

## Review of paper gmd-2016-153

An unusual way to validate regional chemistry-transport models, L. Menut et al.

Dear Editor and reviewers,

We acknowledge the reviewers for the time spent to evaluate our work. We also acknowledge the Editor and we made all proposed changes in the revised manuscript.

This is the second revised version for this manuscript. As many points were already discussed, we report the first set of answers to the reviewers at the end of this document.

All our answers are detailed in this letter, but they can be summarized as:

- The English was checked and improved. Some parts of text were moved, shortened or deleted.
- As requested, we added the values of nRMSE for the discussion.
- The fact that the proposed methodology is focused on the model variability (and not on the model errors) is better explained in the text
- The introduction was simplified and some parts are now in the "methodology" section.
- The title was changed and is now: "An alternative way to evaluate chemistry-transport models variability". The word "variability" is explicitly in the title to show we are working more on correlation than on RMSE.
- The principle of the  $I_{mv}$  indicator is better explained: this is an indicator corresponding to a weight on the correlation.
- The presentation of the results was improved: the section 4 presents the time series of the correlations and nRMSE for each studied variables, when the section 5 summarizes the results with the Table containing the  $I_{mv}$  values and the colored figures representing all the  $I_{mv}$  values on a single plot.

Finally, please note that our answers are in blue in the text and after each reviewers remark.

Best regards,  
Laurent MENUT  
February 18, 2017

## 1 Message from the Editor

Topical Editor Decision:

Reconsider after major revisions (12 Feb 2017) by Slimane Bekki

### Comments to the Author:

Anonymous Referee #2 has provided a number of corrections and specific suggestions that the authors should take into account.

## 2 Second revised version: Answers to Anonymous Referee #1

### 2.1 Major comments

The methodology proposed by the authors is original and has potential to complement the traditional approach. Unfortunately, as presented in this publication, the added value of this work remains limited because subjective judgements must be made to interpret the results. I also feel that my previous comments have been accounted for, only very partially.

*Answer:*

We thank the reviewer to consider the study as original. We regret his feeling about his comments.

In our first answer (in section ??, p.?? of this document), we wrote detailed pages to answer all his points. We consider that a review, and more in a journal called "discussion", is an exchange of ideas. Here, we have the feeling we have to write answers about the same questions. It probably means that our explanations were not enough clear. So, we answer again, trying to be more clear.

1. The bibliography has been extended but those references are not used much in the text. For example the decomposition of the main indicator RMSE (or MSE) into its three components (Solazzo and Galmarini, Thunis et al., Taylor et al.) could be used as starting point to identify the different indicators and justify the choice of the correlation as central one for this study. The decomposition into a systematic and unsystematic error has already been done in other works that are not referenced.

*Answer:*

The main goal of the introduction is to present previous works done in the field of "model validation" in a general manner. But it is not to describe in details all previous results. It is to present the general choices of these studies and to clearly show that we are doing differently and explain why. The authors cited by the reviewer have an impressive background in regional CTM validation. They develop tools widely used in the community. In several articles, they showed that they consider that the MSE is the best indicator for model validation. We fully understand this point of view but this is not the goal of our study: we want to focus on correlation only because we think this indicator is sufficient for our observations to model comparisons in the framework of the **model variability**. Our choice is better explained and justified. For the references, there is certainly more to add. The reviewer wrote that "the decomposition... has already been done" but doesn't propose reference. If the reviewer thinks to specific works, why did he not cite them?

2. Even if not applicable in a diagram, information on the bias or on the standard deviation could be provided. Values of SN and D could be calculated and added to the analysis. The fact that the bias is low and does not show any variability (e.g. for T2m) is an interesting information per se. It could also be that the bias shows more variability for other species and then become the crucial parameter to analyse.

*Answer:*

As previously explained, the RMSE could be used to calculate differences between years. We added the values in this revised version for the discussion. But as already explained in the previous revision, their addition in the  $I_{mv}$  indicator needs to completely change the calculation. We already tried this and this did not provide results having a clear added value compared to the indicators based on the correlations. We added a few explanations about this in the text.

3. My point about the use of the 'score' terminology has not been addressed. This term is used (55 times in the whole document) indifferently to describe the indicator (e.g. p2 l5-6, p3 l63, p3 l87) and the value taken by this indicator (p4 l55), which makes it confusing to understand. Could the Author give a definition of "score" and check that it is consistent though the document?

*Answer:*

We agree with this point and this is now corrected.

4. The title does not reflect what the methodology is about. I agree with the Authors that the methodology is unusual but many things can be defined as unusual. I believe that the title should

provide insight on the novel aspects discussed in the document (the use of different meteorological years)

*Answer:*

This is the second time that the reviewer wants to change the title. For this second revised version, we propose: "**An alternative way to evaluate chemistry-transport models variability**" or "Evaluation of model variability using measurements data from other years than the modeled one".

5. Regarding the qualitative aspects of the methodology. I agree that the indicators are calculated quantitatively but the judgement made on whether the results are good or bad remains subjective. This judgement is based on expert knowledge (e.g. that a correlation of 0.5 is very good for a given species). In my view, this limits the benefit of the methodology, as users need to know a-priori what a good behaviour is. I understand that the key point is in the use of several years of data but if at the end the interpretation of the indicator depends on expert judgement, this is a limitation. Examples (p5 125-27; p5 147-50; p6 111)

*Answer:*

This point is about the indicators in a general way. For the correlations, for example, yes, the fact to have a correlation of 0.5 would be good or not, depending on the modelled variables. We agree that the user has to know that a correlation of 0.8 for temperature is the state-of-the-art, when this is closest to 0.5 for  $PM_{10}$ . The values are different and probably will evolve (and increase) with time and model improvements. In fact, our methodology has to be used by researchers already knowing the atmospheric composition modeling, as well as the current performances for regional air quality modeling. For the  $I_{mv}$  indicator, this is the same. It evolves between 0 and 1, considering this indicator is a weighted value of the correlation usually used.

Thus, from our point of view, there is no problem of 'qualitative' discussion in our paper. If the reader is able to know if a correlation is correct or not, he is also able to know if  $I_{mv}$  is correct or not. Mainly because the  $I_{mv}$  indicator is just a weighed value of the correlation.

6. I still do not understand why observations cannot be used to fix a minimum threshold. According to me, values of SN and D calculated on the only basis of the set of observations (substituting the model value by the observation of the reference year) could be calculated to make the approach a little bit less subjective.

*Answer:*

We already developed an answer about this point. But it was probably not clear enough.

The  $I_{mv}$  indicator is dedicated to the model variability: this is why this is based on the correlations only. This is already written in p3. 166-69 (where the reviewer found this is unclear and this is corrected in this revised version). The main goal of this indicator is to have only one value: the correlation and a weight depending on the other years. This is the originality of the study.

Thus, this indicator has exactly the same meaning than the statistical indicator used. If we use the temporal correlation, the  $I_{mv}$  value is just the result of this indicator multiplied by a weight function (between 0 and 1) and relative to the differences between the years. For example, if you have  $R_t=0.5$  for aerosol, you can consider at the regional scale that your model gives satisfactory results. But, if all years are modelled in the same way, the differences are low, the indicator will automatically decrease.

The use of observations is dedicated to know if the studied year is very different from the others or not. This is more a problem of trends along years and this has no link with the model quality. In fact the observations are already contained in our calculations: we use the correlation, i.e the way the model is able to reproduce the observed variability. If the difference  $D$  is low for one year, it means that the observations evolve in the same way for this another year than for the reference year. To make directly the differences between the observations would give exactly the same result and does not provide an additional 'objective' information.

7. English has been improved but many misspells and unclear sentences remain (only few examples provided below).

*Answer:*

OK. A second round was done for English.

## 2.2 Minor comments

8. P2 132 require
9. P2 164 A couple of lines would be needed to indicate that the Authors now start the description of the methodological approach.

*Answer:*

The introduction was simplified and this text was putted in another section, more dedicated to the methodology.

10. P3 166-69: unclear
11. P3 172: The bias is an indicator, not a score!

*Answer:*

All sentences with score/indicator were checked and corrected.

12. P3 173: I disagree with the Authors, the RMSE is not driven by bias. Depending on the variable and the period of time considered, the RMSE can be dominated by correlation, bias or by standard deviation.

*Answer:*

This sentence was deleted.

13. P3 178: unclear formulation, please re-phrase.

*Answer:*

A large amount of the article was rewritten, including this part.

