An alternative way to evaluate chemistry-transport models variability

Laurent MENUT[1], Sylvain MAILLER[1], Bertrand BESSAGNET[2], Guillaume SIOUR[3], Augustin COLETTE[2], Florian COUVIDAT[2], and Frédérik MELEUX[2]

[1]Laboratoire de Météorologie Dynamique, Ecole Polytechnique, IPSL Research University, Ecole Normale Supérieure, Université Paris-Saclay, Sorbonne Universités, UPMC Univ Paris 06, CNRS, Route de Saclay, 91128 Palaiseau, France
[2]INERIS, National Institute for Industrial Environment and Risks, Parc Technologique ALATA, F-60550 Verneuil-en-10 Halatte, France
[3]Laboratoire Inter-Universitaire des Systèmes Atmosphériques, UMR CNRS 7583, Université Paris Est Créteil et Université Paris Diderot, Institut Pierre Simon Laplace, Créteil, France

Correspondence to: Laurent Menut, menut@lmd.polytechnique.fr

Abstract. A simple and complementary model evaluation technique for regional chemistry-transport is discussed. The methodology is based on the concept that we can learn on models performances by comparing the simulation results with observational data available for other time periods than the period originally targeted. First, the statistical indicators selected in this study (spatial and temporal correlations) are computed for a given time period, using colocated observation and simulation data in time and space. Second, the same indicators are used to calculate scores for several other years yet conserving the spatial locations and Julian days of the year. The difference between the results provides useful insights on the model capability to reproduce the observed day to day and spatial variability. In order to synthesise the large amount of results, a new indicator is proposed, designed to compare the several error statistics between all the years of validation and to quantify if the period and area being studied was well captured by the model for the good reasons.

1 Introduction

Chemistry transport models (CTM) aim at simulating the atmospheric composition where humans and the environment can be affected by air pollution. Air pollution results from the presence of chemical compounds emitted into the atmosphere due to anthropogenic activities and natural sources (biogenic emissions from vegetation, soil erosion, sea salts, volcanic activity, and wild fires). CTMs are used to represent the dynamical and chemical processes that drive spatial and temporal features of the atmospheric composition.

To estimate the quality of CTMs, model output results are usually compared with available observations. These comparisons are performed since the models exist, they are crucial to quantify the ability of models to reproduce particular events or a general behaviour. The quantification of the model quality is performed in every research work. It dependis on the case being studies, the modelled variables, and the spatial and temporal resolution. The comparison between observations and model outputs is a complex task and has to take into account numerous factors such as, for example, the spatial representativeness of the monitoring stations (Valari and Menut, 2008; Solazzo and Galmarini, 2015). From many years, the best approach to

evaluate a model results has been discussed and, in the field of atmospheric composition, numerous methods were proposed. This is not possible to give an exhaustive list of all validation studies and we present here some examples.

Baldridge and Cox (1986) and Cox and Tikvart (1990) proposed the use of error statistics like correlation, bias, Root Mean Squared Error (RMSE) in the specific framework of air quality, *i.e.* the atmospheric composition when criteria pollutant concentrations exceed pre-defined limit values. Chang and Hanna (2004) also proposed an evaluation framework dedicated to air quality model performance and explained there is not "*a single best evaluation methodology*" and how important it is to use as much as possible evaluation criteria to really well understand model results. Later, and in order to ensure the use of systematic procedures in the evaluation process, dedicated tools were developed for the model evaluation. For example, Appel et al. (2011) and Galmarini et al. (2012) proposed complex statistical modules to extract all possible information related to the capability of a model to reproduce an observed event. In parallel, some studies were dedicated to revisit the way to evaluate models such as Thunis et al. (2012), dedicated to air quality in a policy framework. In this study, they proposed the "Target diagram" to have on the same plot the bias and the RMSE. Complementary to the definition of performance indicators to be used, Simon et al. (2012) use these indicators to compile photochemical models performances over a large set of data over several years of simulation. This kind of evaluation may also be done in dedicated projects such as the recent AQMEII (Air Quality Model Evaluation International Initiative), comparing chemistry-transport models running both in Europe and Northern America, (Vautard et al., 2012; Campbell et al., 2015) or the EURODELTA project, (Bessagnet et al., 2016) and in the EMEP (European Monitoring and Evaluation Programme) context in the frame of the United Nation Convention on Long-range Transboundary Air Pollution, (Prank et al., 2016). Using comparisons between observations and models outputs, some studies proposed methodologies to decompose the statistical scores in order to estimate the main source of errors, (Solazzo and Galmarini, 2016). Finally, other studies also use observations to adjust the result by implementing methods to unbias simulation without changing the model, as in Porter et al. (2015) for ozone over the United States. The common point of all these studies is that they are always using, as best as possible, the observations corresponding in time and location to the model grid cell.

In the present study, a simple method is proposed to add information about the model performances with a focus on its spatial and temporal variability. To reach this objective, we propose to use observations corresponding to the modelled period and geographical domain but also, to use observations for the same domain but other periods. By this way, we want to extract the information about the model variability and to answer the question: *Are the performances of the model satisfactory because the model is accurate or just because the model is able to reproduce a situation which is recurrent from year to year?* The issue to be solved and the tools developed are presented in section 2. The new methodology with the presentation of the indicator developed for this study are presented in section 3. The results and discussions to point out the drivers of model errors are presented in section 4 and section 5 for the new indicator.

2 Methodology

In the present study, a simple method is developed to improve the evaluation of models variability and to identify the processes responsible for discrepancies of models outputs *versus* observations. The methodology is general and could be applied to

all types of model. In this study, the methodology is presented for the specific case of the regional atmospheric composition modelling: a topic mixing meteorology and chemistry, with a high spatial and temporal variability, thus having a good potential to test the relevance of our methodology.

2.1 Regional chemistry-transport modelling

In chemistry-transport modelling, several processes are involved, some of them directly influencing the others. When studying both meteorological and chemical variables, the dependencies between all variables are helpful to better interpret the model results.

