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.

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: [Baldridge and Cox, 1986], [Cox and Tikvart, 1990], [Chang and Hanna, 2004] [Appel et al., 2011], [**?**], [Galmarini et al., 2012], [Vautard et al., 2012], [Bennett et al., 2013], [Schaap et al., 2015], [Campbell et al., 2015], [Bessagnet et al., 2016].

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.

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

Best regards,
Laurent MENUT
December 19, 2016

Message from the Editor

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, l4: "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, l8: "The spatial correlation, noted Rs, 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, l14: "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 minus 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 that 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.

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:

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

where α and β 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 α=0, β=-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 α=1, β=1 and we impose to have MYV_s=0 when negative values are estimated.

The value for δ 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 that the differences: in this case, we select δ=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 1 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.

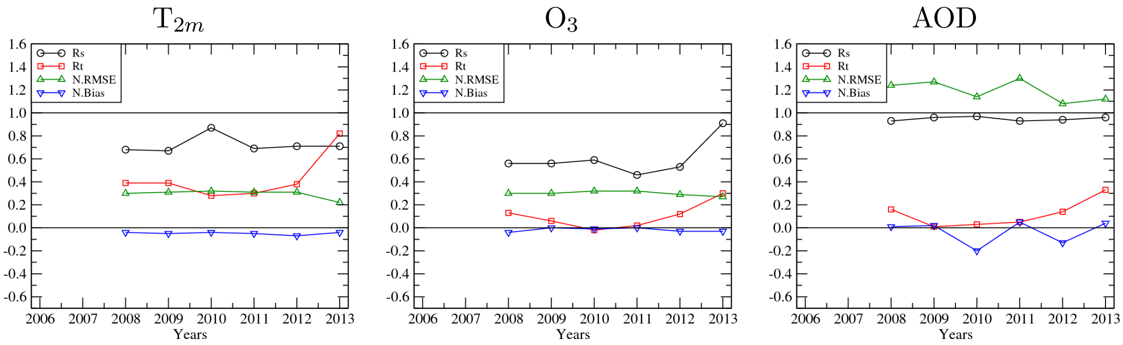

Figure 1: *Multi years scores for the 2m temperature, ozone and AOD. The reference year is 2013.*

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 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 δ=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 δ 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 [Valari and Menut, 2008, Solazzo and Galmarini, 2015]. 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, l9: 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, l10: 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, l9-15: this all paragraph is unclear and should be rewritten

    This was rewritten.

38. P6, l18-19: 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, l1: from Figure 3

    Ok, corrected.

41. P7, l1: we can consider that

    Ok, corrected.

42. P7, l1-2: 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 δ 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, l6: done → calculated

    OK corrected.

44. P7, l6: 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, l12: vary a lot → vary significantly

    This is P7L13 and this was corrected.

46. P7, l13: 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_2 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, photochemistry... 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.

References

[Appel et al., 2011] Appel, K. W., Gilliam, R. C., Davis, N., Zubrow, A., and Howard, S. C. (2011). Overview of the atmospheric model evaluation tool (AMET) v1.1 for evaluating meteorological and air quality models. *Environmental Modelling and Software*, 26(4):434 – 443.

[Baldridge and Cox, 1986] Baldridge, K. and Cox, W. (1986). Evaluating air quality model performance. *Environmental Software*, 1(3):182 – 187.

[Bennett et al., 2013] Bennett, N. D., Croke, B. F., Guariso, G., Guillaume, J. H., Hamilton, S. H., Jakeman, A. J., Marsili-Libelli, S., Newham, L. T., Norton, J. P., Perrin, C., Pierce, S. A., Robson, B., Seppelt, R., Voinov, A. A., Fath, B. D., and Andreassian, V. (2013). Characterising performance of environmental models. *Environmental Modelling and Software*, 40:1 – 20.

[Bessagnet et al., 2016] 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. (2016). 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(19):12667–12701.

[Campbell et al., 2015] 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. (2015). A multi-model assessment for the 2006 and 2010 simulations under the Air Quality Model Evaluation International Initiative (AQMEII) phase 2 over North America: Part I. Indicators of the sensitivity of O3 and PM2.5 formation regimes. *Atmospheric Environment*, 115:569 – 586.

[Chang and Hanna, 2004] Chang, J. and Hanna, S. (2004). Air quality model performance evaluation. *Meteorology and Atmospheric Physics*, 87:167–196.

[Cox and Tikvart, 1990] Cox, W. M. and Tikvart, J. A. (1990). A statistical procedure for determining the best performing air quality simulation model. *Atmospheric Environment. Part A. General Topics*, 24(9):2387 – 2395.

[Galmarini et al., 2012] Galmarini, S., Bianconi, R., Appel, W., Solazzo, E., Mosca, S., Grossi, P., Moran, M., Schere, K., and Rao, S. (2012). {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. AQMEII: An International Initiative for the Evaluation of Regional-Scale Air Quality Models - Phase 1.

[Schaap et al., 2015] Schaap, M., Cuvelier, C., Hendriks, C., Bessagnet, B., Baldasano, J., Colette, A., Thunis, P., Karam, D., Fagerli, H., Graff, A., Kranenburg, R., Nyiri, A., Pay, M., Roul, L., Schulz, M., Simpson, D., Stern, R., Terrenoire, E., and Wind, P. (2015). Performance of european chemistry transport models as function of horizontal resolution. *Atmospheric Environment*, 112:90 – 105.

