Response to Referee #1: We would like to thank the referee for the careful review and thoughtful suggestion, especially the suggestion of replacing the t=1 bias forecast with that of t=0, which helps us further improve the quality of the manuscript.

Our response follows (*the reviewer's comments are in italics*)

General comments:

The main purpose of this paper is to explore the impact of observational biases on dust assimilation. These biases exist because surface PM10 contains both dust and non-dust contributions. However, the authors discuss two methods to separate out dust PM10 and then use it in a data assimilation scheme. Novel is their use of a machine learning technique (a neural net called LSTM) to do this. The authors show that, if not corrected, such biases can seriously (and negatively) affect dust forecasts but that use of the LSTM allows for significantly better forecasts. The paper is fairly technical and may be better suited to GMD but is appropriate for ACP too. The paper is well-written, properly organised with good graphics. Its conclusions are warranted by the data presented.

My main comment would be to add more detail to the section on the LSTM bias correction.

Reply: We agree with the referee that more details about LSTM are necessary, more information is now added to describe the LSTM bias correction in ***Section. 3.2 Machine learning for non-dust PM10 simulation*** on page 10-12.

Question 1: *p 2, l 21: please consider adding "Inverting East Asian Dust Emissions of a Severe Winter Dust Event Using the Ensemble Kalman Smoother and Himawari-8 AODs" by Tie Dai, currently in review.*

Reply: We are very interested in this relevant paper. We would appreciate if the referee could provide us the link of source paper/citation (at the present we are unable to find the article).

Question 2: *p 3, l 25: it is a pity the authors do not compare the correction scheme in Eq (1) also. Maybe it is too much work to do the full assimilation work but at least a comparison of non-dust PM10 estimates should be possible?*

Reply: Thanks for pointing out this issue. We already used this method (Eq.1) in our previous paper (Jin et al., 2018). This method worked fine when the relation between PM2.5 and PM10 in a given observation site exhibited a simple linear correlation. However, in many sites the application of Eq.1 failed when the linear correlation (***R***) between PM2.5 and PM10 is weak. Actually, in that study we only performed the Eq.1 for the site when ***R*** ($PM_{2.5}$, PM_{10}) > 0.8 (accuracy is promised). In that case, observations in 45% sites are abandoned. Unlike the machine learning method which can be used almost everywhere.

To point out the weakness of Eq.1, we added the remarks on page 3, line 30-32 "***For instance, in the dust event studied in Jin et al. (2018), the application of Eq.1 in many sites failed since R is weak. To have a quality-assured bias correction, Eq.1 is performed only when the Pearson correlation coefficient R > 0.8. Consequently, measurements in around 45% sites are rejected in that case.***"

Question 3: *p 7, l 5: What sort of uncertainty in emissions results from this: 1%, 10%, 100%? What is the spatial structure of delta f? Does i vary freely from grid-box to grid-box or is some correlation imposed?*

Reply: The emission uncertainty is assumed to result from the uncertainty of friction velocity threshold in the windblown parametrization. We add more explanation on page 7, line 5-10 "***Following Jin et al. 2018, the errors in dust emission field were assumed to be only caused by the uncertainty in the friction velocity threshold in the dust windblown parametrization, and similar assumptions on the uncertainty are used to build an emission error covariance B. The friction velocity threshold is perturbed with a spatially varying multiplicative factor β with a mean of 1 and a standard deviation of 0.1. In addition, an exponential profile of distance-based spatial correlation is posed on βs (Jin et al., 2018)***.

Question 4: *p 10, l 10: why are PM25 measurements at nearby sites used? The authors already used PM25 from the AQ network. What sort of error do they think is introduced by using measurements taken at different locations?*

Reply: Yes, the reason why the air quality observations (not only PM2.5) at nearby sites are used are indeed not clearly clarified. We now add some contexts to explain the necessities by saying "***For the non-dust PM10 machine learning forecasts in a given site, observations from its nearby sites are also vital and are used in two ways. First, missing data records are unavoidable in an air quality monitoring network, while the LSTM model training requires an uninterrupted time series of features. In this study, data interpolations of air quality measurements (PM10, PM2.5, SO2, NO2, O3 and CO) are performed using both a linear interpolation and a k-Nearest-Neighbor algorithm (Zhang, 2012) if a site has no more than 30% of missing data. Otherwise, all the measurements in the given sites are abandoned. Generally, more information available from the nearby sites will result in a more accurate interpolation. Second, learning in the presence of data errors is pervasive in machine learning, and the measurements from nearby stations are used to limit their influence. Data errors occur due to incorrect sensor readings, software bugs in the data processing pipeline, or even the inaccurate data interpolation. Statistical analysis tests have been conducted which did not only indicate a strong correlation between the non-dust PM10 and air quality measurements in the given sites, but also show that the predictor (non-dust PM10) is correlated to the observation indices (especially the PM2.5) at its nearby sites. In order to eliminating errors caused by incorrect inputs at the modeling site, the measurements at the nearby stations are considered as the essential indices. In this study, a data instance will only be selected for training the LSTM model if there is at least one nearby site within an empirical radius 0.8\degree (approx 80 km), and a maximum of 3 nearby sites will be randomly selected where observation stations are densely distributed. To save the computation costs on machine learning model training, only the PM2.5 from the nearby sites are included as one of the inputs in this study.***" on page 11, line 20-34.

Question 5: *p 10, l 20: I would not say the "actually are" non-dust PM10 but rather that they may be assumed to be very close to non-dust PM10.*

Reply: We agree with the referee. "***actually are***" is now changed to "***very close to***" on page 11, line 12.

Question 6: *p 10, l 22: "measurements at the nearby sites are considered as the essential input". Maybe the authors can clarify?*

Reply: See the ***Reply*** to ***Question 4***.

Question 7: *p 10, Sect 3.2: since this is arguably the new contribution by this paper, more details seems required. E.g. 1) how large where training and testing datsets; 2) how where they separated?; 3) how where LSTM's hyper-parameters set? 4) Please provide error statistics for training & testing. I am familiar with LSTM use in time-series but in this paper the authors also claim it is useful for spatial*

variations (p 10, l 1). Was the LSTM adapted to allow for this? It seems random forests might equally be useful. Did the authors consider this? Random forest is easier to train and would also provide feedback on what are the most relevant features for non-dust PM10 prediction.

Reply: Regarding the training and testing dataset: It was not clearly illustrated, we now added more remarks on page 11, line 4-6 "***The training dataset covers the period from January 2013 to March 2015. In other words, the LSTM model L is trained to best fit the samples from this period. The two months April and May 2015 in which the studied dust event occurred is set as the testing period.***"

Regarding how training/testing period are separated: We did not consider this separation too much, like the 8:2 (or 7:3) ratio between the training and testing dataset. The air quality measurements that we have started from 2013. The dust storm we studied is in 2015 April. So, we set the training period from 2013 Jan to 2015 Mar, while the following two months in which the dust event occurred in set as the testing periods.

Regarding the hyper-parameters set: Hyper-parameters are listed in Table.1 LSTM hyper-parameters on page 12.

Regarding the error statistics for training and testing: Both statistic errors from the CTM and Machine learning are now presented. RMSE of CTM are given on page 12, line 21-22 by saying "***The CTM LOTOS-EUROS/non-dust in general underestimates the non-dust PM10. The forecast results in a relatively large root mean square error (RMSE) 89.4 ug/m3.***"; RMSE of the two LSTM are given on page 12, line 24-27 by saying "***The two LSTM forecasts show on average a good agreement with the observations. The RMSEs of the forecasts of the two machine learning models in the two years of training period are reduced to 55.9 and 60.7 ug/m3, and in the two months of test period (excluding the dust event April 14 to 16) they also stay at comparable low levels of 58.6 and 60.2 ug/m3.***"; These RMSEs are also shown in Fig.3 on page 13.

Regarding "***the LSTM is useful for spatial variation***": It is a typo error, actually we performed the LSTM training and testing/forecast site by site as we mentioned on page 12, line 1-3 "***The machine learning model for non-dust PM10 forecast is trained site by site, with the hyper-parameters shown in Table.1***". Also on page 10, line 16 "***in predicting spatial and temporal varying time series problem***" is corrected to "***in predicting time series problem***".

Regarding the option of random forests: Thanks for the suggestion. The random forests will be valuable when there are various input features but only parts of them are relevant to the output. We did not consider it yet, but will take it into account in our future work.

Question 8: *p 11, l 22: why use a t=1 forecast to estimate non-dust PM10. Can't you train the LSTM to estimate non-dust PM10 at t=0?*

Reply: Thanks for this suggestion. We agree with the referee that t=0 simulation could indeed replace the t=1 forecast, and are expected to give a better non-dust PM10 simulation for bias correction. New bias simulations "t=0" have been performed, as well as assimilation runs with the new bias-corrected dust observations. Both of the bias-correction as well as the posterior dust forecast using t=0 indeed showed some improvements. New results are now used to replace the previous ones, they can be mainly found in Fig.3b (non-dust simulation vs. PM10); ***Fig.4d-4e*** (spatial distribution of the non-dust level and bias corrected dust observations); ***Fig.5*** (time series in 6 cities); ***Fig.7d*** (posterior emission field by assimilating bias-corrected data using LSTM); ***Fig.8d*** (posterior of dust simulation using the bias-corrected measurement by LSTM); ***Fig.9*** (posterior dust forecast in 3 cities) and ***Fig.10*** (RMSE). From the perspective of RMSE, t=0 provides a slightly better result.

Question 9: *p 16, l 1: Would R vary due to bias correction scheme? Afterall that scheme will affect remaining errors in dust PM10. In particular, I wonder whether the LSTM scheme (which trains on all data throughout a domain) would not lead to off-diagonal elements to be non-zero (i.e. correlations amongst errors at nearby stations)? Please discuss.*

Reply: All the three types of assimilation tests used the same ***R***, and all the observation assimilated are assumed to be independent (since the LSTM is performed site by site). We agree with the referee that the three types of data might have different ***R*** since they have different remaining biases. However, it will then be difficult to judge/evaluate whether the adding values of new assimilation is achieved by using the bias correction or new observation error covariance.

We added the remarks to point out the possibility of using different ***R*** by saying "***Note that the raw PM10 and the bias-corrected dust measurements might have different uncertainties in representing the real dust storm level. This is not yet taken into account in our study, and the three types of the assimilated measurements, raw PM10, bias-corrected dust observation either using the CTM or using the machine learning, are all configured with the same observation error in Eq.9. In addition, all the measurements are assumed to be independent, hence, the observation error covariance R is diagonal.***" on page 18, line 12-16.

Question 10: *p 16, l 5: Can the authors discuss the largest sources for the errors of 10%? Measurement error? non-dust correction? Spatial representation? I see the authors suggest (!) representation is the most important but they give no reason for this. non-dust correction errors are of course available from the LSTM training.*

Reply*:* We assume the spatial representation error is the largest source in this study. More remarks are given to explain this point by saying "***Theoretically, the observation uncertainties are due to the representation errors as well as the measurement errors, while the former one is widely considered as the largest source. Limited by the computation resources, our dust model uses a spatial resolution of 25 km, while the in-situ measurements cover the much less of atmosphere surrounding them (Schutgens et al. 2016). This of course limits our capability of resolving the fine-scale fields that are reflected in observation spaces. Therefore, the spatial representation error is assumed to be the dominant error source and taken into the account in approximating the observation uncertainties. In addition, the error due to the different bias correction term is another source. It is not considered yet in this study but will be exploited for have a more accurate assimilation test in our future work.***" on page 17, line 14-21.

Question 11: *p 16, l 15: "and thus the representation uncertainty" I do not agree with this. The representation error is the difference between a point measurement and a grid average. Here the authors consider variation among point measurements, which is likely to over-estimate representation error. Also, representation error will can show significant variation depending on location of point measurement in grid box, see e.g. Schutgens et al. ACP 2016.*

Reply: Yes, the representation error could be calculated by comparing the model simulation at different spatial resolutions. We have now pointed this out on page 17, line 22-25, "***The spatial representation error quantification itself is a complex task. It could be calculated through comparing the model simulations at different scales of resolutions. In this study, the availability of multiple measurement sites in a single model grid cell provides an alternative way to quantify the representation error. When multiple observations are present, the statistical error in the observed values reflects the spatial representation uncertainty.***"

What we did in Fig.6 is to present the hourly average vs. stand deviation of hourly observations in the grid cell of Beijing. The standard deviation is then used to approximate the observation representation error.

We also agree with the reviewer that the representation error might vary in space and time. But in most of the grid cells, we only have one or two monitoring sites, which make it difficult to parameterizing the representation error for other grid cells in a similar way as for Beijing. To explain this, we added some new remarks on page 18, line 8-11 by saying "***The representation uncertainty has already been validated to fluctuate both in space (Schutgens et al., 2016). However, for most other grid cells the number of observations sites is simply one, which makes it difficult to parametrize a representation error in a similar way. Therefore, the representation error parametrized for Beijing is used for all other locations too.***"

Question: p 16, l 16: Can the authors surmise why relative standard deviation (std/mean) is almost constant for non-dust and dust event? Of course, PM10 levels increase but wouldn't one expect a smoother field in case of a dust plume (long range transport) vs local pollution?
Reply: Yes, it is true that the relative **std** of dust period and non-dust period in Fig.6 looks quite constant. But it is not easy to draw a general conclusion since we only have a few samples.

Question: p 16, l 27: Would the authors not expect the representation error in an Urban area to be significantly larger than in the countryside? Also, shouldn't representation error really be calculated for the dust PM10, instead of the total PM10?
Reply: Regarding the different representation error in the urban area and countryside: In the future, we would use a model of higher resolutions that helps to quantify the dynamic spatial representation error.
Regarding the different representation errors: It comes to the same question (***Question 9***) that whether we should use the same ***R*** for different assimilated data (PM10, bias-corrected dust observation either using CTM or machine learning).

Response to Referee #2: We would like to thank the referee for the careful review and the thoughtful comments, which help us to further improve the quality of the manuscript.

Our response follows (*the reviewer's comments are in italics*)

General comments:

This article looks at the effect of applying a bias correction term to PM10 observations in order to assimilate them in to a dust model as 'dust' observations. The bias correction is needed since PM10 observations account not just for dust, but also other types of aerosol. The paper compares two different bias correction methods to just assimilating the PM10 observations directly. The first bias correction method is to use a CTM to calculate the non-dust part of the PM10 observations and the second more novel method is a machine learning approach.

The paper is in general well written with clear figures, although there are a few grammatical English errors (those that I could easily comment on are listed under typos below). The structure is straight forward and the logic easy to follow. It is also fairly comprehensive and so provides an in-depth examination of the two different bias correction methods. However, it does not feel that this is too much information but rather provides context as to why the assimilation results are seen.

I have two main comments to make on the article. The first is that it wasn't clear to me, from reading through the article, how such a bias correction would be applied in a real life scenario. There is some comment on this in the summary and conclusions but it would be interesting and very relevant to understand how either bias correction scheme might be applied in practice. It is not obvious to me how this would be possible. My second comment is that I sometimes struggled to understand the detail of what had actually been done in each of the two schemes and this is addressed in the more specific comments below:

Comments:

Question 1: *Pg. 1, line 20. I don't believe that the conclusion can be drawn in general that the best results are obtained when using a machine learning model. This is true in this article using this CTM, but the results that show an under-estimation of the non-dust PM10 from the CTM would, to me, provide evidence that the CTM is a slightly flawed model for this aspect and hence why the machine learning approach performs better. It would be enough to change this sentence to 'The best results are obtained when using the machine learning model...'*

Reply*:* Accepted. *"**a machine learning model**" is now changed to "**the machine learning model**"* on page 1, line 20.

Question 2: *Pg. 7, line 7. What do you mean to release the efforts in updating the tangent linear model? To me this would imply the effort to keep the tangent linear model up to date with the full non-linear model, but I'm assuming that what you really mean is to reduce the time spent in calculating the tangent linear model?*

Reply*:* Thanks for point out this misleading sentence. *"**To release the efforts in updating the full tangent linear model**" is now changed to "**To reduce the computational costs in calculating the tangent linear model**"* on page 7, line 12.

Question 3: *Pg. 9, Section 3.1. I assume this chemical transport model was chosen because it produces a full operational forecast over the modeling domain? Does this therefore make the application of the bias correction possible for a real life scenario? If true, then this would provide a good explanation for why this model is used to evaluate between the two different schemes when it is clearly unable to estimate the non-dust PM10 observations.*

Reply: Yes, this CTM we used to model non-dust aerosol levels is the same as the operational one, but the desert dust emissions are disabled.