The boundary conditions prescribe the concentrations of chemical species which may enter the simulation domain. Usually for large domains, they are issued from global models as monthly climatologies. They correspond to averaged values suitable to characterize the background concentrations of long-lived species such as ozone, carbon monoxide, mineral dust. Anthropogenic emissions are prescribed from databases and the influence of meteorology is limited in the model. Vegetation, fires and mineral dust emissions depend both on landuse data and meteorology. These emissions are not measurable, it is almost impossible to directly assess their quality.

The meteorological variables influence transport and mixing processes, with a direct effect on gas and aerosol plumes locations and their vertical distribution. Cloudiness and temperature impact the photolysis efficiency, the boundary layer height impact the surface mixing of pollutants, rainfall impact the wet deposition. Moreover, meteorology has also an impact on emissions: wind variability is the prevalent driver for dust emissions, and it has also a major impact on wildfires emissions. Both temperature and solar irradiance influence the magnitude of biogenic emissions from vegetation. The spatial variability of landuse data has also a strong impact on all these natural emissions.

The chemistry-transport model is a numerical integration tool of all forcings and processes. The chemical mechanism handles the life cycle of chemical species (production and loss) when the deposition processes are the only net sink of species. In the model, the spatial (horizontal and vertical) and temporal resolutions are also prescribed, directly impacting the simulation representativeness and thus the quality of the modelled air pollutant concentrations when they are compared to available observations.

2.2 The studied case

The study focuses on the summer 2013 period (1st May to 31 August) over the Euro-Mediterranean region. This period is called "reference period" in this paper. This case has already been modelled (using the same models, WRF and CHIMERE) and the results were discussed in Menut et al. (2015). The same simulation is used in this study, all parameters are identical.

The observational data come from different sources depending on the variables, Table 1 . In this region, where the monitoring network are dense enough, comparisons are performed with observations from surface stations that provide hourly O_3, NO_2 surface concentrations for gases and $PM_{2.5}$ and PM_{10} (particulate matter with mean mass median diameter lower than 2.5 and 10μm, respectively) for particles. Complementary to surface concentrations data, evaluated using the EBAS database, (Tørseth et al., 2012), the meteorology is also evaluated for 2m temperature, T_{2m}, 10m wind speed, U_{10m}, and precipitation

Variable	Network	Spatial coverage	Vertical coverage	Temporal frequency	Unit
O_3, NO_2	EBAS/EMEP	Europe	Surface	Hourly	ppb
$PM_{2.5}$, PM_{10}	EBAS/EMEP	Europe	Surface	Hourly	μg m^{-3}
AOD, Angström	AERONET	Global	Column	Hourly	ad.
T_{2m}	BADC	Global	Surface	Tri-hourly	^{o}C
U_{10m}	BADC	Global	Surface	Tri-hourly	m s^{-1}
Precipitation	BADC	Global	Surface	Daily	mm day^{-1}

Table 1. *List of measurements data used for the statistical comparison with the model results. All data used are issued from surface stations, representative of their own environment. Originally provided hourly or three-hourly, they are used as daily averaged in this work.*

rates (in mm day^{-1}) from the BADC (British Atmospheric Data Centre). In order to quantify the transport of aerosols in dense plumes aloft, observations from AERONET (AErosol RObotic NETwork) program are used for the optical depth, AOD, and the Angström exponent. In this study, all variables are used as daily mean (except for precipitation corresponding to daily cumulated values) in order to (i) have homogeneous scores between the variables, (ii) be able to separate the systematic and the day-to-day variabilities. The use of an hourly time frequency was ruled out to avoid a too strong weight of the diurnal cycle in the temporal variability.

3 Proposed methodology

As discussed in the introduction, many Statistical Indicators (SI) exist to quantify the model ability to simulate observed pollution events. The correlations (temporal and spatial), the Root Mean Squared Error (RMSE), its normalized expression nRMSE, and the bias (the difference between observations and modelled values) are widely used in regional air pollution modelling. The correlations are able to split the relative contributions of systematic meteorology or sources related variability and day-to-day variability. The RMSE and the bias are a direct quantification of the model error.

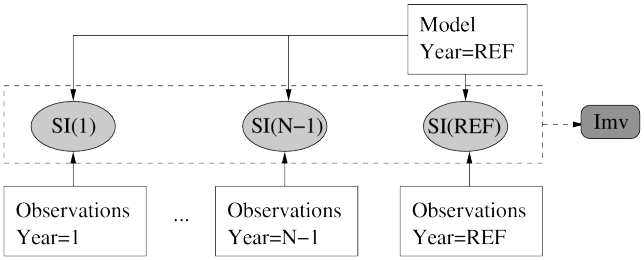

Figure 1. *Principle of the multi-year variability indicator (I_{mv}) calculation, using one modelled year and several year of observations. SI stands for "Statistical Indicator" and is related to spatial and temporal correlation.*

The main goal of this study is to separate the contributions due to systematic and sporadic events. The systematic events correspond to yearly phenomena when the sporadic events correspond to event observed during one year but not the others. In addition, complementary to the model variability quantification, the model error is also important to estimate. The key point of this study is to (i) study the model variability which is statistically represented by the correlations, and (ii) add complementary information on the model errors, here that could be represented by the RMSE (or the nRMSE).

First, as presented in Figure 1, the SI are calculated between observations data and model outputs for the simulation year (*i.e.* the reference year). Second, the SI are calculated between the observations data for other years and the model output for the reference year. Logically, the scores calculated for the reference year for observations and model outputs would give the better results. By difference with the scores calculated for other years (with the observations only), we expect to conclude if the model is able to catch the observed variability and for the good reasons. Using this approach, the goal is to give complementary information to those usually obtained when using only SI calculated for a single year, the studied year.