[Solazzo and Galmarini, 2015] Solazzo, E. and Galmarini, S. (2015). Comparing apples with apples: Using spatially distributed time series of monitoring data for model evaluation. *Atmospheric Environment*, 112:234 – 245.

[Valari and Menut, 2008] Valari, M. and Menut, L. (2008). Does increase in air quality models resolution bring surface ozone concentrations closer to reality? *Journal of Atmospheric and Oceanic Technology*.

[Vautard et al., 2012] Vautard, R., Moran, M., Solazzo, E., Gilliam, R., Matthias, V., Bianconi, R., Chemel, C., Ferreira, J., Geyer, B., Hansen, A., Jericevic, A., Prank, M., Segers, A., Silver, J., Werhahn, J., Wolke, R., Rao, S., and Galmarini, S. (2012). Evaluation of the meteorological forcing used for the Air Quality Model Evaluation International Initiative (AQMEII) air quality simulations. *Atmospheric Environment*, 53:15–37.

An unusual way to validate regional chemistry-transport models

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

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

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

Abstract. A simple and ~~exhaustive~~ complementary model evaluation technique for regional chemistry-transport is discussed. ~~It~~ The methodology is based on the concept that we can learn more on models performances by comparing the ~~results to in situ measurements~~ simulation results with observational data available for other time periods than the period originally targeted~~in the simulation~~. First, the usual scores selected in this study (spatial and temporal ~~correlation~~correlations) are computed for a ~~reference~~ given period, using ~~the actual temporal synchronization and spatial location of measurements~~ co-localised observation and simulation data in time and space. Second, the same scores are calculated for several other years by conserving only the ~~actual~~ spatial locations and Julian days of the year. The difference between the two score provides complementary insights to the following questions: (i) is the model performing well only because the situation is ~~persistent~~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? In order to ~~synthetize~~ synthesise the large amount of results, a new ~~score~~ indicator is proposed: the "multi-year variability", designed to compare the several ~~indicators~~ error statistics between all the years of validation and to quantify if the studied period was ~~well modelled and, if yes,~~ 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 where humans and the environment can be affected ~~when air quality is poor~~by air pollution. Air pollution results from the presence of chemical components emitted into the ~~air~~ atmosphere due to anthropogenic activities and ~~by~~ natural sources (biogenic emissions from vegetation, soil erosion, sea salts, volcanic activity, and wild-land fires). CTMs are used to represent the dynamic 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. ~~In areas where the monitoring network are dense enough, such as in Europe, comparisons can be made with observations from surface stations that provide hourly concentrations of O_3, NO_2 for gas and $PM_{2.5}$ and PM_{10} for aerosols. In order to quantify transport of aerosols in dense plumes aloft, observations from lidar or the AERONET network (to have the optical depth) are increasingly used with regional models .~~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 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 many years, comparisons between observations and models outputs were performed using complex statistical approaches. A non exhaustive list of validation studies are provided hereafter, Baldridge and Cox (1986) and Cox and Tikvart (1990) proposed the use of error statistics like correlation, bias, Root Mean Squared Error 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.

~~But there can me 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,~~ Dedicated tools to model evaluation have been developed such as Appel et al. (2011) and ~~it is challenging to easily identify why a modelwould exhibit an inappropriate behavior~~Galmarini et al. (2012), to ensure the use of systematic procedures in the evaluation process. 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 to be used, Simon et al. (2012) use these scores to compile photochemical models performances over a large set of data over several years of simulation. This kind of evaluation may also be done in dedicated projects such as the recent AQMEII (Air Quality Model Evaluation International Initiative), comparing chemistry-transport models running both in Europe and Northern America, Vautard et al. (2012); Campbell et al. (2015) or the EURODELTA project, Bessagnet et al. (2016) and in the EMEP (European Monitoring and Evaluation Programme) context in the frame of the United Nation Convention on Long-range Transboundary Air Pollution, Prank et al. (2016). Using comparisons between observations and models outputs, some studies proposed methodologies to decompose the statistical scores in order to estimate the main source of errors, Solazzo and Galmarini (2016). Finally, other studies also use observations to adjust the result by implementing methods to unbias simulation without changing the model, as in Porter et al. (2015) for ozone over the United States.

A fundamental difference between ~~models and observations is the spatial representativeness~~ observations data and models results is the coherence of the spatial representativeness of the monitoring stations compared to the model cell (Valari and Menut, 2008; Solazzo and Galmarini, 2015). To ~~isolate problems intrinsic to the models~~quantify the model errors due to mis-representation of physics and chemistry from those only due to representativeness, several methodologies have been developed~~to extract the relevant information in the simulations, particularly to identify what could be the processes most responsible for model discrepancies~~. These methods are effective but often ~~with huge~~ required important computation time. Among these approaches, ensemble ~~modeling~~ modelling is used in analysis of case studies and forecasting, (Kioutsioukis and Galmarini, 2014; Marécal et al., 2015; Lemaire et al., 2016). By performing several perturbed simulations, ~~one can identify if there is~~ 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 ~~modeling~~ modelling allows tracking the ~~behavior~~ 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, (Menut, 2003; Pison et al., 2007). In practice, the validity of this approach is limited to chemical species with a long lifetime as presented in Kopacz et al. (2010); Mao et al. (2015).