14. Figure 1: I guess MYV should be  $I_{mv}$

*Answer:*

Following this reviewer recommendation, we changed the MYV term by  $I_{mv}$  in the first revised version. But, sorry for that, we forgot to change also the first two Figures. This is now corrected.

15. p5 l 23: have → has

16. p5: why a subscript "s" in Ds

*Answer:*

Sorry, it was an error and this is now corrected.

17. Figure 2; MYV should become Imv I guess *Answer:*

Yes, right and corrected.

18. P5 l70: unclear, please re-phrase

19. P6 l14: year → years

20. P6 l16-18: please re-phrase

21. P6 l27-28: disagree: see point 12

### **3 Second revised version: Answers to Anonymous Referee #2**

This reviewer proposed to accept the paper as is.

*Answer:*

We thank this reviewer to have accepted our corrections.

## 4 First revised version: Answers to Editor

There is some common remarks which can be synthesized:

1. *The bibliography could be improved:* this was done and the state of the art regarding the current ways to validate CTMs was rewritten. In brief, the following references were added: [?], [?], [?] [?], [?], [?], [?], [?], [?], [?].
2. *The scores could use RMSE and bias:* This is right, and in fact, we did it during the preparation of the manuscript. This was removed for the submission because we considered that the added-value was low. A long explanation for this choice is proposed in the answer to the reviewer #2.
3. *The interest to have a MYV score:* There is two kinds of novelties in this paper. First, the fact to use data from other years than the studied year is the most important novelty. This is why the title is "unusual way", because this is the first time that such way to estimate the model realism is used. Second, the MYV score. This is also new and the goal is to have a quantified link between the "differences" and the scores (correlation, RMSE, etc.). The constant value is arbitrary, this is true. But the user can select another value. In the case of CTM, this is subjective, but knowing the state of the art of CTM modelling, a correlation of 0.5 is considered as "very good" for some species (such as inorganics or PM, for example). Thus, this is important to put this subjective information on a plot to show that the results are not perfect, but may appear as good, knowing the current capabilities of CTMs.

## 5 First revised version: Answers to Anonymous Referee #1

Received and published: 3 August 2016

In this paper, the authors present an extension of the evaluation of (atmospheric chemistry) models by using measurements from other years than the year which was simulated by the model. New scores are introduced to quantify the ability of the model to capture the day to day variability as opposed to persistent patterns.

### General comments:

While reading the paper I asked myself the question if the approach presented by the authors has a real added value as compared to a more traditional model evaluation based on bias, RMS and (one type of) correlation, and may be adopted by other groups. In the end I decided that it probably does, for the following reasons:

- The approach proposed quantifies the importance of day-to-day, weather dominated variability versus systematic patterns which are repeated from year to year.
- The approach naturally leads to an overview of the performance for multiple species in one graph (e.g. Fig.5), which is especially also useful (maybe even more useful) for comparisons between different models. This include both trace species as well as meteorological variables. This is a bit similar to the use of Taylor diagrams.
- The approach explicitly exploits both spatial and temporal correlations, which bring complementary information.
- The approach provides new insight into the performance of the WRF-CHIMERE model.

Because of this I am in favour of publication. However, to my opinion there are several major and minor points to be addressed before the paper can be considered by GMD.

These are listed below:

- Is this approach really new? The authors provide a few interesting references in the paper, but I would like to see a more systematic overview of the model evaluation approaches and techniques/scores adopted in the past (e.g. including several European/American CTM intercomparison exercises) to better understand the added value of the approach proposed.

We think the approach is really new: we never see before a comparison between a model simulation and data from other years. We made a complete bibliography, improved in this revised version. This is the novelty of the paper: considering that using other years is the way to split results between "climatological" events and sporadic events and, thus, the model's ability to catch sporadic events.

- The formulation is incomplete, and mathematical formulas are not well defined. In particular, the authors should provide the equations for  $R_s$  and  $R_t$ , and the mathematical formula for the MYV needs more discussion, see my comments below. Also, the authors should motivate why the  $R_s, t$  scores are chosen.

The part with the mathematical formulas was rewritten and is now more complete. More arguments are proposed for the choice of the MYV formulation. We understand the reviewer comments and, clearly, the score as it proposed may be discussed. In fact, we tried several scores before submitting the publication and we found that the proposed one corresponds to the best choice regarding the type of result we want. The choice of the correlations is detailed below. The bibliography added in the introduction showed that the models are usually validated using three scores: correlation, RMSE and bias. For regulatory purposes, the bias and RMSE are important scores. The bias is certainly the most important to catch the annual mean difference between the model and the observations. But this does not reflect the model variability, i.e the ability of the model to reproduce the real physico-chemical variability. The RMSE is strongly influenced by the bias. For these reasons, we focus on the correlations, spatial and temporal, in this study because we are more interested by the processes evaluation.

- The MYV is not really a model score to my opinion, but rather an indicator of how much the score is influenced by day-to-day variability. In particular one can argue that  $R=1$  and  $D=0$  is a good result. Also, I wonder if a formula for MYV is really needed. Showing  $D$  and  $R$  is maybe enough (see e.g. Fig. 5)? This should be more carefully presented/discussed.

We agree with the concept of indicator in place of score. This was changed accordingly in the whole text. We think a formula for MYV is really needed because this is the only numerically way to link  $D$  and  $R$  and to propose a unique value to analyze. Showing  $D$  and  $R$  is a good way, but mainly a graphical way. In addition, we want values for the discussion, and possibly, inter-model comparisons. Is  $R=1$  and  $D=0$  a good score? Not really, because it means that the model is good to reproduce something easy to model (being every years).

#### Detailed comments:

- p4, l13: "they are used as daily averaged in the present study": why this choice to focus on daily averages instead of hourly values? Please motivate.

There is two reasons for the use of daily averaged measurements and model outputs: (1) as shown in the table 1, some data are hourly and some others are tri-hourly. In fact, even if we are presented as tri-hourly, the precipitation data are correct to use only in a daily way. As we want to have the same score for all measurements, we then chosen to use daily averaged data. Another reason: we want to split the high temporal frequency variability and the systematic patterns. The day-to-day is the best frequency for that. If we had used the hourly measurements, we certainly added a false variability due to "systematic daily" behaviours such as the diurnal cycle for temperature or  $NO_x$  emissions.

- p4, table: Provide also the full names of the variables, e.g. "Temperature at 2m above ground" etc. The full name of all variables was added in the text.

- p4, last line: replace "same day for another is" by "same day for another year is"  
OK corrected.
- p5, 14: "The correlation is the more appropriate statistical metric for such analysis." Please explain and motivate this statement in detail. This is important for the rest of the paper!  
This point is similar to a reviewer #2 remark and a long discussion is proposed below. The correlations are able to split the relative contributions of systematic weather or sources dominated variability and day-to-day variability. The key point of this study is the study of the model variability and the variability is statistically represented by the correlations. The mean bias (or the normalized bias) is not a score to quantify the variability. And the RMSE is a score containing a part of variability but is mainly driven by the bias. This was added in the revised version.
- p5, 18: "The spatial correlation, noted  $R_s$ , is calculated from the temporal mean averaged values of observations and model for each location where observations are available." Please provide a detailed mathematical formula/recipe to be clear. Are observations and model first collocated for individual observations, or are means computed and then compared. Are these means of daily means or means of hourly values? It is important to define precisely how the correlations are computed: the devil is in the details.  
All correlations are calculated using mean daily values. Using these daily values, the spatial correlation is the correlation using all data, for all sites. The formula for the correlation was added in the revised version.
- p5, 13: Also for the temporal correlation: be more precise. Is it based on daily means, hourly values or something else.  
All scores values are estimated using daily averages values. This was added in the text.
- p5, 114: "The longer the atmospheric lifetime of the species, the lower the relevance of temporal correlation" I would dispute this. For long-lived tracers the transport (wind direction) and location/strength of the sources becomes crucial, directly influencing temporal correlations. I suggest to remove this remark.  
We agree, this remark was removed.
- p5, eq.1: Why is there an absolute value introduced. Instead of  $absolute(s_i - s_N)$  I would suggest  $(s_N - s_i)$  assuming higher values of "s" (or "s" close to 1) indicate better performance, which is the case for correlations.  
There is an absolute value because all values are not always positive: for some variables and some years, you may have a positive correlation for the year N and a negative one for another year. More difficult, in some cases, you may have a better correlation for another year than for the studied year.
- p6, eq.2: Remove the "X" (multiplication) from the formula. This is not needed (in eq.1 there is also no X). Please introduce a one character symbol for the "Multi Year Variability" instead of writing "MYV" in eq 2, which, in mathematical formula's means M times Y times V. "D\_s" has not been introduced: is it the same as "D" ?  
The formulas were cleaned and the MYV is now noted  $I_{mv}$ , for "Indicator of Model Variability".
- p6, eq.2: Why this complicated exponential form?? It seems that you ideally would have the MYV to be =1 for (s=1 and D=1), and =0 for (s=0 or D=0). A much simpler form  $s_{MYV} = s_N D$  would do the trick. In fact, eq.2 is not =1 for  $s_N=D=1$ . Where does this formula come from? Is there a reference to a paper introducing this form? Also, it would be good if the formula has clear limits, e.g. 0 (very bad) and 1 (very good). This is not the case when D=1.  
The exponential form is really complicated? We think this is easy to implement and to use it. The form was chosen to have a non-linear indicator in order to give more weight to the high values and to take into consideration that the scores (correlation, RMSE or bias) may have a different weight than the differences between years. Of course, the modeller may just use the values of the score and the difference (two values), but the indicator is able to provide just one synthetic value for the discussion.