Question 4: *Pg. 10, line 10. Are the input observations of PM25, SO2, NO2, O3 and CO all coming from the same station as the PM10 observations used to compare the output? Is this why the nearby sites of PM25 are included as a separate line. Does the machine learning model therefore come up with a different value of PM10 for each station separately or is all the data included together so that given the input observations from any station, a generic PM10 output would be generated?*

Reply*:* We actually estimate the non-dust aerosol simulation site by site. The inputs are the PM25, SO2, NO2, O3 and CO coming from the same site, plus the PM2.5 from the nearby sites. The output is the non-dust PM10 in the given site. The machine learning model is trained to output the value that can best fit the PM10 records in the training period.

To make it clear, we added the new remarks that on page 12, line 3, "***The machine learning model for non-dust PM10 forecast is trained site by site***". In addition, we added a new paragraph on page 11, line 20-34 to explain why the measurements at the nearby site are important.

Question 5: *Pg. 10, eqn 8. What does the 'm' stand for in this equation. Is it observation across the full domain and time period or just observations at one station over the time period?*

Reply: Following the **Question 4**, we actually train and use the machine learning model for each site separately, the "***m***" here only represents the observations at the given station over the past "***m*=18**" hours.

Question 6: *Pg. 11, line 22-line 28. The description in this paragraph needs further clarification. I believe that you are subtracting the 1-hr forecast made from April 15, 19:00 (so valid at 20:00?) from all the PM10 observations between 8:00 and 19:00 for assimilation. Or is it the one hour forecast valid at each specific time of the observation? In which case does each observation from each station require different 18hour input? Why is the 1-hr forecast chosen rather than the actual value coming from the machine-learning method? How is a forecast made from a machine learning method? Similarly, is the 12-hr forecast from April 15, 19:00 (so valid at April 16, 07:00) added to all the forecast dust PM10 values to compare to the observations?*

Reply: Thanks for pointing out this issue, the original context is indeed ambiguous.

Regarding the 0h forecast: Actually 1h forecast is valid at each specific time of the observation, e.g., forecast of 20:00 is valid at 19:00. In addition, there is no need to use 1h forecast as the bias level, 0h forecast (simulation of 20:00 is also valid at that moment) could make our system work and give a better performance. We have conducted the non-dust bias simulation again but now with "t=0", as well as the corresponding assimilation test. Indeed, both of the bias-correction as well as the posterior dust forecast are somewhat improved when using t=0. New results can be mainly found in Fig.3b (non-dust simulation vs. PM10); ***Fig.4d-4e*** (spatial distribution of the non-dust level and bias corrected dust observations); ***Fig.5*** (time series in 6 cities); ***Fig.7d*** (posterior emission field by assimilating bias-corrected data using LSTM); ***Fig.8d*** (posterior of dust simulation using the bias-corrected measurement by LSTM); ***Fig.9*** (posterior dust forecast in 3 cities) and ***Fig.10*** (RMSE). From the perspective of RMSE, t=0 provides a slightly better result.

Regarding the 12 hours bias forecast: It means the non-dust PM10 forecast of April 16, 07:00 is available at April 15, 19:00. To make these 0h and 12h terms clear, we added more explanation on page 12, line 32; on page 13, line 1-2, "***Note that here t=0 forecasts denote the forecasts valid at each specific snapshot of the observations, while the 12 hour forecasts are the forecasts valid 12 hours in advance, e.g., the non-dust PM10 forecast at April 16 07:00 are valid at April 15 19:00.***"

Question 7: *Pg. 9-11, Section 3.2. How would such a machine learning algorithm be used in practice? Would it need the constant evaluation of running 18hrs of data? Or if a dust storm was forecast, would the calculations be switched on. At what point would the PM10 observations be assimilated. Or is this envisaged mode for a reanalysis style product of PM10 observations?*

Reply*:* It is indeed important to clarify how the machine learning for bias simulation can be used in practice. We added new remarks on page 12, line 13-15 "***Our two bias models, LOTOS-EUROS/non-dust and LSTM, could both be used for air quality forecast operationally when there is no dust storm. Once a dust storm is obsereved, the dust emission inversion system will be enabled, the two non-dust PM10 models will then be used in dust observation bias correction.***"

Regarding when to start the dust storm data assimilation: The emission inversion method we used here is 4DVar, we perform the assimilation test at April 15, 19:00. The assimilation window sets are the same as the ones in Jin et al., 2018. More details are given in ***Fig. 2*** on page 9.

Question 8: *Pg. 12, line 9-10. You state that the according to the LOTOS-EUROS bias corrected observations, the dust storm seems to have already reached central China which was in reality not the case? How do you know this? Nothing from this figure provides evidence of what the reality was? It would be worth referencing how you know this to be true.*

Reply: The LSTM bias estimation is validated to give the best performance in general. The bias-corrected dust observations by LSTM (in Fig 4.e1 on page 15) indicate that there is no dust in the central China, we consider it is more close to the reality than Lotos-Euros bias corrected observations in Fig 4.c1.

However, we agree with the reviewer that it is incorrected (100% sure) to say "***in reality***". So, "***was in reality not the case***" is now changed to "***was probably not the case***" on page 14, line 14-15.

Question 9: *Pg. 13, line 17-19. This to me provides evidence that the LOTOS-EUROS model is not doing a good job of predicting the non-dust PM10 and so is never going to match the performance of the machine learning algorithm. Is there a reason for choosing this model or not exploring CTMs that may provide a better match?*

Reply: As all CTM's, also Lotos-Euros is under permanent development. Its current performance with respect to aerosols is similar to other CTM models in the Marcopolo-Panda project (Brasseur et al., 2019; Petersen et al., 2019). The ultimate goal is a very accurate simulation. In the current state of development, the machine learning algorithm is doing a better job in the measurement sites as a computationally efficient method. However, machine learning can give prediction at the measurement sites while the CTM can provide simulation results over the entire 3D fields. Machine learning could therefore never fully replace a CTM, and when both are available, it makes sense to compare their ability for use in a bias correction.

The errors in forecast-observation are mainly attributed to inaccuracy in weather forecast and errors in the adopted surface emission schemes. There is definitely space to improve it using nudging techniques like data assimilation. However, it takes lots of efforts and is not exploited in this study yet. To explain this, we add new remarks on page 10, line 3-9 "***The operational CTM Lotos-Euros over the China is in its early phase of development as well as the other six CTMs used in the MarcoPolo-Panda project.***

The purpose of the project is to diagnose statistical differences between the ensemble model simulations and observations. An important objective is to determine ways by which the models could be improved. These differences are mostly attributed to inaccuracy in the weather forecast and errors in the adopted surface emissions (Brasseur et al. 2019, Petersen et al. 2019). Indeed, there is room for minimizing the forecast-observation differences using nudging methods like data assimilation, which takes considerable efforts and not yet exploited in this study."
Also in the Section of ***Future work*** on page 25, line 2-8, we mentioned that "***Both our CTM and machine learning based bias correction methods have room for improvements. It might be useful to improve the CTM simulations by assimilating PM10 observations during the hours where no dust storms are present, and use these improved simulations to remove the non-dust part of the observations during an event. These additional assimilations would then involve repeated forward ensemble bias-model runs which could be computationally expensive.***"

Question 10: *Pg. 18, line 20. Again, you state that the LOTOS-EUROS observations over-estimate the observations and yet I can't see anything in Figure 8 that shows the observations, so that we know this.*
Reply: "***they still overestimate the dust observations***" is now changed to "***they are still likely to overestimate the real dust levels***" on page 20, line 9-10.

Question: *Pg. 18, line 31-33. It is interesting to me that although the peak is a better match with the machine learning algorithm, the forecast is actually slightly better with the LOTOSEUROS model. Do you have any feeling for why this may be?*
Reply: It could be explained by the context on page 20, line 24-29 "***This can be explained from the fact that the dust storm is a strong flow-dependent phenomenon in which concentrations at a certain location are strongly correlated to earlier concentrations at upwind locations. For Hohhot, only a limited number of observation sites is located upwind, and therefore hardly any data is available to constrain the concentrations at this location. To improve the forecast at Hohhot it will be necessary to have additional observation data, for example from sites actually within the source region, or from satellites observing the aerosol load over the source region (Jin et al. 2019)***".

Typos:

Pg. 1, line 13. I think this should read 'The latter is trained by learning using two years of historical samples'.
Reply: Accepted

Pg. 2, line 9. Should be 'progress has also been made'
Reply: Accepted

Pg. 2, line 19. Should be 'A wide variety of data assimilation techniques have been used...'
Reply: Accepted

Pg. 2, line 28. Should be 'In the presence of biases...'
Reply: Accepted
Pg. 10, line 25. Should be 'Earlier studies showed that the...'
Reply: Accepted

Response to Referee #3: We would like to thank the referee for the careful review and the insightful comments, which helps us to further improve the quality of the manuscript.

Our response follows (*the reviewer's comments are in italics*)

General comments:

The aim of this work is to develop a methodology able to use PM10 as a tracer of dust with the last end of assimilating PM10 to improve the forecast of dust storms. The deterrent of direct assimilation of PM10 is that it does not only encompass dust but also other aerosol flavors that are more significant where anthropogenic sources of aerosols are present. To isolate dust and non-dust aerosols the idea is to consider non-dust as bias in the PM10 records. Then the problem is reduced to apply bias correction methods to PM10. Here the authors propose to use machine learning and chemical transport model for bias correction. Results show a more accurate forecast of dust storm if the PM10 assimilated is isolated form non-dust aerosols using the machine learning method.

The paper is well written, structured and results support the conclusion archived. The topic is innovative since it is one of the first works to apply machine learning to bias correction to isolate non-dust aerosols from PM10. In addition, the methodology developed here goes beyond to improve the forecast of dust storm and it can benefit the assimilation of other magnitudes. This study match with the ACP topics and its quality is close to its standards.

However, there are some general comments that need to be addressed or clarified along with the manuscript.

Question: *1) How this approach guarantees the "bias" in PM10 that you are correcting is non-dust and it is not a "real" bias present in your PM10 records?*

Reply: We agree with the reviewer that the PM10 observation itself might have "real" (or "native") bias. However, it is unknown, nor can be quantified using any method as we know. We hence assume that the PM10 measurements are unbiased, but they are biased when they are used as proxy as dust observations. To point out the possibility of "real" bias in PM10, we add some remarks on page 8, line 2-4 "***Note that the PM10 measurements themselves might also contain 'native' biases due to the incorrect sensor reading or systematic errors. However, this part of the bias in the PM10 observations is unknown and not considered in this study.***"

Question: *2) How do you know a better performance of LSTM for bias correction (or dust isolation from PM10) and also better dust forecast if you are not comparing against dust records? I am aware of the lack of availability of them but at least it will be necessary to actually know that your PM10 without bias is dust by comparing with some station or a satellite image.*

Reply: Yes, it will be better if we would have "pure" dust observations with which we could validate our dust forecasts. However, almost all the observations available measure the sum of dust and non-dust aerosols, e.g., PM10 or aerosol optical depth. To fully evaluate our system, we validate the bias forecast as well as the dust storm forecasts using the bias-corrected data.

The bias simulation is actually the operational full aerosol forecast when there is no dust storm. We compare it to the real PM10 observations in test period (April and May) in ***Fig.3*** on page 13. Since the test period includes the dust storm event (April 14 to 16), we now plot the scatters using different marker for dust and non-dust period. The new ***Fig.3b*** and ***Fig.3c*** show that our method can accurately simulate

the non-dust PM10. The bias correction performance can also be seen in the time series at 6 different cities in ***Fig.5*** on page 16. The comparison indicates that our bias simulation is very close to the non-dust aerosol level hence can be used to calculate the "real" dust level.
In addition, since we only have PM10 measurements and don't have the pure dust observation during the dust storm, to evaluate the final dust storm forecast, we calculate the difference between full aerosol simulation (dust forecast + non-dust forecast) and PM10 measurement. The RMSE in ***Fig.10*** on page 24 also indicates our forecast has a good and constant performance in the full aerosol modeling during the dust storm.

Question: *3) Also, the main caveat of the machine learning methods is the availability of a long series of records and the computational time. Do you think that this method will be able to correct the PM10 bias obtained a few hours (e.g. 12 h) before a dust storm and be ready to assimilate and perform the forecast a few hours later? The real applicability of LSTM is not clear. This point needs a little more discussion.*
Reply: We have machine learning based non-dust aerosol forecast in this study, they are 0 and 12 hours in advance, respectively. The former one is used for bias correction, hence the bias simulation only need be ready at the specific moment that we want to have dust storm data assimilation. When the dust forecast is ready after the assimilation analysis, the later one (12 h) will be added to the posterior dust storm forecast and then treated as the full aerosol forecast. To clarify this, we added extra remarks on page 12, line 31-32, on page 13, line 1-2 by saying "***When we perform the assimilation analysis at April 15, 19:00, the short period of t=0 h forecast will be treated as the non-dust levels in the bias correction of the original PM10 measurements. Note that here t=0 forecasts denote the forecasts valid at each specific snapshot of the observations, while the 12 hours forecasts are valid 12 hours in advance, e.g., the non-dust PM10 forecast (12 h) at April 16 07:00 is valid at April 15 19:00.***"

We also add more remarks to explain the computational efficiency regarding to our machine learning work on page 12, line 3-5 "***The machine learning model for non-dust PM10 forecast is trained site by site, with the hyper-parameters shown in Table.1. With the following hyper-parameters, the machine learning model training costs several minutes for each site. The model training in each site is independent hence the whole workload is highly parallelizable.***"

Question: *4) Finally, the conclusion about what method is better for bias correction looks like that is the same method that better "simulate" the non-dust aerosols. Then, the worst bias correction with LOTOS-EUROS model is just the worst performance in reproducing non-dust aerosols. This simplification is not straightforward needs to be clarified.*
Reply*:* We assume that the PM10 measurements themselves are not biased, but if we use them as proxy of dust observations, then they are biased. The real dust level is calculated by subtracting the non-dust level from the PM10 observations. Therefore, we make the conclusion that the method that better simulate the non-dust aerosol is the better bias correction method.
To make the conclusion more justified, we added more statistical descriptions on page 24, line 1-5 "***The two methods to estimate the non-dust part of the PM10 load have been validated. The simulations by the LOTOS-EUROS/non-dust model in general underestimate the PM10 concentrations. The root mean square error stays at a relative high level of 89.4 ug/m3. It is mainly caused by missing emissions and aerosol components such as secondary organic matter. In comparison, the data-driven machine learning model agrees more closely with the real measurements, the RMSE declines to 58.6 ug/m3.***"

Also on page 24, line 14-19 "***The dust emissions estimated using the assimilation can be used to drive a dust forecast. When the original PM10 observations were used in the assimilation, the forecast skill of the system actually decreased due to the strong overestimation of dust concentrations, the RMSE rose from averagely 230 (prior forecast) to 300 ug/m3. Better forecasts are obtained when using the model-based and especially the machine learning based bias-corrected observations. The RMSE of the former one was reduced to 200 ug/m3 while the RMSE of the latter one further declined to 150 ug/m3.***"

Minor comments
P4-L29. to large -> too large
Reply: "***to large extent***" is changed to "***to a large extent***"

P5-L2. Can easily applied -> can easily be applied
Reply: Accepted

P7-L3. Following Jin et al., 2018
Reply: Accepted

P9-L15. Anthropogenic emissions are from MEIC but it is not clear where natural emissions are obtained. Are these natural emissions; sea salt, biogenic emissions, and wildfires, but not dust?
Reply: To clarify the source of the nature emission, we added new remarks on page 9, line 16-17 "***Natural emissions included are the sea salts that are calculated online, biogenic emissions that are calculated online using the MEGAN model (Guenther et al., 2018), and wild fires which were taken from the operational GRAS product (Kaiser et al. 2012).***"

P10-L11 How LSTM is used needs little more details. It is not clear why these elements of the input vector have been chosen. Do these observations correspond to blue dots in Fig. 1? Why do you use nearby sites for PM2.5 but not for the other species? Is the training period only past 18 h or from January 2013 to March 2015 as it is said in P11-L4? Are these data available hourly? P11-L23 What criteria do you use to consider that a series has high data missing?
Reply: Yes, the LSTM indeed needs more clarifications.
Regarding to "***why these elements of inputs are selected***": We added new explanation on page 11, line 28-31 "***Statistical analysis tests are conducted which not only indicate a strong correlation between the non-dust PM10 and air quality measurements in the give sites, but also show that the predictor (non-dust PM10) is correlated to the observation indices (especially the PM2.5) at nearby sites.***"
Regarding to site map in Fig. 1: We added new remarks on page 6, line 4-5 "***At the present, the monitoring network has grown to 1,500 field stations covering all over China as shown in Fig. 1***"
Regarding to training period: Remarks are added on page 11, line 4-6 "***The training dataset covers the period from January 2013 to March 2015. In other words, the LSTM model L is trained to best fit the samples from this period. The two months April and May 2015 in which the studied dust event occurred is set as the testing period.***"
Regarding to *"Are these data available hourly"*: Yes, all the measurements are the hourly averages. We added new remarks to clarify this on page 10, line 25 "***hourly observations of PM2.5, SO2, NO2, O3 and CO from the ……***"
Regarding to the inputs of PM2.5 at nearby sites, data missing rate: We added more remarks on page 11, line 20-34 "***For the non-dust PM10 machine learning forecasts in a given site, observations from its***

nearby sites are also vital and are used in two ways. First, missing data records are unavoidable in an air quality monitoring network, while the LSTM model training requires an uninterrupted time series of features. In this study, data interpolations of air quality measurements (PM10, PM2.5, SO2, NO2, O3 and CO) are performed using both a linear interpolation and a k-Nearest-Neighbor algorithm (Zhang, 2012) if a site has no more than 30% of missing data. Otherwise, all the measurements in the given sites are abandoned. Generally, more information available from the nearby sites will result in a more accurate interpolation. Second, learning in the presence of data errors is pervasive in machine learning, and the measurements from nearby stations are used to limit their influence. Data errors occur due to incorrect sensor readings, software bugs in the data processing pipeline, or even the inaccurate data interpolation. Statistical analysis tests have been conducted which did not only indicate a strong correlation between the non-dust PM10 and air quality measurements in the given sites, but also show that the predictor (non-dust PM10) is correlated to the observation indices (especially the PM2.5) at its nearby sites. In order to eliminating errors caused by incorrect inputs at the modeling site, the measurements at the nearby stations are considered as the essential indices. In this study, a data instance will only be selected for training the LSTM model if there is at least one nearby site within an empirical radius 0.8\degree (approx 80 km), and a maximum of 3 nearby sites will be randomly selected where observation stations are densely distributed. To save the computation costs on machine learning model training, only the PM2.5 from the nearby sites are included as one of the inputs in this study."