We apply this methodology for the simulation of the year 2013 and using observations data for years ranging from 2008 to 2013. In order to give some synthetic answers, the different SI scores are aggregated into a single indicator, called I_{mv} and presented in detail in the next section. Of course it seems apparently awkward to evaluate day by day a model with observational data from another year. For a given station at a given day of the reference year air concentrations will be affected by a different local meteorology, emissions and also long range transport of chemical species. But we can consider that to take the same date for another year is strictly the same that to choose randomly a date in the same season. This trivial method can emphasize how a model is affected by large scale patterns and long term temporal cycles.

3.1 Calculation of correlations and nRMSE

In this study, we focus on three Statistical Indicators: the spatial correlation, the temporal correlation and the normalized RMSE. For these three indicators, it is important that, for all years of validation, the same list of stations with valid measurements is used.

The correlation used in this study is the Pearsons' correlation. Each correlation provides specific information on the quality of the simulation. The temporal correlation, noted R_t, is estimated station by station and using daily averaged data in order to have homogeneous comparisons between all variables. This correlation is directly related to the variability from day to day, for each station. $O_{t,i}$ and $M_{t,i}$ represent the observed and modelled values, respectively, at time t and for the station i, for a total of T days and I stations. The mean time averaged value $\overline{X_i}$ is:

$$\overline{X_i} = \frac{1}{T}\sum_{t=1}^{T} X_{t,i} \tag{1}$$

The temporal correlation $R_{t,i}$ for each station i is calculated as:

$$R_{t,i} = \frac{\sum_{t=1}^{T}(M_{t,i} - \overline{M_i})\,(O_{t,i} - \overline{O_i})}{\sqrt{\sum_{t=1}^{T}(M_{t,i} - \overline{M_i})^2 \sum_{t=1}^{T}(O_{t,i} - \overline{O_i})^2}} \tag{2}$$

The mean temporal correlation, R_t, used in this study is thus:

$$R_t = \frac{1}{I}\sum_{i=1}^{I} R_{t,i} \tag{3}$$

with I the total number of stations. The spatial correlation, noted R_s, uses the same formula type except it is calculated from the temporal mean averaged values of observations and model for each location where observations are available. A good correlation shows that the model correctly locates the largest horizontal gradients as known sources and long range transport plumes.

The spatio-temporal mean averaged value is estimated as:

$$\overline{\overline{X}} = \frac{1}{I}\sum_{i=1}^{I} \overline{X_i} \tag{4}$$

and the spatial correlation is thus expressed as:

$$R_s = \frac{\sum_{i=1}^{I}(\overline{M_i} - \overline{\overline{M}})\,(\overline{O_i} - \overline{\overline{O}})}{\sqrt{\sum_{i=1}^{I}(\overline{M_i} - \overline{\overline{M}})^2 \sum_{i=1}^{I}(\overline{O_i} - \overline{\overline{O}})^2}} \tag{5}$$

The normalized Root Mean Square Error is expressed as:

$$nRMSE = \sqrt{\frac{1}{T}\frac{1}{I}\sum_{t=1}^{T}\sum_{i=1}^{I}\left(\frac{O_{t,i} - M_{t,i}}{O_{t,i}}\right)^2} \tag{6}$$

for all stations i and all times t.

3.2 Definition of the I_{mv} indicator

For the specific purpose of the model variability (and not the model error), we define an indicator, I_{mv}, dedicated to express in one value the results obtained with the temporal and spatial correlations. The goal of this indicator is to quantify how the correlation between measurements data (for different years) and model outputs (for the reference year) evolves from a year to another one. This indicator does not replace the usual statistical indicators but aims at providing complementary information about the variability between years.

We first define the differences, D, between all years as:

$$D = \frac{1}{N-1}\left(\sum_{i=1}^{N-1} |s_i - s_N|\right) \tag{7}$$

with s_N the score of the indicator for the reference year being modelled and s_i the score of the indicator computed using observations corresponding to other meteorological years (from 1 to $N-1$ if there is $N-1$ other available years for the observations).

We now aim to develop a simple indicator, called I_{mv}, which is a combination of the statistical indicator for the reference year and the differences between years. This I_{mv} corresponds in fact to the SI itself weighted by the differences between the SI scores of all years. We want that I_{mv} follows these rules:

- I_{mv} has the same evolution than the studied SI. If the correlation increases, I_{mv} also increases.
- I_{mv} is bounded between 0 and 1, like the correlation. This enables to compare the results for different variables (with different metrics).
- In case of high correlation value found for the studied year i.e ideally s_N tends to 1:
  - If the differences between the other years are low (D tends to 0), it means that the model is correct for the studied year, but possibly because it reproduces a recurrent phenomena. In this case, we want that I_{mv} decreases and tends to 0.
  - If the differences between the other years are high (D tends to 1): in this case, the model gives good results for the studied year but it is not because it simulates a systematic event. In this case, we want that I_{mv} remains close to the indicator value. With $s_N \approx 1$ and $I_{mv} \approx 1$, we can conclude that the model is very good for the studied year and this is not due to a recurrent process.
- In case of low correlation value, and whatever the magnitude of differences between years, the model is not correct. I_{mv} must be low, as the indicator value.

These constraints induce to define an indicator having this kind of formulation:

$$I_{mv} = s_N \left(1 - exp(-D_s)^4\right) \tag{8}$$

s

This means that I_{mv} has always, as maximum value, the value of the indicator itself. The power 4 is here defined to have a specific shape for I_{mv} respecting the rules presented below. Finally, this expression gives an indicator variability presented in Figure 2. Considering the state-of-the art of chemistry-transport modelling, the model is considered as accurate and to have an acceptable variability for $I_{mv} > 0.4$: this means that the correlation is at least 0.5 and the differences are also at least greater than 0.5.