- p7, l2: Where does the number 0.3 come from? It will depend a lot on how the score "s" is defined. The number seems arbitrarily chosen.  
Yes, the value was arbitrarily selected and this is explained in the text, page 6 - line 5. This is a tunable parameter and its only role is to provide a weight on the scores and their differences. The user can change this value as a function of the studied problem. In our case, we found that 0.3 is a good proxy to have values representative of the state-of-the-art of chemistry-transport modelling and validation. As we said, this value is not really important and has no impact on the discussion: this is just a way to highlight the good performances (or not) of the model simulations compared to the observations.
- p7, l14: " ... is challenging because several uncertainties ... "  
We agree, we corrected in the revised version.
- Table 2: It would be helpful to remind the reader that these are Summer periods (1-5 to 1-9) and that the scores are based on daily mean values. Please also highlight the special situation for 2013 (I would suggest to start with 2013, add a thick line, and continue with 2008 2009 ... Perhaps it can be stressed once more in the caption that observations for 2008-2012 (and 2013) are compared with 2013 model results.  
Yes, we agree with that and for the whole paper, the captions were extended and are now more precise. For the order of the lines, we prefer to keep the increasing order for the years. But the new caption will help to well understand this Table.
- Figure 4: Caption is incomplete.  
The caption was completely rewritten and is now more clear.
- Table 3: "... Values of MYV above 0.3 are shown in bold..."  
OK this was corrected.
- p11, l18: ... with differences above 0.5...  
OK this was corrected.

## 6 First revised version: Answers to Anonymous Referee #2

This work addresses the important issue of the validation of chemistry transport models. The authors present a new methodology in which the traditional approach consisting of comparing measurements with model results for a given time period is extended to comparisons of the same model results with measurements from other years. The authors develop then a specific indicator on this basis that allows discriminating results that are good for the good reason from those that are good only because of highly persistent pattern present in the observations from year to year. While the proposed methodology is original and has a potential to complement the traditional approach, the authors remain unfortunately superficial and qualitative in their way of presenting and applying this methodology. As a consequence, the proposed examples are qualitative as well and are not helpful. Finally, the document is poorly written: (1) English would need revisions throughout the whole document and (2) many sections would need to be re-written (some suggestions are proposed below).

We thank the reviewer for the interesting suggestions in this review. The English was completely revised and the proposed sections were rewritten.

### Major points:

1) The authors mention Solazzo and Galmarini (2016) for their decomposition of the error but they finally focus on the correlation only. As noted by these two Authors but also by many others (the referencing to other works relating to model evaluation should be improved), it is important to look at all three possible source of errors because focusing on the only correlation may lead to the wrong conclusions (see comments below). I'm wondering why the Authors make this choice as the proposed

methodology could easily be developed for other indicators that are more representative of the overall model performance (e.g. MSE).

This remark is an important and interesting point. Why the scores are done for the correlations (spatial and temporal) and not for the RMSE and the bias? In fact, we did this work in a preliminary version of the paper. Finally, after discussion between all authors, we decided to present only scores for the correlations. We understand this choice may appear surprising but there are several reasons for that:

1. The main goal of this paper is to separate the contributions due to systematic events (i.e the model seems good but finally is only able to model the same thing every day and every year) and due to sporadic events (i.e the model is good because able to retrieve day to day variability). For this goal, the correlations (spatial and temporal) are the most interesting indicators. We agree that RMSE and bias are also important indicators but **the goal of this study is not to replace already existing approaches but to give a complementary insight on the results.**
2. The behaviour of correlations and bias and RMSE is not the same. The correlations are always between 0 and 1. More the correlation is high more the indicator is high. This is the contrary for RMSE and bias: More the score is high more the indicator is low (a large bias indicates a wrong simulation). In addition, the RMSE and bias are not bounded between 0 and 1, may have large values or negative values. Thus, in a previous version of this paper, we tried to combine the formulation of the indicator with only one formula as:

$$MYV_s = (\alpha - \beta s_N) \times (1 - \exp(-D_s)^\delta) \quad (1)$$

where  $\alpha$  and  $\beta$  are arbitrarily chosen constants, to define differently depending on the score (correlation or based on absolute values), as:

- For the correlations (Rs and Rt), we want an indicator increasing when the correlations increase. We thus select  $\alpha=0$ ,  $\beta=-1$ .
- For the bias and RMSE, we want an indicator increasing when the values decrease. We also want that the score is only between 0 and 1 for readability. But, RMSE and bias may be very large. We thus use  $\alpha=1$ ,  $\beta=1$  and we impose to have  $MYV_s=0$  when negative values are estimated.

The value for  $\delta$  is arbitrary but has just to be larger than 1. This tuning parameter enables to adapt the relative weight we want between the absolute value of the scores for the studies year and the differences between all years. In general, we want that a good score for the studied year have a largest weight than the differences: in this case, we select  $\delta=4$ . By adding RMSE and bias, we are **obliged to have a more complicated formula, with more tuning parameters.**

3. Last: when using these scores with the data presented in the paper, we found no benefit when using RMSE and bias for the discussion. For this letter, we add some results previously found (but not submitted in the paper). This is to show to the reviewer that the use of RMSE and bias is, with this specific approach, not a real benefit for the interpretation of the results. Examples are proposed in Figure ?? for 2m temperature, AOD and O3. But the conclusion is the same for all studied variables: **there is no variability for the RMSE and bias able to help to conclude on the model quality.** A new paragraph is now added in the manuscript to explain this point and why we decided to focus on correlations only.

2) The approach proposed by the Authors remains qualitative and the interpretation depend on the setting of an arbitrary threshold (e.g.  $MYV=0.3$  in Figure 3). Throughout the text, the Authors

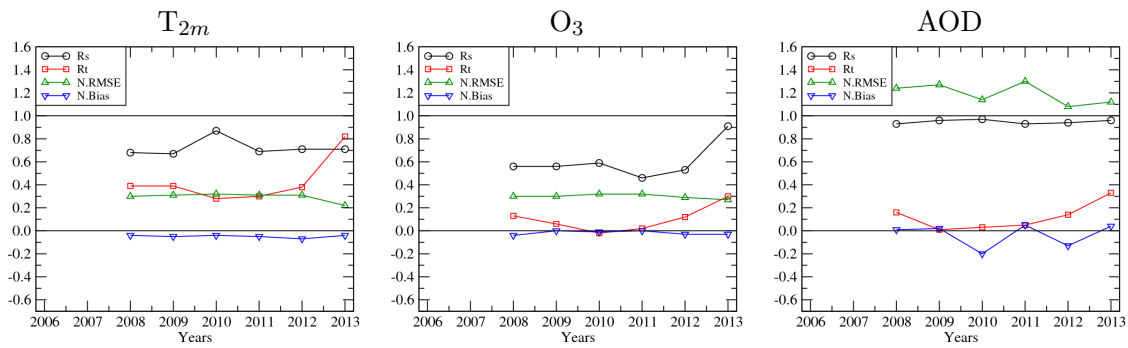


Figure 1: Multi years scores for the 2m temperature, ozone and AOD. The reference year is 2013.

make qualitative judgements (0.6 is good, 0.5 is poor...). This limits the usefulness of the proposed methodology as we never know what a good value of the indicator is. I do not understand this limitation as it would seem relatively straightforward to calculate a value of the MYV indicator in a similar way but on the only basis of measurements. This observation-based MYV value could then serve as the threshold beyond which model results would be considered good enough.