~~In chemistry transport modeling, emissions are well known to constitute one of the most uncertain forcings. There are therefore studies devoted to scenario simulations in order to quantify the relative weight of each pollutant emitted in the final calculated concentration budget,~~ Finally, the common point of all these studies is that they are always using the observations corresponding in time and location to the model cell.

~~More recently, proposed to decompose statistical scores to better understand the errors in surface ozone modeling. Finally, other studies also use observations to adjust the result by implementing methods to unbias simulation without changing the model , as in for ozone on the United States.~~

In the present study, ~~we try to provide~~ a simple method is developed to improve the ~~validation of a simulation~~ evaluation of models and to identify the processes responsible for ~~the differences between the model and the available observations . For this, we compute several correlation scores,~~ 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 ~~accuracy~~ capacity compared to observations. ~~Afterwards, we apply these scores to a model simulation and several different observations datasets.The originality of the approach presented here is that we do not compare the simulation of a case study only to the~~

~~corresponding observational dataset (in time and space) but we use all available data of the other years . The new dataset of scores will highlight the differences between specific and systematic errors.~~

~~Therefore, we want to elaborate scores that provides answer to the subsequent~~ 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 ~~persistent~~ recurrent from year to year? For a given variable, does the model have a good spatial representativeness compared to ~~available observationsfor a given variable~~the corresponding observations?*, and *Are the biases introduced by meteorological or emissions variability or by the ~~processes parameterized~~ formulation of processes in the chemistry-transport model itself?*

The issue to ~~solve~~ be solved and the tools developed are presented in section 2. The new methodology with the presentation of the ~~score~~ 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.

2 The problem to solve

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

2.1 Regional chemistry-transport ~~modeling~~modelling

~~Figure ?? presents the several forcing and processes involved in a typical~~ In chemistry-transport ~~model (CTM). The objective of this simple figure is to remind the dependencies between each "geophysical compartment" involved in such modeling tools.~~

~~The several processes taken into account in a regional chemistry-transport model.~~

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

The boundary conditions prescribe the ~~concentration in~~ concentrations of chemical species which may enter the ~~modeled area during the simulation~~ 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.

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 ~~a strong~~ also a major impact on wildfires emissions. Both temperature and solar irradiance influence the magnitude of biogenic emissions from vegetation. The spatial variability of landuse data has also a strong impact on all these natural emissions.

Anthropogenic emissions are prescribed from databases and the influence of meteorology is limited in the model. ~~On the other hand, biogenic,~~ Vegetation, fires and mineral dust emissions also depend both on landuse ~~and meteorology~~data and meteorology variables. These emissions are difficult to measure~~; this is not possible~~ , it is almost impossible to quantify their realism.

The chemistry-transport model is a numerical integration tool of all the forcings and processes. The ~~chemistry mechanism prescribes the amount of the chemical species~~ chemical mechanism handles the chemical species life cycle (production and loss) when the deposition ~~is~~ processes are the only 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 ~~modeled~~ modelled air pollutant concentrations when they are compared to ~~the~~ available observations.

2.2 The studied case and the models

~~We focus on a case study for the summer of~~ The case study focuses on the summer 2013 period (1st May to 31 August) ~~in~~ over the Euro-Mediterranean region, this period is called "reference period" in this paper. This case has already been ~~modeled~~ modelled (using 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 ~~variable and they are presented in Table 1. Originally provided hourly or three-hourly, they are~~ variables, Table 1.

Ozone (O_3) and nitrogen dioxide (NO_2) are the main pollutants targeted in this study. $PM_{2.5}$, PM_{10} are the surface concentrations of particulate matter with mean mass median diameter lower than 2.5 and 10μm, respectively. Surface concentrations of pollutants are issued from the EBAS database, (Tørseth et al., 2012). AOD and Angström are the Aerosol Optical Depth and the Angström exponent. T_{2m} is the 2m temperature above ground, U_{10m} 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 ~~averaged in the present study.~~ 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

~~The proposed methodology~~ As discussed in the introduction, many scores exist to quantify the model ability to realistically simulate observed pollution events. The correlations scores (temporal and spatial), the Root Mean Squared Error (RMSE) 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 bias.

The goal of this study is to separate the contributions due to systematic 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.

The methodology follows three steps: (i) compute the correlation scores (spatial and temporal) between the measurements and the model ~~and during~~ for the whole reference period, (ii) recalculate these scores between the ~~modeled~~ 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 if the model ~~had~~ has high scores for good reasons or not. This is summarized in Figure 1.

