Ensemble Forecasts of Air Quality in Eastern China

Part 2. Evaluation of the MarcoPolo-Panda Prediction System, Version 1.

Anna Katinka Petersen[1], Guy P Brasseur[1,2], Idir Bouarar[1], Johannes Flemming[3], Michael Gauss[4], Fei Jiang[5], Rostislav Kouznetsov[6], Richard Kranenburg[7], Bas Mijling[8], Vincent-Henri Peuch[3], Matthieu Pommier[4], Arjo Segers[7], Mikhail Sofiev[6], Renske Timmermans[7], Ronald van der A[8,9], Stacy Walters[2], Ying Xie[10], Jianming Xu[10], Guangqiang Zhou[10]

[1] Max Planck Institute for Meteorology, Hamburg, Germany
[2] National Center for Atmospheric Research, Boulder, CO, USA
[3] European Centre for Middle Range Weather Forecasts, Reading, UK.
[4] Norwegian Meteorological Institute, Oslo, Norway
[5] Nanjing University, Nanjing, China
[6] Finnish Meteorological Institute, Helsinki, Finland.
[7] TNO, Utrecht, The Netherlands
[8] Royal Netherlands Meteorological Institute (KNMI), De Bilt, The Netherlands
[9] Nanjing University of Information Science and Technology, Nanjing, China
[10] Shanghai Meteorological Service, Shanghai, China

Abstract:

An operational multi-model forecasting system for air quality has been developed to provide air quality services for urban areas of China. The initial forecasting system included seven state-of-the-art computational models developed and executed in Europe and China (CHIMERE, IFS, EMEP MSC-W, WRF-Chem-MPIM, WRF-Chem-SMS, LOTOS-EUROS and SILAMtest). Several other models joined the prediction system recently, but are not considered in the present analysis. In addition to the individual models, a simple multi-model ensemble was constructed by deriving statistical quantities such as the median and the mean of the predicted concentrations.

The prediction system provides daily forecasts and observational data of surface ozone, nitrogen dioxides and particulate matter for the 37 largest urban agglomerations in China (population higher than 3 million in 2010). These individual forecasts as well as the multi-model ensemble predictions for the next 72 hours are displayed as hourly outputs on a publicly accessible web site (www.marcopolo-panda.eu).

In this paper, the performance of the predictions system (individual models and the multi-model ensemble) for the first operational year (April 2016 until June 2017) has been analysed through statistical indicators using the surface observational data reported at Chinese national monitoring stations. This evaluation aims to investigate a) the seasonal behavior, b) the geographical distribution and c) diurnal variations of the ensemble and model skills. Statistical indicators show that the ensemble product usually provides the best performance compared to the individual model forecasts. The ensemble product is robust even if occasionally some individual model results are missing.

Overall and in spite of some discrepancies, the air quality forecasting system is well suited for the prediction of air pollution events and has the ability to provide alert warning (binary prediction) of air pollution events if bias corrections are applied to improve the ozone predictions.

1. Introduction

With the rapid development of its economy, China has been experiencing repeated intense air pollution episodes (e.g. *Guo et al., 2014, Huang et al., 2014, Wang et al., 2014*) with a wide range of health effects (*Kampa and Castanas 2008; Wu et al., 2012; Hamra et al. 2015; Boynard et al., 2014; WHO, 2018*) and serious consequences on ecosystems (*Fowler et al., 2008, Ashmore, 2005; Leisner et al., 2012; Sinha et al., 2015*) and on climate (*Sitch et al*. 2007; *Brasseur et al., 1999; Akimoto, 2003*). High concentrations of particulate matter often cover a large area of eastern China during winter when air remains stagnant for several days and chemical compounds emitted by power plants, industrial complexes, traffic and domestic infrastructures remain trapped near the surface (e.g. *Wang et al., 2014; Zhao et al., 2013*). During summer, photochemical processes convert nitrogen oxides (NO_X) and volatile organic compounds (VOCs) into tropospheric ozone (O_3) (e.g. *Xu et al., 2008, Sun et al., 2016*).

Long-term solutions to mitigate air pollution require a fundamental transformation of the energy system, which may require decades to be fully implemented. Short-term actions to avoid severe air pollution episodes, however, can be put in place immediately if such episodes can be reliably predicted a few days prior to their occurrence. Comprehensive air quality models that capture meteorological, chemical and physical processes in the troposphere and predict the fate of air pollutants are key tools to forecast the likelihood of air pollution episodes and hence to inform the authorities.

Within the EU projects MarcoPolo and Panda, that include European as well as Chinese partner organizations, an operational multi-model forecasting system for air quality including a number of different chemical transport models has been developed, and is providing daily forecasts of ozone, nitrogen oxides, and particulate matter for the 37 largest urban areas of China (population higher than 3 million in 2010). These individual forecasts as well as the mean and median concentrations for the next 3 days are posted on a dedicated website (www.marcopolo-panda.eu/forecast) together with the hourly observational data from local measurements reported by the Chinese monitoring network of the China National Environmental Monitoring Centre (CNEMC) (data available at www.pm25.in). This operational air quality analysis and forecasting system is presented in detail in a companion paper (*Brasseur et al, 2018*), where the individual models contributing to the MarcoPolo-Panda prediction system are described, and details about the individual models and their individual settings are provided. Information about selected parametrization options for the physical processes, including boundary layer, radiation, convection and surface processes, and about the emissions adopted in MarcoPolo-Panda prediction system are also provided.

In the present study, we evaluate the prediction system of the MarcoPolo and Panda projects that have been in operation for more than one year. We concentrate on the period April 2016 to June 2017 and analyse the model forecasts (7 individual models and the ensemble median) and observational data for 34 cities (covered by most of the models, depending on the extent of the domains, for two models only 31 and 32 cities).

We evaluate the performance of the individual models involved in the present study, and to examine the performance of the overall forecasting system by comparing the predicted surface concentrations to values reported by the Chinese air pollution monitoring network. Section 2 of the paper provides a brief description of the forecasting system, while Section 3 investigates the performance of the system using different statistical indicators including the mean bias (BIAS), the root mean square error (RMSE), the modified normalised bias (MNBIAS), the fractional gross error

96(FGE) and the correlation coefficient. We derive in particular (a) statistical indicators for each
97model over the time of the year (on a monthly basis) in order to analyse seasonal characteristics, (b)
98the geographical distribution of the statistical indicators for the ensemble median in order to derive
99regional characteristics and issues, (c) the statistical indicators of all models and of the ensemble
100median over the time of the day (considering all model-observation pairs of all cities and for the
101whole time period) and for a specific city (Beijing) together with the diurnal variation of the
102pollutants during the whole time period. In Section 4, we assess the impacts of missing forecasts
103from one or more models on the production of the ensemble. As the prediction system intends to
104provide warning of air pollution episodes to the general public, the system performance has been
105evaluated regarding its ability to predict the exceedence of air quality thresholds (binary prediction
106of pollution events). This analysis is presented in Section 5. We conclude with a summary and
107outlook in Section 6.
108
109

1102. Description of the Analysis and Forecasting System

111Within the EU projects MarcoPolo and Panda, a number of chemistry transport models have been
112applied to provide daily air quality forecasts for a selection of 37 large Chinese agglomerations
113(population over 3 million, 2010 census). Initially, seven models, CHIMERE (Royal Netherlands
114Meteorological Institute (KNMI)), IFS (European Centre for Medium Range Weather Forecast
115(ECMWF)), WRF-chem-SMS (Shanghai Meteorological Service (SMS)), SILAMtest (Finish
116Meteorological Institute (FMI)), WRF-chem-MPIM (Max Planck Institute for Meteorology
117(MPIM) in Hamburg), EMEP MSC-W (hereafter referred to as 'EMEP', Norwegian Meteorological
118Institute (MET Norway)) and LOTOS-EUROS (The Netherlands Organisation for Applied
119Scientific Research (TNO)) were providing daily forecasts every day at 0:00 UTC for the next 72
120hours (three days) for NO_2, O_3, PM10 and PM2.5 (see Figure 1). WRF-CMAQ and WRMS-CMAQ,
121both used by Chinese institutions (Nanjing University and SMS), have joined recently the
122prediction system, but are not considered in the present analysis.
123

124We should note that the models considered in the present study may have significantly evolved
125since the present analysis was performed. This is the case, for example, of the SILAM model
126developed by the Finish Meteorological Institute, whose configuration was still in a test mode, and
127is therefore referred to as SILAMtest. Several of the models considered here have been involved in
128a previous intercomparison summarized by Bessagnet et al. (2016).
129

130The individual models are executed independently on the computing systems available in each
131partner institution. The surface concentrations of the key chemical species are extracted locally
132from the model outputs and forwarded to a central database operated by the Royal Netherlands
133Meteorological Institute (KNMI).
134

135Hourly predictions of surface concentrations (expressed in $\mu g/m^3$), are provided by the models as
136grid values, which are bi-linearly interpolated to city center coordinates. The average for the data
137provided by the urban network (usually around 5-12 stations), is posted together with the
138corresponding standard deviation and the number of contributing stations. In the present analysis,
139we consider only the model simulations corresponding to 34 cities, since the cities of Ürümqi (most
140western, only covered by three models), Changchun and Harbin (most northern cities), are located
141outside of the domains covered by most individual models, which are indicated in the companion
142paper (*Brasseur et al., 2018*).
143

144In addition to the forecasts provided by the individual participating models, a multi-model ensemble
145was constructed from which the median and the mean were derived. To process the ensemble
146median, all seven individual models are first interpolated to a common horizontal grid. For each
147grid point, the ensemble model is calculated as the median value of the individual model forecasts.
148The median is relatively insensitive to outliers in the forecasts. The method is also less vulnerable to
149occasionally missing data from individual models, as the minimum number of model results needed
150to calculate a meaningful ensemble mean or median is almost always available. This will be
151discussed in detail in Section 4. The multi-model approach also provides more accurate forecasts
152and thus reduces the underlying uncertainties (as will be shown in the following section). More
153advanced methods, e.g. based on individual model skills, are discussed in the literature (e.g.
154*Galmarini et al, 2013*). They are significantly more costly from a computational point of view and
155therefore not well suited for daily operations.

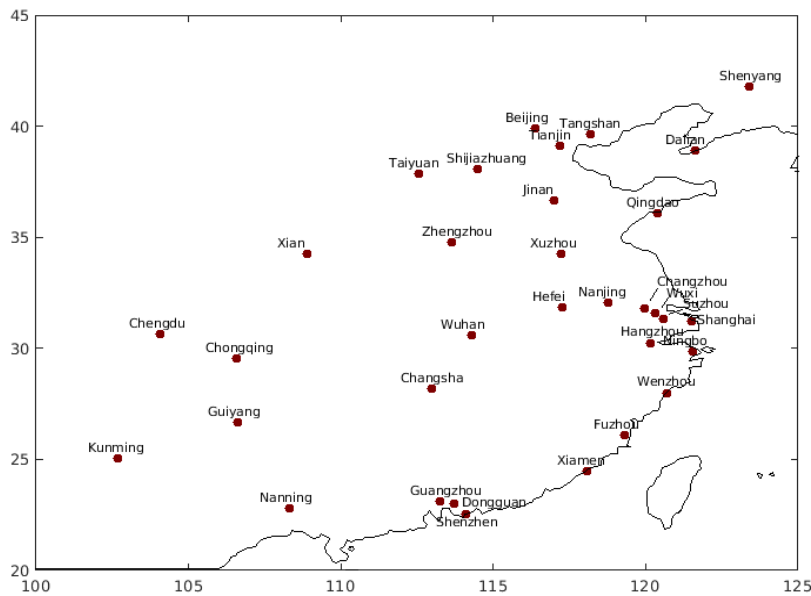



Figure 1: Map of the 34 cities/urban clusters (population over 3 million (2010 census)) with available data (observational and model ensembles), used in this evaluation.

1573. Evaluation of the performance of the system

158
159The evaluation of the performance of a forecasting system is a necessary step for assessing the
160quality of the predictions and demonstrating its usefulness. It also provides important information
161that can lead to the improvement of the forecasting system and to further model development. The
162comparison between model output and in situ measurements is not straightforward because of the
163different nature of the respective quantities: air quality models provide volume averaged quantities
164over each model grid cell and time averages over the modeling time step. Observations are available
165at fixed measurement sites and at a fixed time. Further, they are influenced by local processes that
166are not necessarily well captured by relatively coarse resolution models. Thus, the
167representativeness of the observational site is not always guaranteed.
168
169The MarcoPolo-Panda forecasting and analysis system uses the surface observations available at the
170web site www.pm25.in for 37 Chinese cities. For a given city, the observational data considered for
171the evaluation of the model consist of an average of the measurements made at the different stations
172of the urban network, usually 5 – 12 stations, which are aggregated to one value for the whole city.
173The model fields are bilinearly interpolated to the city center coordinates.