There is two different things in the paper: (1) the idea to compare a simulation for a specific year with data from another year. This is not qualitative but fully quantitative. (2) the proposal for an indicator, linking the differences between the years and the correlation values, in order to have only one indicator (and not two). This may appear as qualitative because we prefer to say that the user may change this value. But, in reality, we tried a large range of values and we conclude that the proposed value is the best for the problem related to regional chemistry-transport modelling. We changed the text to be clearer: "using  $\delta=4$ , we consider that the relative weight of the correlation value against the difference reflects well the state-of-the-art of CTMs regional modelling. Using this value, we consider that the model is good enough and for the well reason if  $MYV > 0.3$ ". In addition, even if this seems a good idea, this is not straightforward to establish an "universal" value of the parameters using only observations. Observations are the reality and to compare several years can not provide the information we need. **But, the important thing is that the choice of  $\delta$  and  $MYV > 0.3$  is not the key point of the paper.** The key point is to use other years that the modelled year to validate the model results. **Please consider these parameters only as an additional help to synthesize and interpret the results.**

3) The document is poorly written. Many sections are unclear and lack sufficient details to be understood. Some suggestions are provided below but the whole document should be thoroughly revised. Ok, thanks. We made all proposed changes. We are happy to see all these corrections showing the reviewer considers the work is interesting to publish. Detailed answers are provided after each reviewer remark.

### Minor points:

1. P1, l1: The title is not very representative of the work

The "unusual way" is the fact that the validation is done using years different from the studied one. To our knowledge (and after an improved bibliography), this is new and unusual.

2. P1, l3: "and by natural" → "and natural"

OK corrected.

3. P1, l19: the transport

OK corrected.

4. P1, l20: or from the QAERONET

OK corrected.

5. P2, l1: can be

OK corrected.

6. P2, l2-3: sentence to be revised

The sentence was too long and was simplified. This is now: *But there can be multiple reasons for a model simulation to agree or disagree with observations. That is because the result of a simulation is the integrated budget of several processes.*

7. P2, l4: "spatial representativeness" → "spatial representativeness of the monitoring stations". In addition, this concept is mentioned for the first time and should be defined. Finally, I do not get the added value of mentioning this here.

The term is now better defined in the new paragraph (see answer just below for P2L5).

8. P2, l5: "to isolate problems intrinsic to the models,". This is unclear and should be re-phrased

We agree and the sentence was rewritten and is now more clear as: *A fundamental difference between observations data and models results is the coherence of the spatial representativeness of the monitoring stations compared to the model cell [?, ?]. To quantify the model errors due to mis-representation of physics and chemistry from those only due to representativeness, several methodologies have been developed. These methods are effective but often required important computation time.*

9. P2, l6: "relevant": which ones?

This word was removed in the new version.

10. P2, l7: "but often with huge" → "but often require important"

OK corrected.

11. P2, l8: references should be within brackets

OK corrected.

12. P2, l15-17 and l18-20: if the authors cite these works, they should explain in a little bit more detail their main aspects and why these are important in the context of their work. All these references are introduced independently from the scope of the work. For example on l18, what is the decomposition about? L17, what did Rea et al. find that is relevant for this work...

This part was completely rewritten and new references were added. The work of Real et al. is just cited to show that some studies are dedicated to split the individual contributions. Of course, this is not the same goal as this paper. The reference was removed.

13. P2, l18: scores is often misused in the text. Sometimes as real score, some times meant as correlation. I guess the authors here refer to indicators.

We agree with this remark and the words "score", "correlation" and "indicator" were harmonized in the paper.

14. P2, l23: "we apply these scores to a model simulation" is unclear. I do not understand how to apply a score to a model simulation. Please check all occurrences of "scores" and check relevance.

This paragraph was also rewritten. This is now: *For all these variables, temporal and spatial correlations are computed to identify the model capacity compared to observations. First, the correlations are calculated between observations data and model outputs for the simulation year (i.e. the reference year). Second, the correlations are calculated between the observations data for other years and the model output for the reference year. Logically, the correlations calculated for the reference year for observations and model outputs would give the better results. By difference with the correlations 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 scores (correlations, bias, RMSE) calculated for a single year, the studied year. It is thus expected to give additional elements to answer these questions: 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? For a given variable, does the model have a good spatial representativeness compared to the corresponding observations?, and Are the biases introduced by meteorological or emissions variability or by the formulation of processes in the chemistry-transport model itself?*

15. P2, l27: provide

OK corrected (rewritten in the new paragraph).

16. P2, l29: spatial representativeness is not yet defined. Is special representativeness really assessed by this method? I do not believe so (see following comments)

This is now done with the new paragraph (see answer for P2L5).

17. P2, l33: Score meant as indicator?

Yes, and it was corrected.

18. P3, figure 1: I do not believe this figure helps understanding. The proposed methodology is quite universal and does not require to enter these details

This figure is very simple and is just here to illustrate the paragraph. This could be important for people not familiar with the impact of some variables errors on other variables in the chemistry-transport modelling system. But if the reviewer considers this is not useful and this can be a limitation for the publication, we accept to remove this figure.

19. P3, l7: forcings

The paragraph was completely rewritten.

20. P3, l9-23: these lines are not necessary to the methodology and application

These lines are not necessary for the methodology application, this is correct. But the knowledge of the several dependencies between the variables helps to the interpretation of the results.

21. P4, l4: unclear

This was rewritten.

22. P4, l9: for → in

OK corrected.

23. P4, l12: variable (Table 1)  
OK corrected.
24. P4, l16: and during → for  
OK corrected.
25. P4, l21: take the same day for another → to re-phrase  
Yes, OK. In fact this is "the same date".
26. P5, l4: why is correlation the more appropriate metric. Why couldn't we say the same for the bias, for example?  
Yes, we understand this remark. The reasons for the use of correlation or bias were explained before in this letter. This line was changed as the complete paragraph was rewritten.
27. P5, l5: What is a usual correlation score? A correlation is a correlation and a score a score!  
There is several types of correlations. We added the definition of the Pearson correlation we used in this study.
28. P5, l11-12: I disagree with the authors. A good correlation score does not indicate that the resolution is adequate, transport is adequate... Correlation could be 1 while keeping a huge bias due to a too coarse resolution.  
The reviewer is right if we are talking about absolute value of the variable. But in our case, as indicated P5L9, we are here talking about the location of pollutants plumes (and not their intensity). Our sentence was dedicated to the day to day variability, independently of the bias value.
29. P5, l16: "particularly": why?  
Yes, this is right, there is no reason. This word was deleted.
30. P5, l20: which differences? Between what?  
The differences between the correlations values. The sentence was corrected. But we are here in the paragraph dedicated to the definition of D.
31. P6, l5: why should it be larger than unity?  
Because, at the end, you want to have an indicator between 0 and 1.
32. P6, l5-6: These lines are totally unclear and should be re-phrased  
Yes, OK. This is probably because these lines are unclear that the reviewer was so critical with the principle of an indicator. The paragraph was thus rewritten.
33. P6, l7: have → has  
Ok, the paragraph was completely rewritten.
34. P6, l7: why do we want that a good score... ": although it may appear straightforward, please give a few words of explanation.  
Ok, the paragraph was completely rewritten.

35. P6, 19: What is an academic value of the score, what is the score meaning here?  
The "academic" value is just because the plot does not contain real data but only the values of the indicator. This was added in the text. And we are OK with the wording; this is not "score" here but "indicator".
36. P6, 110: absolute score but also variable: unclear  
OK this was corrected. The text is now: *Ideally we would hope that the model performs well for the correlation scores but also be able to reproduce the observed variability.*
37. P6, 19-15: this all paragraph is unclear and should be rewritten  
This was rewritten.
38. P6, 118-19: 5 times scores in these sentences!  
This was also rewritten.
39. Figure 3 and Figure 6 seems to be inconsistent in terms of X axis labeling.  
There is "correlation" and "score". We replaced "correlation" by "score" in fig 3 for consistency.
40. P7, 11: from Figure 3  
Ok, corrected.
41. P7, 11: we can consider that  
Ok, corrected.
42. P7, 11-2: This means that all conclusions will remain subjective because of this arbitrarily fixed delta parameter. I believe that a measurement based threshold value for delta can be fixed, withdrawing this arbitrary aspect (see major comment above).  
As discussed before, this is not really subjective: the correlations values and the differences values are completely objective. The way to link these two values using the  $I_v$  may appear as subjective (because we are fixing a  $\delta$  value, but the reviewer has to consider that this is our choice to define an indicator as we want. For the second point, we don't know how to do the same job for observations: the indicator is defined to characterize the model ability to simulate real observed events. The observations alone have not the same meaning: what can we conclude if an observations for the 12 May 2013 is different or not that the same observations for the 12 May of 2008, 2009, 2010... etc? This is not the goal of this paper.
43. P7, 16: done → calculated  
OK corrected.
44. P7, 16: MYV scores  
This was replaced by the new name of the indicator: *To better understand the relevance of  $I_v$ , two examples are detailed in this section.*
45. P7, 112: vary a lot → vary significantly  
This is P7L13 and this was corrected.
46. P7, 113: is challenging because  
This is P7L14 and this was corrected.