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 ~~day for another~~ date for another year is strictly the same that to choose randomly a ~~day~~ date in the same season. This trivial method can emphasize how a model is affected by large scale patterns and long term temporal cycles. ~~The correlation is the more appropriate statistical metric for such analysis.~~

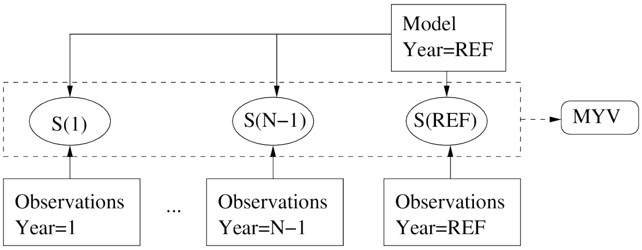

Figure 1. *Principle of the multi-year variability score's calculation, using one modelled year and several observations years.*

3.1 Calculation of ~~usual~~ the correlation scores

To compute the correlation ~~coefficient~~coefficients, 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 $\overline{X_i}$ is:

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

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

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

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

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

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

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

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 ~~plumes during~~ long range transport ~~. For processes leading to large plumes(dust, fires, volcanoes), this indicator indicates that the model is using realistic emissions and is able to reproduce a correct transport. For all the studied parameters, it is also an indicator that the resolution of the model is adapted to the variable considered.~~ plumes.

The ~~temporal correlation, noted R_t, is estimated station by station. This indicator is directly related to the variability from day to day, for each station. The longer the atmospheric lifetime of the species, the lower the relevance of temporal correlation .~~ spatio-temporal mean averaged value is estimated as:

$$\overline{\overline{X}} = \frac{1}{I}\sum_{i=1}^{I} \overline{X_i} \qquad (4)$$

and the spatial correlation is thus expressed as:

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

(

(5)

For the correlations, obviously better scores are expected for the reference year compared to the other~~, particularly for the temporal correlation~~. This would confirm that during the transport of pollutants, the model is able to correctly model the day to day variability.

3.2 The multi-year variability ~~'MYV' score~~I_{mv} indicator

~~We aim to develop a simple indicator that would increase with correlation but would be moderated if the differences with other yearsare low. We thus first estimate~~ The goal of this indicator is to quantify how the correlation between measurements data (for different years) and model output (for the reference year) evolves from a year to another one. We first define the differences, D, between all years as:

$$D = \frac{1}{N-1}\left(\sum_{i=1}^{N-1} |s_i - s_N|\right) \qquad (6)$$

with s_N the score for the actual year being modelled and s_i the score 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:

1. The indicator increases with the correlation: More the correlation is high, better the model is.
2. 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.
3. 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" ~~, noted MYV~~ indicator, noted I_{mv} as:

$$\underline{MYV}I_{mv} = s_N \underline{\times} \left(1 - exp(-D_s)^{\delta}\right) \quad (7)$$

The value for δ is arbitrary but it should be larger than unity~~.~~, in order to have an indicator I_{mv} 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 the differences between all years. In general, we want that a good score for the studied year have a ~~largest weight that the differences : in the present case , we choose to select~~ larger weight than the differences between several years. Using δ=~~4.~~ 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.

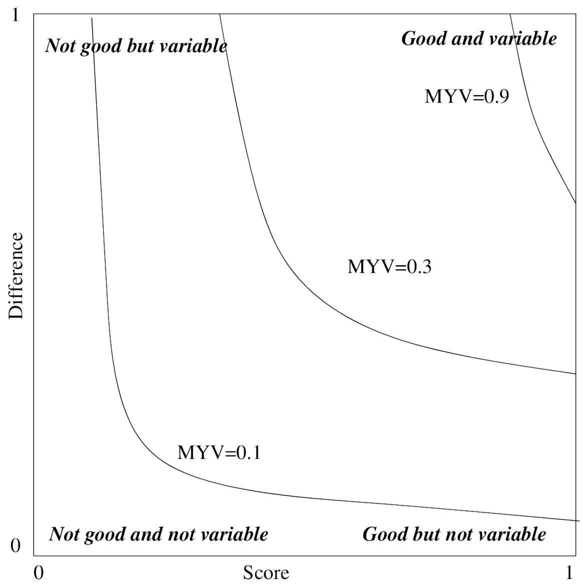

Figure 2. *Scheme of the ~~MY~~ I_{mv} score as a function of the studied year correlation and the multi-years differences.*

The ~~behavior of MYV~~ behaviour of I_{mv} is plotted on Figure 2 for ~~academic~~ values of the scores and the differences ~~. Ideally we would~~ ranging from 0 to 1. Ideally we hope that the model performs well for the ~~absolute~~ correlation scores but also ~~variable i.e.~~ be able to reproduce ~~very different situations. When MYV~~ the observed variability. When I_{mv} tends to 1 this means that the correlation value is close to 1 and the differences of the ~~modeled~~ modelled studied year compared to the other years are also close to 1. In reality, this ideal situation is rarely obtained since we are ~~modeling~~ modelling a very complex atmospheric system, based on processes with different variabilities and uncertainties. ~~On the other hand~~Moreover, if the correlation is close to zero, the model is definitely poor. ~~And~~ Finally, if the difference is ~~also~~ close to zero, ~~we~~ one can conclude that ~~the low performance of the model is independent from~~ model performances are independent of the selected year: in that case, ~~MYV~~ I_{mv} is also close to 0.

~~It is also possible to have a good score but a low variability: in that case, MYV is also close to zero. This is because the MYV score has~~ The role 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 ~~score (~~correlation score and not replacing ~~the score). Thus, if we already know that the score is good, we can rely on MYV to ensure that the score was also good for other years.~~

it. From a subjective point of view, ~~and from the~~ considering the state-of-the art of chemistry-transport modelling and from Figure 2, we ~~can consider~~ consider that the model is accurate and has an acceptable variability for ~~MYV >~~ $I_{mv} >$ 0.3: this means that the correlation is at least 0.5 and the differences are also at least greater than 0.5. Of course, this ~~directly depends of~~ value may change if the δ ~~parameter value~~ value is different.