174
175The mean bias
176

$$BIAS = \frac{1}{N} \sum_i \left(m_i - o_i \right),$$

177

178
179where m_i and o_i are the model forecast value and the observation value, and N the number of model-
180observation pairs, the root mean square error
181

$$RMSE = \sqrt{\frac{1}{N} \sum_i \left(m_i - o_i \right)^2},$$

182

183
184the modified normalized bias
185

$$MNBIAS = \frac{2}{N} \sum_i \frac{\left(m_i - o_i \right)}{\left(m_i + o_i \right)},$$

186

187
188the fractional gross error
189

$$FGE = \frac{2}{N} \sum_i \left| \frac{m_i - o_i}{m_i + o_i} \right|$$

190

191
192and the correlation coefficient between the model forecast and observed values
193

$$R = \frac{\frac{1}{N} \sum_i \left(m_i - \bar{m} \right)\left(o_i - \bar{o} \right)}{\sigma_m \sigma_o}$$

194

195
196are used to measure the system performance. Here $\bar{m}$ and $\bar{o}$ are the mean values of the model
197forecast and observed values, and σ_m and σ_o are the corresponding standard deviations.
198
199The evaluation presented here aims to investigate a) the statistical indicators for each model over
200the time of the year (on a monthly basis) so that the seasonal features can be characterized and
201related issues of individual models can be identified (Section 3.1); b) the geographical distribution
202of the statistical indicators of the ensemble median to highlight regional characteristics and related
203issues (Section 3.2); c) statistical indicators of all models and the ensemble median over the time of
204the day (considering all model-observation pairs of all cities and for the whole time period) and for
205a specific city (Beijing) together with the diurnal variation of the pollution species over the whole
206time period (Section 3.3).
207
208

2093.1 Evaluation of the Seasonal Behavior of the Models

210
211We start our evaluation of the multi-model prediction system by examining the seasonal behavior of
212the predicted concentrations of key chemical species. The statistical indicators mentioned above
213have been calculated separately for each month from April 2016 to June 2017 and for the entire
214period during which the forecasting system was operational. Due to storage issues, only the

215predictions for the first 24 hours (0-23h) were saved while the predictions from 24h-72h were not
216retained and not analyzed in this work.
217
218
219Figure 2 shows the RMSE, BIAS, MNBIAS and FGE of NO_2 (left panel) and O_3 (right panel) for
220each of the seven individual models included in the system and for the model ensemble median, for
221each individual month between April 2016 and June 2017. The same results are also provided for
222the whole period ("all"). It can be seen, that there is a wide spread of the results produced by the
223seven models. The individual models are continuously improving during the first months because
224many changes have been applied by the different modeling groups in order to improve their
225individual predictions. In the case of NO_2, most individual models slightly overestimate the
226concentrations compared to observations. In the EMEP model, it may be explained by the larger
227nitrogen emissions used in comparison with the other models (Brasseur et al., 2018). This results in
228a positive BIAS and MNBIAS for most models and the ensemble median. The RMSE of the model
229ensemble is highest in July/August/September 2016 and remains relatively constant after October
2302016. It can be seen, that the median of the model ensemble has the lowest RMSE for NO_2, the
231smallest BIAS and MNBIAS (slightly positive) and the lowest FGE. This demonstrates the
232advantage of adopting a model ensemble rather than the prediction provided by individual models.
233
234Most models underestimate O_3 (likely as a result of the overestimated NO_2 because the O_3
235production is not NOx-limited) during the whole period under consideration. For O_3, the CHIMERE
236model shows slightly better performance (lowest RMSE) than the model ensemble median. The
237median BIAS for O_3 is relatively constant (slightly negative). For this particular species, the model
238ensemble median does not provide the best results regarding the BIAS. In fact, in this case, the
239model LOTOS-EUROS gives the best performance for ozone, Interestingly, this particular model
240has the largest negative BIAS for NO_2. The median BIAS of O_3 remains relatively constant during
241the period, while the MNBIAS exhibits higher negative values during the winter months, as a result
242of the relative low O_3 concentrations during winter time.
243
244As stated above, the MarcoPolo-Panda prediction system has the tendency to overestimate surface
245NO_2, which leads to O_3 titration especially during night time. The emission injection height is also a
246relevant factor here since it can largely influence the results in the planetary boundary layer. During
247night-time, emissions from stacks may be take place above the mixing layer and explain model-data
248discrepancies since the models often assume that the injection of primary pollutants takes place in
249the first layer above the surface.
250
251Anthropogenic emissions of primary pollutants are changing extremely rapidly in China. The
252adopted emissions inventories usually reflect to the situation a few years before the period during
253which the model simulations were performed. Since the recent NO_X emissions have decreased
254significantly in some urban areas of China in response to measures taken by the local authorities (*F.*
255*Liu et al., 2017*), the anthropogenic emissions used for the current forecasts may be overestimated
256in some areas. Some models use reduced NO_X and SO_X anthropogenic emissions (for details see
257*Brasseur et al., 2018*), however, daytime concentrations of ozone are generally underestimated in
258most models, even when the level of NO_2 is in reasonable agreement with the observational values.
259The discrepancy could be caused by an underestimation of the emissions of some VOCs, especially
260in the center of urban areas where ozone is often VOC-limited.
261
262For PM10 and PM2.5, the model ensemble median shows the best performance compared to all
263individual models during the time period under consideration (see Figure 3). For PM10, there is an
264overall slight underestimation by all models except by CHIMERE and hence, by the median of the

265model ensemble. For PM2.5, the BIAS is relatively constant (apart in the WRF-Chem-SMS model
266which exhibits a lot of variation in the BIAS of PM10 and PM2.5). In this case, the BIAS is slightly
267overestimated, but close to zero.
268
269Figure 4 shows the temporal correlation coefficients for NO_2, O_3, PM10 and PM2.5 for each
270individual month, and for the whole time period. It can be seen, that there is a wide spread between
271the individual models: the calculated correlations range from 0.2 to 0.7 for NO_2, PM10 and PM2.5
272and from 0.3 to 0.8 for O_3. The model ensemble median and CHIMERE are characterized by high
273correlation coefficients in the case of NO_2, O_3 and PM2.5. For PM10, the model ensemble median
274and the LOTOS-EUROS model provide the highest correlation coefficients. In general, the model
275ensemble median gives the best performance.
276
277The correlation coefficient of O_3 for the ensemble median remains relatively unchanged during the
278whole time period, and ranges between 0.6 and 0.8. Considering the whole time period, it is of the
279order of 0.75, with CHIMERE providing a slightly higher correlation coefficient for the whole time
280period, and also for each individual months. All models exhibit low correlation coefficients in
281March 2017. High correlation coefficients are found during the early summer months (June/July).
282For PM10 and PM2.5 the correlation coefficients exhibit more variability, starting with very low
283correlation for all models and for the ensemble during April and May 2016, high correlation from
284June 2016 to March 2017, and again low correlation during April and May 2017. These differences
285may be due to missing sources of biomass burning or dust or to individual model tunings. An
286important difference between the models included in the ensemble is the formulation of dust
287mobilization (see Table 3 of the companion paper by Brasseur et al., 2018). Note that the
288CHIMERE and EMEP models do not include dust in their calculation of particulate matter and that
289the emissions provided by the IFS-ECMWF are substantially higher than in other models. For the
290entire time period, the correlation coefficient of the ensemble mean is higher than for each
291individual models (~0.58 for PM10 and ~0.78 for PM2.5). The correlation between the model
292ensemble and the observations is therefore relatively satisfactory.
293

2943.2 Evaluation of the Geographical Distribution

295The statistical indicators, described above for all contributing cities, have also been calculated for
296the individual cities. The purpose here is to assess regional characteristics and to identify model
297issues. Figure 5 shows the statistical indicators (RMSE, BIAS and correlation coefficient) for O_3,
298NO_2 and PM2.5 of the Ensemble Median for each city during the time period under consideration
299(April 2016 until June 2017). In the upper most left panel, the BIAS of ozone for each city is
300shown. It can be seen, that the ensemble median is underestimating the ozone concentrations in the
301north and northeastern regions of China, while no significant bias compared to the observations is
302found in cities in the southern part of the country. RMSE in the northern/northeastern cities are
303higher (around 40 µg m^{-3}) than in southern and western cities (around 20-30 µg m^{-3}).
304
305The temporal correlation coefficients for ozone calculated for each city over the whole period under
306consideration are slightly higher in the northern part of the country and slightly smaller in the
307southern regions. This indicates that the day-to-day variability is well simulated, even though the
308models are slightly underestimating the ozone pollution in the north. NO_2 concentrations (see the
309middle panels of Figure 5) are overestimated in some cities and underestimated in other cities.
310There is, however, no systematic geographical characterization of the bias. When considering
311individual cities, it can be seen that the NO_2 concentrations are slightly overestimated in most urban
312areas including Beijing, Shanghai, Chengdu, Wuhan and Changsha. The missing urban
313parameterization could be one of the reason due to less vertical mixing in the model. The RMSE for

314NO$_2$ in the middle panel of Figure 5 is very uniform (around 20 µg m^{-3}) in the whole country. The
315correlation coefficients of NO$_2$ (between 0.5 and 0.7) are smaller than those of O$_3$, as NO$_2$ exhibits
316more temporal variability than O$_3$. In the case of PM2.5, (see upper most right panel), the
317concentrations are well simulated in the northern and southern parts of China, but there are a few
318city clusters in the middle of the domain (Chengdu, Chongqing, Wuhan and Changsha) in which the
319PM2.5 concentrations are overestimated by more than 50µg m^{-3}. These cities also show an
320overestimation of NO$_2$ concentrations. The overestimation of PM2.5 may therefore be related to
321the errors in precursor emissions, e.g. NO$_X$, SO$_2$. The RMSE of PM2.5 is smaller in the southern
322part of the domain and along the coastline of China, while the model results are less satisfactory in
323the city clusters located in the central part of the domain, with very high RMSE of 60-80µg m^{-3} in
324three cities. The correlation coefficients for the individual cities are relatively constant around 0.7
325with few cities characterized by lower correlation coefficients (mostly in the central part of the
326domain).

327

328 3.3 Evaluation of the diurnal variation

329We now examine the ability of the models to reproduce the diurnal variations of the chemical
330species' concentrations. We first provide a general view based on all observations in China and then
331examine the particular situation in the city of Beijing.
332

333 3.3.a Analysis based on all observations in China

334The RMSE, BIAS, MNBIAS, and FGE of O3, NO2, PM10 and PM2.5 for the seven models and the
335ensemble median for all available observations in China are displayed over the forecasting time (0-
33623h) (Figure 6 and 7). Due to storage limitations, only the predictions for the first 24 hours (0-23h)
337were saved while the predictions for the 24h-72h period performed by all models were not retained.
338Unfortunately, this does not allow the investigation of a day to day degradation of the statistical
339indicators (from day1 to day3). Only the diurnal behavior of the statistical indicators can be
340assessed, which provides important hints for possible model issues.
341

342It can be seen in the left panels of Figure 6 that the statistical indicators of NO$_2$ for the ensemble
343median is relatively stable over the time of the day, with slightly higher RMSE and higher
344BIAS/MNBIAS during the night time hours. For the individual models, the variability of the RMSE
345is somewhat higher during daytime, while some models exhibit very high RMSE and BIAS during
346the night time hours. Most models show a positive BIAS of NO$_2$ during the night, but a few of them
347exhibit a negative bias; this results in a relatively small BIAS for the ensemble median, showing
348good results with respect to the BIAS throughout the day.
349

350In the case of ozone, the statistical indicators exhibit a variation over the time of the day. The
351RMSE is smallest between 7:00 and 9:00 local time, after which it increases until 18:00 in the
352evening to become constant at about 30 µg m^{-3} during the night.
353

354An examination of the BIAS and MNBIAS for O$_3$ over the day shows that O$_3$ is underestimated by
355nearly all models, apart from WRF-Chem-SMS. This might result from the slight overestimation of
356NO$_2$ concentrations by most models. Especially during nighttime when the height of the boundary
357layer is low, near surface NO$_2$ concentrations are high, and ozone is underestimated by 50% – 100%
358by most models. In the first hours of the day, only SILAMtest, WRF-Chem-SMS and LOTOS-

359EUROS exhibit slightly positive O_3 BIAS. The same models produce a negative BIAS for NO_2
360during the first hours of the day.
361
362Figure 7 shows that the BIAS and MNBIAS of both PM10 and PM2.5 stay relatively constant over
363the time of the day. PM10 is slightly underestimated by the ensemble median (-5 to -10%), while
364PM2.5 is slightly overestimated (10 to 25%). In most cases, the models overestimate the PM2.5
365observations, while for PM10 there are stronger differences between the individual models.
366
367For PM10 and PM2.5, the ensemble median exhibits a better performance than the individual
368models: the RMSE BIAS, MNBIAS and FGE of the ensemble are on average lower than the
369corresponding statistical parameters of the individual models. This demonstrates again the
370advantage of using the ensemble median for the prediction of PM10 and PM2.5.
371
372Figure 8 presents the diurnal variation of the concentrations of O_3, NO_2, $O_3 + NO_2$ and PM2.5 from
373the individual models (and the ensemble median) and from the observations at a specific location
374(Beijing). The RMSE and the BIAS are also provided during the whole period under consideration.
375
376It can be seen that the ensemble median (black line) underestimates the O_3 observations (red line)
377throughout the day, especially during the nighttime hours and in the late afternoon. Only WRF-
378Chem-SMS reproduces the amplitude of the O_3 diurnal cycle, but it also underestimates the O_3
379concentrations after 18:00 when the height of the boundary layer is rapidly decreasing. All models
380and the ensemble median reproduce the diurnal cycle with a maximum in the late afternoon, but this
381maximum produced by the model appears about 2 hours earlier than observed. When considering
382the RMSE, the models produce the best results during the morning, and with increasing O_3
383concentrations as the day progresses, the RMSE is also increasing. The negative BIAS is increasing
384for all models and for the model ensemble throughout the day.
385

386 3.3.b Analysis for the specific case of Beijing

387
388In Beijing, the diurnal variation of the NO_2 concentrations is overestimated by the individual
389models as also reflected by the ensemble median. During the nighttime, for example, the observed
390concentrations are about 20-30 µg m^{-3} lower than the concentrations associated with the ensemble
391median. The individual models and the ensemble median show a much stronger diurnal behavior
392than the observations. Atmospheric measurements suggest that the concentrations of NO_2 are
393relatively constant over the time of the day. This might be due to applied temporal profiles of the
394anthropogenic emissions or issues in the vertical mixing of the individual models. Also, the models
395with their spatial resolution may not capture the details seen in the observations by the ground
396network. The RMSE of all models and for the ensemble median is highest in late afternoon and
397during the night. The MarcoPolo-Panda prediction system has thus a tendency to overestimate
398surface NO_2, which leads to an overestimation of the O_3 titration especially at night.
399
400To further analyze the chemical coupling between ozone and NO_2, we have added at each time step
401the mixing ratios of O_3 and NO_2. The resulting variable, called Ox and expressed here in ppbv, has
402the advantage of not being affected by the fast interchange (null cycle) and the resulting partitioning
403between ozone and NO_2 produced by reactions $NO + O_3$, $NO_2 + hv$ and $O + O_2 + M$. If only these
404three rapid photochemical reactions are considered, Ox is a conserved quantity. In other words,
405even when a more comprehensive chemical scheme is adopted, the diurnal cycle of Ox should be
406considerably less pronounced than the diurnal cycle of NO_2 and O_3.
407

408In fact, in the model forecasts, the sum of O_3 and NO_2, is nearly constant during the day, but 409exhibits nevertheless some diurnal variation, which appears to be weaker than in the observation. 410The calculated O_X is slightly too high at night and too low during daytime, suggesting an 411overestimation in photochemical activity by the majority of the models. The partitioning of O_X into 412NO_2 and O_3 is not well reproduced despite the simple chemistry that determines this partitioning: 413NO_2 is generally too high and O_3 too low, especially in the afternoon and early night. The simple 414partitioning approach does not seem to work properly under high NO_X loading. As a result, the 415diurnal cycle of O_3 is not well reproduced by the forecasting ensemble and high ozone events are 416generally underestimated. This issue is discussed in more detail in the companion paper by 417*Brasseur et al., 2018.*
418
419The observed diurnal variation of PM2.5 is not well reproduced by the models and by the ensemble 420median. The calculated variability in Beijing is substantially higher than suggested by the 421observations (which are characterized by relatively constant concentrations throughout the day). 422The models show a maximum in PM2.5 concentrations around 8-9 a.m., and a second maximum 423during nighttime hours. This morning maximum is not present in the observations. The model 424ensemble is overestimating the observations in the morning and underestimating them in the early 425afternoon, resulting in a diurnal variability of the BIAS, shown in the lowest panel. Again, this 426might be related to the adopted diurnal profiles of the anthropogenic emission sources or might be 427due to errors in the formulation of vertical mixing in the PBL. Specifically, one should note that the 428models do not include a detailed formulation if small scale urban canopy effects, which could 429generate some mechanic and thermal turbulence with related vertical mixing during nighttime. With 430increased nighttime ventilation from the boundary layer to the free troposphere, the calculated 431amplitude of the diurnal variation of gases and particulates would be reduced and become closer to 432the observation.
433
434
435

4364. The impact of missing model data on the ensemble performance

437To assess the impact on the ensemble forecast of occasionally missing results from one or several 438models, we compare the following ensembles during a given test period (1-30 May 2017), 439separately for O_3, NO_2 and PM2.5: This approach has already been adopted by *Marécal et al., 2015,* 440to evaluate European air quality predictions. We consider the following cases*:*
441
442- "MEDIAN 7", the median provided by the operational ensemble method, which includes all seven 443models;
444- "MEDIAN 5", the median built on five individual models, excluding the "best" and the "worst" 445models;
446- "MEDIAN 3", the median built on three individual models, excluding the two "best" and the 447"two" worst models;
448- "BEST", the model with the highest performance;
449- "WORST", the model with the lowest performance.
450
451Since the relative performance of individual models varies in time and space, the criterion to order 452the seven individual models from "worst to best" is provided by the value of their respective RMSE 453over the test period. For ozone, the criterion is measured by the RMSE over the 30 days between 45412:00 and 18:00 LST (ozone peak time) (this criterion is based on the fact that the "best" model 455refers to the best forecast of daytime ozone levels). RMSE is seen as the most objective criterion 456since MB and MNMB can include compensating effects.

