Frohn, L.M., Hov, O., Jacobsen, I., Reimer, E., Stern, R., Zimmermann, J., 2002. Comparison of five Eulerian air pollution forecasting systems for the summer of 1999 using the German ozone monitoring data. Journal of Atmospheric Chemistry 42, 91–121.

Van Loon, M., Roemer, M.G.M., Builtjes, P.J.H., 2004. Model inter-comparison in the framework of the review of the unified EMEP model. TNO-Report R2004/282, Apeldoorn, The Netherlands.

Vautard, R., Van Loon, M., Schaap, M., Bergström, R., Bessagnet, B., Brandt, J., Builtjes, P.J.H., Christensen, J.H., Cuvelier, K., Graf, A., Jonson, J.E., Krol, M., Langner, J., Roberts, P., Rouil, L., Stern, R., Tarrasón, L., Thunis, P., Vignati, E., White, L., Wind, P., 2006. Is regional air quality model diversity representative of uncertainty for ozone simulation? Geophysical Research Letters, in press.

Vautard, R., Builtjes, P.H.J., Thunis, P., Cuvelier, K., Bedogni, M., Bessagnet, B., Honoré, C., Moussiopoulos, N., Pirovano, G., Schaap, M., Stern, R., Tarrason, L., VanLoon, M., 2007. Evaluation and intercomparison of Ozone and PM10 simulations by several chemistry-transport models over 4 European cities within the CityDelta project. Atmospheric Environment 41, 173–188.

Vestreng, V., 2003. Review and revision of Emission data reported to CLRTAP. EMEP Status Report, July 2003.

Further reading

Carter, W.P.L., 1996. Condensed atmospheric photooxidation mechanisms for isoprene. Atmospheric Environment 24, 4275–4290.

Schaap, M., van der Gon, H.A.C.D., Dentener, F.J., Visschedijk, A.H.J., van Loon, M., ten Brink, H.M., Puteaud, J.P., Guillaume, B., Liousse, C., Builtjes, P.H.J., 2004. Anthropogenic black carbon and fine aerosol distribution over Europe. Journal of Geophysics Research 109D, 18201.

Toth, Z., Kalnay, E., 1997. Ensemble forecasting at NCEP and the breeding method. Monthly Weather Review 125, 3297–3319.