47. P7, l13: again spatial representativeness needs to be defined  
This is now defined in the new paragraph in a previous section.
48. P7, l17: "The spatial correlation is good for all years". I do not understand which arguments the Authors use to state that the score is good. If the spatial pattern is easy to reproduce, it could well be that a correlation of 0.7 should be considered as bad. This seems to be confirmed by the next sentence: "the model reproduces fairly well a spatial patter observed every year". One way forward is to calculate the correlations on the only basis of measurements to get some indicative threshold of what is good or not.  
This remark is close to previous remarks and we rewritten several paragraphs to make it clearer.
49. P8, l2: Are we sure this is for the good reasons?  
If the correlation and the differences are high, we can conclude this is for the good reasons, i.e a correct modelling of the day-to-day variability. In general, the temperature is one of the variables the most well modelled. The result is not surprising.
50. P8, l6: "This species is secondary" seems to contradict p7, l12.  
NO<sub>2</sub> is both a primary and a secondary species. This was corrected here.
51. P8, l6,7: I do not agree that a good score for correlation is indicating a good transport, photo-chemistry... Correlation is indeed only one of the indicators to assess model performances and it only provides a partial vision of model performances. Correlation could be perfect even with a very large bias.  
We agree with that, but here we focus on the emissions and transport in the text. And the correlation is a good indicator for that. The bias is related to the intensity of the source and not to its location or to the transport.
52. P8, l8: low → coarse  
OK corrected.
53. P8, l8: less good → worse  
OK corrected.
54. "Its spatial extent of its representativeness": totally unclear, this should be rephrased  
OK, this was corrected with: *...being more spatially limited (emissions...*
55. P8, l18: "The scores": The correlations are calculated, not the scores which are the correlation values  
OK, this was corrected.
56. P8, l20: "each score type". I do not understand what the Authors mean.  
OK. The part "each score type" has no interest since we already defined  $I_v$ . This was removed.
57. P8, l20: "Results are presented in Table 3. These results..." → Results (Table 3) are discussed...  
OK corrected
58. P8, l24: why only?  
Yes, Ok not "only".



59. P8, l24: Which arguments are used to state that the spatial correlation is not correct?  
Because the value in the Table is  $R_s=0.09$ . This was added in the text.
60. P8, l24: for one year → from one year  
OK corrected.
61. P8, l26, 27 and 28: "very good spatial", "less good", "well retrieved". The Authors should explain how they come to these statements.  
We followed the criteria we defined to help the interpretation. Now that the paragraph about the indicator definition is clearer, we think that this part would be also clearer.
62. P8, l31: A few words to explain what the AOD and ANG are would be helpful  
Also following the Reviewer #1, the acronyms were extended. We already removed the figure explaining how a CTM works because the reviewer considers this is too simple and there is no need to remind this in this paper. This is probably the same for the aerosol optical properties, the basis for anyone studying aerosols.
63. Figure 4 caption: Should include explanations of the two curves represented  
Yes, that's right, more informations are added in the caption.
64. P10, l9,10,11: Again I do not agree with these conclusions which cannot be drawn from the only correlation values. Please see all our answers in this letter about the use of the correlations.
65. P11, l19-20: this sentence is unclear  
Ok, the sentence was changed. This is now: *The low values of correlations show that some variables are systematically badly 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.*
66. P12, l29: dued → due  
Oups. OK, thanks, this was corrected.

# An ~~unusual~~ alternative way to ~~validate regional~~ evaluate chemistry-transport models variability

Laurent MENUT<sup>1</sup>, Sylvain MAILLER<sup>1</sup>, Bertrand BESSAGNET<sup>2</sup>, Guillaume SIOUR<sup>3</sup>, Augustin COLETTE<sup>2</sup>, Florian COUVIDAT<sup>2</sup>, and Frédéric MELEUX<sup>2</sup>

<sup>1</sup>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

<sup>2</sup>INERIS, National Institute for Industrial Environment and Risks, Parc Technologique ALATA, F-60550 Verneuil-en-10 Halatte, France

<sup>3</sup>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](mailto: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 ~~more~~ on models performances by comparing the simulation results with observational data available for other time periods than the period originally targeted. First, the ~~usual scores~~ statistical indicators selected in this study (spatial and temporal correlations) are computed for a given period, using co-localised observation and simulation data in time and space. Second, the same ~~scores are calculated~~ indicators are used to calculate scores for several other years by conserving only the spatial locations and Julian days of the year. The difference between the ~~two score results~~ provides complementary insights ~~to the following questions: (i) is the model performing well only because the situation is recurrent? (ii) is the model representative enough of the measurements for all variables? (iii) if the pollutants concentrations are not well modelled, is it due to meteorology or chemistry?~~ on the model capability to reproduce well the observed variability.

10 In order to synthesise the large amount of results, a new indicator is proposed: ~~the "multi-year variability"~~, designed to compare the several error statistics between all the years of validation and to quantify if the studied period was fairly modelled for the good reasons.

## 1 Introduction

Chemistry transport models (CTM) aim at simulating the ~~air pollutants concentrations in the lowest layers of the atmosphere~~ atmospheric composition where humans and the environment can be affected by air pollution. Air pollution results from the presence of chemical components emitted into the atmosphere due to anthropogenic activities and natural sources (biogenic emissions from vegetation, soil erosion, sea salts, volcanic activity, and wild-land fires). CTMs are used to represent the ~~dynamic~~ 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: this is crucial to quantify the ability of models to reproduce particular observed events or a general behaviour. ~~Depending on the model resolution and domain size, the comparison between model outputs and~~

~~observations data may be tricky due to~~ The quantification of the model quality is performed in every research study: depending on the studied case, the modelled variables, 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). ~~All modelling studies takes into account this problem of model representativeness and, for~~ From many years, ~~comparisons between observations and models outputs were performed using complex statistical approaches. A non~~ the way to evaluate a model results is discussed and in the field of atmospheric composition, numerous methods were proposed. This is not possible to give an exhaustive list of ~~validation studies are provided hereafter, all validation studies and we present here some examples.~~

Baldrige 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.

~~Dedicated tools to model evaluation have been developed such as and, 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 ~~scores indicators~~ to be used, Simon et al. (2012) use these ~~scores 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 unbiased simulation without changing the model, as in Porter et al. (2015) for ozone over the United States.

~~A fundamental difference between observations data and models results is the coherence of the spatial representativeness of the monitoring stations compared to the model cell. To quantify the model errors due to mis-representation of physics and chemistry from those only due to representativeness, several methodologies have been developed. These methods are effective but often required important computation time. Among these approaches, ensemble modelling is used in analysis of case studies and forecasting. By performing several perturbed simulations, a general tendency on the error can be identified. But if the case study consists of a complex real situation, the analysis can be challenging. Adjoint modelling allows tracking the behaviour of chemical species with respect to model input parameters. But it requires tedious model developments and the~~

result is generally valid for an infinitesimal perturbation since the problem to solve was linearized. In practice, the validity of this approach is limited to chemical species with a long lifetime as presented in . Finally, the 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 cell.

In the present study, a simple method is ~~developed to improve the evaluation of models and to identify the processes responsible for discrepancies of models outputs versus observations~~. In areas where the monitoring network are dense enough, like in Europe, comparisons are performed with observations from surface stations that provide hourly  $O_3$ ,  $NO_2$  concentrations for gases and  $PM_{2.5}$  and  $PM_{10}$  for particles. Complementary to surface concentrations data, the meteorology is evaluated using meteorological networks providing 2m temperature, 10m wind speed and precipitation rates. In order to quantify the transport of aerosols in dense plumes aloft, observations from lidar or from the AERONET (AERosol RObotic NETwork) program for the optical depth are increasingly used to assess regional models.