P11-L4 I would like to see (or at least discuss) the sensitivity of the results to the training period. For instance, in the manuscript it is argued that from 01-2013 to 03-2015 the frequency of dust storm was low then it could be considered that all PM10 were nondust. However, yb will be significantly larger during the days with a dust storm and this could affect the regression model. Have you checked if the regression model changes if you remove those days with dust storm?

Reply: Yes, the dust records in the training period would indeed influence the machine learning model. However, the dust storm event we studied is reported to be the most severe one since 2002. Such a large-scale dust events has not been recorded in the training period (2013 to 2015). It is true that cities close to the deserts might have experienced several dust events with less pollutant levels. However, the learning rate on this single sample is set as 0.0001 in our machine learning algorithm. Therefore, these records would not have too much influence on the final machine learning model, whereas to identify these records and remove them takes a lot of efforts.

To clarify this, we also add some new remarks on page 11, line 7-14 by saying "***Dust storms themselves occur with very low frequency. To our knowledge, the studied dust event is the most severe one since 2002, and there are no such large-scale dust events recorded in our training period. Note that cities that are close to the Gobi and Mongolia deserts might have experienced several small- scale dust events with limited increase of dust concentrations. However, the machine learning tries to find the global best fits for the whole training dataset. The default learning rate, which determines the weights are updated during training, on a simple sample is 10^{-4} in our machine learning algorithm. Therefore, the PM10 records y^b are very close to the non-dust PM10 concentrations, and the rare dust event records are not excluded from the training dataset for convenience and for the expected little impact on the training result.***"

P12-L12 In non-dust PM10 evaluation, how do you know that they are dust in Fig. 3? Is it only based on larger numbers?

***Reply**:* We have modified the Fig.3 on page 13. Different markers are used to represent scatters in the dust period (April 14-16) and the rests in April 01 to May 30. The new plots indicate that most of the scatters in the bottom right corner are from the 3 days of the dust period while few parts are form the rest 57 days of non-dust period.

P11-L1 How do you know LE/non-dust underestimate non-dust PM10? Is it based in Fig. 3 and lower values of PM10?

Reply: Yes, according to the revised ***Fig.3*** and ***Fig.5***. It is clear that the LE/non-dust model underestimates the non-dust PM_{10}.

P11-L5 To show the better agreement of LSTM than LE/no-dust, could you give some number such as correlations?

Reply: The RMSEs of all the three non-dust simulations vs. measurements are given in the revised Fig.3 on page 13. They are also described when evaluating the non-dust simulation performance on page 12, line 19-27 by saying "***The CTM LOTOS-EUROS/ non-dust in general underestimates the non-dust PM10. The forecast results in a relatively large root mean square error (RMSE) 89.4 ug/m3. This could be explained from the fact not all types of particulate matters, such as secondary organic aerosols, are included in the model, and some aerosol emissions are very difficult to estimate (e.g., wood burning by households). The two LSTM forecasts show on average a good agreement with the observations. The RMSEs of the forecasts by the two machine learning models in the two years of training period are reduced to 55.9 and 60.7 ug/m3, and in the two months of test period (excluding the dust event from April 14 to 16) they also stay at comparable low levels of 58.6 and 60.2 ug/m3.***"
Also on page 24, line14-19 "***The two methods to estimate the non-dust part of the PM10 load have been validated. The simulations by the LOTOS-EUROS/non-dust model in general underestimate the PM10 concentrations. The root mean square error stays at a relative high level of 89.4 ug/m3. It is mainly caused by missing emissions and aerosol components such as secondary organic matter. In comparison, the data-driven machine learning model agrees more closely with the real measurements, the RMSE declines to 58.6 ug/m3.***"

P13-L25. Is a high level of pollution the reason why these sites are chosen to test the performance of LE/non-dust?

Reply: To clearly explain why we choose these six cities, we added some new remarks on page 14, line 25-27 "***These cities were selected because they all experienced a severe pollution and illustrated the general performance of the LOTOS-EUROS/non-dust and LSTM methods. In addition, each of these cities have at least 4 monitoring sites which assured a high accuracy.***"

P20-L15 In addition to the orography, a reason why in Beijing the PM10 is underestimated could be related to higher PM10 non-dust values? Did you find any relationship.

Reply: We agree with the reviewer about this point. Explanation is added on page 20, line 34; on page 22, line 1-2 "***This suggests that the simulation model simply is not able to increase the dust concentrations here, for example because of uncertainties in the meteorological data, a removal of dust that is too efficient, or because some local sources of dust are absent (equally, non-dust PM10 levels are underestimated).***"

In the section of the evaluation of the forecast skill, I miss comparing with "real" dust records instead of PM10.

Reply: Since we never have the "real" dust records/measurements. What we compare is the difference between the full aerosol simulation (dust + non-dust aerosol simulation) and PM10 observations. It is now pointed out on page 22, line 10-12 by saying "***To evaluate the forecast skill of the assimilation(s), the root mean square error (RMSE) of the reference and three posterior full aerosol simulations (dust forecasts plus non-dust predictions) with respect to the observed PM10 over the whole observation sites has been computed for each hour.***"

Figures, in general, are clear and the captions properly describe them.
Fig.1 It is difficult to distinguish the star symbols which denote the cities location. Also,
Reply: We now use double-size markers for the six cities.

Is N=1352 number of stations used in LSTM?
Reply: Yes. We also add "***LSTM based non-dust PM10 forecast are performed only in stations of blue dot (N=1351)***" in the caption of Fig.1 on page 6.

Fig. 3 Use properly notation for float numbers (e.g. 2 10-6 instead of 0.00002).
Reply: Fig. 3 is modified. See new version on page 13.

Fig. 6 Unify the legends and try it does not overlap with the time series.
Reply: Accepted. Legends in ***Fig.5*** on page 16 are modified.

Machine learning for observation bias correction with application to dust storm data assimilation

Jianbing Jin[1], Hai Xiang Lin[1], Arjo Segers[2], Yu Xie[1], and Arnold Heemink[1]
[1]Delft Institute of Applied Mathematics, Delft University of Technology, Delft, the Netherlands
[2]TNO, Department of Climate, Air and Sustainability, Utrecht, The Netherlands
Correspondence: Jianbing Jin (J.Jin-2@tudelft.nl)

Abstract. Data assimilation algorithms rely on a basic assumption of an unbiased observation error. However, the presence of inconsistent measurements with nontrivial biases or inseparable baselines is unavoidable in practice. Assimilation analysis might diverge from reality, since the data assimilation itself cannot distinguish whether the differences between model simulations and observations are due to the biased observations or model deficiencies. Unfortunately, modeling of observation biases or baselines which show strong spatiotemporal variability is a challenging task. In this study, we report how data-driven machine learning can be used to perform observation bias correction for data assimilation through a real application, which is the dust emission inversion using PM_{10} observations.

PM_{10} observations are considered as unbiased, however, a bias correction is necessary if they are used as a proxy for dust during dust storms since they actually represent a sum of dust particles and *non-dust* aerosols. Two observation bias correction methods have been designed in order to use PM_{10} measurements as proxy for the dust storm loads under severe dust conditions. The first one is the conventional chemical transport (CTM) model that simulates life cycles of *non-dust* aerosols. The other one is the machine learning model that describes the relations between the regular PM_{10} and other air quality measurement. The latter is trained by learning using two years of historical samples. The machine learning based *non-dust* model is shown to be in better agreements with observations compared to the CTM. The dust emission inversion tests have been performed, either through assimilating the raw measurements, or the bias-corrected dust observations using either the CTM or machine learning model. The emission field, surface dust concentration and forecast skill are evaluated. The worst case is when we directly assimilate the original observations. The forecasts driven by the posterior emission in this case even results in larger errors than the reference prediction. This shows the necessities of bias correction in data assimilation. The best results are obtained when using the machine learning model for bias correction, with the existing measurements used more precisely and the resulting forecasts close to reality.

1 Introduction

For centuries, East Asia experienced regular dust storms in the spring time. Those dust events mainly originated from the dust source regions of the Gobi and Taklamakan deserts. Annually, thousands tons of "yellow sands" are blown

eastward over the densely populated areas in China, the Korean peninsula, and Japan by the prevailing winds. Dust storms can also carry irritating spores, bacteria, viruses and persistent organic pollutants (WMO, 2017). Next to the human health, the resulting low visibility can cause a severe disruption of transportation systems. For example, more than 1,100 flights have been delayed/canceled in Beijing after the city was struck by a choking dust storm in early May 2017.

A large number of dust simulations models has been developed over the past decades (Wang et al., 2000; Gong et al., 2003; Liu et al., 2003). These chemical transport models help to understand the life cycles of the dust storms, and are also used for dust forecasts and to aid early warning systems. Apart from advances in simulation of dust storms, progress has also been made in the monitoring of dust or general aerosol loads. Field station networks are constructed to observe the in-situ particulate matter (PM) levels over densely populated regions (Li et al., 2017a). Ground-based sun photometers, e.g., the global Aerosol Robotic Network (AERONET) (Cesnulyte et al., 2014), are widely used to monitor column-integrated aerosol profiles. Satellite onboard instruments such as Moderate Resolution Imaging Spectroradiometer (MODIS) (Remer et al., 2005), Cloud-Aerosol Lidar and Infrared Pathfinder Satellite Observations (CALIPSO) (Sekiyama et al., 2010), and Advanced Himawari Imager/Himawari-8 (Yoshida et al., 2018) provide the measurements of airborne particles with further wide coverages. These measurements could be used to calibrate the parametrization in the dust simulation models and to evaluate their ability to forecast dust concentrations. Moreover, the observations could be combined with a dust modeling system through data assimilation to improve the forecast skills.

A wide variety of data assimilation techniques have been used with dust simulation models, including variational methods (Yumimoto et al., 2008; Niu et al., 2008; Gong and Zhang, 2008; Jin et al., 2018) and ensemble-based sequential methods (Lin et al., 2008; Sekiyama et al., 2010; Khade et al., 2013; Di Tomaso et al., 2017). In these systems, the available observations are either used to estimate the model states (dust concentrations) or to reduce uncertainties in the emissions and/or other model parameters. Challenges for dust assimilations include development of more and more accurate dust simulations, and use of new types of observations including vertical profiles from Lidars and latest satellite observations. A further challenge for any assimilation system is the proper definition of the observation and representation errors, as well as characterization of biases.

In general, the commonly used data assimilation schemes all rely on the basic assumption of an unbiased observation. In real applications, however, measurement biases are often unavoidable. In the presence of biases, it is impossible to determine whether a difference between an *a prior* simulation and an observation are due to the biased observations or model deficiencies. The biases might lead to assimilations that diverge from reality (Lorente-Plazas and Hacker, 2017). A well known example of observation biases is in radiance observation assimilation systems in presence of clouds (Eyre, 2016; Berry and Harlim, 2017). To avoid problems with these biases, up to 99% of cloudy observed measurements are discarded although they may also contain valuable information. If dust storms are coincident with clouds, it is also possible that in satellite retrieval algorithms clouds are mistaken for dust, leading to strong biases in the data to be assimilated (Jin et al., 2019).

Another example where observation biases are important is when ground-based PM_{10} measurements are assimilated in dust simulation models. Due to the high temporal resolutions and the rather dense observation network, the ground-based air quality observing network has become a powerful source of measurements on dust aerosols. The records, mainly the PM_{10} feature, were widely used to calibrate, assess or estimate the dust model (Lin et al., 2008; Wang et al., 2008; Huneeus et al., 2011; Yumimoto et al., 2016; Benedetti et al., 2018). However, the observed PM_{10} concentrations do not only consist of dusts, but are actually the sum of the dust and other regular particles. The latter one are emitted not only from anthropogenic activities such as industries, vehicles, and households, but also from natural sources such as wild fires and sea spray. In this paper we will simply refer to these particles as the *non-dust* fraction of the total PM_{10}. The concentrations of *non-dust* aerosols in urbanized areas could be substantial, reaching values up to 500 $\mu g/m^3$ (Shao et al., 2018).

Although PM_{10} observations include a nontrivial bias, the wide spread availability makes them still useful in dust storm assimilation system. During dust storm events, extreme high peaks of more than 1000-2000 $\mu g/m^3$ PM_{10} are recorded which can be attributed mainly to dust. If these would be assimilated directly in dust simulation model, ignoring the fact that at least some part represents *non-dust*, the assimilation system would diverge to states that overestimate the dust load. In case of less severe dust events, the dust analysis divergence would then become extremely critical.

However, modeling of observation biases is very challenging when they have strong spatial and temporal variabilities. Little progress has been made in bias correction of full-aerosol measurements for their use in dust storm data assimilation. Lin et al. (2008) selected only PM_{10} observations for assimilation when at least one occurrence of dust clouds was reported by the local stations. In Jin et al. (2018), it was found that on sites with both PM_{10} and $PM_{2.5}$ observations, only the PM_{10} concentration increased during a dust episode, while the $PM_{2.5}$ concentrations were not affected and remained at a constant level. Besides, Xu et al. (2017) and Jin et al. (2018) suggested a strong correlation between $PM_{2.5}$ and *non-dust* PM_{10}. Therefore, a very simple *non-dust* PM_{10} baseline removal (called observation bias correction) was proposed, in which the available $PM_{2.5}$ was used to approximate the *non-dust* PM_{10} (or baseline) during a dust event by:

$$PM_{10}^{non\text{-}dust} = b + r \times PM_{2.5} \tag{1}$$

where the b and $r > 1$ are linear regression parameters based on a 24-hour history of measurements before arrival of the dust storm. The aforementioned methods either exclude a selection of the measurements, which may still contain useful information, or work under ideal circumstance only when a simple correlation $\mathcal{R}$ between PM_{10} and $PM_{2.5}$ is valid. For instance, in the dust event studied in Jin et al. (2018) the application of Eq.1 in many sites failed since R is weak. To have a quality-assured bias correction, Eq.1 is performed only when the *Pearson* correlation coefficient $\mathcal{R} > 0.8$. Consequently, measurements in around 45% sites are rejected in that case. To fully exploit the dust information present in total PM observations, a more advanced method is needed. In this paper we proposed two methods, either using a conventional chemistry transport model, or a machine learning model.

A chemistry transport model (CTM) implements all available knowledge on emission, transport, deposition, and other physical processes in order to simulate concentrations of trace gases and, important here, aerosols. Daily air quality forecasts are often provided using such CTMs. A simulation model for dust storm events is usually just a CTM with all tracers removed except dust; by using the full CTM, an estimate of the *non-dust* part of the aerosol load could be made. In this study, the LOTOS-EUROS CTM is used to simulate the dust as well as the *non-dust* aerosol concentrations. If the non-dust model was perfect, the difference between simulation and observed PM_{10} would be unbiased, and assimilation could be applied to the combined dust and *non-dust* concentrations. In case of a dust storm event, it remains necessary to distinguish between the dust and *non-dust* part of the simulations since the two parts will have very different error characteristics. The dust part is quickly varying and has a large uncertainty, while the *non-dust* part is more smooth but very persistent in time and has a relatively small uncertainty. An assimilation system on the combined simulations should be able to handle these differences. However, the error attribution to their proper sources (dust and *non-dust* error) then becomes extremely critical as explained in Section 2.4. Since this paper focuses on dust during a severe event only, we will not explore the error characteristics of the *non-dust* part of the model. Therefore we will not apply an assimilation on the combined aerosol (dust and *non-dust*) model. Instead, the *non-dust* simulations will solely be used to remove the *non-dust* baseline from PM_{10} observations.