457
458Figure 9 shows the statistical indicators for May 2017 as a function of the forecasting time (0-23h)
459of the ensemble median based on all 7 models (MEDIAN7, shown in red), 5 models (MEDIAN5,
460shown in blue), and 3 models (MEDIAN3, shown in black). The results are also shown for the
461"best" and the "worst" model (BEST (magenta) and WORST (light blue)). For all three species, the
462ensemble median based on 7 models is of highest quality (based on the statistical indicators used in
463this analysis), and generally surpasses the results provided by the "best" model. When only 5
464models (excluding the best and the worst) are available to calculate the ensemble, all statistical
465indicators show only very small differences with the more inclusive MEDIAN7 case based on seven
466models. Reducing the ensemble calculation further to three models (MEDIAN3), the statistical
467scores degrade slightly compared to the MEDIAN7 and MEDIAN5 for all three species, but remain
468higher or at least similar to the score of the "best" model (BEST).
469
470It is interesting to note that the "best" model (BEST) is not the same model for the different months
471that are investigated, nor the same model for all species. For example, in August 2016, the "best"
472model for O_3 and PM2.5 is IFS, while LOTOS-EUROS shows the best performance for NO_2. In
473May 2017, the best model for PM2.5 is LOTOS-EUROS and the worst model is IFS, but the results
474remain the same: the ensemble product performs better than (or at a similar level as) the best model.
475Since the "BEST" model can change depending on time period and species, the ensemble product is
476particularly valuable for the sustained quality of the forecasting system. This study shows therefore
477that using the ensemble product (median) of models, even if occasionally based on fewer models, is
478more useful than using a single model, even if the performance of this individual model is high. The
479ensemble product is still robust compared to the observations if the output of some contributing
480models is occasionally missing. It also shows that an ensemble product remains valuable even if
481only few models are available for the production of the forecast.
482
483

484## 5. Performance of the Forecasting System for Alert Warnings

485The prediction system has been designed to support the development of policies and the calculation
486of air quality indices. One of the applications of the system is to provide alerts to the general public
487when acute air pollution episodes are expected. Thus, the performance of the forecast system has
488been tested regarding the likelihood to predict air pollution events. We will refer to this type of
489forecast as binary prediction of events (*Brasseur and Jacob, 2017*).
490
491A model prediction of a specific event such as an air pollution episode at a given location (e.g.
492concentration of pollutants exceeding a regulatory threshold) is evaluated by considering a binary
493variable and by distinguishing between four possible situations: (1) the event is predicted and
494observed, (2) the event is not predicted and not observed, (3) the event is predicted but not
495observed, (4) the event is not predicted but is observed. Cases (1) and (2) are regarded as successful
496predictions (hits), while (3) and (4) are considered to be failures (misses). The skill of the model for
497binary prediction (event or no event) is measured by the fractions of observed events that are
498correctly predicted (probability of detection (POD)). The fraction of predicted events, that did not
499occur is measured by the false alarm rate (FAR), both POD and FAR as defined in *Brasseur and*
500*Jacob, 2017*.
501
502We have calculated the POD and the FAR for the ensemble median for the cities of Beijing,
503Shanghai and Guangzhou between April 2016 and June 2017, specifically for ozone (based on the 8
504hour and the daily maximum value), NO_2 and PM2.5. Based on the 1-hourly time series of ozone,
505NO2 and PM2.5, the time series for 1) 1-hour ozone, 2) 8-hour ozone concentrations 3) 24-hour

506mean NO_2 concentrations, 4) 1-hour NO_2 concentrations and 5) 24-hour PM2.5 concentrations have
507been constructed and the thresholds of the air quality indices (AQI) have been applied for each
508definition. The definitions breakpoints for the individual air quality indices (AQI) are shown in
509Table 1 and Table 2; they are based on current definitions of AQI from the Chinese government.
510
511
512**Table 1:** Chinese AQI categories
513

Index values	AQI levels	AQI categories
0-50	1	Good
51-100	2	Moderate
101-150	3	Lightly polluted
151-200	4	Moderately polluted
201-300	5	Heavily polluted
>300	6	Severely polluted

514
515
516**Table 2:** Individual AQI for 1-hour and 8-hour Ozone, 24-hour and 1-hour NO_2 and 24-hour PM2.5
517

IAQI	1-hour O_3 [µg m^{-3}]	8-hour O_3 [µg m^{-3}]	24-hour NO_2 [µg m^{-3}]	1-hour NO_2 [µg m^{-3}]	24-hour PM2.5 [µg m^{-3}]
0	0	0	0	0	0
50	160	100	40	100	35
100	200	160	80	200	75
150	300	215	180	700	115
200	400	265	280	1200	150
300	800	800	565	2340	250
400	1000	Use hourly	750	3090	350
500	1200	Use hourly	940	3840	500

518
519
520In order to highlight the presence of thresholds violated during the time period under consideration,
521Figure 10-12 show the time series for the period April 2016 – July 2017 of the 1) daily maximum
522ozone concentrations, 2) 8-hour moving average of ozone, 3) the 24-hour mean NO_2 concentrations,
5234) the daily maximum NO_2 concentrations and 5) the 24-hour mean PM2.5 concentrations for
524Beijing (Figure 10), Shanghai (Figure 11) and Guangzhou (Figure 12) derived from the model and
525from the observations at each location. Pink lines indicate the thresholds for the air quality indices
526for moderate (line), lightly polluted (dashed line) and moderately polluted (dotted line) conditions
527for each pollutant.
528
529In Beijing and Shanghai, the daily maximum ozone concentrations exceeded the thresholds of 160
530(moderate) and 200 (lightly polluted) within the considered time period only during the months of
531April to September 2016. During the months of October 2016 to March 2017, the ozone
532concentrations remained below the threshold of 160, highlighting fair air quality conditions with
533regard to ozone in wintertime. In Beijing, the ensemble median has a probability of detection of air
534pollution events for moderate 1-hour ozone AQI of 0.44 (55 out of 126 events of 1-hour ozone

535breaking the threshold of 160 µg m^{-3} have been detected). The False Alarm Rate (FAR) is 0.05 (the 536model ensemble predicted 58 events where ozone exceeds the threshold of 160 µg m^{-3}, where 3 out 537of these 58 events were false alarm (observations below the threshold). Lightly polluted events (1-538hour ozone exceeding 200 µg m^{-3}) were correctly predicted only 14 times, while the observations 539exceeded the threshold 79 times. The FAR for lightly polluted ozone events is 0.12 (2 out of 16).
540
541For moderately polluted ozone events (1-hour ozone exceeding 300 µg m^{-3}), the POD is 0, the 542model ensemble was not able to predict the 4 observed events (FAR is not applicable, (0 out of 0)). 543Looking at the 8-hour ozone predictions for Beijing, the model ensemble is very similar, with a 544POD of 0.45 (864 out of the 1921 observed events have been predicted correctly) and a FAR of 5450.06 (56 counts are false alarm out of 920 events). For lightly polluted ozone conditions, the POD is 5460.18 (118 out of 657 observed events) with a FAR = 0.06 (7 out of 125 are false alarm). For 547moderately polluted conditions, the model ensemble predicted 7 out of 150 observed events 548correctly with a FAR of 0.22 (2 out of 9 alarms are false).
549
550For Shanghai, the PODs for ozone predictions are lower than in Beijing: for moderate air quality 551conditions, the POD is 0.16 (15 out of 92 observed events are predicted correctly) with a FAR of 0 552(no false alarm) for 1-hour ozone predictions, and POD = 0.21 (488 out of 2346 observed events) 553with a FAR of 0.01 (7 false alarms relative to 495 counts) for 8-hour ozone predictions. For lightly 554polluted conditions, the POD is decreasing: POD = 0.08 (3 correct predictions out of 38 observed 555events) with FAR of 0 (no false alarm, 3 correct predictions) for 1-hour ozone, and POD = 0.07 (27 556out of 398 observed) with a FAR of 0.10 (3 false alarms out of 30) for 8-hour ozone. For 557moderately polluted conditions (1-hour ozone exceeding 300 µg m^{-3} or 8-hour ozone exceeding 215 558µg m^{-3}), the POD for 1-hour ozone is not applicable (no predicted, no observed events), and for 8-559hour ozone POD = 0 (0 predicted out of the 29 observed), FAR = 1 (2 false alarms out of 2 560predicted, but not observed).
561
562In Guangzhou, there is no clear difference between ozone conditions in summer or wintertime 563during the considered time period. Ozone observations regularly exceed the threshold of 160 564(moderate) and 200 µg m^{-3} (lightly polluted) during the whole time period, and 5 times 1-hour 565ozone is exceeding the threshold of 300 µg m^{-3}.
566
567The POD of 1-hour ozone in Guangzhou is 0.16 (15 correct predictions out of 94 observed) with 568FAR = 0.21 (4 false alarms out of 19 predicted) for moderate conditions, and POD = 0.03 (1 569predicted out of 36 observed) with FAR = 0 (0 out of 1 predicted) for lightly polluted conditions, 570and POD = 0 (0 predicted out of 5 observed events) for moderately polluted ozone conditions. For 5718-hour ozone, the POD is 0.31 (315 correct predicted out of 1032 observed) with FAR = 0.28 (122 572false alarms of 437 predicted events) for moderate conditions, POD = 0.06 (12 out of 217 observed) 573with FAR = 0 (no false alarm out of 12 predicted events) for lightly polluted ozone conditions, and 574POD = 0 (0 out of 47 observed events) for moderately polluted ozone conditions.
575
576In general, the ability of the model ensemble to predict correctly ozone air pollution events is best 577for light ozone pollution, while it fails to predict correctly the ozone pollution events for moderately 578polluted situations. This is mostly a result of the model ensemble being too low compared to the 579observations. The predictions can be improved by applying a bias correction to the ozone 580predictions. This is investigated in the following Section 5.1.
581
582The NO$_2$ predictions of the ensemble median are in general too high compared to the observation, 583especially in Beijing and Shanghai. Especially, in summertime (June/July/August/September), the 584model predictions are sometimes twice as high as the observations, which might be a result of

585uncertainties in the emissions. In all three cities under consideration, the NO_2 concentrations are 586only exceeding the thresholds of 40 µg m^{-3} for 24-hour NO_2 (100 for 1-hour NO_2) and 80 µg m^{-3} for 58724-hour NO_2 (200 µg m^{-3} for 1-hour NO_2) during the considered period (moderate and lightly 588polluted conditions for NO_2). During wintertime (November/December/January), the observations 589are slightly higher than in summer and the ensemble system is in better agreement with the 590observations.

591

592In Beijing, the POD for 24-hour NO_2 is 1 (214 of 214 observed events are predicted) for moderate 593conditions with a FAR of 0.46 (180 false alarms relative to 394 predicted events). This indicates 594that NO_2 is generally overestimated by the model ensemble. For lightly polluted events, the POD is 5950.79 (27 predicted out of 34 observed events) with FAR = 0.70 (63 false alarms out of 90 596predicted). For the 1-hour NO_2, the POD for moderate conditions is 0.61 (36 out of 59 observed 597events) with FAR = 0.80 (141 false alarms out of 177 predicted). For lightly polluted conditions, no 598events have been observed nor predicted for 1-hour NO_2 in Beijing during the considered period. In 599Beijing, the threshold for moderately polluted NO_2 conditions has not been exceeded neither by 1-600hour NO_2 nor by 24h- NO_2 during the considered period.

601

602In Shanghai, the numbers are very similar to those in Beijing: POD for 24-hour NO_2 is 1 (208 of 603208 observed events are predicted) for moderate conditions with a FAR of 0.42 (152 false alarms of 604360 predicted events). There is also a general overestimation by the model ensemble compared to 605the observations. For lightly polluted conditions, the POD for 24-hour NO_2 is 0.67 (10 out of 15 606observed) and a FAR of 0.86 (60 false alarms of 70 predicted), which is a clear result of the 607overestimated NO_2. For the 1-hour NO_2, the POD is 0.91 (48 predicted out of 53 observed) with a 608FAR of 0.70 (111 false alarms out of 159 predicted) for moderate conditions. The thresholds for 609lightly polluted and moderately polluted conditions for 1-hour NO_2 have not been exceeded in 610Shanghai during the considered period, but there was 1 false alarm (1 out of 1) for lightly polluted 611conditions.

612

613In Guangzhou, the model ensemble and the observations for NO_2 are in better agreement. There is 614slight overestimation of the NO_2 concentrations from May to September 2016, and in May 2017, 615but in general, there is a good agreement between the model time series and the observations. The 616POD for 24h-NO_2 exceeding the threshold for moderate conditions is 0.94 (208 predicted out of 222 617observed) with a FAR of 0.35 (110 false alarms of 318 predicted events), for lightly polluted 618conditions POD is 0.56 (15 predicted out of 27 observed) with 32 false alarms out of 47 predicted 619events (FAR = 0.69). Stronger polluted events have not been observed nor predicted for NO_2 in 620Guangzhou. For the 1-hour NO_2, 58 events have been predicted out of 76 observed for moderate 621conditions (POD = 0.76, FAR = 0.63 (97 false alarms out of 155 predicted). For lightly polluted 622conditions, there was 1 false alarm (1 out of 1), with neither observed nor correctly predicted 623events.

624The thresholds for moderately polluted conditions for 24-hour NO_2 and 1-hour NO_2 have not been 625exceeded in Guangzhou during the considered period, no events have been predicted nor observed.

626

627The predictions of PM2.5 concentrations (24-hour PM2.5) of the model ensemble are in very good 628agreement with the observations in all three cities during the considered period.