For all these variables, temporal and spatial correlations are computed to identify the model capacity compared to observations. First, the correlations are calculated between observations data and model outputs for the simulation year (i.e. the reference year). Second, the correlations are calculated between the observations data for other years and the model output for the reference year. Logically, the correlations calculated for the reference year for observations and model outputs would give the better results. By difference with the correlations 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 scores (correlations, bias, RMSE) calculated for a single year, the studied year. It is thus expected to give additional elements to answer these questions proposed to add information about the model performances more focussed on its 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? For a given variable, does the model have a good spatial representativeness compared to the corresponding observations?, and Are the biases introduced by meteorological or emissions variability or by the formulation of processes in the chemistry-transport model itself?

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 ~~The problem to solve~~Methodology

~~The problem to solve is presented in a general way by presenting the principle of chemistry-transport modelling. Then, the studied case and the models used are presented~~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 interest of our methodology.

## 2.1 Regional chemistry-transport modelling

5 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 know how to better interpret the model results. ~~These processes may be broken down into four categories: (i) boundary conditions, (ii) dynamics, (iii) emissions, and (iv) chemistry.~~

10 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 quantify their realism.

15 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 impact 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  
20 also a strong impact on all these natural emissions.

~~Anthropogenic emissions are prescribed from databases and the influence of meteorology is limited in the model. Vegetation, fires and mineral dust emissions also depend both on landuse data and meteorology variables. These emissions are difficult to measure, it is almost impossible to quantify their realism.~~

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

## 2.2 The studied case ~~and the models~~

30 The ~~case~~-study focuses on the summer 2013 period (1st May to 31 August) over the Euro-Mediterranean region, ~~this~~. 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,~~

Ozone (O<sub>3</sub>) and nitrogen dioxide (NO<sub>2</sub>) are the main pollutants targeted in this study. PM<sub>2.5</sub>, PM<sub>10</sub> are the surface concentrations of particulate matter with mean mass median diameter lower than 2.5 and 10 μm, respectively. Surface concentrations of pollutants are issued from the EBAS database, AOD and Angström are the Aerosol Optical Depth and the Angström exponent. T<sub>2m</sub> is the 2m temperature above ground, U<sub>10m</sub> the wind speed module at 10m above ground and "Precipitation" is the amount of precipitation in millimetres cumulated during a whole day. 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.

Variable	Network	Spatial coverage	Vertical coverage	Temporal frequency	Unit
O <sub>3</sub> , NO <sub>2</sub>	EBAS/EMEP	Europe	Surface	Hourly	ppb
PM <sub>2.5</sub> , PM <sub>10</sub>	EBAS/EMEP	Europe	Surface	Hourly	μg m <sup>-3</sup>
AOD, Angström	AERONET	Global	Column	Hourly	ad.
T <sub>2m</sub>	BADC	Global	Surface	Tri-hourly	°C
U <sub>10m</sub>	BADC	Global	Surface	Tri-hourly	m s <sup>-1</sup>
Precipitation	BADC	Global	Surface	<del>Tri-hourly</del> Daily	mm day <sup>-1</sup>

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 the present study.

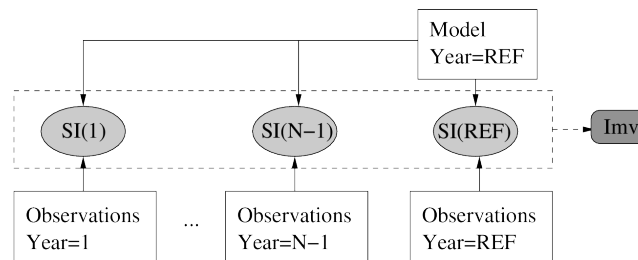
**Table 1.** List of measurements data used for the statistical comparison with the model results. All data used are issued from surface stations, representative of their own environment. Originally provided hourly or three-hourly, they are used as daily averaged in this work.

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<sub>3</sub>, NO<sub>2</sub> surface concentrations for gases and PM<sub>2.5</sub> and PM<sub>10</sub> (particulate matter with mean mass median diameter lower than 2.5 and 10 μm, 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<sub>2m</sub>, 10m wind speed, U<sub>10m</sub>, and precipitation rates (in mm day<sup>-1</sup>) 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 The proposed methodology

As discussed in the introduction, many ~~scores~~ Statistical Indicators (SI) exist to quantify the model ability to realistically simulate observed pollution events. The correlations ~~scores~~ (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 ~~key point of this study is the study of model variability which is statistically represented by the correlations. The mean bias (or the normalized bias) is not a score able to quantify the variability. And the RMSE is a score containing a part of variability but remains driven by the~~ RMSE and the bias are a direct quantification of the model error.

The-



**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 (i. e. when the model seems good, but simulate the same thing every day and every year) and due to sporadic events ((i. e. when the model is good because and able to retrieve the day to day variability). This is why the proposed methodology is based on the calculation of the temporal and spatial correlations only 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).

The methodology follows three steps: (i) compute the correlation scores (spatial and temporal) between the measurements. 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 for the whole reference period, (ii) recalculate these scores between the modelled reference period and the observed data for the similar period in 2008, 2009, 2010, 2011 and 2012, (iii) build and use a synthetic score to quantify 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 has high scores for good reasons or not. This is summarized in [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.

~~Principle of the multi-year variability score's calculation, using one modelled year and several observations years.~~

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 ~~the correlation scores~~ [correlations and nRMSE](#)

~~To compute the correlation coefficients,~~ 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.

The  $O_{t,i}$  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  $\bar{X}_i$  is:

$$\bar{X}_i = \frac{1}{T} \sum_{t=1}^T X_{t,i} \quad (1)$$

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

$$R_{t,i} = \frac{\sum_{t=1}^T (M_{t,i} - \bar{M}_i)(O_{t,i} - \bar{O}_i)}{\sqrt{\sum_{t=1}^T (M_{t,i} - \bar{M}_i)^2 \sum_{t=1}^T (O_{t,i} - \bar{O}_i)^2}} \quad (2)$$

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

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



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:

$$\bar{X} = \frac{1}{I} \sum_{i=1}^I \bar{X}_i \quad (4)$$

and the spatial correlation is thus expressed as:

$$R_s = \frac{\sum_{i=1}^I (\bar{M}_i - \bar{M})(\bar{O}_i - \bar{O})}{\sqrt{\sum_{i=1}^I (\bar{M}_i - \bar{M})^2 \sum_{i=1}^I (\bar{O}_i - \bar{O})^2}} \quad (5)$$

~~For the correlations, obviously better scores are expected for the reference year compared to the other. This would confirm that during the transport of pollutants, the model is able to correctly model the day to day variability. The normalized Root Mean Square Error is expressed as:~~

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

15 ~~for all stations  $i$  and all times  $t$ .~~

### 3.2 ~~The multi-year variability $I_{mv}$ indicator~~

#### 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 ~~output-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:

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

with  $s_N$  the score ~~for the actual 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 ~~that would follow these rules:-~~

~~The indicator increases with the correlation: More the correlation is high, better the model is. The indicator increases with the differences  $D$ : more the differences are important more the studied year was different from the others, more the system has a variability. The indicator is moderated if the differences  $D$  are low. For example, we want that a correlation of 0.8 has not the same meaning if  $D=0$  or  $D=1$ : the indicator has to give a higher value for  $(R=1, D=1)$  than for  $(R=1, D=0)$ .~~

~~We can thus estimate a "Multi-Year Variability" indicator, noted called  $I_{mv}$  as:-, 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:~~

$$I_{mv} = s_N (1 - \exp(-D_s)^\delta)$$

~~The value for  $\delta$  is arbitrary but it should be larger than unity, in order to have an indicator-~~

- ~~-  $I_{mv}$  has the same evolution than the studied SI. If the correlation increases,  $I_{mv}$  also increases.~~
- ~~-  $I_{mv}$  is bounded between 0 and 1. This tuning parameter enables to adapt the relative weight we want to attribute to the absolute value of the scores for the selected year and 1 such as 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 all years. In general, we want that a good score the other years are low ( $D$  tends to 0), it means that the model is correct for the studied year have a larger weight than the differences between several years. Using  $\delta=4$ , we consider that the relative weight of the correlation value against the difference reflects well the fact that the model has correct scores and variability.~~~~

~~Scheme of the  $I_{mv}$  score as a function of the studied year correlation and the multi-years differences:~~

~~The behaviour of, but possibly because it reproduces a recurrent phenomena. In this case, we want that  $I_{mv}$  is plotted on for values of the scores and the differences ranging from 0 to 1. Ideally we hope that the model performs well for the correlation scores but also be able to reproduce the observed variability. When  $I_{mv}$  decreases and tends to 0.~~

- ~~- If the differences between the other years are high ( $D$  tends to 1 this means that the correlation value is close to 1): in this case, the model gives good results for the studied year and this 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 the differences of the modelled studied year compared to the other years are also close to 1. In reality, this ideal situation is rarely~~

obtained since we are modelling a very complex atmospheric system, based on processes with different variabilities and uncertainties. Moreover, if the correlation is close to zero and  $I_{mv} \approx 1$ , we can conclude that the model is very good for the studied year and this is not due to an easy phenomenon to model.