Similar to the air quality forecast, the accuracy of a CTM for *non-dust* aerosols is hampered by lack of accurate input data. For example, the timely update of anthropogenic emission inventories is always a key issue for air quality forecasts. With the ever-increasing complexity and resolution, the CTMs are now becoming highly nonlinear and time-consuming. However, they may still not be able to identify explicit representations of the non-dust aerosol dynamics, especially regarding fine-scale processes.

In addition to the conventional CTM, we propose a new method for removing the *non-dust* part of the PM_{10} observations which is based on machine learning (ML). Data-driven methods have already been proved to be a powerful tool to provide air quality forecasts for horizons of a few days, (e.g., Li et al. (2016); Fan et al. (2017); Li et al. (2017b); Chen et al. (2018)). Different from the chemical transport models which simulate the aerosol physical processes, machine learning models describe mathematical relations of input-output and trained by learning a large number of samples from historical records. Our machine learning system used a neural network, namely *long short term memory* (LSTM). The input is formed by air quality indices for a number of relevant tracers (PM2.5, SO_2, NO_2, CO, and O_3), as well as meteorology data. The output of the system is an estimate of the *non-dust* PM_{10} concentration. The input features are to a large extent independent of the dust storms, even the $PM_{2.5}$ concentrations as shown in Jin et al. (2018); observations of PM_{10} are excluded since excessive dust loads are visible mainly in this component. Recent development and the availability of open source machine learning tools provide a good opportunity to estimate the air quality indices using a data-driven machine learning models.

Whereas these are previous studies on dust storm data assimilation using various kinds of combined aerosol measurements, we are the first to investigate the necessities of bias correction for these full-aerosol observations in order to use them as 'real' dust measurements in a dust storm assimilation system. The adding values of observation

bias correction in dust emission inversion is explored through the ground-based PM_{10} measurement assimilation. It can easily be applied to others general applications, e.g., remote sensing data assimilation. Our contributions are threefold. Firstly, we present and examine the conventional CTM for removing the *non-dust* part from PM_{10} observations. Secondly, we design and examine a novel machine learning based bias correction which is data-driven and free of the time-consuming numerical CTMs. Thirdly, we evaluate the two *non-dust* aerosol model simulations by comparing to the PM_{10} measurements during regular periods (rare dust events involved); we evaluate dust emission fields, surface dust concentration simulation and forecast skills which are obtained by either assimilating the raw PM_{10} data, or bias-corrected measurements either using the CTM or machine learning model.

The paper is organized as follows. A brief description of our dust simulation model (LOTOS-EUROS/Dust) and the four dimensional variational data assimilation method for emission inversion are presented in Section 2. The biased observation representing error and its influence on the assimilation system are also explained. The two bias correction methods, the *non-dust* aerosol regional chemical transport model and a machine learning model, are discussed and the bias simulation is evaluated in Section 3. Section 4 reports the assimilation results using the two bias correction methods, and evaluates the forecast skills using independent measurements. Section 5 discusses the necessities of observation bias correction in assimilation works, highlights our key contributions.

2 Dust storm data assimilation system

2.1 Dust model

The dust storm event studied in this paper took place in East Asia in April 2015, and has already been used as a test case for assimilation experiments in Jin et al. (2018). The LOTOS-EUROS/Dust simulation model is used with similar configurations to our previous studies, which is configured on a domain from 15°N to 50°N and 70°E to 140°E, but with a higher model resolution of 0.25°. The model is driven by European Center for Medium-Ranged Weather Forecast (ECMWF) operational forecasts for horizons of 3-12 hours. The dust load is described by 5 aerosol bins within a diameter range 0.01 μm $< D_p <$ 10 μm. Physical processes included are emission, advection, diffusion, dry and wet deposition, and sedimentation. The dust emission scheme implemented in LOTOS-EUROS is mainly based on the formulation of horizontal saltation flux (Marticorena and Bergametti, 1995) and sandblasting efficiency (Shao et al., 1996). A terrain preference parameter F_{ps} was used in the dust emission in Jin et al. (2018). This geographic dependent parameter was first introduced by Ginoux et al. (2001), and used to approximate the probability of having accumulated sediments that can be resuspended. In this work, F_{ps} is disabled since the preference factor was found to limit the emission rate in some regions where the fine-scale topographic feature is actually unknown. Snapshots of a reference simulation of the dust episode has been performed and is shown in Fig.8(a).

2.2 Observation network

The China Ministry of Environmental Protection (MEP) has commenced to release the hourly-average measurements of atmospheric constituents including $PM_{2.5}$, PM_{10}, CO, O_3, NO_2 and SO_2 since 2013. A huge number of ground stations measuring these air quality indices have been established in densely populated areas. At the present, the monitoring network has grown to 1,500 field stations covering all over China as shown in Fig.1.

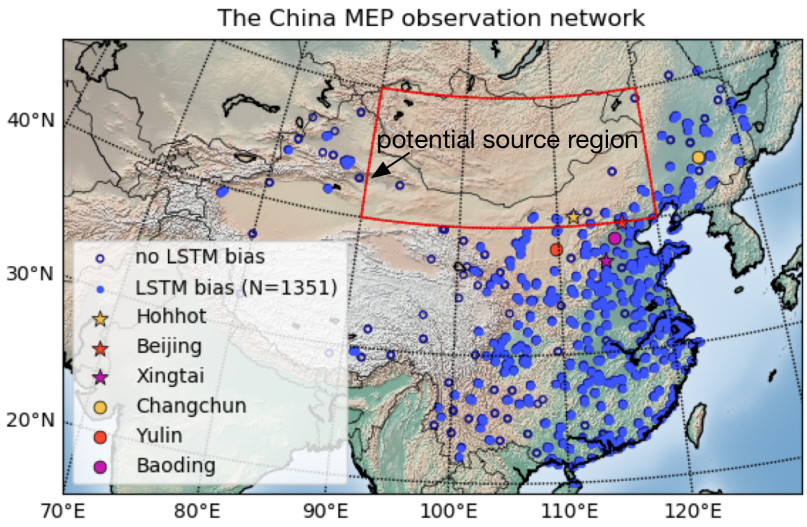

Figure 1. The China MEP air quality monitoring network and the potential dust storm source region. LSTM based *non-dust* PM_{10} forecast are performed only in stations of blue dot (N=1351), while ones of black circles are skipped.

2.3 Reduced tangent linearization 4DVar

The assimilation system, which will be used to combine bias-corrected PM_{10} observations with simulations, is based on a reduced-tangent-linearization four dimensional variational (4DVar) data assimilation. The goal of a 4DVar technique is to find the maximum likelihood estimation of a state vector, which is here the dust emission field $\boldsymbol{f}$, given the available observations over a time window. A common approach is to use an incremental formulation, which aims to find the optimal emission deviation $\delta\boldsymbol{f}$ as the minimum of the cost function:

$$J(\delta\boldsymbol{f}) \;=\; \frac{1}{2}\,\delta\boldsymbol{f}\;\mathbf{B}^{-1}\;\delta\boldsymbol{f} \;+\; \frac{1}{2}\sum_{i=1}^{k}\left(\mathbf{H}_i\mathbf{M}_i\delta\boldsymbol{f}+\boldsymbol{d}_i\right)^T\mathbf{R}_i^{-1}\left(\mathbf{H}_i\mathbf{M}_i\delta\boldsymbol{f}+\boldsymbol{d}_i\right) \tag{2}$$

where k is the number of time steps within the assimilation window. The vector $\delta\boldsymbol{f}$ denotes a perturbation of the emissions with respect to the background one. For an observation time i, the innovation vector (length m_i) is defined

as the difference between the simulations and observations:

$$\boldsymbol{d}_i = \mathcal{H}_i(\ \mathcal{M}_i(\boldsymbol{f})\ )\ -\ \boldsymbol{y}_i \tag{3}$$

where $\mathcal{M}_i$ denotes the LOTOS-EUROS/Dust transport model, $\mathcal{H}_i$ is the operator that converts state variables into observation space, and $\boldsymbol{y}_i$ is the vector with dust observations at this time step i. The operators $\mathbf{H}_i$ and $\mathbf{M}_i$ denote linearizations of $\mathcal{H}_i$ and $\mathcal{M}_i$ around the reference emission vector $\boldsymbol{f}_b$. Following Jin et al. (2018), the errors in dust emission field were assumed to be only caused by the uncertainty in the friction velocity threshold in the dust wind-blown parametrization, and similar assumptions on the uncertainty are used to build an emission error covariance $\mathbf{B}$. The friction velocity threshold is perturbed with a spatially varying multiplicative factor β. β is configured with a mean of 1 and a standard deviation of 0.1. In addition, an exponential profile of distance-based spatial correlation is posed on βs (Jin et al., 2018). The observation error term is weighted by an observation error covariance $\mathbf{R}$, for which the individual elements will be described in Section 4.1.

To reduce the computational cost in calculating the *tangent linear model* $\mathbf{M}_i$, a reduced-tangent-linearized 4DVar (Jin et al., 2018, 2019) is used. The simplified method is based on proper orthogonal decomposition (POD) of the background covariance $\mathbf{B}$ which efficiently carries out model reduction by identifying the few most energetic modes:

$$\begin{aligned} \mathbf{B}\ &=\ \mathbf{U}\mathbf{U}^T\ \approx\ \tilde{\mathbf{U}}\tilde{\mathbf{U}}^T \\ \delta\boldsymbol{f}\ &\approx\ \tilde{\mathbf{U}}\ \delta\boldsymbol{w} \end{aligned} \tag{4}$$

where $\mathbf{U} \in \mathbf{R}^{P\times P}$ is the background emission covariance square root, with P the size of the emission field of $O(10^4)$ elements. while $\tilde{\mathbf{U}} \in \mathbf{R}^{P\times p}$ is the truncation of $\mathbf{U}$ based on POD, with p the reduced rank size of $O(10^2)$. The vector $\delta\boldsymbol{w} \in \mathbf{R}^p$ stores the transformed control variables.

The cost function of the reduced-tangent-linearization 4DVar is formulated as:

$$J(\delta\boldsymbol{w})\ =\ \frac{1}{2}\delta\boldsymbol{w}^T\delta\boldsymbol{w}\ +\ \frac{1}{2}\sum_{i=1}^{k}\left(\mathbf{H}_i\tilde{\mathbf{M}}_i\tilde{\mathbf{U}}\delta\boldsymbol{w}+\boldsymbol{d}_i\right)^T\mathbf{R}_i^{-1}\left(\mathbf{H}_i\tilde{\mathbf{M}}_i\tilde{\mathbf{U}}\delta\boldsymbol{w}+\boldsymbol{d}_i\right) \tag{5}$$

where $\tilde{\mathbf{M}}_i$ denotes the reduced tangent linear model with a rank p, which is approximated using the perturbation method. More details about the reduced-tangent-linearization 4DVar algorithm can be found in Jin et al. (2018).

2.4 Biased observation representing error

In real applications, the observations inevitably have biases which cannot be attributed to the model simulation, as following:

$$\boldsymbol{y}_i = \mathcal{H}_i(\ \mathcal{M}_i(\boldsymbol{f})\ )\ +\ \boldsymbol{b}_i\ +\ \boldsymbol{\sigma}_i \tag{6}$$

where $\boldsymbol{\sigma}_i$ is the vector of Gaussian distributed observation errors which have zero means and a known covariance matrix $\mathbf{R}_i$, and $\boldsymbol{b}_i$ denotes the vector of observation bias. In our application, the vector $\boldsymbol{y}_i$ contains the observed

PM_{10} concentrations, while the aerosols released in the local anthropogenic activities and other *non-dust* related processes are referred as $\boldsymbol{b}_i$. Note that the PM_{10} measurements themselves might also contain 'native' biases due to the incorrect sensor reading or systematic errors. However, this part of the bias in the PM_{10} observations is unknown and not considered in this study.

In the course of data assimilation, it is impossible to determine whether the departures ($\boldsymbol{d}_i$) of the prior simulations from the observations are due to the biased observations $\boldsymbol{b}_i$ or emission errors $\delta\boldsymbol{f}$. Thus, the assimilation result will diverge from the true state when a bias is present. In complex dynamic models as the atmospheric transport model, the biases (*non-dust* aerosols) could have high spatial and temporal variabilities and is therefore difficult to quantify.

In this work, we proposed two methods to quantify the bias levels for the observation bias correction. The first one is the *non-dust* parts of LOTOS-EUROS chemical transport model (CTM) which simulates the aerosol life cycles including emission, transport and deposition. The second method is to describe the *non-dust* aerosol levels using a data-driven machine machine model. Details of these two methods are illustrated in Section 3.

In fact, both LOTOS-EUROS CTM and machine learning model are imperfect, and some biases might still exist after the correction. The former one is known to be limited by errors in the emission inventories, meteorological forecasts and all kinds of input sources. The latter is then hampered by the deficiency of the type model (e.g., insufficient to represent the complexity of the phenomenon), inadequate amount of training data. However, by combining the bias-corrected observation with the dust model, the assimilation will adapt to posteriors which are more close to reality.

There were a few studies that addressed both the model deficiency and uncertainty in observation bias simultaneously using either variational data assimilation (Dee and Uppala, 2009) or sequential filters (Dee, 2005; Lorente-Plazas and Hacker, 2017). Those assimilation schemes not only require a formulation of a model for the bias, but also need a quality-assured reference to describe the uncertainty of the bias model. The need to attribute errors to their proper sources is obviously a key part in any assimilation systems, but becomes especially critical when it involves bias correction. This is because a wrong error attribution will force the assimilation to be consistent with a biased source. If the source of a known bias is uncertain, assimilation without considering the uncertainty of bias model is the safest option (Dee, 2005). Therefore, these two *non-dust* models are solely set as references for the bias, and the uncertainties are not explored here.

2.5 Assimilation Window

Fig.2 shows a time line for the assimilation experiment around the April 2015 dust event, which is very similar to what was used in Jin et al. (2018). The dust event has a short duration, and therefore only a single assimilation window with a length of 36 hours is used. The dust emissions take place at the start of the window, while the observations become available at the end of the window since they are located downwind from the source region (see Fig.1). A long assimilation window is therefore necessary in order to estimate the correct emission parameters given the observations.

When we perform the assimilation analysis at April 15, 19:00, only the dust observations from April 15, 08:00 to 19:00 will be assimilated and they are calculated by subtracting the *non-dust* part (CTM based or ML based) from the PM_{10} observations. After the analysis, the simulation model is used to perform a dust forecast for the next 12 hours using the newly-estimated emission parameters. A full-aerosol PM_{10} forecast will then be calculated by adding the dust forecast and *non-dust* aerosol forecast, where the later again originates from either the CTM and machine learning model.

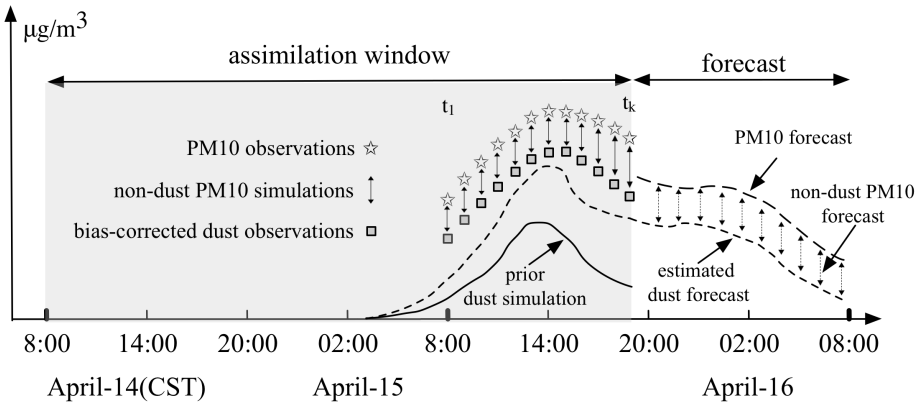

Figure 2. Timeline of observation availability, assimilation cycles and forecasts

3 Observation bias correction methods

Two systems are introduced to correct the *non-dust* bias when using PM_{10} observations in a dust assimilation. The first one is CTM LOTOS-EUROS/*non-dust* model that simulates the physical processes of the *non-dust* aerosols. The latter is the machine learning model that estimates the *non-dust* aerosol based on historical records. The following sections describe the two methods in more detail.