629

630In Beijing, the POD for the prediction of moderate condition for 24-h PM2.5 is 0.95 (268 correctly 631predicted events out of 283 observed) with a FAR of 0.19 (61 false alarms out of 329 predicted 632events). For lightly polluted conditions, the POD is 0.76 (111 correct predicted events of 146 633observed events) with a FAR of 0.28 (43 false alarms for 154 predicted events). Moderately

634polluted PM2.5 events have been correctly predicted 33 times out of 64 observed events (POD =
6350.52) with a FAR of 0.35 (18 false alarms out of 51 predicted events).
636
637In Shanghai, 191 moderate condition-events for PM2.5 have been correctly predicted out of 220
638observed events (POD = 0.87, FAR = 0.19), with 46 false alarms out of the 237 predicted events.
639For lightly polluted events, the POD is 0.84 (32 out of 38 observed events) with a FAR of o.47 (28
640false alarms of 60 predicted events). For moderately polluted conditions of PM2.5, the POD is 0.50
641(3 correctly predicted events out of 6 observed) with a relatively high FAR (0.67, 6 false alarms out
642of 9 predicted).
643
644In Guangzhou, the POD for moderate conditions of PM2.5 is 0.85 (149 correctly predicted out of
645175 observed) with 65 false alarms out of 214 predicted events (FAR = 0.30). Lightly polluted
646events have been observed only 7 times, the ensemble median predicted 4 of them correctly (POD =
6470.57), but with a very high false alarm rate (16 false alarms out of 20 predicted events, FAR =
6480.80), this indicates a slight overestimation of the PM2.5 concentrations of the models compared to
649the observations. In Guangzhou, no moderately polluted events of PM2.5 have been observed nor
650predicted during the considered period.
651
652Only in Beijing, and only with regard to 24-hour PM2.5, heavily polluted conditions have been
653observed and predicted during the considered period in the winter months 2016/2017: The POD is
6540.5 (18 correct predicted out of 36 observed events) with a FAR of 0.28 (7 false alarms out of 25).
655
656These investigations show, that the model ensemble is well suited to be used in air quality
657predictions of PM2.5. For ozone, due to biases of the model ensemble compared to observations,
658the model ensemble is not able to predict ozone pollution in an appropriate way. Although the FAR
659is very low for ozone predictions, the POD of model ensemble is not very high. In the following
660Section, we apply bias correction to improve the predictions for ozone pollution events.
661

6625.1 Bias Correction for Ozone Predictions

663Bias corrections can be applied to improve the predictions of an individual model or a model
664ensemble. In our case, we have calculated the summertime bias of the time series of the hourly
665ozone concentrations from the model ensemble with respect to the hourly observations, and
666subtracted the bias from the hourly time series. For predictions of ozone air pollution, the
667summertime is an appropriate season to consider since the ozone thresholds are exceeded only
668during this season. As the bias between the observations and the model might not be the same for
669each month, and our goal is to obtain the best improvement in the ozone predictions for
670summertime, we have subtracted the mean summertime bias (mean of the bias of June/July/August/
671September 2016) from the original time series. The daily maximum ozone values and the 8-hour
672moving average for the corrected time series have then been calculated. The resulting, POD and
673FAR for 1-hour ozone and 8-hour ozone under different air quality conditions are shown in Table 3.
674This table shows that, for bias-corrected predictions, the POD in all three cities is larger than for the
675non-corrected time series, especially in the case of moderate and lightly polluted conditions of
676ozone. Thus, the predictions of air pollution events are significantly improved when the bias
677correction is applied in the case of ozone. Only for the predictions of moderately polluted
678conditions of ozone, the POD is not changing. The FAR is also slightly decreasing for all cities, but
679the improvement is small.
680
681In Beijing, the POD air pollution events represented by a moderate AQI for 1-hour ozone increased
682from 0.44 for Beijing (55 out of 126 observed events) before bias correction to 0.69 (87 out of 126

683events) after bias correction. The False Alarm Rate (FAR) also increased from 0.05 (3 false alarms
684out of these 58 events) to 0.10 (10 false alarms out of 97 predicted events). Lightly polluted events
685(1-hour ozone exceeding 200 µg m^{-3}) have been predicted correctly 31 times (14 times without the
686corrections), while the observations exceeded the threshold 79 times. The FAR for lightly polluted
687ozone events also slightly increased from 0.125 (2 out of 16) to 0.2 (8 false alarms out of 40).
688
689For moderately polluted ozone events (1-hour ozone exceeding 300 µg m^{-3}), the POD for the bias-
690corrected prediction is still 0. The model ensemble was not able to predict the 4 observed events
691(FAR is not applicable, (0 out of 0)).
692
693Looking at the 8-hour ozone predictions for Beijing, the POD of 0.45 (864 out of the 1921 observed
694events have been predicted correctly) increased to 0.76 (1452 out of 1921) after bias corrections,
695and the FAR from 0.06 (56 counts are false alarm out of 920) to 0.23 (424 false alarms out of 1876
696predictions) for moderate ozone pollution. For lightly polluted ozone conditions, the POD increased
697to 0.44 (291 out of 657) and FAR = 0.22 (81 false alarms of 372 predicted) for the bias corrected
698predictions compared to POD = 0.18 (118 out of 657 observed events) with a FAR = 0.06 (7 out of
699125 are false alarm). For moderately polluted conditions, the model ensemble with bias corrected
700predicted 27 (instead of only 7) out of 150 observed events correctly with a FAR of 0.28 (13 false
701alarms of 47 predictions) compared to FAR of 0.22 (2 out of 9 are false alarm).
702
703For Shanghai, for moderate air quality conditions of ozone, the POD increased from 0.16 to 0.51
704(47 (15 for non-corrected) out of 92 observed events are predicted correctly); the FAR increased
705from 0 (no false alarm) to 0.10 (5 false alarms out of 52) for 1-hour Ozone predictions. For 8-hour
706ozone predictions, the POD increased from 0.21 to 0.66 (1554 (non-corrected: 488) out of 2346
707observed events), the FAR increased from 0.01 (7 false alarms of 495 predicted events) to 0.32
708(726 false alarms of 2280 counts) for 8-hour ozone predictions. For lightly polluted ozone
709conditions, the POD increased from 0.08 (3 correct predictions out of 38 observed) with FAR of 0
710(no false alarm, 3 correct predictions) to POD = 0.34 (13 out of 38) with FAR = 0.07 (1 false alarm
711of 14 predicted events) for 1-hour ozone, and for 8-hour ozone, the POD increased from 0.07 to
7120.27 (109 (non-corrected: 27) out of 398 observed) and the FAR increased from 0.10 (3 false alarms
713out of 30) to 0.13 (16 false alarms in 125 predicted events). For moderately polluted ozone
714conditions, the POD for 1-hour ozone is not applicable for both non-corrected and bias-corrected
715predictions (no predicted, no observed events), but for the bias-corrected prediction, one false alarm
716is observed (FAR = 1, 1 false alarm in 1 predicted event), and for 8-hour ozone POD increased
717from 0 to 0.10 (3 (non-corrected: 0) predicted out of the 29 observed), the FAR decreased from 1 (2
718false alarms out of 2 predicted, but not observed) to 0.8 (12 false alarms of 15 predicted events).
719
720In Guangzhou, the predictions are not as accurate as in Beijing and Shanghai, and the bias
721corrections result only in slight improvements of the ozone forecasts for Guangzhou. The POD of 1-
722hour ozone in Guangzhou increased from 0.16 to 0.32 (30 (non-corrected: 15) correct predictions
723out of 94 observed) and the FAR slightly increased from 0.21 (4 false alarms out of 19 predicted) to
7240.33 (15 false alarms out of 45 predicted events) for moderate conditions. For lightly polluted ozone
725conditions, the POD increased from 0.03 to 0.14 (5 (non corrected: 1) predicted out of 36 observed)
726and the FAR increased from 0 (0 out of 1 predicted) to 0.29 (2 false alarms of 7 predicted events).
727For moderately polluted ozone predictions, the POD and FAR did not change with bias corrections
728(POD = 0 (0 predicted out of 5 observed events), FAR not applicable).
729
730For 8-hour ozone of moderate conditions, the POD increased from 0.31 to 0.49 (508 (non-corrected:
731315) correct predicted out of 1032 observed) and the FAR increased from 0.28 (122 false alarms of
732437 predicted events) to 0.37 (296 false alarms for 804 predictions). For lightly polluted ozone

733conditions the POD increased from 0.06 to 0.13 (29 (non-corrected: 12) out of 217 observed) and
734the FAR increased from 0 (no false alarm out of 12 predicted events) to 0.19 (7 false alarms for 36
735predicted events). For moderately polluted ozone conditions, the POD and FAR did not change with
736bias corrections (POD= 0 (0 out of 47 observed events), FAR not applicable).
737
738Figure 13 a–c shows the time series of the model ensemble, the bias corrected time series of the
739model ensemble and the observations. For the daily maximum ozone, the bias correction results in a
740better agreement with the observations, which also results in better event predictions. For 8-hour
741ozone, there is better agreement during summertime, while during the wintertime, the bias-corrected
742ozone time series are too high compared to the observations (both correcting for the bias derived
743from the total time series, or only from the summertime time series). This shows (as we have seen
744in Section 3.1), that the bias is not the same during the whole year, and also that the diurnal cycle of
745ozone is not well captured by the model ensemble. While the bias corrected daily maximum ozone
746is in better agreement with the observations, the 8-hour bias corrected moving average is too high
747during winter time (with very low ozone concentrations). As the ozone is too low in winter to
748exceed the lowest threshold (moderate conditions) for air quality index calculations, this is not
749affecting the quality of the event prediction. A more sophisticated bias-correction (bias correction
750with diurnal and annual variation included) could be applied to further improve the predictions,
751provided that a longer time series (more than one year of data) is available. The statistical bias
752correction can then be used for the improvement of future predictions.
753
754

755 6. Conclusions and Future Developments

756
757In this paper, we evaluate the forecasting system developed and implemented as part of the EU
758Panda and MarcoPolo projects after a little more than one year of operation. The forecasting system
759is based on an ensemble of seven state-of-the-art chemistry-transport models (CHIMERE, EMEP,
760IFS, LOTOS-EUROS, WRF-Chem-MPIM, WRF-Chem-SMS, SILAMtest). Each model is
761executed on a computer platform hosted by individual institutes in China and Europe. Input for
762meteorological forcing, emissions and boundary conditions have been carefully chosen and adopted
763for the specific situation of China, but vary from model to model. The forecasting system provides
764every day hourly forecasts for 3 days ahead for four major chemical pollutants (O_3, NO_2, PM10 and
765PM2.5) together with hourly observational data provided by the Chinese observational network
766(www.pm25.in).
767
768The models, whose predictions are strongly influenced by the adopted weather forecast, reproduce
769in general the regional features and capture many air pollution events. In most cases, the model
770ensemble reproduces satisfactorily the day-to-day variability of the concentrations of the primary
771and secondary air pollutants and in particular, predicts the occurrence of pollution events a few days
772before they occur. Overall, and in spite of some discrepancies, the air quality forecasting system is
773well suited for the prediction of air pollution events and has the ability to be used for alert warning
774(binary prediction) of the general public, specifically if bias corrections are applied to improve the
775ozone forecasts.
776
777In most cases, the ensemble approach provides more accurate forecasts and reduces the
778uncertainties in comparison with the individual models results. The calculation of the median of all
779models is also relatively insensitive to model outliers, and is computationally efficient. Using the
780ensemble median based on all models provides the best performance for all species, as the relative
781performance of any individual model may vary in time, space and species. We showed, that the

782ensemble product, even if occasionally based on fewer models, is more useful than a single model 783of good quality, and that the ensemble product is still robust compared to the observations if data 784from some contributing models are occasionally missing.
785
786Despite the fact that the prediction system is in its development phase and that the resources 787available to improve the system are limited, the MarcoPolo and Panda forecasting system can be 788viewed as already quite successful. The inter-comparison presented in the companion paper by 789*Brasseur et al., 2018* and the present evaluation were performed to diagnose differences between 790models, identify problems and contribute to individual model improvements. Specifically, the 791underestimation of ozone under high NO_X conditions and the resulting errors in the diurnal cycle of 792ozone need to be addressed in an effort to improve the model forecasts in China. Although major 793efforts are ongoing to improve emission inventories for China, the remaining uncertainties, 794especially in regard to local emissions, may partly explain the differences between models and 795observations. This is subject of further investigation. Furthermore, data assimilation of satellite and 796in situ observations should significantly improve the performance of the forecasting system (see 797e.g., Mizzi et al., 2016). Finally, a more advanced approach to extract observations provided by the 798Chinese network is expected to improve the model-data comparison.

799Data Availability

800
801The models described here are used operationally by the participating research and service 802organizations involved in the present study. The data produced by the multi-model forecasting 803system are available from the Royal Dutch Meteorological Institute (KNMI).
804

805Author contributions

806AKP were in charge of the WRF-Chem simulations and performed the evaluation of the 807simulations. GPB coordinated the Panda Project and contributed to the analysis of the results 808provided by the WRF-Chem model. RvdA coordinated the MarcoPolo Project and was involved in 809the analysis of the results provided by the Chimere model. IB and SW were in charge of the WRF-810Chem simulations. YX, JX and GZ developed and used the WRF-Chem-SMS model. VHP and JF 811performed the simulations with the IFS model. MG and MP were in charge of the simulations 812performed by the EMEP model. FJ is in charge of the WRF-CMAQ model. MS and RK were 813responsible for the forecasts made the SILAM model, while RT, AS and RK were using the 814LOTOS-EUROS model. BM developed the MarcoPolo and Panda website and collected all the 815model results and observational data.