- In case of low correlation value, and whatever the magnitude of differences between years, the model is definitely poor. Finally, if the difference is close to zero, one can conclude that model performances are independent of the selected year: in that case, not correct.  $I_{mv}$  is also close to 0. must be low, as the indicator value.

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

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

§

This means that  $I_{mv}$  has always, as maximum value, the value of the indicator  $I_{mv}$  is to provide complementary information than the correlation and the differences separately analysed. This indicator has thus to be viewed as complementary to the correlation score and not replacing it. From a subjective point of view, considering 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 and from, we consider that, the model is accurate and has considered as accurate and to have an acceptable variability for  $I_{mv} > 0.304$ ; this means that the correlation is at least 0.5 and the differences are also at least greater than 0.5. Of course, this value may change if the  $\delta$  value is different.

### 3.3 Detailed examples of $I_{mv}$ calculation

To better understand the relevance of  $I_{mv}$ , two examples are detailed in this section. The scores are calculated for 2m temperature,  $T_{2m}$ . 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 of the studied year compared to the other years. In this case, the correlations represent statistical indicators more adapted 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 surface concentration of nitrogen dioxide,  $\text{NO}_2$  years 2008 to 2013. For each year, it is reminded that only the May to August period is considered. Results are presented in 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.

These two variables are presented here because they represent very different variables in a CTM simulation:-

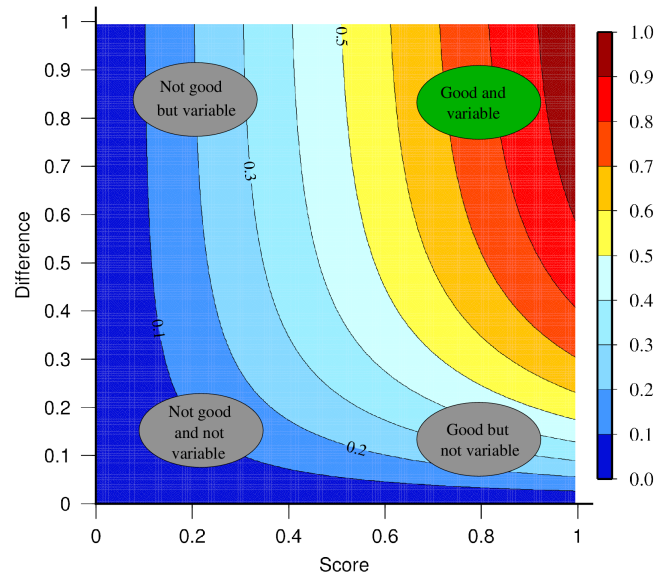


Figure 2. Scheme of the  $I_{mv}$  values as a function of the studied year correlations values and the multi-years differences  $D$ .

#### 4.1 Meteorological variables

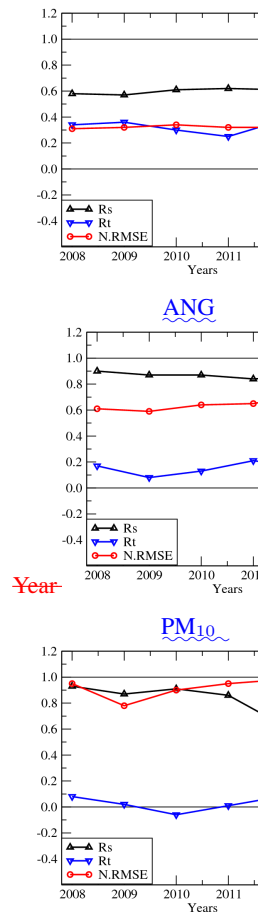
The meteorological variables are  $T_{2m}$ ,  $u_{10m}$  and the precipitation rate. The values of the Statistical scores are provided, year by year, in Figure 3. As an example, the same values are reported for  $T_{2m}$  in Table 2.

Year	$R_s$	$R_t$	$nRMSE$
2008	0.58	<del>0.36</del> <u>0.34</u>	<del>2008-0.44</del> <u>0.00</u> <u>0.31</u>
2009	0.57	<del>0.38</del> <u>0.36</u>	<del>2009-0.42</del> <u>-0.04</u> <u>0.32</u>
2010	<del>0.60</del> <u>0.61</u>	0.30	<del>2010-0.66</del> <u>-0.04</u> <u>0.34</u>
2011	0.62	<del>0.26</del> <u>0.25</u>	<del>2011-0.79</del> <u>-0.03</u> <u>0.32</u>
2012	0.61	<del>0.40</del> <u>0.37</u>	<del>2012-0.76</del> <u>0.04</u> <u>0.32</u>
2013	<del>0.61</del> <u>0.60</u>	<del>0.94</del> <u>0.91</u>	<del>2013-0.88</del> <u>0.22</u>
$\overline{D}$	0.02	<del>0.60</del> <u>0.59</u>	<del>0.27</del> <u>0.23</u> <u>0.10</u>
$I_{mv}$ <del>0.04</del> <u>0.85</u> $I_{mv}$ <del>0.58</del> <u>0.13</u>			

Table 2. Scores for  $T_{2m}$  and  $NO_2$ . The correlations and  $nRMSE$  are calculated between the observations (2008-2013) and the model results (2013).

##### 4.1.1 Analysis of $T_{2m}$ scores

$T_{2m}$  is a meteorological variable, constraining processes both for meteorology and chemistry. Its diurnal cycle is well marked as its latitudinal variability



**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.

$T_{2m}$  is a meteorological variable, constraining processes both for meteorology and chemistry. Its diurnal cycle is well marked 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 0.61±0.06, 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  $D=0.02$ , and this yields to  $I_{mv}(R_s)=0.04$ . 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 has a latitudinal variability larger than between each measurements stations. This The temporal

correlation ranges from 0.26 to 0.94. And the best score is for 0.25 to 0.91 (2013) leading to a good score of  $I_{mv}(R_t)=0.85$ . The model is thus performing well in capturing the day to day variability for T2m and for the good reasons.

#### 5 4.1.1 Analysis of NO<sub>2</sub> scores

Nitrogen dioxide is both a primary and secondary species quickly produced by oxidation of NO and the scores show if the sources are properly placed and if the photochemistry and transport processes have been well simulated. In general, at coarse model resolution, the scores for this species are worse than for ozone. NO<sub>2</sub> is very dependent on the quality of emission inventories, however the measurements stations considered in this study are background sites.

10 The spatial correlation gives a score of  $R_s=0.88$ ). The nRMSE is less evolving than the correlations, with values ranging from 0.22 (2013) to 0.34 (2010). The lowest value is found for 2013. Being the best comparison, we obtain  $I_{mv}(R_s)=0.58$ . This shows the importance of NO<sub>x</sub> emission source location that is the main driver of spatial performances. The temporal correlation is low for 2013,  $R_t=0.22$ , but is close to 0 for other years. In the end, we have a low score with  $I_{mv}(R_t)=0.13$  even if the simulated year is better. These two scores show highlighting the fact that the model certainly captures the right location  
 15 of emission sources (low variability of  $R_s$ ). For the temporal variability, the model is not able to reproduce the day to day variability, but it remains significantly better error is the lowest for the reference year compare to the others.

### 5 Results and discussion

The correlations are calculated for all variables described in and for the years 2008 to 2013, it is reminded that only the May to August 2013 period was modelled. Results are presented as time series in . Using all correlations and differences values, a  
 20  $I_{mv}$  is estimated for each variable. Results ( ) are discussed in the following sections.