3.1 Chemistry transport model (LOTOS-EUROS/*non-dust*)

The regional CTM LOTOS-EUROS/*non-dust* is configured similar to the LOTOS-EUROS/Dust used in the assimilation, but now includes all trace gases and *non-dust* aerosols. The configuration is similar to what is used for daily air quality simulations over China as described in (Timmermans et al., 2017). Anthropogenic emissions are taken from the Multi-resolution Emission Inventory for China (MEIC) inventory (*http://www.meicmodel.org*). Natural emissions included are the sea salts that are calculated online, biogenic emissions that are calculated online using the MEGAN model (Guenther et al., 2006), and wild fires which were taken from the operational GRAS product

(Kaiser et al., 2012). The LOTOS-EUROS full aerosol operational forecast over this modeling domain are released via the MarcoPolo-Panda projects through *(www.marcopolo-panda.eu)*.

The operational CTM Lotos-Euros over China is in its early phase of development as well as the other six CTMs used in the MarcoPolo-Panda project. The purpose of that project is to diagnose statistical differences between the ensemble model simulations and observations. An important objective is to determine ways by which the models can be improved. These differences are mostly attributed to inaccuracy in the weather forecast and errors in the adopted surface emissions (Brasseur et al., 2019; Petersen et al., 2019). Indeed, there is room for minimizing the forecast-observation differences using nudging methods like data assimilation, which requires considerable efforts and not yet exploited in that study.

3.2 Machine learning for *non-dust* PM_{10} simulation

Given a set of training data, a machine learning algorithm attempts to find the relation between input and output. When a proper model is used, the machine learning algorithm can lean to reproduce the complex behaviors of a dynamic system. The description is purely based on the data, physical knowledge is not included. Machine learning algorithms are popular tools to forecast the air quality indices using the history records (Li et al., 2016; Fan et al., 2017; Chen et al., 2018; Lin et al., 2019). In this study, the machine learning algorithm used is the *long short term memory* (LSTM) neural network, which has demonstrated its ability in predicting time series problems (Li et al., 2017b).

The LSTM operator $\mathcal{L}$, which is configured with parameters $\boldsymbol{\theta}$, for predicting *non-dust* PM_{10} can be described as:

$$\boldsymbol{b}^{t_0+t} = \mathcal{L}_{\boldsymbol{\theta}}(\boldsymbol{x}^{t_0},\ \boldsymbol{x}^{t_0-1},\ ...,\ \boldsymbol{x}^{t_0-m+1}) \tag{7}$$

where $\boldsymbol{b}^{t_0+t}$ represents the predictor, which is in this study the *non-dust* PM_{10} concentration forecast t hours in advance. The temporal correlation between the input and output features declines when t increases. In our system, the maximum forecast period t is 12 hours. The input vectors $\boldsymbol{x}^{t_0}, \boldsymbol{x}^{t_0-1}, ..., \boldsymbol{x}^{t_0-m+1}$ are the observed data of the past m hours, which is set as 18 hours empirically. The input vectors consist of:

- hourly observations of $PM_{2.5}$, SO_2, NO_2, O_3, and CO from the ground based air quality network described in Section 2.2;
- observations of $PM_{2.5}$ at the nearby sites;
- local meteorological data (temperature and dew point at 2 m, wind speed at 10 m) which are taken from the LOTOS-EUROS model input and originate from the European Center for Medium-Ranged Weather Forecast (ECMWF).

The LSTM neural network parameters $\boldsymbol{\theta}$ are determined by minimizing the objective function $J_{\boldsymbol{\theta}}$ that represents the mean squared error of predictors $\boldsymbol{b}$ with respect to the measured values $\boldsymbol{y}^b$:

$$J_\theta = \frac{1}{m} \sum_{i=1}^{m} (\boldsymbol{b}_i - \boldsymbol{y}_i^b)^2 \tag{8}$$

The training dataset covers the period from January 2013 to March 2015. In other words, the LSTM model $\mathcal{L}$ is trained to best fit the samples from this period. The two months April and May 2015 in which the studied dust event occurred is set as the testing period.

Dust storms themselves occur with very low frequency. To our knowledge, the studied dust event is the most severe one since 2002, and there are no such large-scale dust events recorded in our training period. Note that cities that are close to the Gobi and Mongolia deserts might have experienced several small-scale dust events with limited increase of dust concentrations. However, the machine learning tries to find the global best fits for the whole training dataset. The default learning rate, which determines the weights are updated during training, on a simple sample is 10^{-4} in our machine learning algorithm. Therefore, the PM_{10} records $\boldsymbol{y}^b$ are very close to the *non-dust* PM_{10} concentrations, and the rare dust event records are not excluded from the training dataset for convenience and for the expected little impact on the training result. The regression model $\mathcal{L}$ is thus assumed to reflect only the relation between input features and the *non-dust* PM_{10}.

Note that including PM_{10} observations in the series of input vectors will certainly improve the skill of the machine learning forecasts. However, the LSTM model would then lack the ability to discriminate between the dust and *non-dust* fractions in PM_{10} during a dust event. Earlier studies showed that the input variables, including $PM_{2.5}$, are independent on the dust storm as illustrated in Jin et al. (2018).

For the non-dust PM_{10} machine learning forecasts in a given site, observations from its nearby sites are also vital and are used in two ways. First, missing data records are unavoidable in an air quality monitoring network, while the LSTM model training requires an uninterrupted time series of features. In this study, data interpolations of air quality measurements (PM_{10}, $PM_{2.5}$, SO_2, NO_2, O_3 and CO) are performed using both a linear interpolation and a k-Nearest-Neighbor algorithm (Zhang, 2012) if a site has no more than 30% of missing data. Otherwise, all the measurements in the given sites are abandoned. Generally, more information available from the nearby sites will result in a more accurate interpolation. Second, learning in the presence of data errors is pervasive in machine learning, and the measurements from nearby stations are used to limit their influence. Data errors occur due to incorrect sensor readings, software bugs in the data processing pipeline, or even the inaccurate data interpolation. Statistical analysis tests have been conducted which did not only indicate a strong correlation between the non-dust PM_{10} and air quality measurements in the given sites, but also show that the predictor (non-dust PM_{10}) is correlated to the observation indices (especially the $PM_{2.5}$) at its nearby sites. In order to eliminating errors caused by incorrect inputs at the modeling site, the measurements at the nearby stations are considered as the essential indices. In this study, a data instance will only be selected for training the LSTM model if there is at least one nearby site within an empirical radius 0.8°(approx 80 km), and a maximum of 3 nearby sites will be randomly selected where observation

stations are densely distributed. To save the computation costs on machine learning model training, only the $PM_{2.5}$ from the nearby sites are included as one of the inputs in this study.

The machine learning model for non-dust PM_{10} forecast is trained site by site, with the hyper-parameters shown in Table 1. With the following hyper-parameters, the machine learning model training takes several minutes for each site. The training in each site is independent, hence, the whole workload is highly parallelizable.

Table 1. LSTM hyper-parameters.

LSTM layers	neurons per layer	epochs	batch size	forecast length (hours)
2	30	50	64	0 or 12

Fig.1 presents the original field observation network (N≈1500) established by the China Ministry of Environmental Protection (MEP) up to 2018, as well as the sites (N=1351) where LSTM based *non-dust* forecasts are performed. It is clear that the LSTM forecast cannot be performed in each monitoring site. A part of the sites is skipped due to the lack of nearby sites, the rest are caused by high data missing rate in the training period.

3.3 Evaluation of *non-dust* PM_{10} bias corrections

Our two bias models, LOTOS-EUROS/*non-dust* and LSTM, could both be used for air quality forecast operationally when there is no dust storm. Once a dust storm is observed, the dust emission inversion system will be enabled, the two non-dust PM_{10} models will then be used in dust observation bias correction. The forecasts are expected to have a good performance when dust is not present, and to underestimate the PM_{10} levels in case of dust storms.

Both the CTM LOTOS-EUROS and LSTM are tested to forecast *non-dust* PM_{10} over April-May 2015. This period includes the 2 to 3 days dust event that is used as test case for the assimilation. Fig.3(a)~(c) show density plots comparing PM_{10} observations with either LOTOS-EUROS/*non-dust* forecasts, or with LSTM forecast 0 hour and 12 hours in advance.

The CTM LOTOS-EUROS/*non-dust* in general underestimates the *non-dust* PM_{10}. The forecast results in a relatively large root mean square error (RMSE) 89.4 $\mu g/m^3$. This could be explained from the fact not all types of particulate matters, such as secondary organic aerosols, are included in the model, and some aerosol emissions are very difficult to estimate (e.g., wood burning by households). The two LSTM forecasts show on average a good agreement with the observations. The RMSEs of the forecasts by the two machine learning models in the two years of training period are reduced to 55.9 and 60.7 $\mu g/m^3$, and in the two months of test period (excluding the dust event from April 14 to 16) they also stay at comparable low levels of 58.6 and 60.2 $\mu g/m^3$. As expected, a smaller forecast period t=0 hour gives a better result than the forecast over 12 hours.

The scatters in the dust period (April 14 to 16) are denoted using different markers in Fig.3. The underestimation of PM_{10} during the dust period (April 14 to 16) is visible in the bottom right corners of these plots.

When we perform the assimilation analysis at April 15, 19:00, the short period of t=0 hour forecast will be treated as the *non-dust* levels in the bias correction of the original PM_{10} measurements. Note that here t=0 forecasts denote

the forecasts valid at each specific snapshot of the observations, while the 12 hours forecasts are valid 12 hours in advance, e.g., the non-dust PM_{10} forecast (12 h) at April 16 07:00 is valid at April 15 19:00. Subsequently, the bias-corrected data are used to estimate the dust emissions over the past 36-hour window. Obviously, one important aim of the assimilation is to make a better forecast, in this study, the forecast skills will be evaluated in the following 12 hours from April 15, 19:00. Besides, the forecast is assessed by comparing the combined PM_{10} forecasts to PM_{10} observations. The LSTM forecast with t=12 hours in advance will be added to the dust storm forecast to build the combined aerosol forecast.

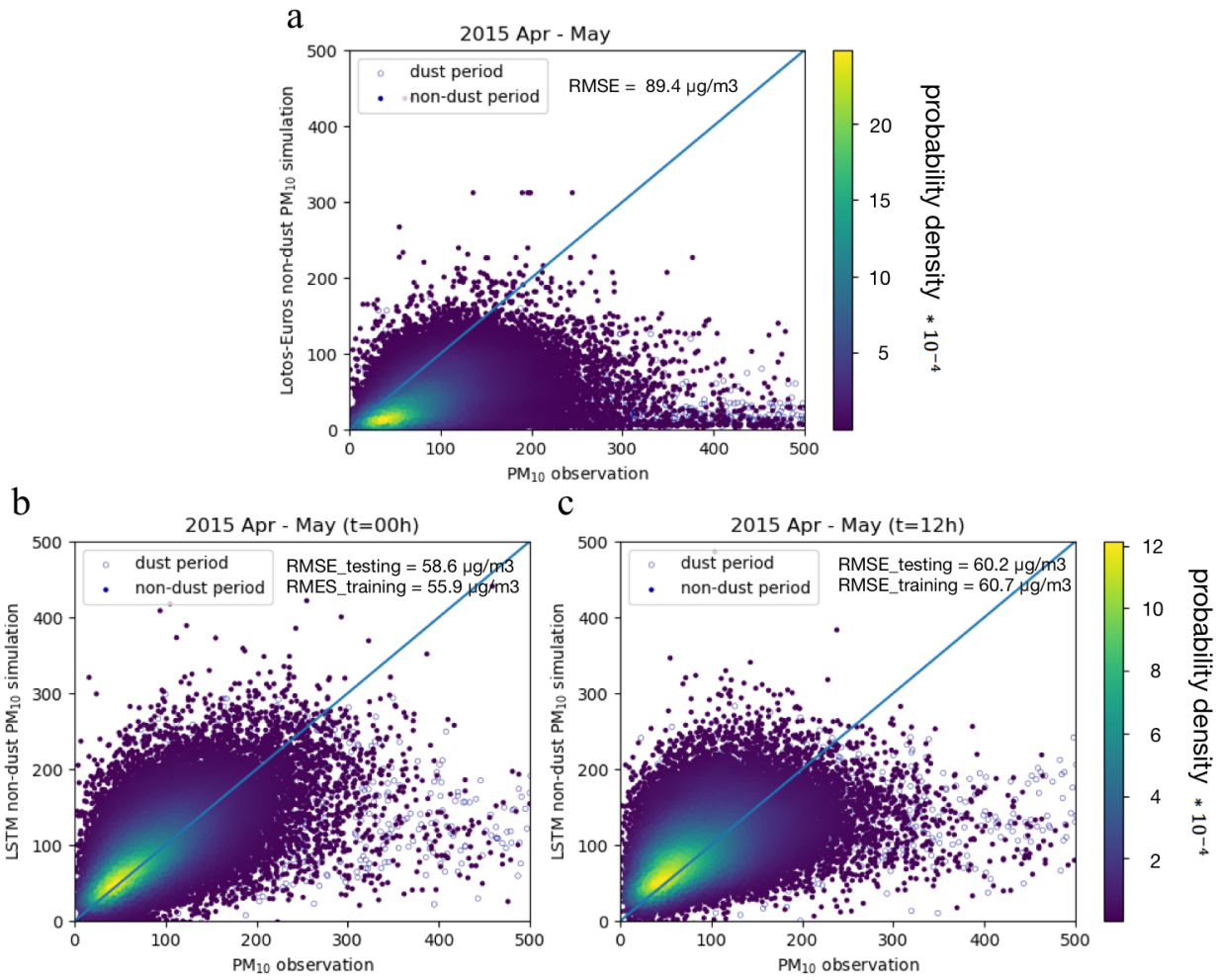

Figure 3. *Non-dust* PM_{10} simulation evaluations. (a): LOTOS-EUROS/*non-dust* forecast vs. PM_{10} measurements; (b): LSTM forecast 0 hour in advance vs. PM_{10} measurements; (c): LSTM forecast 12 hour in advance vs. PM_{10} measurements; (NOTE: the solid circles show the 5% random samples over the non-dust period from April to May 2015 while the hollow ones denote the 5% random ones from the dust period (April 14 to 16).

3.3.1 Spatial patterns at observation sites

To assess our two *non-dust* PM_{10} models, Fig.4 shows the snapshots of the PM_{10} measurements, LOTOS-EUROS/*non-dust* simulations, LSTM forecasts, and the corresponding bias-corrected dust observations at three timestamps: April 15 08:00, 19:00 and 22:00. These first two moments are the start and end of the observation interval in the assimilation window (only observations from the last 12 hours of the assimilation window are assimilated as shown in Fig.2), and observations at 22:00 is treated as independent data for cross-validation. At 08:00, actually only few stations close to the dust source area have already observed the dust storm. Some of the sites in central China observed high PM_{10} concentrations which are believed to be caused by presence of *non-dust* aerosols. Nearly all the stations in north China reported this dust storm at 19:00 and 22:00, as a band covering central and northeast China, see Fig.4 (a.2)~(a.3). Fig.4 (b.1)~(b.3) shows that the LOTOS-EUROS/*non-dust* model forecasts quite stable and constant *non-dust* PM_{10} levels, most of the simulated values are less than 100 $\mu g/m^3$. Subsequently, the corresponding bias-corrected dust measurements (see Fig.4 (c.1)~(c.3)) are very similar to the original PM_{10} observations. This could be problematic when trying to measure the dust storm from the PM_{10} observations; for instance at 08:00 in Fig.4 (c.1), according to the bias-corrected observations the dust storm seems to have already reached central China which was probably not the case. In comparison, the LSTM based bias-corrected dust observations (see Fig.4 (e.1)~(e.3)), which is calculated by subtracting the LSTM *non-dust* part (see Fig.4(d.1)~(d.3)) from the raw PM_{10} measurements, are close to our expectations. Only for sites that are very close to the source regions high dust concentrations are derived at 08:00, while for the other sites hardly any dust is derived. At 19:00, thus 11 hours later, at half of the stations in the north of the domain high dust concentrations are derived. In the southeast of the domain, the derived dust concentrations remain almost zero since the dust plume did not arrive there yet. At 22:00, the plume is moved further south, and the dust load closer to the source region started to decrease.

3.3.2 Time series

To further evaluate the two bias correction methods, Fig.5 shows the time series at the following selected cities: Hohhot, Changchun, Beijing, Baoding, Xingtai and Yulin. The location of these cities/sites can be found in Fig.1. These cities were selected because they all experienced a severe pollution and illustrated the general performance of the LOTOS-EUROS/*non-dust* and LSTM methods. In addition, each of these cities have at least 4 monitoring sites which assured a high accuracy.