816Acknowledgements

817
818The model inter-comparison presented in the present study has been conducted during a workshop 819organized in May 2017 by the Shanghai Meteorological Service (SMS) in China. The authors thank 820Dr. Jianming Xu for hosting this meeting and providing support to the participants. The ensemble of 821models described here has been produced under the Panda and MarcoPolo projects supported by the 822European Commission within the Framework Program 7 (FP7) under grant agreements n°606719 823and n°606953. The National Center for Atmospheric Research (NCAR) is sponsored by the US 824National Science Foundation. We thank the two anonymous reviewers whose comments helped 825improve and clarify this manuscript.
826
827

828**Table 3**: POD and FAR for Beijing, Shanghai and Guangzhou
829

	Probability of Detection (POD)			False Alarm Rate (FAR)		
Beijing	**AQI 2** (moderate)	**AQI 3** (lightly poll.)	**AQI 4** (moderately poll.)	**AQI 2** (moderate)	**AQI 3** (lightly poll.)	**AQI 4** (moderately poll.)
1-hour O_3 [µg m^{-3}]	0.44 (55/126)	0.18 (14/79)	0 (0/4)	0.05 (3/58)	0.12 (2/16)	NA (0/0)
Bias corrected 1-hour O_3 [µg m^{-3}]	0.69 (87/126)	0.41 (32/79)	0 (0/4)	0.10 (10/97)	0.20 (8/40)	NA (0/0)
8-hour O_3 [µg m^{-3}]	0.45 (864/1921)	0.18 (118/657)	0.05 (7/150)	0.06 (56/920)	0.06 (7/125)	0.22 (2/9)
Bias corrected 8-hour O_3 [µg m^{-3}]	0.76 (1452/1921)	0.44 (291/657)	0.23 (34/150)	0.23 (424/1876)	0.21 (81/372)	0.28 (13/47)
24-hour NO_2 [µg m^{-3}]	1 (214/214)	0.79 (27/34)	NA (0/0)	0.46 180/394)	0.70 (63/90)	NA (0/0)
1-hour NO_2 [µg m^{-3}]	0.61 (36/59)	NA (0/0)	NA (0/0)	0.80 (141/177)	NA (0/0)	NA (0/0)
24-hour PM2.5 [µg m^{-3}]	0.95 (268/283)	0.76 (111/146)	0.52 (33/64)	0.19 (61/329)	0.28 (43/154)	0.35 (18/51)
Shanghai						
1-hour O_3 [µg m^{-3}]	0.16 (15/92)	0.08 (3/38)	NA (0/0)	0 (0/15)	0 (0/3)	NA (0/0)
Bias corrected 1-hour O_3 [µg m^{-3}]	0.51 (47/92)	0.34 (13/38)	NA (0/0)	0.10 (5/52)	0.07 (1/14)	1 (1/1)
8-hour O_3 [µg m^{-3}]	0.21 (488/2346)	0.07 (27(398)	0 (0/29)	0.01 (7/495)	0.10 (3/30)	1 (2/2)
Bias corrected 8-hour O_3 [µg m^{-3}]	0.66 (1554/2346)	0.27 (109/398)	0.10 (3/29)	0.32 (726/2280)	0.13 (16/125)	0.80 (12/15)
24-hour NO_2 [µg m^{-3}]	1 (208/208)	0.67 (10/15)	NA (0/0)	0.42 (152/360)	0.86 (60/70)	NA (0/0)
1-hour NO_2 [µg m^{-3}]	0.91 (48/53)	NA (0/0)	NA (0/0)	0.70 (111/159)	1 (1/1)	NA (0/0)
24-hour PM2.5 [µg m^{-3}]	0.87 (191/220)	0.84 (32/38)	0.50 (3/6)	0.19 (46/237)	0.47 (28/60)	0.67 (6/9)
Guangzhou						
1-hour O_3 [µg m^{-3}]	0.16 (15/94)	0.03 (1/36)	0 (0/5)	0.21 (4/19)	0 (0/1)	NA (0/0)
Bias corrected 1-hour O_3 [µg m^{-3}]	0.32 (30/94)	0.14 (5/36)	0 (0/5)	0.33 (15/45)	0.29 (2/7)	NA (0/0)
8-hour O_3 [µg m^{-3}]	0.31 (315/1032)	0.06 (12/217)	0 (0/47)	0.28 (122/437)	0 (0/12)	NA (0/0)
Bias corrected 8-hour O_3 [µg m^{-3}]	0.49 (508/1032)	0.13 (29/217)	0 (0/47)	0.37 (296/804)	0.19 (7/36)	NA (0/0)
24-hour NO_2 [µg m^{-3}]	0.94 (208/222)	0.56 (15/27)	NA (0/0)	0.35 (110/318)	0.68 (32/47)	NA (0/0)
1-hour NO_2 [µg m^{-3}]	0.76 (58/76)	NA (0/0)	NA (0/0)	0.63 (97/155)	1 (1/1)	NA (0/0)
24-hour PM2.5 [µg m^{-3}]	0.85 (149/175)	0.57 (4/7)	NA (0/0)	0.30 (65/214)	0.80 (16/20)	NA (0/0)

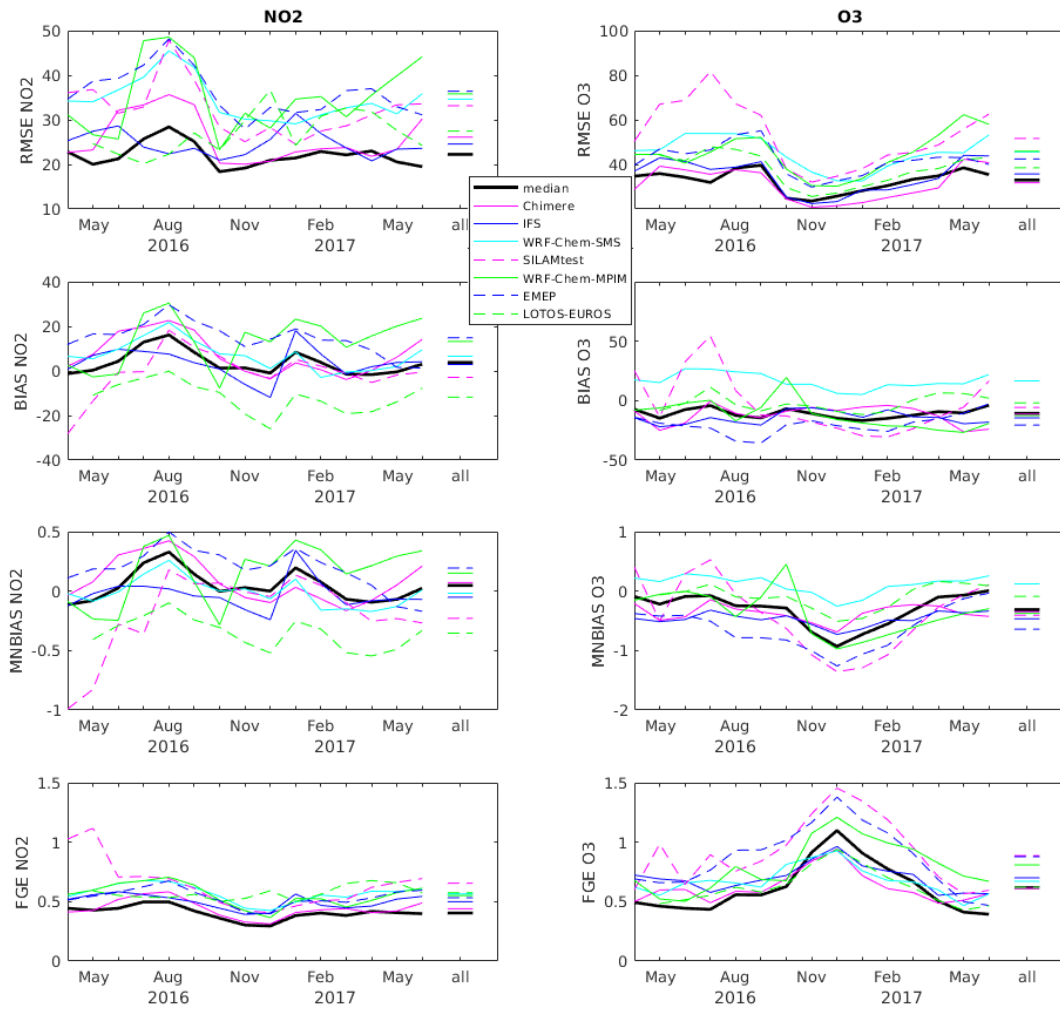
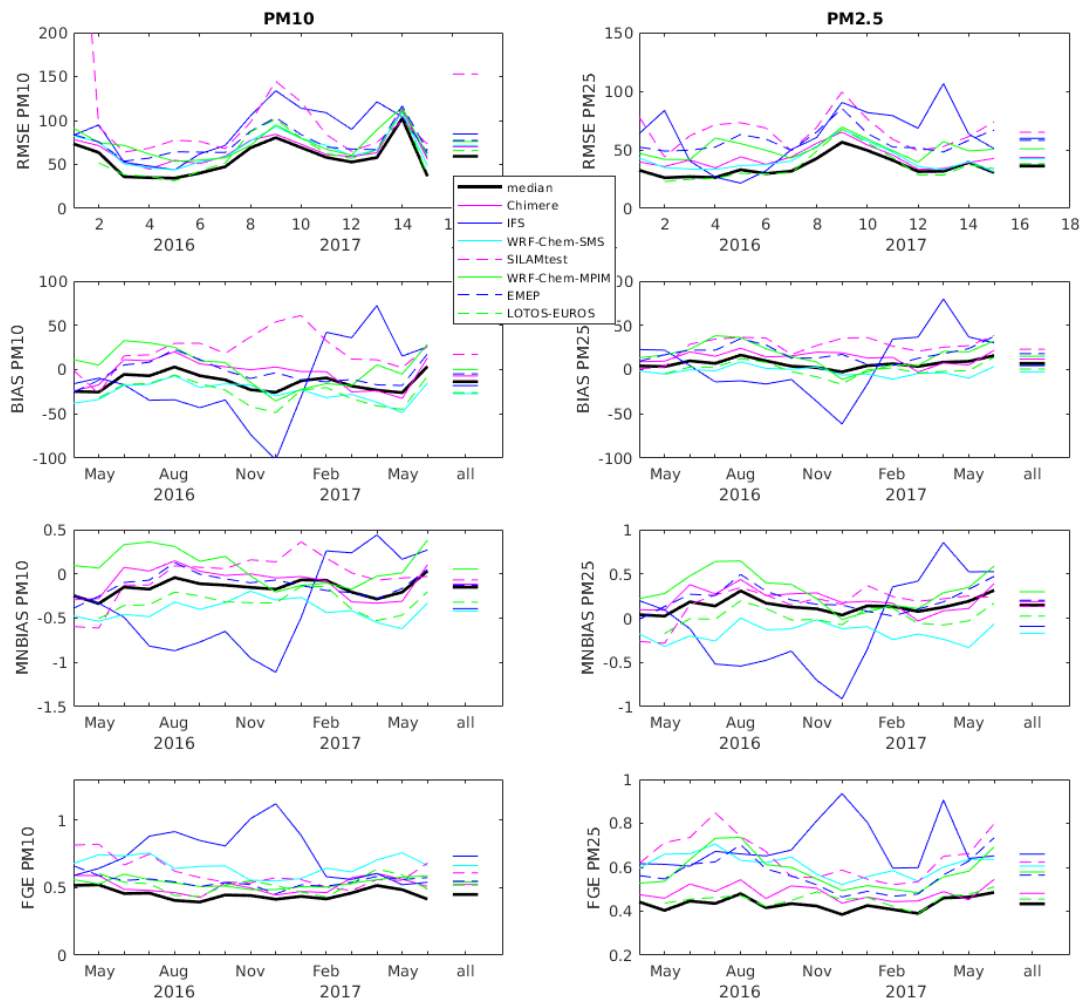
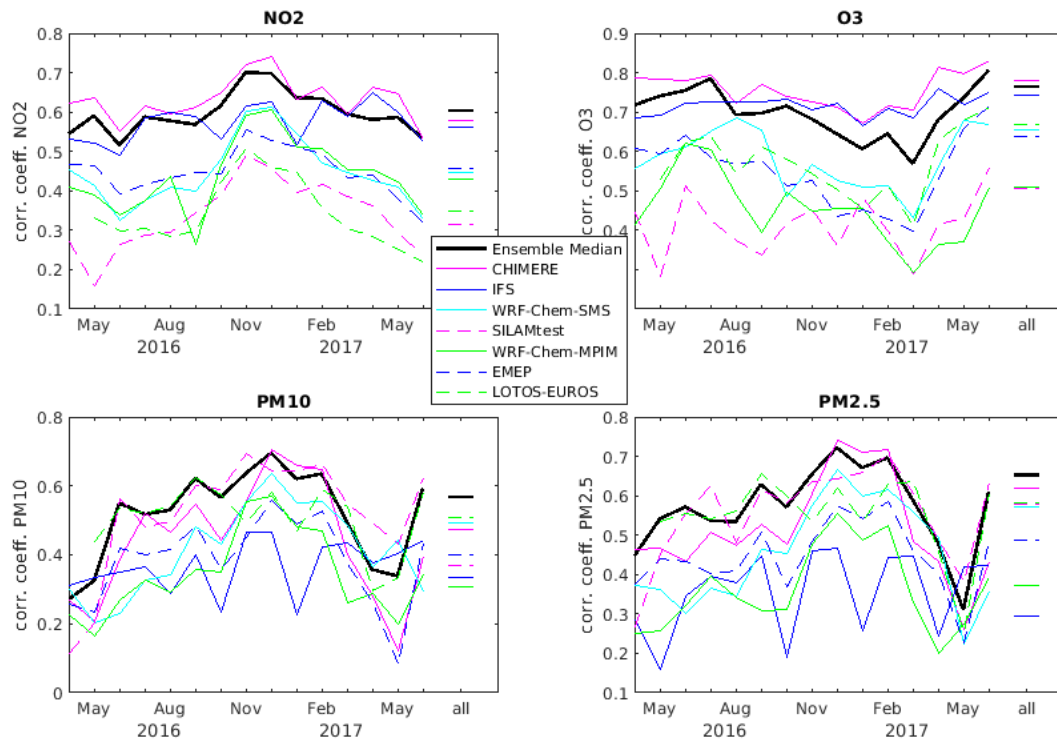
831
832
833

834
835 **Table 4**: POD and FAR for PM2.5 for Beijing under heavily polluted conditions.
836

Beijing AQI heavily polluted	POD	FAR
24-hour PM2.5 [μg m^{-3}]	0.50 (18/36)	0.28 (7/25)

837
838

839

840
Figure 2: RMSE (in µg/m3), BIAS (in µg/m3), MNBIAS and FGE of NO₂ and O₃ for each month and for the entire time period (April 2016 – June 2017, lines on the right side of each panel).