Finally, this indicator is not calculated for nRMSE and bias. Two reasons explain this choice: (i) contrarily to correlations, RMSE and bias are not bounded between 0 and 1. This leads to indicators values possibly varying a lot between several years and thus difficult to compare between years. (ii) The goal of the indicators is to extract a message from the model variability

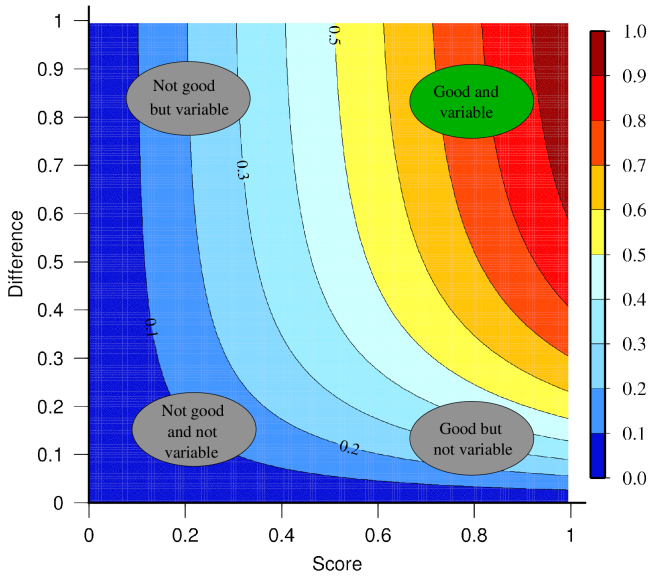

Figure 2. *Scheme of the I_{mv} values as a function of the studied year correlations values and the multi-years differences D.*

of the studied year compared to the other years. In this case, the correlations constitute a statistical indicator wchich is more appropriate to this evaluation.

4 Time series of statistical indicators

The calculations of differences are performed for the correlations and the nRMSE. These values are calculated for all variables described in Table 1 and for the years 2008 to 2013. For each year, it is reminded that only the May to August period is considered. Results are presented as time series in Figure 3 and discussed in the following sections. Note also that some values discussed in these sections are also reported in the synthetic Table 4 .

4.1 Meteorological variables

The meteorological variables are T_{2m}, u_{10m} and the precipitation rate. The values of the Statistical scores are provided, year by year, in Figure 3 . As an example, the same values are reported for T_{2m} in Table 2 .

T_{2m} is a meteorological variable, constraining processes both for meteorology and chemistry. Its diurnal cycle is strong, as well as its latitudinal variability (for large model domains), often ensuring a good spatial correlation. In general, this variable is the less uncertain of all modelled meteorological parameters. The spatial correlation is good for all years, ranging from 0.57 (2009) to 0.62 (2011). For the studied year (2013), the score is 0.60, slightly lower than for 2011. Even if the correlation for the selected year is good, it is not significantly better than for the other year, with D=0.02. This means that the model reproduces fairly well a spatial pattern that is observed every year. Indeed, the simulation domain is large and the temperature