The LOTOS-EUROS grid cells with the selected sites all include other observation sites as well, and to illustrate the spread in the observations the maximum and minimum observed values in the grid cell are added to the time series too. Similarly, the LSTM *non-dust* PM_{10} simulation is given together with the spread within the grid cell.

Before the dust storm arrived at these cites, the LSTM model reproduces the variations in PM_{10} rather well. Some errors are present, for example as can be seen on April 14 from 12:00 to 23:00 in Yulin. After the arrival of the dust storm, the PM_{10} observations strongly increase, while the LSTM *non-dust* fraction remains at a low level since it is

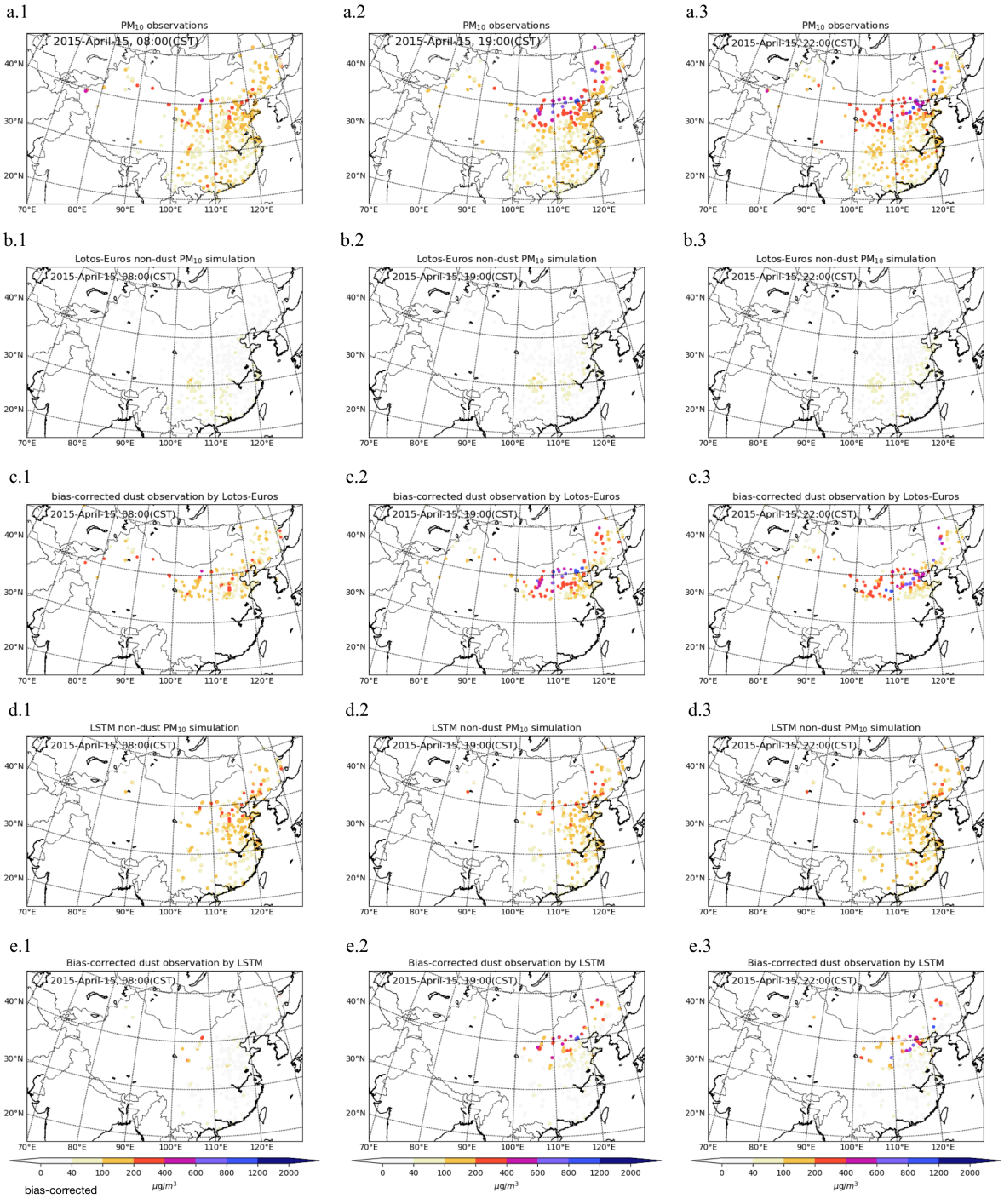

Figure 4. Original PM_{10} measurements (a.1~a.3), LOTOS-EUROS/*non-dust* simulated PM_{10} (b.1~b.3) and the corresponding bias-corrected dust observations (c.1~c.3), LSTM predicted *non-dust* PM_{10} (d.1~d.3) and the derived dust observations (e.1~e.3) at three time snapshots: April 15, 08:00 (a.1~e.1), 19:00 (a.2~e.2) and 22:00 (a.3~e.3)

independent of the dust storm. The real dust measurement is then calculated by subtracting the *non-dust* part from the raw PM_{10} observations.

The LOTOS-EUROS/*non-dust* simulations underestimate the *non-dust* PM_{10} at all the six locations. Thus, the derived bias-corrected dust observations overestimate the actual dust load, and this will affect the dust assimilation results.

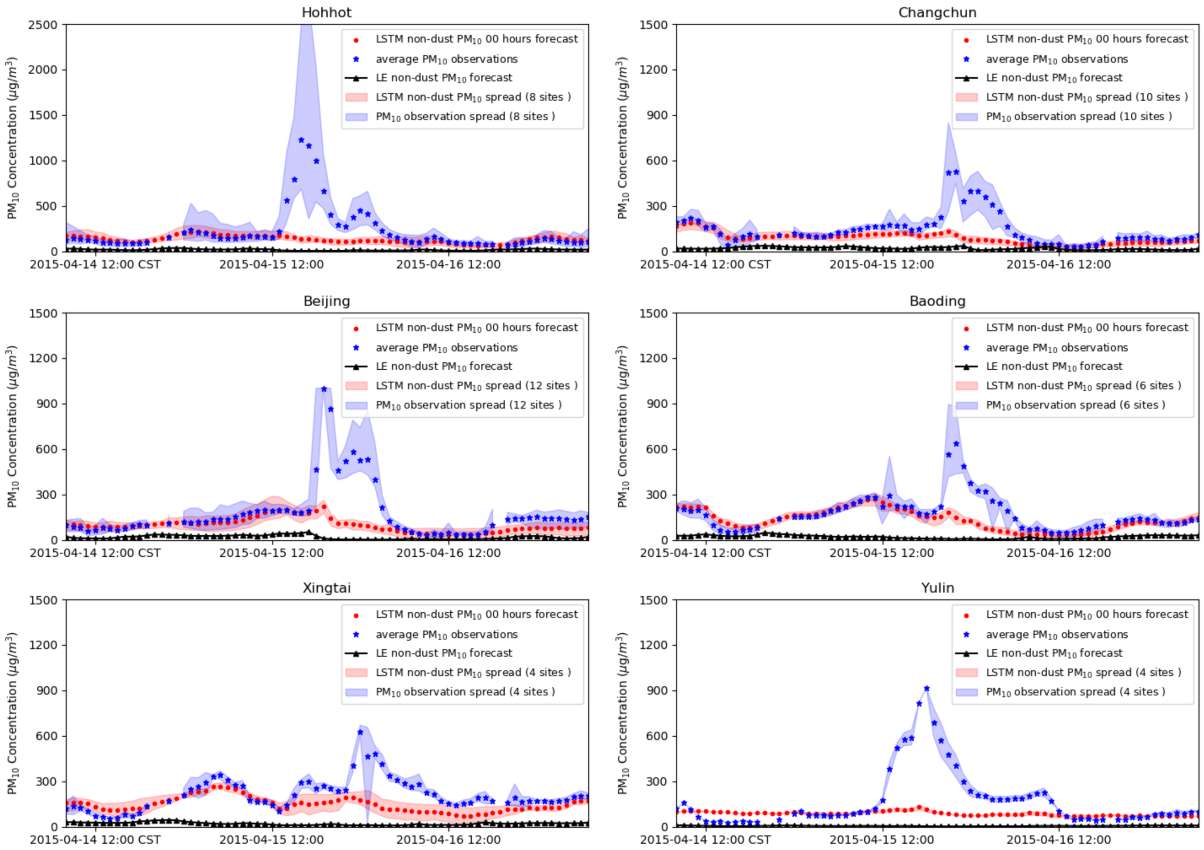

Figure 5. Time series of PM_{10} measurements, LOTOS-EUROS/*non-dust* and LSTM predicted PM_{10} levels at six cities: Hohhot, Changchun, Beijing, Baoding, Xingtai and Yulin. LE: LOTOS-EUROS; LSTM: long short-term memory.

4 Data assimilation experiments

Three different sets of observations are now available for assimilation in the dust model: the original PM_{10} observations, the PM_{10} observations with LOTOS-EUROS bias correction, and the PM_{10} observations with machine learning bias correction. The results have been compared in terms of the posterior dust emission fields and surface dust concentrations.

A practical use of assimilated concentrations is to use them as a start point for a forecast. This could be used to provide early information about the arrival of the dust plume and the expected dust level. The dust forecast after the end of the assimilation window at April 15 19:00 uses the newly estimated emissions. Apart from the dust concentrations, the forecast will also be evaluated in terms of skill scores for the total PM_{10} concentrations in Section 4.3.

4.1 Observation error configuration

A key element of the data assimilation system is the observation error covariance matrix $\mathbf{R}$. This covariance quantifies the possible difference between simulations and observations. The observations with a smaller error have a higher weight in the assimilation process.

In related works, the dust observation errors were usually empirically quantified. Lin et al. (2008) assumed that the observation error is proportional to the measurement with a constant factor of 10%. Jin et al. (2018) used a similar error setting but also assigned a larger error to low valued measurements since the model might easily results in relative large errors when simulating minor dust loads.

Theoretically, the observation uncertainties are due to the representation errors as well as the measurement errors, while the former one is widely considered as the largest source. Limited by the computation resources, our dust model uses a spatial resolution of 25 km, while the in-situ measurements cover the much less of atmosphere surrounding them (Schutgens et al., 2016). This of course limits our capability of resolving the fine-scale fields that are reflected in observation spaces. Therefore, the spatial representation error is assumed to be the dominant error source and taken into the account in approximating the observation uncertainties. In addition, the error due to the different bias correction terms is indeed another source. It is not yet considered in this study but will be exploited for a more accurate assimilation operation in our future work.

The spatial representation error quantification itself is a complex task. It could be calculated through comparing the model simulations at different scales of resolutions. In this study, the availability of multiple measurement sites in a single model grid cell provides an alternative way to quantify the representation error. When multiple observations are present, the statistical error in the observed values reflects the spatial representation uncertainty. An example is the grid cell covering the city of Beijing, where observations from 12 different field stations are available. Note that it is the grid cell which has the most monitoring stations. The spread of the hourly measurements is shown in Fig.5(c). For each hour, the standard deviation of the measured PM_{10} values is plotted against the mean in Fig.6, where the red markers represent 'regular' polluted conditions, and the blue markers the dust event. The result shows that the spread in the observations closely agrees with the average pollution level during the dust event. Based on this result, a simple linear regression is used to obtain a parametrization for the observation representation error:

$$\sigma = \max(\ a \cdot y\ +\ b\ ,\ \ \sigma_{min}\ ) \qquad [\mu g/m^3] \tag{9}$$

where $a = 0.12$ and $b = 55.7$ are the linear regression parameters based on the dust event data (blue markers). It should be noted that the observation sites in Beijing truncate observations at a maximum of 1000 $\mu g/m^3$, and therefore observations close to this number are not used since the true values might have been much higher. A minimum observation representation uncertainty of $\sigma_{min} = 100$ $\mu g/m^3$ is used for the 'dust' observations (PM_{10} with bias correction) to avoid a too strong impact of low valued observations (hardly dust) on the estimation of dust emissions. In case the simulation model estimates dust concentrations at the surface while in reality the plume is elevated, the low valued observations might lead to an unrealistic strong decrease of the dust emissions.

The representation uncertainty has already been validated to fluctuate in space (Schutgens et al., 2016). However, for most other grid cells the number of observations sites is simply one, which makes it difficult to parametrize a representation error in a similar way. Therefore, the representation error parametrized for Beijing is used for all other locations too.

Note that the raw PM_{10} and the bias-corrected dust measurements might have different uncertainties in representing the real dust storm level. This is not yet taken into account in our study, and the three types of the assimilated measurements, raw PM_{10}, bias-corrected dust observation either using the CTM or using the machine learning, are all configured with the same observation error in Eq.9. In addition, all the measurements are assumed to be independent, hence, the observation error covariance $\mathbf{R}$ is diagonal.

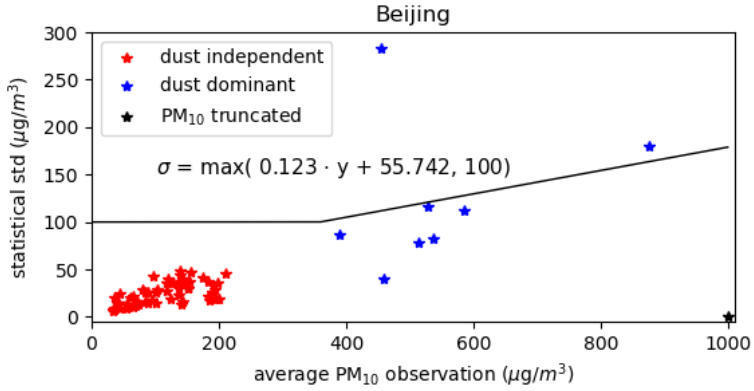

Figure 6. Average vs. Standard deviation of the hourly PM_{10} observations range from April 14 08:00 to April 17 07:00 in the grid cell of Beijing. See Fig.5(c) for the time series.

4.2 Dust emission estimation

To evaluate the posterior dust emission field that is obtained by assimilation of the bias corrected 'dust' observations, an emission index $\mathcal{F}_i$ (g/m^2) is defined as in (Jin et al., 2018). The index represents the accumulated dust emission in a cell i between April 14 08:00 and April 15 19:00. Fig.7 shows the emission index map of the *a prior* model, and

posteriori emissions obtained from assimilation of either the original PM_{10} observations, or the LOTOS-EUROS or LSTM based bias-corrected 'dust' measurements.

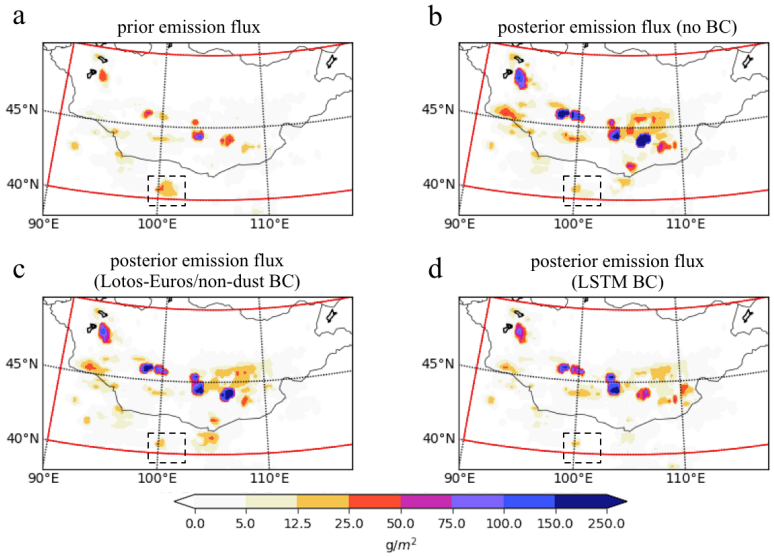

Figure 7. Accumulated dust emission map $\mathcal{F}$ between April 14 08:00 and April 15 19:00 of *a priori* model (a), or (b) *a posteriori* estimates using the original PM_{10} observations, (c) LOTOS-EUROS or (d) LSTM based bias-corrected dust measurements. BC: bias correction

As shown in Fig.7(a), the *a prior* emission was in general rather weak, which resulted in an underestimated surface dust concentration simulation as can be seen for example in Fig.8(a.1)~(a.2). The *posteriori* emissions are almost everywhere higher than the *a prior*. An exception is the black marked region, where the *a prior* emissions are higher. The emissions from this black-dashed region contributed to a too-early arrival of the dust peak in the model cells over Hohhot and Xingtai as shown in Fig.9(a) and (c).