3.3 Detailed examples of ~~MYV~~ I_{mv} calculation

To better understand the relevance of ~~the score MYV~~I_{mv}, two examples are detailed in this section. The scores are ~~done~~ calculated for 2m temperature, T_{2m}, and for the surface concentration of nitrogen dioxide, NO_2. Results are presented in Table 2.

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

- 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), ensuring a good spatial correlation. In general, it is the less uncertain of modelled meteorological variables.
- NO_2 is both a primary and secondary species. Mostly emitted in urbanized areas, the diurnal cycle of this species is well constrained. Depending on meteorological conditions, its lifetime may vary ~~a lot~~significantly, from hours to days. Modelling this species with CTMs is challenging because several uncertainties are acting at the same time, including the spatial ~~representativity~~ representativeness of the model cell.

3.3.1 Analysis of T_{2m} scores

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, ~~thus~~ 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 ~~MYV(R_s)~~$I_{mv}(R_s)$=0.04. This means that the model reproduces fairly well a spatial pattern that is observed every year.

T_{2m}			NO_2		
Year	R_s	R_t	Year	R_s	R_t
2008	0.58	0.36	2008	0.44	0.00
2009	0.57	0.38	2009	0.42	-0.04
2010	0.60	0.30	2010	0.66	-0.04
2011	0.62	0.26	2011	0.79	-0.03
2012	0.61	0.40	2012	0.76	0.04
2013	0.61	0.94	2013	0.88	0.22
D	0.02	0.60	D	0.27	0.23
~~MYV~~ I_{mv}	0.04	0.85	~~MYV~~ I_{mv}	0.58	0.13

Table 2. *Scores for T_{2m} and NO_2. The ~~reference year is 2013.~~correlations are calculated between the observations (2008-2013) and the model results (2013).*

Indeed, the simulation domain is large and the temperature has a latitudinal variability larger than between each measurements stations. This temporal correlation ranges from 0.26 to 0.94. And the best score is for 2013 leading to a good score of ~~MYV(R_t)~~$I_{mv}(R_t)$=0.85. The model is thus performing well in capturing the day to day variability for T2m and for the good reasons.

3.3.2 Analysis of NO_2 scores

~~The second example is related to the surface concentrations of NO_2. This species is a~~ Nitrogen dioxide is both a primary and secondary species quickly produced by oxidation of NO and the scores show ~~at the same time~~ if the sources are properly placed ~~,~~ and if the photochemistry and transport processes have been well simulated. In general, at ~~low~~ coarse model resolution, the scores for this species are ~~less good~~ worse than for ozone~~, its spatial extent of its representativeness being more limited (emissions from traffic in urban environments etc.), even if~~. NO_2 is very dependent on the quality of emission inventories, however the measurements stations considered in this study are ~~all~~ background sites.

~~We can see that the~~ The spatial correlation gives a score of R_s=0.88 for 2013. Being the best comparison, we obtain ~~MYV(R_s)~~$I_{mv}(R_s)$=0.58. This shows the importance of NO_x 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 ~~MYV(R_t)~~$I_{mv}(R_t)$=0.13 even if the simulated year is better. These two scores show that the model certainly captures the right location 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 for the reference year compare to the others.

4 Results and discussion

The ~~scores~~ correlations are calculated for all variables described in Table 1 and for the years 2008 to 2013, it is reminded that only the May to August 2013 period was ~~modeled~~modelled. Results are presented as time series in Figure 3. Using all ~~scores~~ correlations and differences values, a ~~MYV~~ I_{mv} is estimated for each ~~score type and each~~ variable. Results ~~are presented in Table 3. These results~~ (Table 3) are discussed in the following sections.

Variable	R_s			R_t		
	Value	D	~~MYV~~ I_{mv}	Value	D	~~MYV~~ I_{m}
T_{2m}	0.61	0.02	0.04	0.94	0.60	**0.**
u_{10m}	0.09	0.09	0.03	0.60	0.56	**0.5**
precip	0.78	0.29	**0.54**	0.30	0.31	0.
AOD	0.97	0.02	0.09	0.45	0.34	**0.**
ANG	0.91	0.04	0.14	0.59	0.44	**0.**
O_3	0.69	0.13	0.29	0.32	0.27	0.
NO_2	0.88	0.27	**0.58**	0.22	0.23	0.
$PM_{2.5}$	0.16	0.15	0.07	0.27	0.32	0.
PM_{10}	0.57	0.43	**0.47**	0.11	0.10	0.0
~~NH_3~~ Ammonium	0.20	0.13	0.08	0.21	0.20	0.
~~H_2SO_4~~ Sulphate	0.51	0.21	0.29	0.31	0.34	0.
~~HNO_3~~ Nitrate	0.15	0.51	0.13	0.09	0.08	0.0

Table 3. *The ~~MYV scores~~ I_{mv} values for all variables: the meteorology with T_{2m}, u_{10m} and precipitation rate, the vertically integrated column of aerosols with the Aerosol Optical Depth (AOD) and the Angström exponent (ANG), the surface concentrations of all aerosols in term of size distribution with $PM_{2.5}$ and PM_{10} and for the inorganic species with D_p < 10 µm. Values of ~~MYV up to~~ I_{mv} above 0.3 are bolded. Units of the variables are detailed in Table 1.*