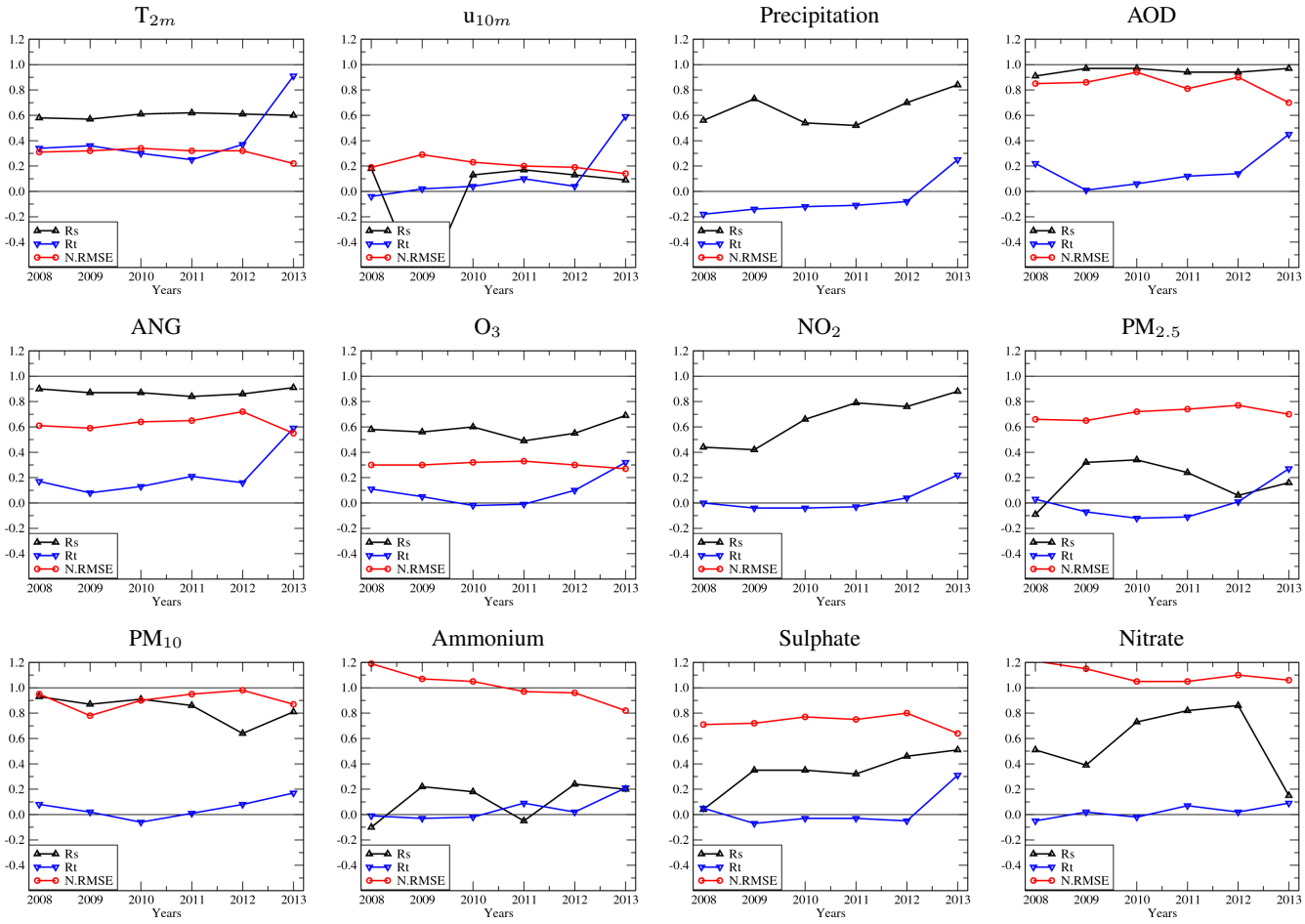

Figure 3. *Multi years scores for T_{2m}, u_{10m}, the precipitation rate, Aerosol Optical Depth (AOD), Angström exponent (ANG), surface concentrations of O_3, NO_2, $PM_{2.5}$, PM_{10}, Ammonium, Sulphate and Nitrate. The correlations and the nRMSE are calculated between the observations (2008-2013) and the model results (2013). The spatial correlation, R_s, is in black, the temporal correlation, R_t in blue, the nRMSE in red..*

has a latitudinal variability larger than between each measurements stations. The temporal correlation ranges from 0.25 to 0.91 (2013). The variability of nRMSE is lower than for the correlations, with values ranging from 0.22 (2013) to 0.34 (2010). The lowest value is found for 2013, highlighting the fact that the model error is the lowest for the reference year. The model is thus performing well in capturing the day to day variability for T_{2m} and for the good reasons.

From Figure 3 , the calculation of u_{10m} also gives satisfactory results with R_t=0.60. The spatial correlation, R_s=0.09, is poor and very variable from one year to another. As for T_{2m}, we also have an effect of the model resolution and the representativeness of the variable.

Year	R_s	R_t	nRMSE
2008	0.58	0.34	0.31
2009	0.57	0.36	0.32
2010	0.61	0.30	0.34
2011	0.62	0.25	0.32
2012	0.61	0.37	0.32
2013	0.60	0.91	0.22
D	0.02	0.59	0.10

Table 2. *Scores for T_{2m}. The correlations and nRMSE are calculated between the observations (2008-2013) and the model results (2013).*

Scores for the precipitation are correct, with a very good spatial correlation, always exceeding 0.6. As for the temperature, the latitudinal effect plays a major role in the variability. Both the spatial and temporal correlations increase significantly for the reference year. The nRMSE is not on the plot, the values being larger than 1.2. The model is biased in absolute values and overestimates the amount of daily precipitation. But the day to day variability is correct and such variability is the most important feature for atmospheric composition modelling (the lower atmosphere is scavenged when a precipitation occurs, whatever its value).

For the meteorological variables, these scores showed that the meteorological forcing is well captured, and always better for the year being considered compared to other years.

4.2 Optical properties

The optical properties are directly linked to the atmospheric composition of aerosol and may be quantified using the Aerosol Optical Depth (AOD) and the Angström exponent (ANG).

For the AOD, the spatial correlation is very good for 2013, R_s=0.97 but it is as good or better for other years. This means that we model a rather recurring phenomenon: every year the same stations are on average exposed to aerosol plumes. The temporal correlation is lower with R_t=0.45 but much better than for other years. This indicates that the model partly reproduces the observed temporal variability but the events are changing from one year to another and the model captures well these changes. In the studied region, the AOD are sensitive to desert dust outbreaks in summer. This means that large scale systems are driving the aerosol plumes; they are spatially recurrent and temporally better captured for the year being considered than for other years.

For the ANG, the spatial correlation is very good, R_s=0.91 but also persistent in time. The temporal correlation is much better for 2013 than other years. This is probably due to a size distribution that is not necessarily well simulated from one day to another (showed by AOD and explained in (Menut et al., 2016)) but the relative contributions of fine and coarse aerosol atmospheric load are fairly reproduced. This feature highlights the high sensitivity of the AOD calculation to the modelled aerosol size distribution, although the overall mass emitted and transported is realistic.

Globally, the AOD and ANG reflect the model's ability to retrieve the long range transport of long-lived aerosols which depends on several processes (emissions, transport, and deposition). These scores show that the model is able to retrieve these yearly recurrent plumes but the model size distribution of particles clearly requires improvements.

4.3 Surface concentrations

For the surface concentrations of gaseous and aerosol species, the variability is much more related to local effects. As an example, the detailed values of the statistical indicators and the differences between years are extensively presented for NO_2.

Year	R_s	R_t	nRMSE
2008	0.44	0.00	1.56
2009	0.42	-0.04	1.76
2010	0.66	-0.04	1.82
2011	0.79	-0.03	2.07
2012	0.76	0.04	2.84
2013	0.88	0.22	1.76
D	0.27	0.23	0.33

Table 3. *Scores for NO_2. The correlations and nRMSE are calculated between the observations (2008-2013) and the model results (2013).*

NO_2 is both primary and secondary in origin. Mostly emitted in urbanized areas, the diurnal cycle of this species is well constrained. Depending on meteorological conditions, its lifetime may vary significantly, from hours to days. Modelling this species with CTMs is challenging because several uncertainties are acting at the same time, including the spatial representativeness of the model cell. The scores show if the sources are properly located and if the photochemistry and transport processes have been well simulated. In general, at coarse model resolution, the model results for this species are worse than for ozone. The spatial correlation gives a score of R_s=0.88 for 2013. This corresponds to the best correlation compared to the other years. The anthropogenic emissions are strongly related to industrial activities and road traffic, and since these activity sectors are fixed in space, the good spatial correlation is more due to anthropogenic sources that vary in space such as biogenic and vegetations fires. The temporal correlation is low for 2013, R_t=0.22, but is closer to 0 for other years, therefore significantly better for the reference year compared to the others. These two correlations values show that the model certainly captures the right location of emission sources (low variability of R_s). The nRMSE is large and shows that the concentrations are overestimated by the model. But this overestimation appears for all years and can be due to the representativeness of the surface measurements compared to the size of model cells.