841*Figure 2: RMSE (in µg/m3), BIAS (in µg/m3), MNBIAS and FGE of NO_2 and O_3 for each month*
842*and for the entire time period (April 2016 – June 2017, lines on the right side of each panel).*
843
844
845
846

848 *Figure 3: RMSE (in µg/m3), BIAS (in µg/m3), MNBIAS and FGE of PM10 and PM2.5 for each*
849 *month and for the entire time period (April 2016 – June 2017, lines on the right side of each panel).*
850
851
852
853
854
855
856
857

859
860
861 *Figure 4: Correlation coefficients based on hourly concentrations of NO₂, O₃, PM10 and PM2.5 for*
862 *each month and for the entire time period between April 2016 and June 2017 (lines on the right*
863 *side of each panel).*
864
865
866
867
868
869
870
871
872
873
874
875
876
877
878
879
880
881
882
883
884
885

886
887
888
889
890
891
892
893
894
895
896
897
898
899
900
901
902
903
904
905
906
907
908
909
910
911
912
913
914
915
916
917
918
919
920
921
922

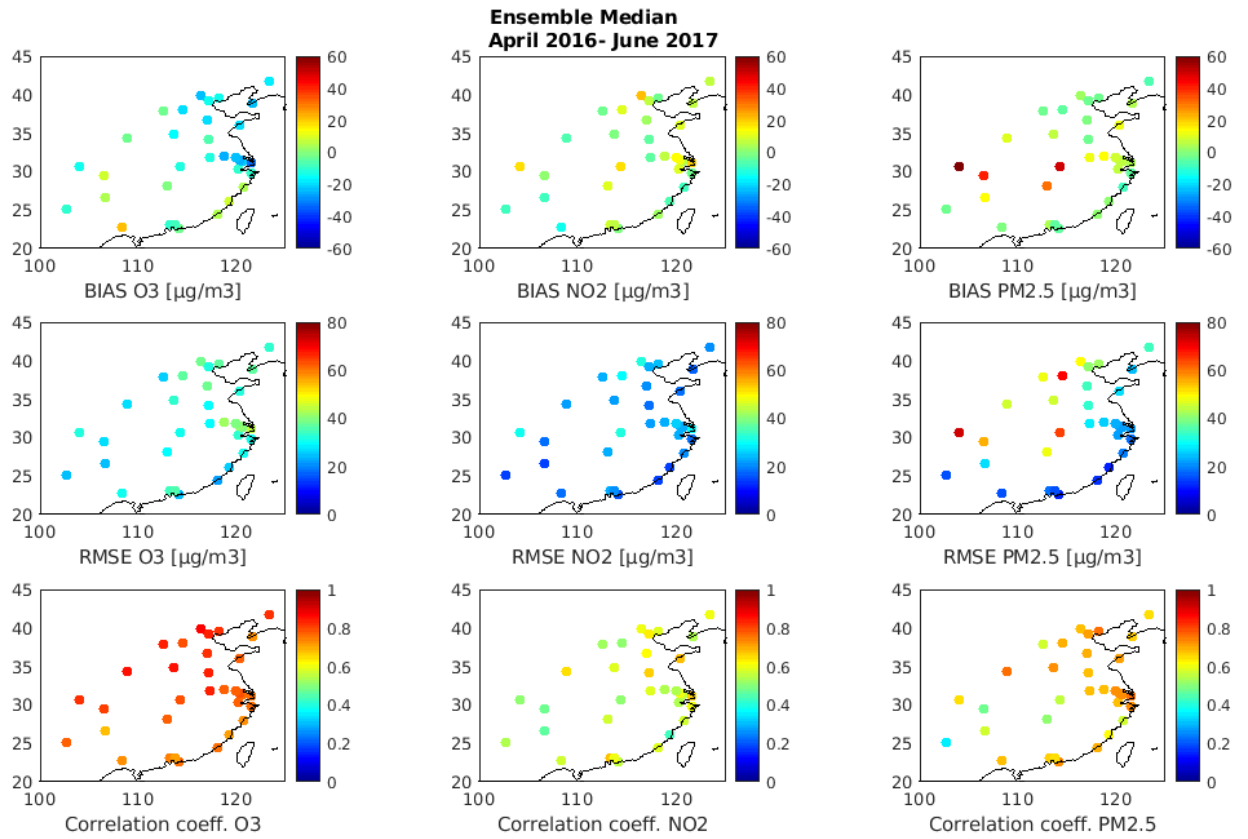
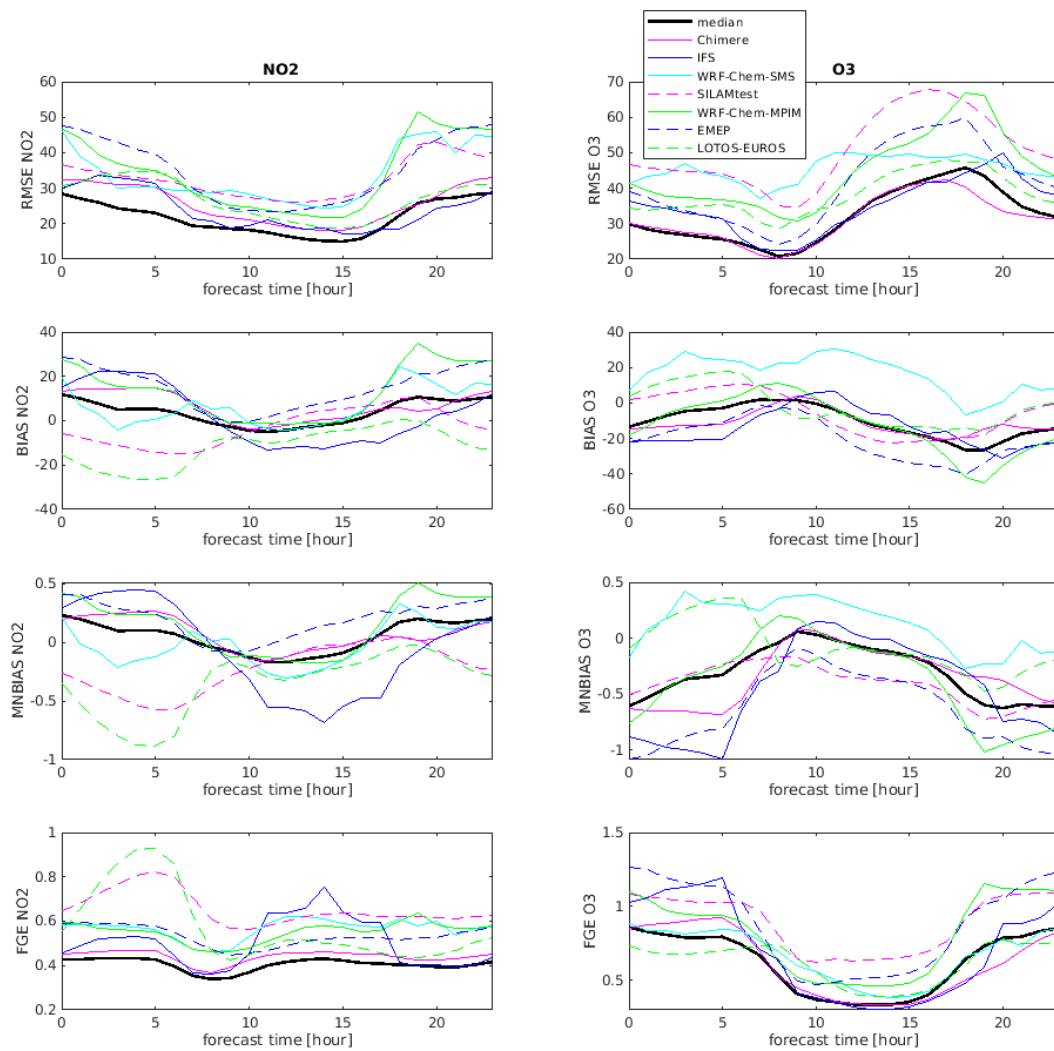
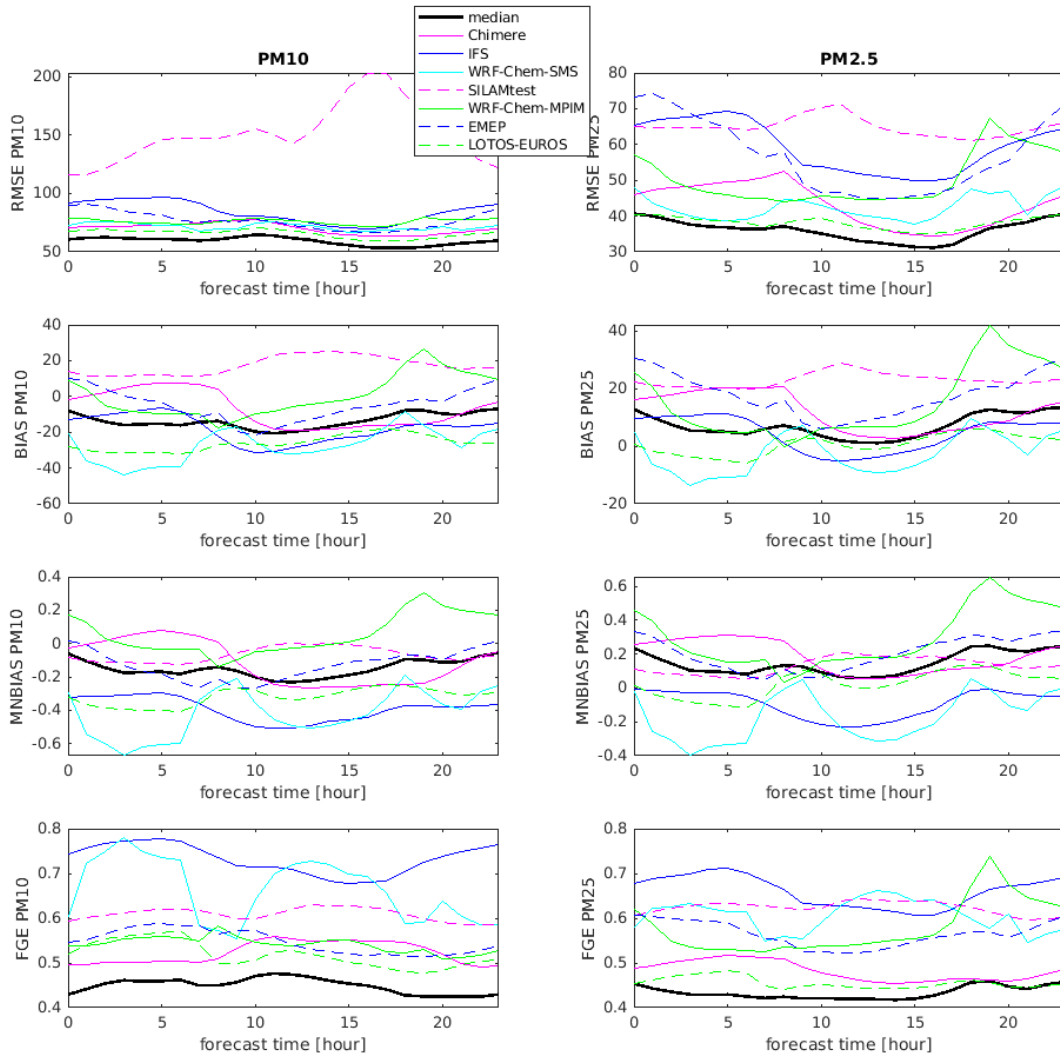
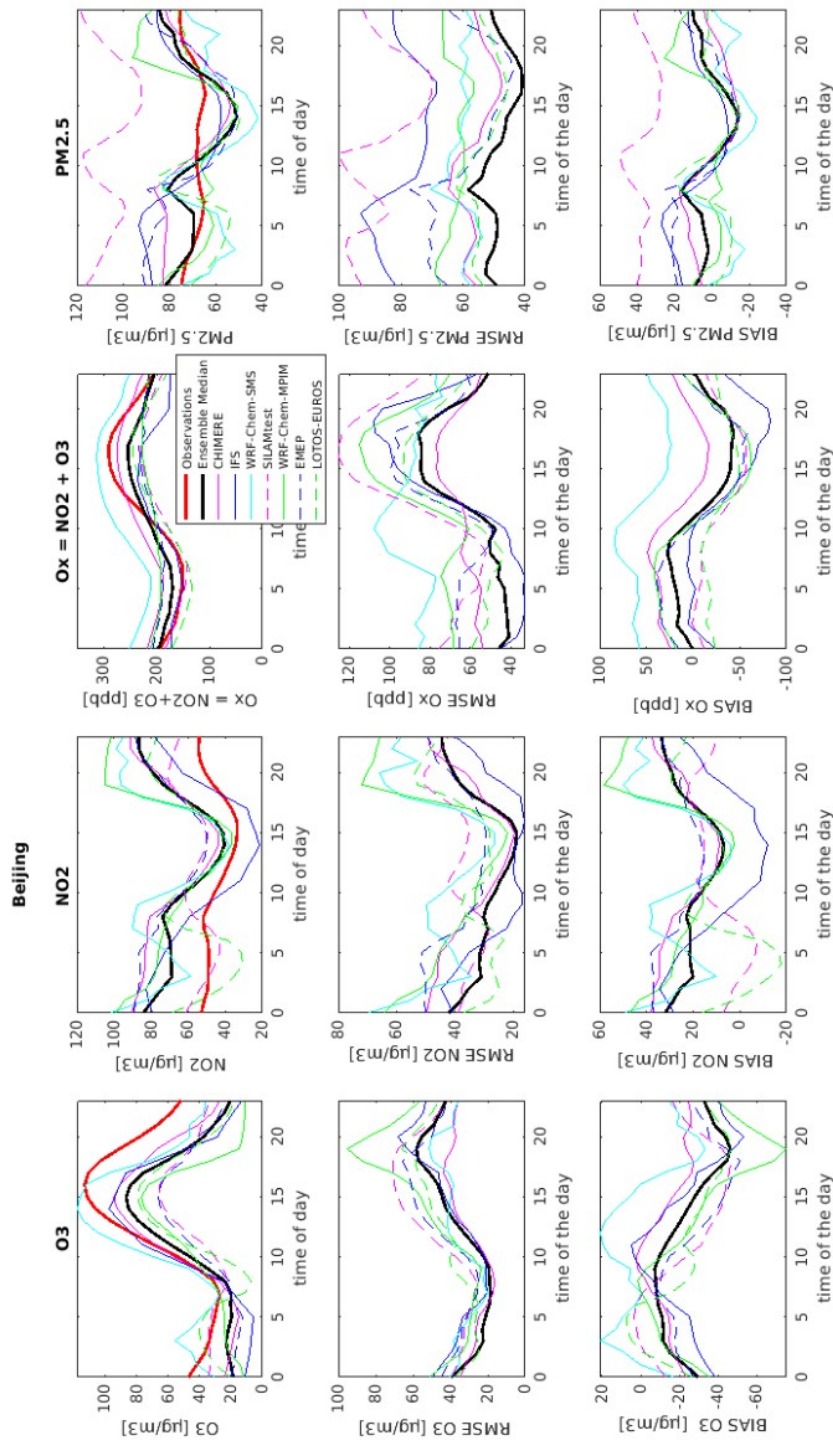
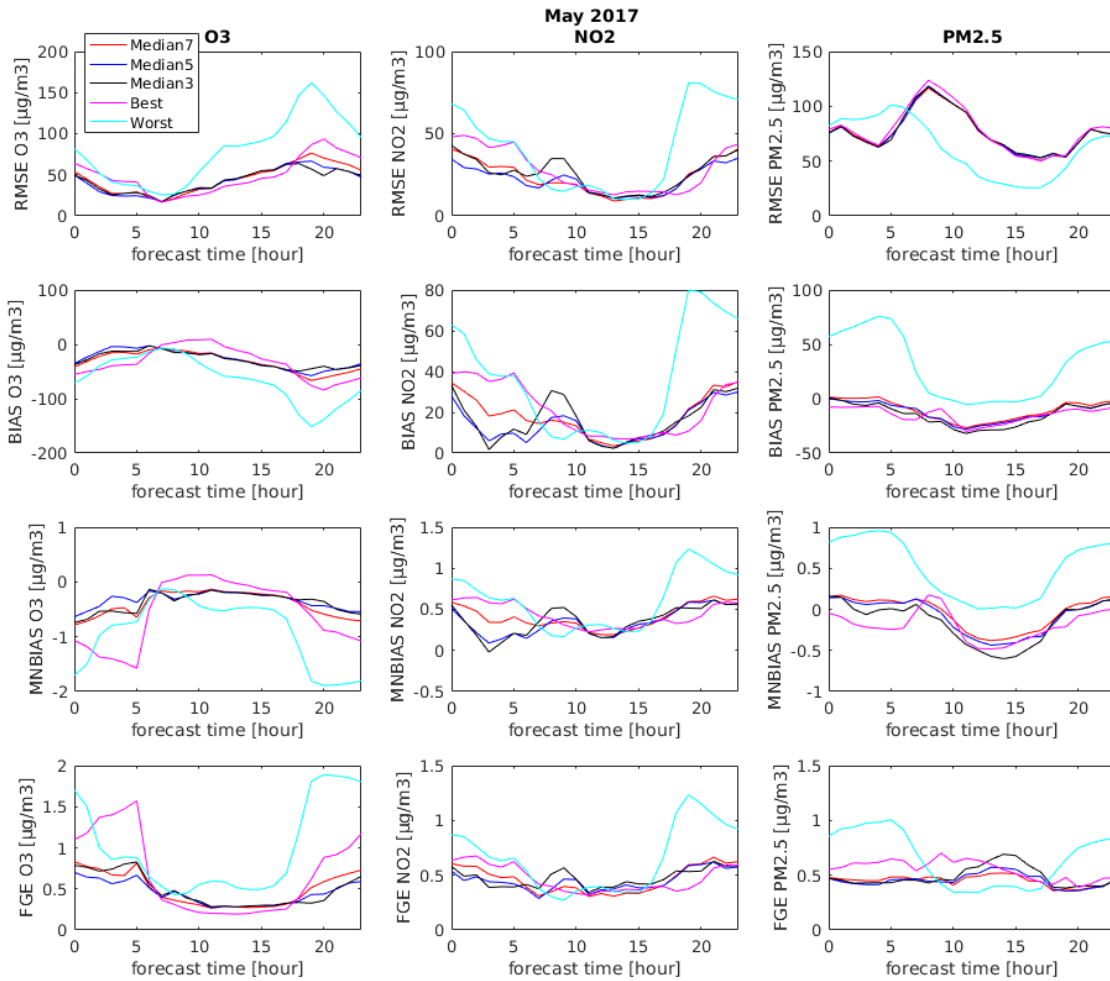



Figure 5: Map of the BIAS, RMSE and temporal correlation coefficient of O₃, NO₂ and PM2.5 for the whole time period (April 2016 until June 2017) for each city.