4.1 Meteorological variables

Scores for T_{2m} were discussed in the previous section. The calculation of u_{10m} also gives satisfactory results ~~but for the temporal correlation is only~~ with R_t=0.60 and ~~MYV~~I_{mv}=0.54. The spatial correlation, R_s=0.09, is not correct and very variable ~~for~~ from one year to another, leading to ~~MYV~~I_{mv}=0.03. As for T_{2m}, we also have an effect of the model resolution and the ~~representativity~~ representativeness of the variable. Scores for the precipitation are correct, with a very good spatial correlation leading to ~~MYV(R_s)~~$I_{mv}(R_s)$=0.54. For the day to day variability, the score is less good with ~~MYV(R_t)~~$I_{mv}(R_t)$=0.21 but significantly higher for 2013. These scores showed that the meteorological forcing is well retrieved, and better for the year being considered compared to other years.

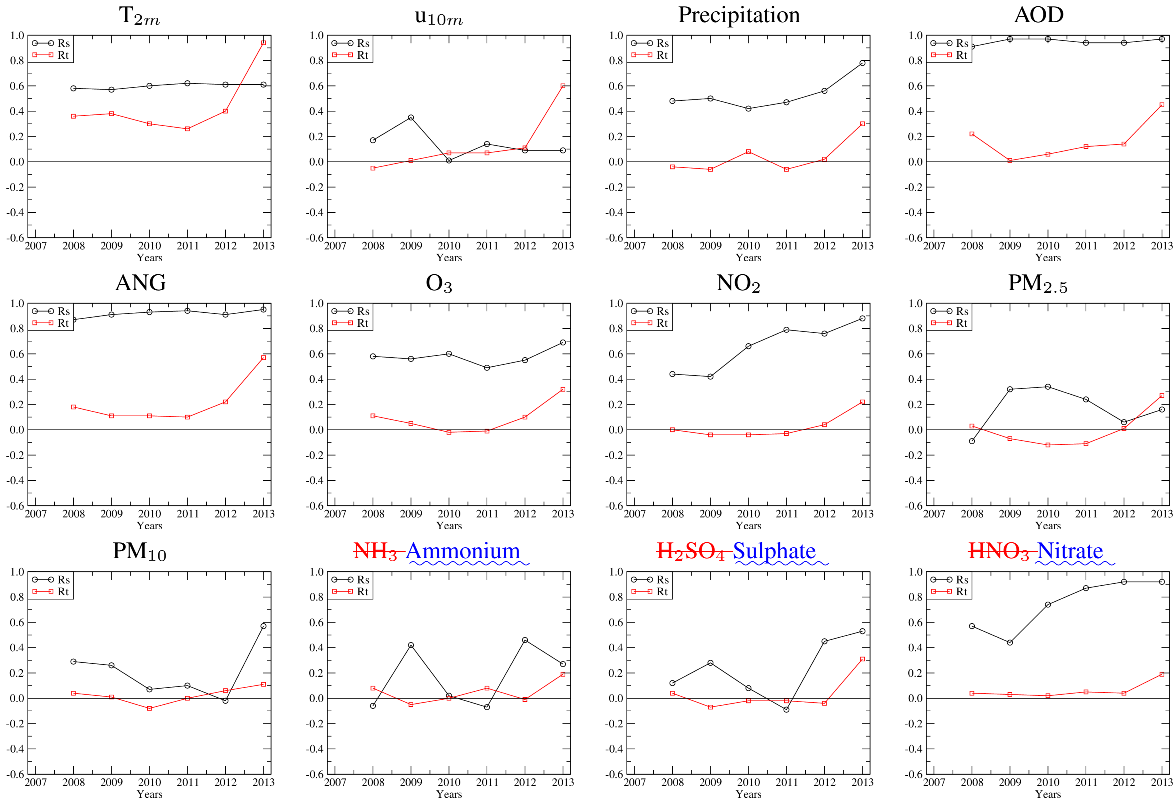

Figure 3. *Multi years scores for ~~the 2m temperature~~T_{2m}, u_{10m}, the ~~10m wind speed~~precipitation rate, Aerosol Optical Depth (AOD), Angström exponent (ANG), surface concentrations of O_3, NO_2, $PM_{2.5}$, PM_{10}, Ammonium, Sulphate and Nitrate. The correlations are calculated between the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

4.2 Optical properties

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

For the AOD, the spatial correlation is very good for 2013, R_s=0.97 but it is as good or better for other years. This means that we model a rather recurring phenomenon: every year the same stations are on average exposed to aerosol plumes: ~~MYV(R_s)~~$I_{mv}(R_s)$=0.09. The temporal correlation is lower with R_t=0.45 but much better than for other years: ~~MYV(R_t)~~$I_{mv}(R_t)$=0.33. This means that the model ~~reproduced partly~~ partly reproduced the observed temporal variability but the events are changing from one year to another and the model captures well these changes. The AOD are sensitive to desert dust outbreaks in summer in that region~~, this~~ . 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 ~~MYV(R_s)~~ $I_{mv}(R_s)$ = 0.14. The temporal correlation is much better for 2013 than other years with ~~MYV(R_t)~~ $I_{mv}(R_t)$ = 0.49. This is probably due to a size distribution that is not necessarily well simulated from one day to another (showed by AOD) but ~~correct~~ the relative contributions of fine and coarse aerosol atmospheric load are fairly reproduced. This feature highlights the high sensitivity of the AOD calculation ~~depending on the modeled~~ to the modelled aerosol size distribution, although the overall mass emitted and transported could be realistic.