The spatial correlation is good for O_3, NO_2 and PM_{10}, with R_s=0.69, 0.88 and 0.81 respectively. For $PM_{2.5}$ this correlation is low with R_s=0.16. The PM_{10} shows that the largest particles are well modelled over the whole domain, and this was also the conclusion for the AOD and ANG. The low score for $PM_{2.5}$ indicates that for the aerosol distribution, the fine mode is not as well modelled as the coarse mode. This is confirmed by the scores of the aerosol inorganic species, Ammonium, Sulphate

and Nitrate that contributes to a large fraction of the fine fraction of particles. Except for Sulphate (with R_s=0.51), the spatial correlations are 0.15 for Nitrate and 0.20 for Ammonium. Thus, the fine part of the aerosol is not well modelled mainly due to a deficiency in the modelling of nitrates.

The temporal correlations have a completely different behaviour that the spatial correlations. The values are generally low, from R_t=0.09 for Nitrate to R_t=0.32 for O_3. Surprisingly, the PM_{10} concentrations display a good spatial correlation but a poor temporal correlation. This is due to the long lifetime in the atmosphere of non-reactive species such as mineral dust: plumes are correctly modelled over large areas but the day to day variability needs improvements. Another point is the good spatial correlation for NO_2 but its low temporal correlation with R_t=0.22. In this case, this means we have a correctly spatialized anthropogenic emissions inventory (mainly for NO_2 sources) but difficulties to model the day to day chemistry.

For the surface concentrations, we can conclude that O_3, NO_2 and PM_{10} concentrations are spatially well modelled and this is not due to a recurrent behaviour. For particles, the problem is more related to the fine mode, where $PM_{2.5}$ concentrations are not well located. This modelling problem is highlighted by the low correlations and I_{mv} values for the inorganic species. For the temporal correlations, the scores are always lower than for the spatial correlation but also always higher for the reference year than for the other years.

5 Estimation of the I_{mv} indicator for all variables

To summarize the results obtained for each statistical indicator and the values of differences between all years, we apply the I_{mv} formulation. This enables to have one values for each SI (R_s and R_t) and each variable. Results are presented in Table 4 and are also displayed on single plots in Figure 4 .

In Table 4 , the I_{mv} larger than 0.4 are highlighted. This threshold is clearly subjective but mentioned here to better highlight the variables being well modelled and with a correct variability from a year to another. As discussed in detail, the best scores are obtained for the meteorological variables, and also better for the temporal variability than for the spatial variability.

In Figure 4 , The x-axis represents the correlation (spatial or temporal), the y-axis represents the differences between all years, D. For each studied variables, their values are reported on the figure where the colours represent the value of I_{mv}. The interpretation of these results follows the quality criteria presented in the academic schematic of Figure 2 . This presentation shows an important spread for the spatial correlation results. If the relative differences D range from 0 to 0.6, the correlations range from 0.09 (for the 10m wind speed) to 0.97 (for AOD). The common point is that there is no variable with differences above 0.5. This means that, spatially, the studied problem shows systematic patterns from year to year. The low values of correlations show that some variables are systematically poorly estimated. This means that some meteorological structures (for u_{10m}) or emission sources (contributing to the $PM_{2.5}$ surface concentrations) are systematically mis-located.

The representation of temporal correlations shows a specific linear pattern. The largest correlation values are positively correlated with differences. This temporal correlation represents the day to day variability at each location. This means that the studied problem is based on high day to day variability without similar consecutive days (in this case, one would have high

Variable	R_s			R_t		
	Value	D	I_{mv}	Value	D	I_{mv}
T_{2m}	0.60	0.02	0.04	0.91	0.59	**0.82**
u_{10m}	0.09	0.23	0.05	0.59	0.56	**0.53**
precip	0.89	0.20	**0.49**	0.08	0.07	0.02
AOD	0.97	0.02	0.09	0.45	0.34	0.33
ANG	0.91	0.04	0.14	0.59	0.44	**0.49**
O_3	0.69	0.13	0.29	0.32	0.27	0.21
NO_2	0.88	0.27	**0.58**	0.22	0.23	0.13
$PM_{2.5}$	0.16	0.15	0.07	0.27	0.32	0.20
PM_{10}	0.81	0.10	0.27	0.17	0.14	0.07
Ammonium	0.20	0.13	0.08	0.21	0.20	0.12
Sulphate	0.51	0.21	0.29	0.31	0.34	0.23
Nitrate	0.15	0.51	0.13	0.09	0.08	0.03

Table 4. *The I_{mv} values for all variables: the meteorology with T_{2m}, u_{10m} and precipitation rate, the vertically integrated column of aerosols with the Aerosol Optical Depth (AOD) and the Angström exponent (ANG), the surface concentrations of all aerosols in term of size distribution with $PM_{2.5}$ and PM_{10} and for the inorganic species with D_p < 10 µm. Values of I_{mv} above 0.4 are bolded. Units of the variables are detailed in Table 1.*

correlations but low differences). This illustrates the fact that the studied problem is primarily an issue of sporadic events and the model is able to correctly find this variability from one day to another.

6 Conclusions

At first glance, using a different year than the simulated one for the day to day evaluation seems awkward. However, we can learn more about the performances of chemistry transport models than using a single year for the usual statistical indicators. Of course, this approach will never replace a strict evaluation of a pollution case analysis using time series, vertical profiles and usual error statistics. However, it offers a very fast and integrated vision of the strengths and weaknesses of a model with very little calculation. This methodology can also be deployed in inter-comparison exercises.