925
926 *Figure 6: RMSE, BIAS, MNBIAS and FGE of NO$_2$ and O$_3$ over the forecasting time (time of the*
927 *day).*
928
929
930
931
932
933
934
935
936
937
938
939
940
941
942
943
944

945
946

947
948*Figure 7: RMSE, BIAS, MNBIAS and FGE of PM10 and PM2.5 over the forecasting time (time of*
949*the day).*
950
951
952
953
954
955
956
957
958
959
960
961
962
963
964

965
966
967
968
969
970
971
972
973
974
975
976
977
978
979
980
981
982
983
984
985
986
987
988
989
990
991
992
993
994
995
996
997
998
999
1000
1001
1002
1003
1004
1005
1006
1007
1008
1009
1010
1011

1012 *Figure 8: Diurnal variations of the concentrations and of the RMSE and BIAS of O₃, NO₂, Oₓ and*
1013 *PM2.5 for Beijing for the whole time period (April 2016 – June 2017).*
1014

1052 *Figure 9: RMSE, BIAS, MNBIAS and FGE of O₃, NO₂ and PM2.5 over the forecasting time*
1053 *(time of the day) for the Median7, Median5, Median3 and the best and the worst model.*

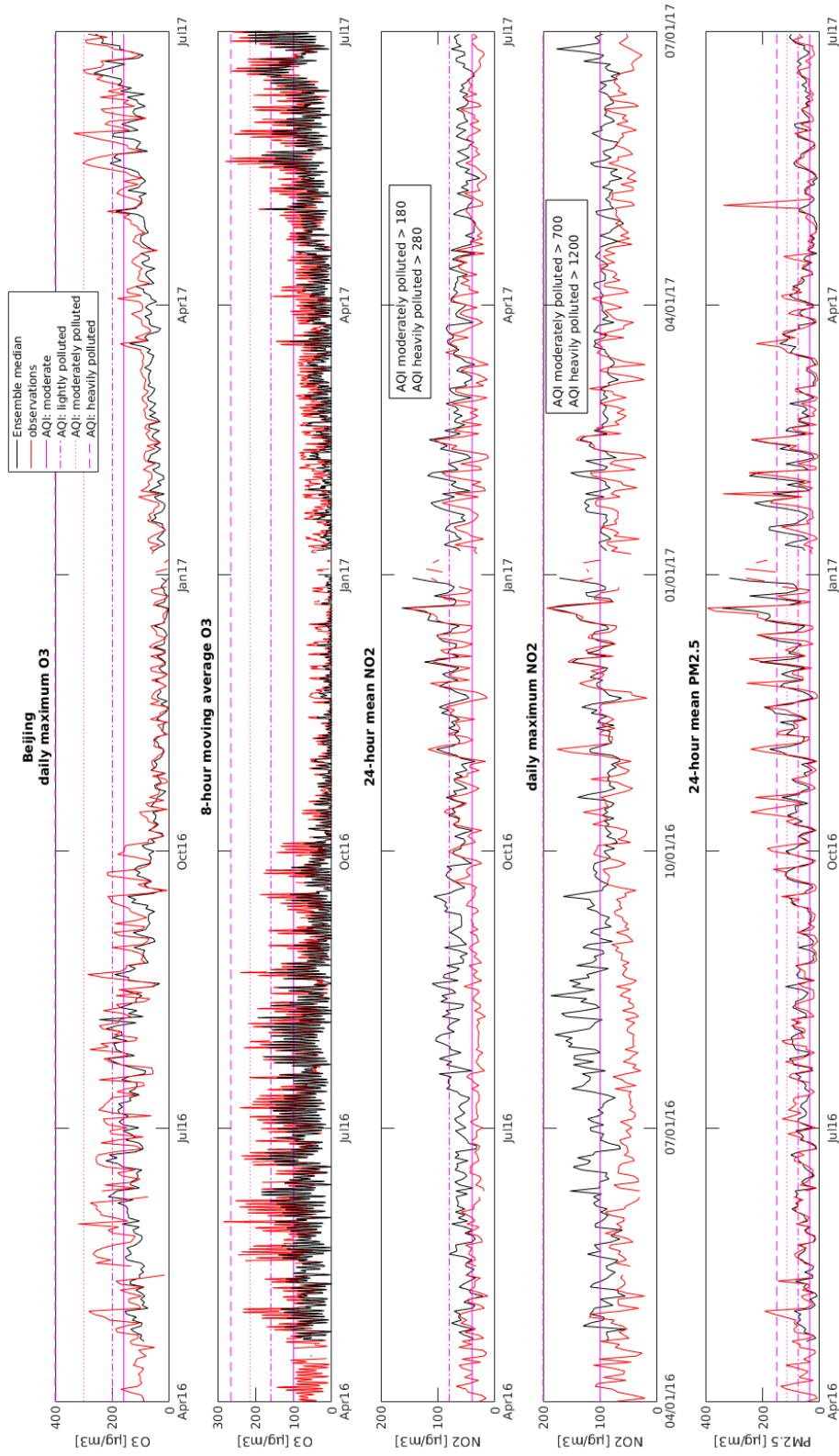
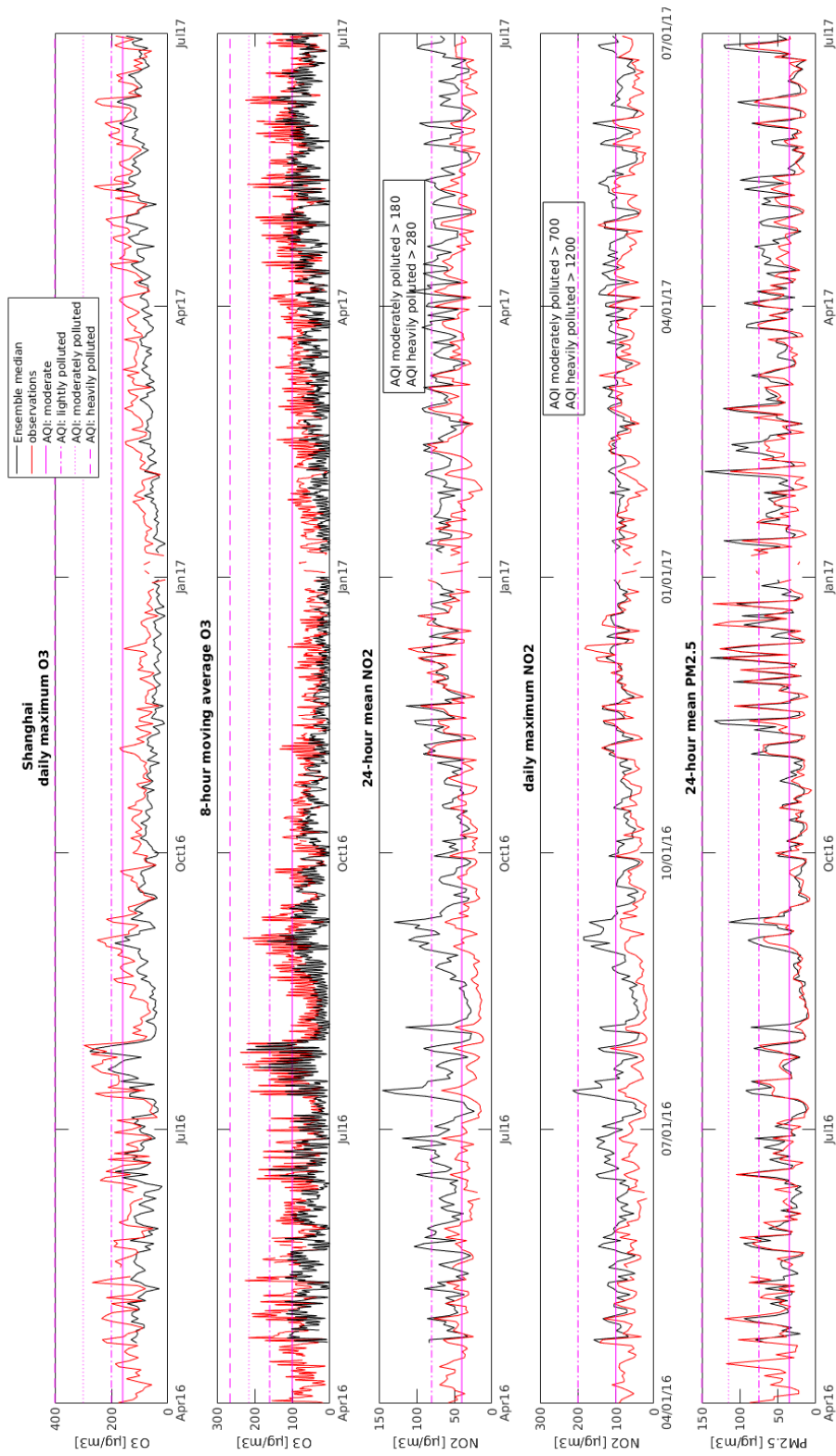



Figure 10: Timeseries of daily maximum O_3, 8-hour moving average O_3, 24-hour mean NO_2, daily maximum NO_2 and 24-hour mean PM2.5 for Beijing from April 2016 until June 2017.

1111
1112
1113
1114
1115
1116
1117
1118
1119
1120
1121
1122
1123
1124
1125
1126
1127
1128
1129
1130
1131
1132
1133
1134
1135
1136
1137
1138
1139
1140
1141
1142
1143
1144
1145
1146
1147
1148
1149
1150
1151
1152
1153
1154
1155

1156*Figure 11: Timeseries of daily maximum O_3, 8-hour moving average O_3, 24-hour mean NO_2, daily*
1157*maximum NO_2 and 24-hour mean PM2.5 for Shanghai from April 2016 until June 2017.*
1158