Globally, the AOD and ANG reflect the model's ability to retrieve the long range transport of long-lived aerosols ~~. This mixes a lot of~~ which depends on several processes (emissions, transport, and deposition). ~~With these scores , we can conclude that~~ These scores show the model is able to retrieve these yearly recurrent plumes but ~~that the mass distributed into the~~ the model size distribution ~~needs~~ of particles clearly requires improvements.

4.3 Surface concentrations

The ~~scores for the surface concentrations of gaseous species (O_3, NO_2) and aerosol ($PM_{2.5}$ and PM_{10}) are very sensitive to the species. The~~ spatial correlation is good for O_3, NO_2 and PM_{10}, with R_s=0.69, 0.88 and 0.57 respectively. For $PM_{2.5}$ this correlation is low with R_s=0.16. The PM_{10} ~~tends to show~~ shows that the largest particles are well modelled over the whole domain, and this was also the conclusion ~~seeing at the scores for~~ for the AOD and ANG. The low score ~~of~~ for $PM_{2.5}$ ~~show that , in~~ indicates that for the aerosol distribution, the fine mode is less well ~~modeled~~ modelled than the coarse mode. This is confirmed by the scores of the aerosol inorganic species, ~~NH_3, H_2SO_4 and HNO_3~~Ammonium, Sulphate and Nitrate. Except for ~~H_2SO_4~~ Sulphate (with R_s=0.51), the spatial correlations are 0.15 for ~~HNO_3~~ Nitrate and 0.20 for ~~NH_3~~Ammonium. Thus, the fine part of the aerosol is not well ~~modeled~~ modelled mainly due to a deficiency in the ~~modeling of nitratesspecies~~modelling of nitrates.

The temporal correlations have a completely different ~~behavior~~ behaviour that the spatial correlations. The values are generally low, from R_t=0.09 for ~~HNO_3~~ Nitrate to R_t=0.32 for O_3. Surprisingly, the PM_{10} ~~have~~ 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 ~~modeled~~ 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 ~~MYV~~I_{mv}=0.58) but its low temporal correlation with R_t=0.22 and a low ~~MYV~~I_{mv}=0.13. In this case, this means we have a correctly localized anthropogenic emissions inventory (main source of NO_2) 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 ~~persistent behavior, the MYV scores~~ recurrent behaviour, I_{mv} having high values. For ~~the aerosol~~particles, the problem is more related to the fine mode, where $PM_{2.5}$ concentrations are not well located. This ~~modeling problem is also obvious with~~ modelling problem is highlighted by the low correlations and ~~MYV~~ I_{mv} values for the inorganic species. For the temporal correlations, the scores are always lower ~~that~~ than for the spatial correlation but also always higher for the ~~year being considered~~ reference year than for the other years.

4.4 ~~All~~ Representation of results on a single plot

Complementary to the Table 3, Figure 4 reports the results on a ~~simple~~ single plot. The x-axis represents the correlation (spatial or temporal), the y-axis represents the differences between all years, D. For each studied variables, their values are reported on the Figure where the ~~colors~~ colours represent the value of ~~the score MYV~~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 ~~up to~~ above 0.5. This means that, spatially, the studied problem shows systematic patterns from year to year. The low values of correlations ~~thus showed~~ show that some variables are systematically ~~on specific parts of the domain and not over the whole region~~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 ~~all~~ temporal correlations shows a specific linear pattern. The largest ~~the correlation values , the most important the~~ correlation values are positively correlated with differences. This temporal correlation represents the day to day variability ~~for~~ 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 correlations but low differences). This illustrates the fact that the studied problem is primarily ~~a problem of sporadic peaks and that~~ an issue of sporadic events and the model is able to correctly find this variability from one day to another.

5 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 ~~when~~ using a single statistical indicator. Of course, this approach will never replace a strict evaluation of a pollution case analysis using time series, vertical profiles and usual error statistics. ~~But it can afford to have~~ However, it offers a very fast and integrated vision of the strengths and weaknesses of a model with very little calculation. This ~~may also allow, for example, to compare different models during~~ 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 model always better simulates~~ , 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 ~~modeled~~ 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 par-