To answer the questions presented in the introduction, and for this particular model and simulated period, the following conclusions can be drawn. The model always simulates better the studied year than any other meteorological year and it is able to reproduce the day to day variability for high concentrations of pollutants.

The spatial correlation is good for 2m temperature and precipitation rate, but not for wind speed: this highlights the fact that the modelled domain is large and the resolution not optimized for small scale processes. The spatial correlation is also very good for the long-range transport of particles as demonstrated with R_s=0.97 and 0.90 for AOD and ANG. But, since this

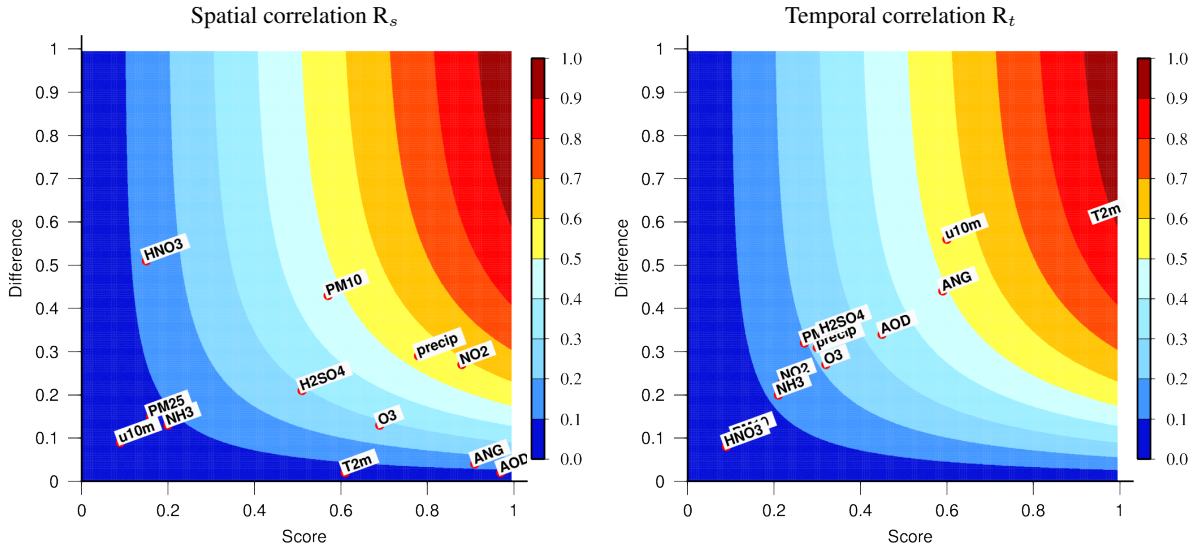

Figure 4. *Results of the I_{mv} scores for the spatial and temporal correlations. For each model variable its value is represented using the correlation on the x-axis and the difference between the studied year and the others on the y-axis. The colours represent the I_{mv} values.*

feature occurs every year, this leads to low I_{mv} values. This means that for a large domain, the main spatial patterns of particle concentrations are recurrent and well modelled. The chemical species that are best modelled are either species with a long atmospheric lifetime (PM_{10}) or species well spatially constrained on the domain (such as NO_2 mainly due to anthropogenic emissions). For particles, the results depend on the size distribution: the coarse particles are better simulated than the fine ones.

The conclusions are different for the temporal correlation. The scores are calculated using daily observations and modelled outputs. Thus, these scores reflect the ability of the model to retrieve the day to day variability. As for the spatial correlation, scores are good for the meteorological variables. For the aerosol, and mainly for the long-lived species (such as mineral dust), the temporal correlation is also correct as the I_{mv} values: I_{mv}=0.33 and 0.49 for AOD and ANG respectively. But for the short-live species the temporal correlation and the I_{mv} values are low. This means that improvements are required in priority for the day to day variability compared to the locations of emissions. This may probably be due to the atmospheric transport, the spatial variability of 10m wind speed being poorly simulated. But, overall, the temporal correlation is better for the studied year than for the others, showing that the problem is highly variable from year to year, but the model is able to capture the evolution of atmospheric composition.

Acknowledgements. This study is partly funded by the French Ministry in charge of Ecology. Thanks to the British Atmospheric Data Centre, which is part of the NERC National Centre for Atmospheric Science (NCAS), for making available the meteorological data, to the EMEP network to provide atmospheric composition measurements, and to the investigators and staff who maintain and provide the AERONET data.

7 Code and/or data availability

This study presenting a methodology using existing data and models, all required information are already included in this article.

References

Appel, K. W., Gilliam, R. C., Davis, N., Zubrow, A., and Howard, S. C.: Overview of the atmospheric model evaluation tool (AMET) v1.1 for evaluating meteorological and air quality models, Environmental Modelling and Software, 26, 434 – 443, doi:doi.org/10.1016/j.envsoft.2010.09.007, 2011.

Baldridge, K. and Cox, W.: Evaluating air quality model performance, Environmental Software, 1, 182 – 187, doi:doi.org/10.1016/0266-9838(86)90023-7, 1986.

Bessagnet, B., Pirovano, G., Mircea, M., Cuvelier, C., Aulinger, A., Calori, G., Ciarelli, G., Manders, A., Stern, R., Tsyro, S., García Vivanco, M., Thunis, P., Pay, M.-T., Colette, A., Couvidat, F., Meleux, F., Rouïl, L., Ung, A., Aksoyoglu, S., Baldasano, J. M., Bieser, J., Briganti, G., Cappelletti, A., D'Isidoro, M., Finardi, S., Kranenburg, R., Silibello, C., Carnevale, C., Aas, W., Dupont, J.-C., Fagerli, H., Gonzalez, L., Menut, L., Prévôt, A., Roberts, P., and White, L.: Presentation of the EURODELTA III intercomparison exercise - evaluation of the chemistry transport models' performance on criteria pollutants and joint analysis with meteorology, Atmospheric Chemistry and Physics, 16, 12 667–12 701, doi:10.5194/acp-16-12667-2016, 2016.