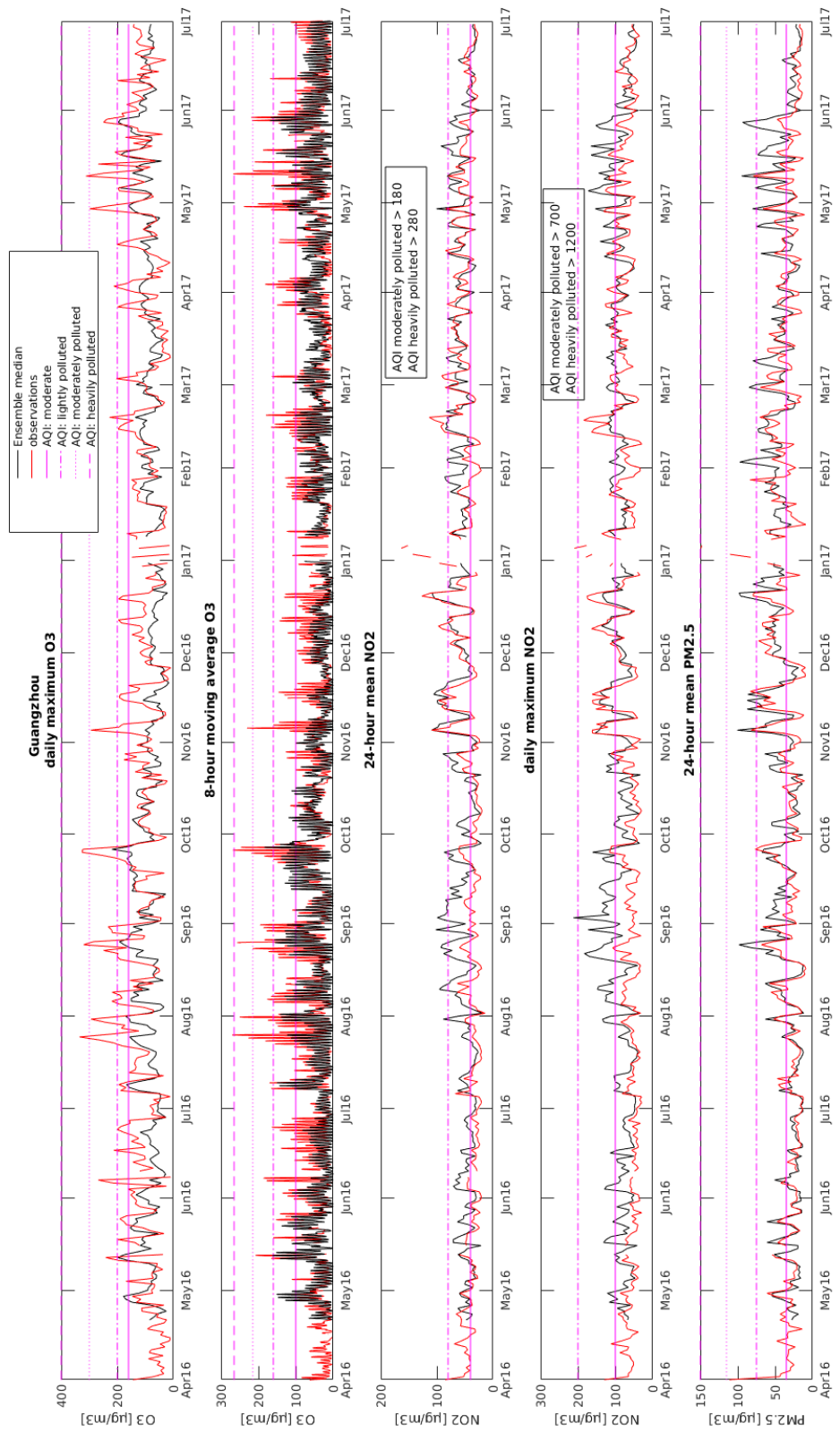
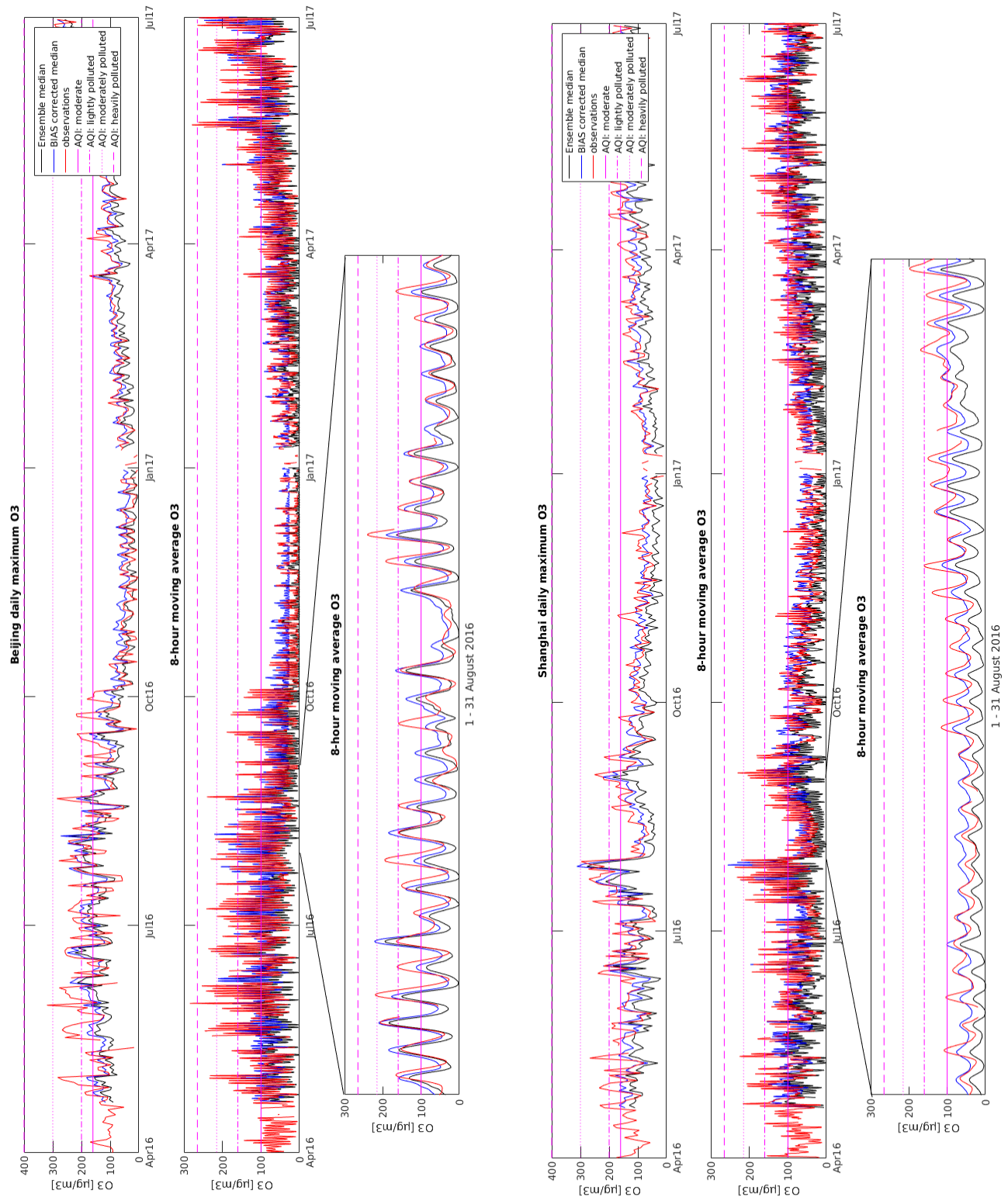



Figure 12: Calculated (ensemble median) and observed timeseries of daily maximum O₃, 8-hour moving average O₃, 24-hour mean NO₂, daily maximum NO₂ and 24-hour mean PM2.5 for Guangzhou from April 2016 until June 2017.

1207 *Figure 13 a and b: Timeseries of calculated (ensemble median) and observed daily maximum and*
1208 *8-hour moving average O₃ for Beijing and Shanghai together with the bias corrected calculated*
1209 *timeseries.*
1210
1211

1212
1213
1214
1215
1216
1217
1218
1219
1220
1221
1222
1223
1224
1225
1226
1227
1228
1229
1230
1231
1232
1233
1234
1235
1236
1237
1238
1239
1240
1241
1242
1243
1244
1245
1246
1247
1248
1249
1250
1251
1252
1253

1254 *Figure 13 c: Timeseries of calculated (ensemble median) and observed daily maximum and 8-hour*
1255 *moving average O₃ for Guangzhou together with the bias corrected calculated timeseries.*

1256
1257
1258
1259
1260
1261

1262
1263**References**
1264
1265
1266Akimoto, H., Global air quality and pollution, *Science, 302*(5651):1716-9, 2003.
1267
1268Ashmore, M. R. (2005), Assessing the future global impacts of ozone on vegetation. *Plant, Cell &*
1269*Environment, 28*: 949–964. doi:10.1111/j.1365-3040.2005.01341.
1270
1271Bessagnet, B., Pirovano, G., Mircea, M., Cuvelier, C., Aulinger, A.,Calori, G., Ciarelli, G.,
1272Manders, A., Stern, R., Tsyro, S., García Vivanco, M., Thunis, P., Pay, M.-T., Colette, A.,
1273Couvidat, F., Meleux, F., Rouïl, L., Ung, A., Aksoyoglu, S., Baldasano, J. M., Bieser, J., Briganti,
1274G., Cappelletti, A., D'Isidoro, M., Finardi, S., Kranenburg, R., Silibello, C., Carnevale, C., Aas, W.,
1275Dupont, J.-C., Fagerli, H., Gonzalez, L., Menut, L., Prévôt, A. S. H., Roberts, P., and White, L.,
1276Presentation of the EU-RODELTA III intercomparison exercise – evaluation of the chemistry
1277transport models'performance on criteria pollutants and joint analysis with meteorology, Atmos.
1278Chem. Phys., 16, 12667-12701, https://doi.org/10.5194/acp-16-12667-2016, 2016.
1279
1280Boynard, A., Clerbaux, C., Clarisse, L., Safieddine, S., Pommier, M., Van Damme, M., Bauduin, S.,
1281Oudot, C., Hadji-Lazaro, J., Hurtmans, D., Coheur, P.-F, First simultaneous space measurements of
1282atmospheric pollutants in the boundary layer from IASI: A case study in the North China Plain,
1283*Geophys. Res. Lett., 41*, 645–651, doi:10.1002/2013GL058333, 2014.
1284
1285Brasseur, G.P., Xie, Y., Petersen, A.K., Bouarar, I., Flemming, J., Gauss, M., Jiang, F., Kouznetsov,
1286R., Kranenburg, R., Mijling, B., Peuch, V.-H., Pommier, M., Segers, A., Sofiev, M., Timmermans,
1287R., Van der A, R., Walters, S., Xu, J., Zhou, G., Ensemble Forecasts of Air Quality in Eastern
1288China, Part 1. Model Description and Implementation, submitted to *Geosci. Model Dev*, 2018.
1289
1290Brasseur, G. P. and D. J. Jacob, *Modeling of Atmospheric Chemistry*, Cambridge University Press,
12912017.
1292
1293Brasseur, G. P., J. Orlando, and G. Tyndall, *Atmospheric Chemistry and Global Change*, Oxford
1294University Press, New York, 1999.
1295
1296Fei, Liu, Beirle, S., Zhang, Q., Van der A, R., Zheng, B., Tong, D., He, K., NOx emission trends
1297over Chinese cities estimated from OMI observations during 2005 to 2015, *Atmos. Chem. Phys., 17*,
12989261–9275, 2017, https://doi.org/10.5194/acp-17-9261-2017, 2017.
1299
1300Fowler, D, Amann, M, Anderson, F, Ashmore, M, Cox, P, Depledge, M, Derwent, D, Grennfelt, P,
1301Hewitt, N, Hov, O, Jenkin, M, Kelly, F, Liss, PS, Pilling, M, Pyle, J, Slingo, J and Stevenson, D
1302(2008) *Ground-level ozone in the 21st century: Future trends, impacts and policy implications.*
1303Royal Society Science Policy Report, 15 (08).
1304
1305Galmarini, S., Kioutsioukis, I., and Solazzo, E.: E pluribus unum*: ensemble air quality predictions,
1306*Atmos. Chem. Phys., 13*, 7153–7182, doi:10.5194/acp-13-7153-2013, 2013.
1307
1308Guo, S., Hu, M., Zamora, M. L., Peng, J., Shang, D., Zheng, J., Du, Z., Wu, Z., Shao, M., Zeng, L.,
1309Molina, M. J. and R. Zhang, Elucidating severe urban haze formation in China, *Proc. Natl. Acad.*
1310*Sci.USA, 111*(49): 17373–17378., 2014.
1311

Hamra, G. B., Laden, F., Cohen, A. J., Raaschou-Nielsen, O., Brauer, M., and D. Loomis, Lung Cancer and Exposure to Nitrogen Dioxide and Traffic: A Systematic Review and Meta-Analysis, *Environ Health Perspect, 123 | 11*, DOI:10.1289/ehp.1408882, 2015.

Huang, K., Zhuang, G., Wang, Q., Fu, J. S., Lin, Y., Liu, T., Han, L., and Deng, C.: Extreme haze pollution in Beijing during January 2013: chemical characteristics, formation mechanism and role of fog processing, *Atmos. Chem. Phys. Discuss., 14*, 7517-7556, doi:10.5194/acpd-14-7517-2014, 2014.

Huang, R.-J., Y. Zhang, et al. (2014). High secondary aerosol contribution to particulate pollution during haze events in China. *Nature 514*(7521): 218-222.

Kampa, M., and Castanas, E., Human health effects of air pollution, *Environmental Pollution, 151*:362–367, DOI: 10.1016/j.envpol.2007.06.012, 2008.

Leisner, C. P. and Ainsworth, E. A.: Quantifying the effects of ozone on plant reproductive growth and development, *Global Change Biol., 18*, 606–616, 2012.

Marécal, V., Peuch, V.-H., Andersson, C., Andersson, S., Arteta, J., Beekmann, M.,Benedictow, A., Bergstroem, R., Bessagnet, B., Cansado, A., Cheroux, F., Colette, A.,Coman, A., Curier, R.L., Denier van der Gon, H.A.C., Drouin, A., Elbern, H.,Emili, E., Engelen, R.J., Eskes, H.J., Foret, G., Friese, E., Gauss, M., Giannaros, C.,Guth, J., Joly, M., Jaumouille, E., Josse, B., Kadygrov, N., Kaiser, J.W., Krajsek, K.,Kuenen, J., Kumar, U., Liora, N., Lopez, E., Malherbe, L., Martinez, I., Melas, D., Meleux, F., Menut, L., Moinat, P., Morales, T., Parmentier, J., Piacentini, A., Plu, M.,Poupkou, A., Queguiner, S., Robertson, L., Rouïl, L., Schaap, M., Segers, A.,Sofiev, M., Thomas, M., Timmermans, R., Valdebenito,A., van Velthoven, P., van Versendaal, R., Vira, J., Ung, A., 2015. A regional air quality forecasting system over Europe: the MACC-II daily ensemble production. *Geosci. Model Dev. 8*, 2777 e 2813. http://dx.doi.org/10.5194/gmd-8-2777-2015.

Mizzi, A.P., A.F. Arellano, D.P. Edwards, J.L. Anderson, and G.G. Pfister: Assimilating compact phase space retrievals of atmospheric composition with WRF-Chem/DART: a regional chemical transport/ensemble Kalman filter data assimilation system, *Geosci. Model Dev., 9*, 965-978, 2016.

Sinha, B., Singh Sangwan, K., Maurya, Y., Kumar, V., Sarkar, C., Chandra, B. P., and Sinha, V.: Assessment of crop yield losses in Punjab and Haryana using 2 years of continuous in situ ozone measurements, *Atmos. Chem. Phys., 15*, 9555-9576, doi:10.5194/acp-15-9555-2015, 2015.

Sitch, S., Cox, P.M., Collins, W.J., Huntingford, C., Indirect radiative forcing of climate change through ozone effects on the land-carbon sink. Nature. 448: 791-794, 2007.

Sun, L., Xue, L., Wang, T., Gao, J., Ding, A., Cooper, O. R., Lin, M., Xu, P., Wang, Z., Wang, X., Wen, L., Zhu, Y., Chen, T., Yang, L., Wang, Y., Chen, J., and Wang, W.: Significant increase of summertime ozone at Mount Tai in Central Eastern China, *Atmos. Chem. Phys., 16*, 10637-10650, https://doi.org/10.5194/acp-16-10637-2016, 2016.

Wang, Y., Yao, L., Wang, L., Ji, D., Tang, G., Zhang, J., Sun, Y., Hu, B., Xin, J., Mechanism for the formation of the January 2013 heavy haze pollution episode over central and eastern China, Sci. China Earth Sci., 57: 14. doi:10.1007/s11430-013-4773-4, 2014.

WHO, http://www.who.int/airpollution/data/cities/en/, 2018.

1362

1363Wu, Q., Wang, Z., Chen, H., Zhou, W., Wenig, M., An evaluation of air quality modeling over the
1364Pearl River Delta during November 2006, *Meteorol. Atmos. Phys., 116*: 113. doi:10.1007/s00703-
1365011-0179-z, 2012.

1366

1367Xu, J., Zhang, Y., Fu, J.S., Zheng, S., Wang, W., Process analysis of typical summertime ozone
1368episodes over the Beijing area, *Science of The Total Environment, 399* (1–3), 147-157,
1369http://dx.doi.org/10.1016/j.scitotenv.2008.02.013, 2008.

1370

1371Zhao, X. J., Zhao, P. S., Xu, J., Meng,, W., Pu, W. W., Dong, F., He, D., and Shi, Q. F.: Analysis of
1372a winter regional haze event and its formation mechanism in the North China Plain, *Atmos. Chem.*
1373*Phys.,13*, 5685-5696, doi:10.5194/acp-13-5685-2013, 2013.

1374

1375Zhang, B., Wang, Y., and Hao, J., Simulating aerosol–radiation–cloud feedbacks on meteorology
1376and air quality over eastern China under severe haze conditions in winter, *Atmos. Chem. Phys., 15*,
13772387-2404, doi:10.5194/acp-15-2387-2015, 2015.