Fig.7 (b) shows the emission index $\mathcal{F}$ that results from directly assimilating the original PM_{10} measurements. As expected the estimated emissions are higher than those obtained by assimilating the bias-corrected observations, since all airborne aerosols observed are attributed to be dust. In comparison, the assimilation with LSTM baseline removed data results in a modest emission level as shown in Fig.7(d). The emissions estimated with LOTOS-EUROS based bias-corrected observations are in between, since the resulting 'dust' observations also overestimate the actual dust loads compared to the LSTM based bias-corrected dust measurements.

4.3 Dust simulation and forecast skill

Fig.8 (a)~(d) show the dust simulations at the surface layer at the end of the assimilation window (April 15 19:00, left column) and the forecast 3 hours later (22:00, right column) using the newly estimated emission field. Note that

the average dust concentration over the affected downwind regions reached at a peak around 22:00. Compared to background simulations in Jin et al. (2018), the *a prior* model simulations have been improved by disabling the topography-based preference factor as mentioned in Section 2.1; however, a large difference from the bias-corrected PM_{10} observations in Fig. 4(e) is still present.

The *posteriori* concentrations in Fig.8(b.1)~(b.2) are the result of assimilating the original measurements PM_{10} observations shown in Fig.4(a.1)~(a.2). As expected, these lead to the highest simulated dust concentrations since all the aerosols observed are assumed to represent dust. Especially in the center of the plume, the dust concentration can be as large as 2000 $\mu g/m^3$. Fig.8(c.1)~(c.2) show the results when using the LOTOS-EUROS/*non-dust* bias-corrected PM_{10} observations as 'dust', and although concentrations are lower, they are still likely to overestimate the real dust levels. The *posteriori* results using the LSTM bias-corrected measurements provide the lowest dust concentrations as shown in Fig.8(d). Only in the grid cells that are close to the source region, the surface dust concentration reach values as large as 2000 $\mu g/m^3$, while in the downwind areas the maximum dust concentrations are usually below 1200 $\mu g/m^3$.

To illustrate the improvements of assimilating bias-corrected measurements, Fig. 9 shows the observed and simulated PM_{10} concentrations in the aforementioned grid cells covering Hohhot, Beijing, and Xingtai. These locations are neither the best nor the worst examples, but illustrate typical results and challenges to be solved in future. For a fair comparison with the PM_{10} observations, the non-dust aerosol concentrations obtained from either LOTOS-EUROS/*non-dust* or LSTM were added to the dust simulations from the inversion system.

Site Hohhot is close to the main dust source region. The *a prior* model simulated the arrival of the dust plume 8 hours before it was actually visible in the PM_{10} observations. The assimilation of the observations is able to produce simulations in which the dust plume arrives at the correct time. The assimilation with LSTM bias-corrected data has the best performance, with the peak of the simulated concentrations (dust plus bias) most close to the observed PM_{10}. During the forecast period (t>April 15, 19:00), all three assimilation based forecasts show a decline in concentrations, which slightly overestimate the observations. This can be explained from the fact that the dust storm is a strong flow-dependent phenomenon in which concentrations at a certain location are strongly correlated to earlier concentrations at upwind locations. For Hohhot, only a limited number of observation sites is located upwind, and therefore hardly any data is available to constrain the concentrations at this location. To improve the forecast at Hohhot it will be necessary to have additional observation data, for example from sites actually within the source region, or from satellites observing the aerosol load over the source region (Jin et al., 2019).

For the grid cell Beijing, which is located further downwind from the dust source region, the arrival of the dust peak is correctly simulated. However, the amplitude of the concentration peak is underestimated compared to the average PM_{10} observations. As can be seen in Fig.8, the dust plume forms a rather small band over central and northeast China. In each of the three assimilations, the dust concentrations in the band are rather low around Beijing. This suggests that the simulation model simply is not able to increase the dust concentrations here, for

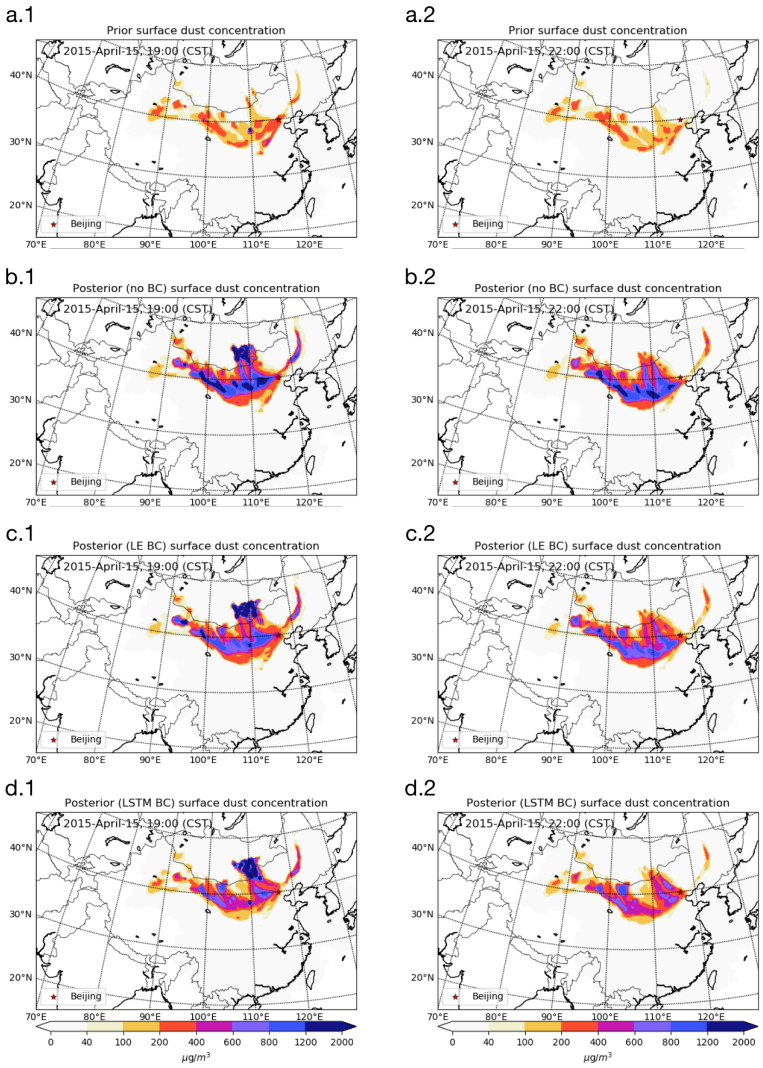

Figure 8. Surface dust concentration of *a prior* (a.1~a.2), *posterior* using no bias-corrected (no BC) data (b.1~b.2), *posterior* using LOTOS-EUROS/*non-dust* bias-corrected (LE BC) data (c.1~c.2), *posterior* using no bias-corrected (LSTM BC) data (d.1~d.2) at April 15, 19:00 (a.1~d.1) and 22:00 (a.2~d.2)

example because of uncertainties in the meteorological data, a removal of dust that is too efficient, or because some local sources of dust are absent (equally, non-dust PM_{10} levels are underestimated).

The grid cell Xingtai is located more to the south, and the model is able to simulate high dust concentrations here. The *a prior* model simulates the arrival of a first dust peak already at 13:00, which is however not visible in the PM_{10} data. The assimilation postpones the arrival of the main dust, which according to the measurements takes place around 22:00 and is already in the forecast period. The forecast simulations all overestimate the amplitude of the peak, especially when using the original PM_{10} data as proxy for dust. The assimilation with the LSTM based baseline removal shows the best agreement with the observations.

4.4 Evaluation of forecast skill

To evaluate the forecast skill of the assimilation(s), the root mean square error (RMSE) of the reference and three posterior full aerosol simulations (dust forecasts plus *non-dust* predictions) with respect to the observed PM_{10} over the whole observation sites has been computed for each hour. A time series of this RMSE is shown in Fig.10; after the assimilation window (marked period), the results are based on the forecast simulations. The *a prior* RMSE values at the end of the assimilation window and during the forecast are about 200-250 $\mu g/m^3$. Direct assimilation of the original PM_{10} measurement actually increases these values to above 300 $\mu g/m^3$ during the forecast, since dust concentrations become strongly overestimated. Assimilation of the LOTOS-EUROS/*non-dust* baseline removed observations nonetheless reduces the RMSE, in particular within the assimilation window. Strongest decrease in RMSE is obtained using the LSTM based baseline removal, with values of 120-200 $\mu g/m^3$ during the forecast.

5 Summary and conclusion

In this study, a dust storm data assimilation experiment has been performed for an event over East Asia in the spring of 2015. PM_{10} observation data from the China Ministry of Environmental Protection observing network were assimilated into a dust simulation model to estimate the dust emissions. The PM_{10} measurements themselves are considered as unbiased. They clearly show the arrival of a dust plume throughout the region due to the high spatiotemporal resolution. However, the data cannot be compared directly to dust simulations since they actually represent a sum of the dust particles and other *non-dust* aerosols. Direct assimilation of these measurements would introduce a bias in the assimilation system, since it cannot distinguish between model and observation errors.

Two methods have been implemented to remove the *non-dust* part the PM_{10} observations during the dust event in order to use them as 'dust' proxy in a dust assimilation system. The first method uses a conventional regional chemical transport model, LOTOS-EUROS/*non-dust*, which simulates the emission, transport, chemistry, and deposition of aerosols mainly related to anthropogenic activities. The second method uses a machine learning model that statistically describes the relations between regular PM_{10} concentrations (outside dust events), and available air quality and meteorological data.

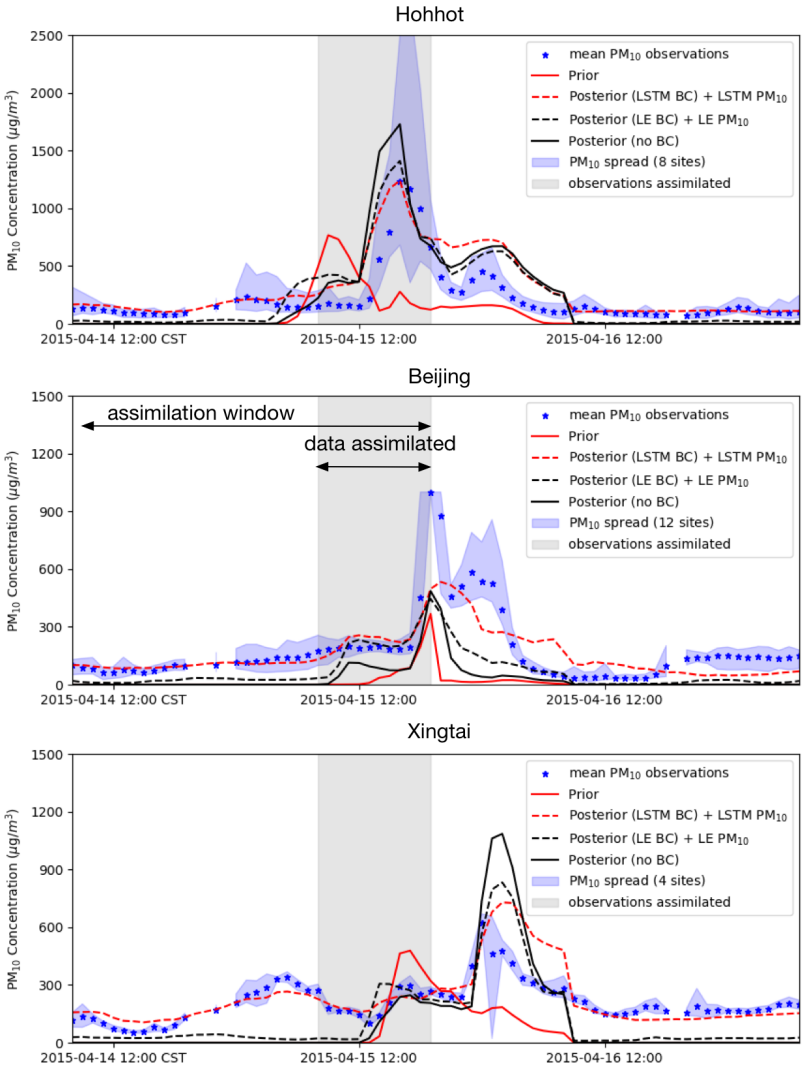

Figure 9. Time series of posterior dust concentration and PM_{10} observations in three cities: Hohhot, Beijing, Xingtai (observations in the gray shadow are assimilated)

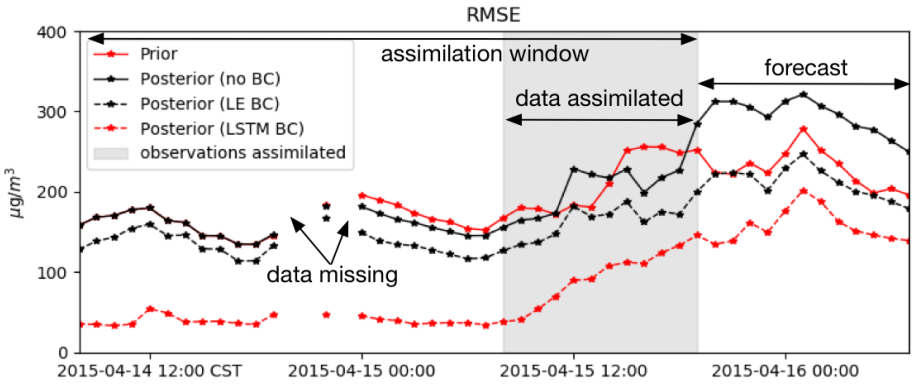

Figure 10. Time series of root mean square error compared to the ground PM_{10}. The assimilation window is set from April 14 08:00 to April 15 19:00, and PM_{10} observation in the gray shadow are assimilated.

The two methods to estimate the *non-dust* part of the PM_{10} load have been validated. The simulations by the LOTOS-EUROS/*non-dust* model in general underestimate the PM_{10} concentrations. The root mean square error stays at a relative high level of 89.4 $\mu g/m^3$. It is mainly caused by missing emissions and aerosol components such as secondary organic matter. In comparison, the data-driven machine learning model agrees more closely with the real measurements, the RMSE declines to 58.6 $\mu g/m^3$.

A variational data assimilation system has been used to estimate the dust emissions that lead to a severe dust storm in April 2015. The system either assimilated the original PM_{10} observations, or the bias-corrected 'dust' observations based on either LOTOS-EUROS/*non-dust* or LSTM model. The posterior simulations using the original observations resulted in a strong overestimation of the dust concentrations, since all PM_{10} are simply attributed to dust. Using the LOTOS-EUROS/*non-dust* bias-corrected observations, a clear improvement on the dust simulation has been obtained, but overestimation of dust concentrations is still present. The best results are obtained when using a LSTM model to remove the *non-dust* part of the PM_{10} observations, with *posterior* concentrations in good agreement with the measurements.

The dust emissions estimated using the assimilation can be used to drive a dust forecast. When the original PM_{10} observations were used in the assimilation, the forecast skill of the system actually decreased due to the strong overestimation of dust concentrations, the RMSE rose from averagely 230 (prior forecast) to 300 $\mu g/m^3$. Better forecasts are obtained when using the model-based and especially the machine learning based bias-corrected observations. The RMSE of the former one was reduced to 200 $\mu g/m^3$ while the RMSE of the latter one further declined to 150 $\mu g/m^3$.

Future work

Both our CTM and machine learning based bias correction methods have room for improvements. It might be useful to improve the CTM simulations by assimilating PM_{10} observations during the hours where no dust storms are present, and use these improved simulations to remove the *non-dust* part of the observations during an event. These additional assimilations would then involve repeated forward ensemble bias-model runs which could be computationally expensive. The machine learning model in our *non-dust* PM_{10} simulation can also be further optimized, such as using a different configuration or deeper neural network, including extra input features like *non-dust* PM_{10} simulation from CTMs (Lin et al., 2019) and other related records.

We will exploited the variabilities of the representation errors comparing the model simulations at different spatial resolutions. The error from the bias correction term will also be taken into account while calculating the observation error.

Data availability

The datasets including measurements and model simulations can be accessed from websites listed in the references or by contacting the corresponding author

Author contribution

JJ and HXL conceived the study and designed the experiments. JJ and YX performed the machine learning based non-dust simulation. AS performed the CTM based non-dust simulation. JJ and AS performed the assimilation tests and carried out the data analysis. AS, HXL, AH provided useful comments on the paper. JJ prepared the manuscript with contributions from all co-authors.

Acknowledgments

The real-time PM_{10} data are from the network established by the China Ministry of Environmental Protection and accessible to the public at http://106.37.208.233:20035/. One can also access the historical profile by visiting http://www.aqistudy.cn/.

Competing interests

The authors declare that they have no conflict of interest.