Campbell, P., Zhang, Y., Yahya, K., Wang, K., Hogrefe, C., Pouliot, G., Knote, C., Hodzic, A., Jose, R. S., Perez, J. L., Guerrero, P. J., Baro, R., and Makar, P.: A multi-model assessment for the 2006 and 2010 simulations under the Air Quality Model Evaluation International Initiative (AQMEII) phase 2 over North America: Part I. Indicators of the sensitivity of O3 and PM2.5 formation regimes, Atmospheric Environment, 115, 569 – 586, doi:doi.org/10.1016/j.atmosenv.2014.12.026, 2015.

Chang, J. and Hanna, S.: Air quality model performance evaluation, Meteorology and Atmospheric Physics, 87, 167–196, doi:10.1007/s00703-003-0070-7, 2004.

Cox, W. M. and Tikvart, J. A.: A statistical procedure for determining the best performing air quality simulation model, Atmospheric Environment. Part A. General Topics, 24, 2387 – 2395, doi:doi.org/10.1016/0960-1686(90)90331-G, 1990.

Galmarini, S., Bianconi, R., Appel, W., Solazzo, E., Mosca, S., Grossi, P., Moran, M., Schere, K., and Rao, S.: {ENSEMBLE} and AMET: Two systems and approaches to a harmonized, simplified and efficient facility for air quality models development and evaluation, Atmospheric Environment, 53, 51 – 59, doi:doi.org/10.1016/j.atmosenv.2011.08.076, aQMEII: An International Initiative for the Evaluation of Regional-Scale Air Quality Models - Phase 1, 2012.

Menut, L., Mailler, S., Siour, G., Bessagnet, B., Turquety, S., Rea, G., Briant, R., Mallet, M., Sciare, J., Formenti, P., and Meleux, F.: Ozone and aerosol tropospheric concentrations variability analyzed using the ADRIMED measurements and the WRF and CHIMERE models, Atmospheric Chemistry and Physics, 15, 6159–6182, doi:10.5194/acp-15-6159-2015, http://www.atmos-chem-phys.net/15/6159/2015/, 2015.

Menut, L., Siour, G., Mailler, S., Couvidat, F., and Bessagnet, B.: Observations and regional modeling of aerosol optical properties, speciation and size distribution over Northern Africa and western Europe, Atmospheric Chemistry and Physics, 16, 12 961–12 982, doi:10.5194/acp-16-12961-2016, 2016.

Porter, P. S., Rao, S. T., Hogrefe, C., Gego, E., and Mathur, R.: Methods for reducing biases and errors in regional photochemical model outputs for use in emission reduction and exposure assessments, Atmospheric Environment, 112, 178 – 188, doi:doi.org/10.1016/j.atmosenv.2015.04.039, 2015.

Prank, M., Sofiev, M., Tsyro, S., Hendriks, C., Semeena, V., Vazhappilly Francis, X., Butler, T., Denier van der Gon, H., Friedrich, R., Hendricks, J., Kong, X., Lawrence, M., Righi, M., Samaras, Z., Sausen, R., Kukkonen, J., and Sokhi, R.: Evaluation of the performance of

four chemical transport models in predicting the aerosol chemical composition in Europe in 2005, Atmospheric Chemistry and Physics, 16, 6041–6070, doi:10.5194/acp-16-6041-2016, 2016.

Simon, H., Baker, K., and Phillips, S.: Compilation and interpretation of photochemical model performance statistics published between 2006 and 2012, Atmospheric Environment, 61, 124–139, doi:10.1016/j.atmosenv.2012.07.012, 2012.

Solazzo, E. and Galmarini, S.: Comparing apples with apples: Using spatially distributed time series of monitoring data for model evaluation, Atmospheric Environment, 112, 234 – 245, doi:doi.org/10.1016/j.atmosenv.2015.04.037, 2015.

Solazzo, E. and Galmarini, S.: Error apportionment for atmospheric chemistry-transport models - a new approach to model evaluation, Atmospheric Chemistry and Physics, 16, 6263–6283, doi:10.5194/acp-16-6263-2016, 2016.

Thunis, P., Pederzoli, A., and Pernigotti, D.: Performance criteria to evaluate air quality modeling applications, Atmospheric Environment, 59, 476 – 482, doi:doi.org/10.1016/j.atmosenv.2012.05.043, 2012.

Tørseth, K., Aas, W., Breivik, K., Fjæraa, A. M., Fiebig, M., Hjellbrekke, A. G., Lund Myhre, C., Solberg, S., and Yttri, K. E.: Introduction to the European Monitoring and Evaluation Programme (EMEP) and observed atmospheric composition change during 1972-2009, Atmospheric Chemistry and Physics, 12, 5447–5481, doi:10.5194/acp-12-5447-2012, http://www.atmos-chem-phys.net/12/5447/2012/, 2012.

Valari, M. and Menut, L.: Does increase in air quality models resolution bring surface ozone concentrations closer to reality?, Journal of Atmospheric and Oceanic Technology, doi:10.1175/2008JTECHA1123.1, 2008.

Vautard, R., Moran, M. D., Solazzo, E., Gilliam, R. C., Matthias, V., Bianconi, R., Chemel, C., Ferreira, J., Geyer, B., Hansen, A. B., Jericevic, A., Prank, M., Segers, A., Silver, J. D., Werhahn, J., Wolke, R., Rao, S., and Galmarini, S.: Evaluation of the meteorological forcing used for the Air Quality Model Evaluation International Initiative (AQMEII) air quality simulations, Atmospheric Environment, 53, 15 – 37, doi:doi.org/10.1016/j.atmosenv.2011.10.065, aQMEII: An International Initiative for the Evaluation of Regional-Scale Air Quality Models - Phase 1, 2012.