The model is thus performing well in capturing the day to day variability for T<sub>2m</sub>, u<sub>10m</sub>, Precipitation AOD ANG O<sub>3</sub> NO<sub>2</sub> PM<sub>2.5</sub> PM<sub>10</sub> Ammonium Sulphate Nitrate Multi years scores for T<sub>2m</sub>, u<sub>10m</sub>, the precipitation rate, Aerosol Optical Depth (AOD), Angström exponent (ANG), surface concentrations of O<sub>3</sub>, NO<sub>2</sub>, PM<sub>2.5</sub>, PM<sub>10</sub>, Ammonium, Sulphate and Nitrate. The correlations are calculated between the observations (2008-2013) and the model results (2013). The spatial correlation,  $R_s$ , is  
 25 in black and the temporal correlation,  $R_t$  is in red. and for the good reasons.

Variable Value  $D I_{mv}$  Value  $D I_{mv}$  T<sub>2m</sub> 0.61 0.02 0.04 0.94 0.60 0.85 u<sub>10m</sub> 0.09 0.09 0.03 0.60 0.56 0.54 precip 0.78 0.29  
 0.54 0.30 0.31 0.21 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<sub>3</sub> 0.69 0.13 0.29 0.32 0.27 0.21  
 NO<sub>2</sub> 0.88 0.27 0.58 0.22 0.23 0.13 PM<sub>2.5</sub> 0.16 0.15 0.07 0.27 0.32 0.20 PM<sub>10</sub> 0.57 0.43 0.47 0.11 0.10 0.04 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 The  $I_{mv}$  values for all  
 30 variables: the meteorology with T<sub>2m</sub>, u<sub>10m</sub> 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<sub>2.5</sub> and PM<sub>10</sub> and for the inorganic species with  $D_p < 10 \mu m$ . Values of  $I_{mv}$  above 0.3 are bolded. Units of the variables are detailed in .

## 4.1 Meteorological variables

Scores for  $T_{2m}$  were discussed in the previous section. The From Figure 3, the calculation of  $u_{10m}$  also gives satisfactory results with  $R_t=0.60$  and  $I_{mv}=0.54$ . The spatial correlation,  $R_s=0.09$ , is not correct and very variable from one year to another, leading to  $I_{mv}=0.03$ . As for  $T_{2m}$ , we also have an effect of the model resolution and the representativeness of the variable.

Scores for the precipitation are correct, with a very good spatial correlation leading to  $I_{mv}(R_s)=0.54$ . For the -, 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, the score is less good with  $I_{mv}(R_t)=0.21$  but significantly higher for 2013. These is correct and this is what is requested in 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 retrieved, and always better for the year being considered compared to other years.

## 4.1 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:  $I_{mv}(R_s)=0.09$ . The temporal correlation is lower with  $R_t=0.45$  but much better than for other years:  $I_{mv}(R_t)=0.33$ . This means. This indicates that the model partly reproduced-reproduces the observed temporal variability but the events are changing from one year to another and the model captures well these changes. The In the studied region, the AOD are sensitive to desert dust outbreaks in summer in that region. This means that large scale systems are driving the aerosol plumes; they are spatially recurrent and temporally better estimated 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 leading to a low score of  $I_{mv}(R_s)=0.14$  in time. The temporal correlation is much better for 2013 than other years with  $I_{mv}(R_t)=0.49$ . 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 could be 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 the model is able to retrieve these yearly recurrent plumes but the model size distribution of particles clearly requires improvements.

## 4.2 Surface concentrations

~~The spatial correlation is~~ 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 fully presented for NO<sub>2</sub>.

5

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<sub>2</sub>. The correlations and nRMSE are calculated between the observations (2008-2013) and the model results (2013).

NO<sub>2</sub> 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 placed 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 being strongly linked to industrial activities and traffic, and these activity sectors being spatially fixed, the good spatial correlation is more due to anthropogenic sources such as biogenic and vegetations fires. The temporal correlation is low for 2013,  $R_t=0.22$ , but is close to 0 for other years, then significantly better for the reference year compare 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 surface of model cells.

10

15

The spatial correlation is good for O<sub>3</sub>, NO<sub>2</sub> and PM<sub>10</sub>, with  $R_s=0.69$ , 0.88 and ~~0.57-0.81~~ respectively. For PM<sub>2.5</sub> this correlation is low with  $R_s=0.16$ . The PM<sub>10</sub> 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<sub>2.5</sub> indicates that for the aerosol distribution, the fine mode is ~~less well modelled than~~ 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.

20



The temporal correlations have a completely different behaviour than 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: large plumes are correctly modelled over regions but the day to day variability needs improvements. Another point is the good spatial correlation for  $NO_2$  (and for the good reasons with  $I_{mv}=0.58$ ) but its low temporal correlation with  $R_t=0.22$  and a low  $I_{mv}=0.13$ . In this case, this means we have a correctly localized-spatialized anthropogenic emissions inventory (main source of mainly for  $NO_2$  sources) but difficulties to model the day to day chemistry.

In conclusion, 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,  $I_{mv}$  having high values. 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.

#### 4.3 Representation of results on a single plot

##### 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 value 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 more temporally than spatially.

Complementary to the, in Figure 4 reports the results on a single plot. The x-axis represents the correlation (spatial or temporal), the y-axis represents the differences between all years,  $D$ . For each studied variable, their values are reported on the Figure figure where the colours represent the value of  $I_{mv}$ . The interpretation of these results follows the quality criteria presented in the academic scheme in 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 badly 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	<b>0.82</b>
$u_{10m}$	0.09	0.23	0.05	0.59	0.56	<b>0.53</b>
precip	0.89	0.20	<b>0.49</b>	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	<b>0.49</b>
$O_3$	0.69	0.13	0.29	0.32	0.27	0.21
$NO_2$	0.88	0.27	<b>0.58</b>	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  $D_p < 10 \mu m$ . Values of  $I_{mv}$  above 0.3 are bolded. Units of the variables are detailed in Table 1.*

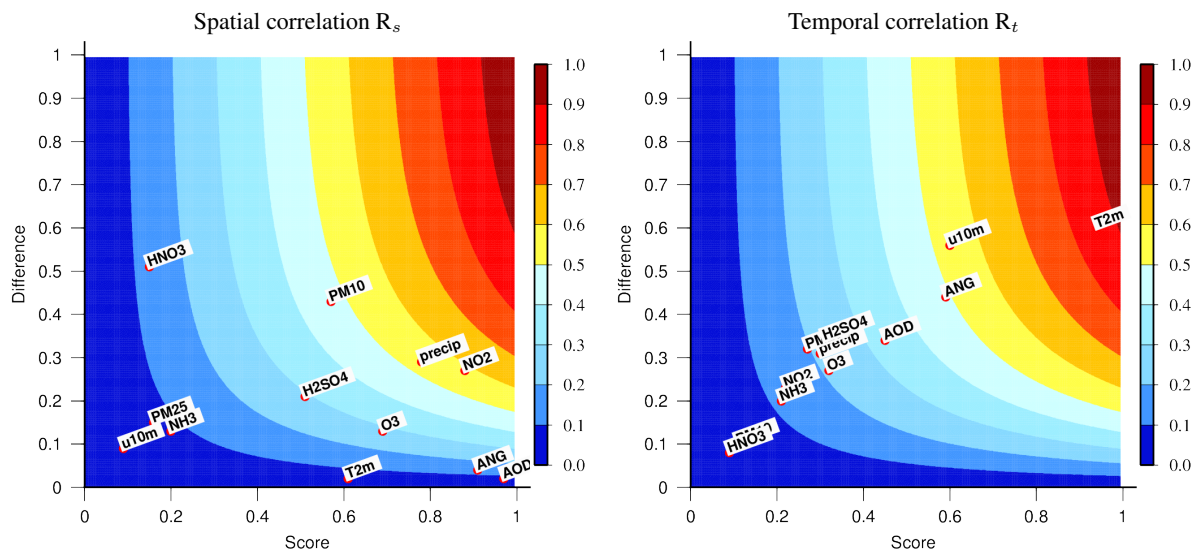
correlations but low differences). This illustrates the fact that the studied problem is primarily an issue of sporadic events and the model is able to correctly find this variability from one day to another.

## 5 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 **statistical indicator** 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  $R_s=0.96-0.97$  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 largest particles are better simulated than the finest 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, on 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 significantly able to catch the evolution of the 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.](#)

## **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.
- Baldrige, 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 O<sub>3</sub> and PM<sub>2.5</sub> 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 I, 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.
- 5 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, 10 *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, 15 *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., 20 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.