References

Benedetti, A., Di Giuseppe, F., Jones, L., Peuch, V. H., Remy, S., and Zhang, X.: The impact of data assimilation on the prediction of Asian desert dust using an operational 4D-Var system, Atmospheric Chemistry and Physics Discussions, pp. 1–17, https://doi.org/10.5194/acp-2018-78, https://editor.copernicus.org/index.php/acp-2018-78-RC1.pdf?_mdl=msover_md&_jrl=10&_lcm=oc108lcm109w&_acm=get_comm_file&_ms=66357&c=144470&salt=614587343304971732, 2018.

Berry, T. and Harlim, J.: Correcting Biased Observation Model Error in Data Assimilation, Monthly Weather Review, 145, 2833–2853, https://doi.org/10.1175/MWR-D-16-0428.1, http://dx.doi.org/10.1175/MWR-D-16-0428.1, 2017.

Brasseur, G. P., Xie, Y., Petersen, A. K., Bouarar, I., Flemming, J., Gauss, M., Jiang, F., Kouznetsov, R., Kranenburg, R., Mijling, B., Peuch, V.-H., Pommier, M., Segers, A., Sofiev, M., Timmermans, R., van der A, R., Walters, S., Xu, J., and Zhou, G.: Ensemble forecasts of air quality in eastern China – Part 1: Model description and implementation of the MarcoPolo–Panda prediction system, version 1, Geoscientific Model Development, 12, 33–67, https://doi.org/10.5194/gmd-12-33-2019, https://www.geosci-model-dev.net/12/33/2019/, 2019.

Cesnulyte, V., Lindfors, A. V., Pitkänen, M. R. A., Lehtinen, K. E. J., Morcrette, J. J., and Arola, A.: Comparing ECMWF AOD with AERONET observations at visible and UV wavelengths, Atmospheric Chemistry and Physics, 14, 593–608, https://doi.org/10.5194/acp-14-593-2014, https://www.researchgate.net/publication/290315319_1-s20-S2352938515000373-main/figures?lo=1, 2014.

Chen, G., Li, S., Knibbs, L. D., Hamm, N. A. S., Cao, W., Li, T., Guo, J., Ren, H., Abramson, M. J., and Guo, Y.: A machine learning method to estimate PM2.5 concentrations across China with remote sensing, meteorological and land use information, Science of The Total Environment, 636, 52–60, https://doi.org/10.1016/j.scitotenv.2018.04.251, http://dx.doi.org/10.1016/j.scitotenv.2018.04.251, 2018.

Dee, D. P.: Bias and data assimilation, Quarterly Journal of the Royal Meteorological Society, 131, 3323–3343, https://doi.org/10.1256/qj.05.137, https://rmets.onlinelibrary.wiley.com/doi/abs/10.1256/qj.05.137, 2005.

Dee, D. P. and Uppala, S.: Variational bias correction of satellite radiance data in the ERA-Interim reanalysis, Quarterly Journal of the Royal Meteorological Society, 135, 1830–1841, https://doi.org/10.1002/qj.493, https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.493, 2009.

Di Tomaso, E., Nick, Jorba, O., and Garcia-Pando, C. P.: Assimilation of MODIS Dark Target and Deep Blue observations in the dust aerosol component of NMMB-MONARCH version 1.0, Geoscientific Model Development, 10, 1107–1129, https://doi.org/10.5194/gmd-10-1107-2017, https://www.geosci-model-dev.net/10/1107/2017/, 2017.

Eyre, J. R.: Observation bias correction schemes in data assimilation systems: a theoretical study of some of their properties, Quarterly Journal of the Royal Meteorological Society, 142, 2284–2291, https://doi.org/10.1002/qj.2819, https://rmets.onlinelibrary.wiley.com/doi/10.1002/qj.2819, 2016.

Fan, J., Li, Q., Hou, J., Feng, X., Karimian, H., and Lin, S.: A Spatiotemporal Prediction Framework for Air Pollution Based on Deep RNN, ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences, IV-4/W2, 15–22, https://doi.org/10.5194/isprs-annals-IV-4-W2-15-2017%7D, https://www.google.nl/_/chrome/newtab?ie=UTF-8, 2017.

Ginoux, P., Chin, M., Tegen, I., Prospero, J. M., Holben, B., Dubovik, O., and Lin, S.-J.: Sources and distributions of dust aerosols simulated with the GOCART model, J. Geophys. Res., 106, 20 255–20 273, https://doi.org/10.1029/2000jd000053, http://dx.doi.org/10.1029/2000jd000053, 2001.

Gong, S. L. and Zhang, X. Y.: CUACE/Dust – an integrated system of observation and modeling systems for operational dust forecasting in Asia, Atmospheric Chemistry and Physics, 8, 2333–2340, https://doi.org/10.5194/acp-8-2333-2008, http://dx.doi.org/10.5194/acp-8-2333-2008, 2008.

Gong, S. L., Zhang, X. Y., Zhao, T. L., McKendry, I. G., Jaffe, D. A., and Lu, N. M.: Characterization of soil dust aerosol in China and its transport and distribution during 2001 ACE-Asia: 2. Model simulation and validation, J. Geophys. Res., 108, 4262+, https://doi.org/10.1029/2002jd002633, http://dx.doi.org/10.1029/2002jd002633, 2003.

Guenther, A., Karl, T., Harley, P., Wiedinmyer, C., Palmer, P. I., and Geron, C.: Estimates of global terrestrial isoprene emissions using MEGAN (Model of Emissions of Gases and Aerosols from Nature), Atmospheric Chemistry and Physics, 6, 3181–3210, https://doi.org/10.5194/acp-6-3181-2006, https://www.atmos-chem-phys.net/6/3181/2006/, 2006.

Huneeus, N., Schulz, M., Balkanski, Y., Griesfeller, J., Prospero, J., Kinne, S., Bauer, S., Boucher, O., Chin, M., Dentener, F., Diehl, T., Easter, R., Fillmore, D., Ghan, S., Ginoux, P., Grini, A., Horowitz, L., Koch, D., Krol, M. C., Landing, W., Liu, X., Mahowald, N., Miller, R., Morcrette, J. J., Myhre, G., Penner, J., Perlwitz, J., Stier, P., Takemura, T., and Zender, C. S.: Global dust model intercomparison in AeroCom phase I, Atmospheric Chemistry and Physics, 11, 7781–7816, https://doi.org/10.5194/acp-11-7781-2011, http://dx.doi.org/10.5194/acp-11-7781-2011, 2011.

Jin, J., Lin, H. X., Heemink, A., and Segers, A.: Spatially varying parameter estimation for dust emissions using reduced-tangent-linearization 4DVar, Atmospheric Environment, 187, 358–373, https://doi.org/https://doi.org/10.1016/j.atmosenv.2018.05.060, https://www.sciencedirect.com/science/article/pii/S1352231018303704, 2018.

Jin, J., Segers, A., Heemink, A., Yoshida, M., Han, W., and Lin, H.-X.: Dust Emission Inversion Using Himawari-8 AODs Over East Asia: An Extreme Dust Event in May 2017, Journal of Advances in Modeling Earth Systems, 11, 446–467, https://doi.org/10.1029/2018MS001491, https://agupubs.onlinelibrary.wiley.com/doi/full/10.1029/2018MS001491, 2019.

Kaiser, J. W., Heil, A., Andreae, M. O., Benedetti, A., Chubarova, N., Jones, L., Morcrette, J.-J., Razinger, M., Schultz, M. G., Suttie, M., and van der Werf, G. R.: Biomass burning emissions estimated with a global fire assimilation system based on observed fire radiative power, Biogeosciences, 9, 527–554, https://doi.org/10.5194/bg-9-527-2012, https://www.biogeosciences.net/9/527/2012/, 2012.

Khade, V. M., Hansen, J. A., Reid, J. S., and Westphal, D. L.: Ensemble filter based estimation of spatially distributed parameters in a mesoscale dust model: experiments with simulated and real data, Atmospheric Chemistry and Physics, 13, 3481–3500, https://www.atmos-chem-phys.net/13/3481/2013/, 2013.

Li, G., Bei, N., Cao, J., Wu, J., Long, X., Feng, T., Dai, W., Liu, S., Zhang, Q., and Tie, X.: Widespread and persistent ozone pollution in eastern China during thenon-winter season of 2015: observations and source attributions, Atmospheric Chemistry and Physics, 17, 2759–2774, http://www.atmos-chem-phys.net/17/2759/2017/, 2017a.

Li, X., Peng, L., Hu, Y., Shao, J., and Chi, T.: Deep learning architecture for air quality predictions, Environmental Science and Pollution Research, 23, 22 408–22 417, https://doi.org/10.1007/s11356-016-7812-9, http://dx.doi.org/10.1007/s11356-016-7812-9, 2016.

Li, X., Peng, L., Yao, X., Cui, S., Hu, Y., You, C., and Chi, T.: Long short-term memory neural network for air pollutant concentration predictions: Method development and evaluation - ScienceDirect, Environmental Pollution, 231, 997–1004, https://doi.org/https://doi.org/10.1016/j.envpol.2017.08.114, https://www.sciencedirect.com/science/article/pii/S0269749117307534, 2017b.

Lin, C., Wang, Z., and Zhu, J.: An Ensemble Kalman Filter for severe dust storm data assimilation over China, Atmospheric Chemistry & Physics, 8, 2975–2983, https://www.atmos-chem-phys.net/8/2975/2008/, 2008.

Lin, H. X., Jin, J., and van den Herik, J.: Air Quality Forecast through Integrated Data Assimilation and Machine Learning, http://insticc.org/node/TechnicalProgram/icaart/presentationDetails/75552, 2019.

Liu, M., Westphal, D. L., Wang, S., Shimizu, A., Sugimoto, N., Zhou, J., and Chen, Y.: A high-resolution numerical study of the Asian dust storms of April 2001, J. Geophys. Res., 108, 8653+, https://doi.org/10.1029/2002jd003178, http://dx.doi.org/10.1029/2002jd003178, 2003.

Lorente-Plazas, R. and Hacker, J. P.: Observation and Model Bias Estimation in the Presence of Either or Both Sources of Error, Monthly Weather Review, 145, 2683–2696, https://doi.org/10.1175/MWR-D-16-0273.1, http://dx.doi.org/10.1175/MWR-D-16-0273.1, 2017.

Marticorena, B. and Bergametti, G.: Modeling the atmospheric dust cycle: 1. Design of a soil-derived dust emission scheme, J. Geophys. Res., 100, 16 415–16 430, https://doi.org/10.1029/95JD00690, http://dx.doi.org/10.1029/95JD00690, 1995.

Niu, T., Gong, S. L., Zhu, G. F., Liu, H. L., Hu, X. Q., Zhou, C. H., and Wang, Y. Q.: Data assimilation of dust aerosol observations for the CUACE/dust forecasting system, Atmospheric Chemistry and Physics, 8, 3473–3482, https://doi.org/10.5194/acp-8-3473-2008, http://dx.doi.org/10.5194/acp-8-3473-2008, 2008.

Petersen, A. K., Brasseur, G. P., Bouarar, I., Flemming, J., Gauss, M., Jiang, F., Kouznetsov, R., Kranenburg, R., Mijling, B., Peuch, V.-H., Pommier, M., Segers, A., Sofiev, M., Timmermans, R., van der A, R., Walters, S., Xie, Y., Xu, J., and Zhou, G.: Ensemble forecasts of air quality in eastern China – Part 2: Evaluation of the MarcoPolo–Panda prediction system, version 1, Geoscientific Model Development, 12, 1241–1266, https://doi.org/10.5194/gmd-12-1241-2019, https://www.geosci-model-dev.net/12/1241/2019/, 2019.

Remer, L. A., Kaufman, Y. J., Tanré, D., Mattoo, S., Chu, D. A., Martins, J. V., Li, R. R., Ichoku, C., Levy, R. C., Kleidman, R. G., Eck, T. F., Vermote, E., and Holben, B. N.: The MODIS Aerosol Algorithm, Products, and Validation, Journal of the Atmospheric Sciences, 62, 947–973, 2005.

Schutgens, N. A. J., Gryspeerdt, E., Weigum, N., Tsyro, S., Goto, D., Schulz, M., and Stier, P.: Will a perfect model agree with perfect observations? The impact of spatial sampling, Atmospheric Chemistry and Physics, 16, 6335–6353, https://doi.org/10.5194/acp-16-6335-2016, https://www.atmos-chem-phys.net/16/6335/2016/, 2016.

Sekiyama, T. T., Tanaka, T. Y., Shimizu, A., and Miyoshi, T.: Data assimilation of CALIPSO aerosol observations, Atmospheric Chemistry and Physics, 10, 39–49, https://www.atmos-chem-phys.net/10/39/2010/, 2010.

Shao, P., Tian, H., Sun, Y., Liu, H., Wu, B., Liu, S., Liu, X., Wu, Y., Liang, W., Wang, Y., Gao, J., Xue, Y., Bai, X., Liu, W., Lin, S., and Hu, G.: Characterizing remarkable changes of severe haze events and chemical compositions in multi-size airborne particles (PM1, PM2.5 and PM10) from January 2013 to 2016–2017 winter in Beijing, China, Atmospheric Environment, 189, 133–144, https://www.sciencedirect.com/science/article/pii/S1352231018304291, 2018.

Shao, Y. P., Raupach, M. R., and Leys, J. F.: A model for predicting aeolian sand drift and dust entrainment on scales from paddock to region, Australian Journal of Soil Research, 34, 309+, https://doi.org/10.1071/sr9960309, http://dx.doi.org/10.1071/sr9960309, 1996.

Timmermans, R., Kranenburg, R., Manders, A., Hendriks, C., Segers, A., Dammers, E., Zhang, Q., Wang, L., Liu, Z., Zeng, L., Denier van der Gon, H., and Schaap, M.: Source apportionment of PM2.5 across China using LOTOS-EUROS, Atmospheric Environment, https://doi.org/10.1016/j.atmosenv.2017.06.003, http://dx.doi.org/10.1016/j.atmosenv.2017.06.003, 2017.

Wang, Y. Q., Zhang, X. Y., Gong, S. L., Zhou, C. H., Hu, X. Q., Liu, H. L., Niu, T., and Yang, Y. Q.: Surface observation of sand and dust storm in East Asia and its application in CUACE/Dust, Atmospheric Chemistry and Physics, 8, 545–553, https://doi.org/10.5194/acp-8-545-2008, http://dx.doi.org/10.5194/acp-8-545-2008, 2008.

Wang, Z., Ueda, H., and Huang, M.: A deflation module for use in modeling long-range transport of yellow sand over East Asia, J. Geophys. Res., 105, 26 947–26 959, https://doi.org/10.1029/2000jd900370, http://dx.doi.org/10.1029/2000jd900370, 2000.

WMO: WMO AIRBORNE DUST BULLETIN: Sand and Dust Storm Warning Advisory and Assessment System, https://library.wmo.int/doc_num.php?explnum_id=3416, 2017.

Xu, L., Batterman, S., Chen, F., Li, J., Zhong, X., Feng, Y., Rao, Q., and Chen, F.: Spatiotemporal characteristics of PM2.5 and PM10 at urban and corresponding background sites in 23 cities in China, Science of The Total Environment, 599600, 2074–2084, http://www.sciencedirect.com/science/article/pii/S0048969717311488, 2017.

Yoshida, M., Kikuchi, M., Nagao, T. M., Murakami, H., Nomaki, T., and Higurashi, A.: Common Retrieval of Aerosol Properties for Imaging Satellite Sensors, Journal of the Meteorological Society of Japan. Ser. II, advpub, https://doi.org/10.2151/jmsj.2018-039, https://www.jstage.jst.go.jp/article/jmsj/advpub/0/advpub_2018-039/_article/-char/en, 2018.

Yumimoto, K., Uno, I., Sugimoto, N., Shimizu, A., Liu, Z., and Winker, D. M.: Adjoint inversion modeling of Asian dust emission using lidar observations, Atmospheric Chemistry and Physics, 8, 2869–2884, https://doi.org/10.5194/acp-8-2869-2008, http://dx.doi.org/10.5194/acp-8-2869-2008, 2008.

Yumimoto, K., Murakami, H., Tanaka, T. Y., Sekiyama, T. T., Ogi, A., and Maki, T.: Forecasting of Asian dust storm that occurred on May 10–13, 2011, using an ensemble-based data assimilation system, Particuology, 28, 121–130, https://doi.org/10.1016/j.partic.2015.09.001, http://dx.doi.org/10.1016/j.partic.2015.09.001, 2016.

Zhang, S.: Nearest neighbor selection for iteratively kNN imputation, Journal of Systems and Software, 85, 2541–2552, https://www.sciencedirect.com/science/article/pii/S0164121212001586?via%3Dihub, 2012.