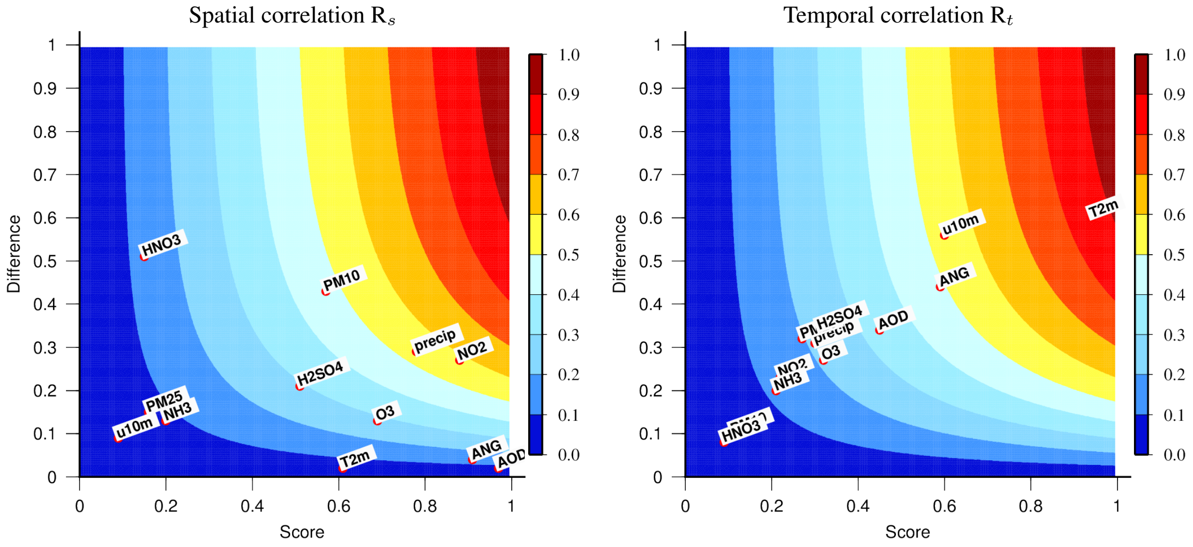

Figure 4. *Results of the ~~MYV~~ I_{mv} scores for the spatial and temporal correlations. For each model variable its value is represented using the correlation on the x-axis and the difference between the studied year and the others on the y-axis. The ~~colors~~ colours represent the ~~MYV~~ I_{mv}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

ticles as demonstrated with R_s=0.96 and 0.90 for AOD and ANG. But, since this feature ~~is recurring~~ occurs every year, this leads to low ~~MYV scores~~ I_{mv} values. This means that for a large domain, the main spatial patterns of particle concentrations are recurrent and well ~~modeled~~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 ~~aerosol~~particles, the results depend on the size distribution: the largest particles are better ~~models~~ simulated than the finest ones.

The conclusions are different for the temporal correlation. The scores are calculated using daily observations and ~~modeled~~ 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 ~~MYV scores: MYV~~ 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 ~~MYV scores~~ I_{mv} values are low. This means that improvements ~~have to be done~~ are required in priority for the day to day variability compared to the locations of emissions. This may probably be ~~dued~~ 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.

6 Code and/or data availability

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

References

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

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

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

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

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

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

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

Kioutsioukis, I. and Galmarini, S.: *De praeceptis ferendis*: good practice in multi-model ensembles, Atmospheric Chemistry and Physics, 14, 11 791–11 815, doi:10.5194/acp-14-11791-2014, 2014.

Kopacz, M., Jacob, D. J., Fisher, J. A., Logan, J. A., Zhang, L., Megretskaia, I. A., Yantosca, R. M., Singh, K., Henze, D. K., Burrows, J. P., Buchwitz, M., Khlystova, I., McMillan, W. W., Gille, J. C., Edwards, D. P., Eldering, A., Thouret, V., and Nedelec, P.: Global estimates of CO sources with high resolution by adjoint inversion of multiple satellite datasets (MOPITT, AIRS, SCIAMACHY, TES), Atmospheric Chemistry and Physics, 10, 855–876, doi:10.5194/acp-10-855-2010, 2010.

Lemaire, V. E. P., Colette, A., and Menut, L.: Using statistical models to explore ensemble uncertainty in climate impact studies: the example of air pollution in Europe, Atmospheric Chemistry and Physics, 16, 2559–2574, doi:10.5194/acp-16-2559-2016, http://www.atmos-chem-phys.net/16/2559/2016/, 2016.

Mao, Y. H., Li, Q. B., Henze, D. K., Jiang, Z., Jones, D. B. A., Kopacz, M., He, C., Qi, L., Gao, M., Hao, W.-M., and Liou, K.-N.: Estimates of black carbon emissions in the western United States using the GEOS-Chem adjoint model, Atmospheric Chemistry and Physics, 15, 7685–7702, doi:10.5194/acp-15-7685-2015, 2015.

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

Menut, L.: Adjoint modelling for atmospheric pollution processes sensitivity at regional scale during the ESQUIF IOP2, J. Geophys. Res., 108, 8562, doi:10.1029/2002JD002549, 2003.

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.

Pison, I., L.Menut, and G.Bergametti: Inverse modeling of surface NOx anthropogenic emissions fluxes in the Paris area during the ESQUIF campaign, Journal of Geophysical Research, Atmospheres, 112, D24 302, doi:10.1029/2007JD008871, 2007.

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

Prank, M., Sofiev, M., Tsyro, S., Hendriks, C., Semeena, V., Vazhappilly Francis, X., Butler, T., Denier van der Gon, H., Friedrich, R., Hendricks, J., Kong, X., Lawrence, M., Righi, M., Samaras, Z., Sausen, R., Kukkonen, J., and Sokhi, R.: Evaluation of the performance of four chemical transport models in predicting the aerosol chemical composition in Europe in 2005, Atmospheric Chemistry and Physics, 16, 6041–6070, doi:10.5194/acp-16-6041-2016, 2016.

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

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

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

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

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

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

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