We would like to thank the reviewers and the editor for the very useful and constructive comments. Following their suggestions, we have done some major changes in the paper to clarify its scope and to improve it according to their recommendations. We summarize the major changes here for clarity. Next, we give our reply and comments for the specific requests.

The major changes for the paper are:

- **We expanded the section on LIRIC, giving a more detailed description of the algorithm. We added a discussion of the major underlying assumptions of the algorithm and the possible sources of uncertainty. We try to make clear to the reader what the current state of the algorithm is. We tried to clarify that our methodology is not limited to LIRIC retrievals, but the specific algorithm is used as an example of a wider class of algorithms. For this we expanded the discussion and comparison of LIRIC with other algorithms to highlight at what extent LIRIC can be considered a representative example.**
- **We have added a description of the quality of our dataset. We have developed a new quality-check procedure, which limited the available profiles to 55 (from 61). The procedure includes checks on change of AOD, fine mode fraction, and qualitative check of back-trajectory analysis. More details are given in the detail responses below. In the new version of the manuscript we present this quality assurance, together with possible limitations of the dataset.**
- **We shortened and moved parts of the text from the "Methodology" and "Retrieval algorithms" sections to the introduction. In this way the sections are more concise and allow us to focus on the main points of the argument.**

Anonymous Referee #1

Obtaining dust/aerosol concentration data information from (ground-based) remote sensing observations is certainly desirable, e.g. for determining aviation safety after volcanic eruptions. It can of course also be used for dust model evaluation, as suggested here. A positive aspect here is the synergetic use routine measurements from the lidar and sunphotometer networks, as the production of such data could possibly be automated at some point. The authors argue that to compare such retrieved concentrations with model concentrations that are directly simulated by the model without assuming particle site distribution (which in fact in the 8-size-bin models are also simulated directly). An alternative approach for direct comparisons of models and observations would be to have all assumptions on the model side – i.e. by simulating the actual measured instrument signals by model forward operators.

This is a good point. Still, a model forward operator used with a dust-only model could not simulate the complete lidar signal, as these signals are affected by other aerosol types. A dust model cannot take into account, for example, the effect of smoke and urban aerosols that frequently affect remote sensing measurements. The model forward operator approach is probably better adapted for complete aerosol models e.g. COSMO-MUSCAT or ECMWF's MACC, or in cases that we are confident that dust is virtually the only aerosol type (e.g. observations near the source).

However even synergetic retrievals such as the LIRIC algorithm need additional assumptions. Here the e.g. assumption of constant microphysical properties is mentioned, also to convert volume concentration to mass concentration requires additional information on particle densities. As dust

modeler I would be hesitant to use such concentration retrievals until such a product is thoroughly evaluated itself, and possible errors und uncertainties are quantified as actually proposed as an outlook in the 'Conclusions' section of this manuscript.

Indeed, a thorough evaluation of LIRIC, or any of the many other similar algorithms developed, is still missing. This is why we use LIRIC profiles here as an example of the dataset that we need to create, and not as the ready tool for model evaluation. Several attempts to evaluate and quantify the uncertainties of such models are under way or planed, and we expect to see improvement in this direction in the following years. We have changes the text in several points to make these facts clear to the reader (see also our general comments).

Further comments:

Page 3608, line 20: Why would 'giving a bird's eye view' be useful? Satellite instruments can do that. I would rather point that models are a useful tool to study the processes and sensitivities controlling the dust distribution, and in particular are used to compute regional and global budgets of dust.

Good point. We updated the text to reflect the reviewer's comments.

New text: "These models simulate the complete 3-D fields of dust concentration and can be used, among others, to study the processes and sensitivities controlling the dust distribution and to compute regional and global budgets of dust."

Page 3608, line 27: 'monitored' – not a good expression here, better: 'tested'

Corrected.

Page 3607, line: 10: Dust models are not only validated with optical thickness data but often measurements of surface concentration and deposition fluxes are used for model evaluation as well.

Correct. We have changed the text to reflect this point.

New text: "Dust model evaluations typically include a combination of surface concentration, deposition fluxes, and remote sensing measurements."

Page 3607, line: 15: 'A better vertical distribution' – this implies that models do not perform well in this aspect, which is in fact contradicted in this manuscript (the vertical structure appears to be quite well represented by the models, even if the total dust concentration is underestimated)

True. This is not what we intended to convey. We update the text to clarify.

New text: "An accurate representation of dust vertical structure is needed to model dust transport and deposition processes, to capture the effects of dust-radiation and dust-cloud interactions, and to properly produce to air quality forecasts."

Page 3610, line 9: 'information about the actual size distribution' – this statement is misleading, at least for some of the retrieval methods. To my understanding LIRIC algorithm retrieves the size distribution from sunphotometer inversions, which I would not label 'actual size distribution ' but

‘retrieved size distribution’.

We updated the text to clarify the meaning. What intended meaning was not that we know the actual size distribution, but that the retrieval contains some – indirect – information of the size distribution through the photometer sky-scanning observations.

New text: “These products are based on indirect observation of the aerosol size distribution --- instead of relying on a modeled size distribution --- further improving the results.”

Page 3611, line 20: ‘is’ -> ‘in’

Corrected.

Page 3613, Table 1: The table is difficult to understand without additional information. I suggest to provide more detailed information on which products are retrieved/available at the different levels described on Page 3613.

The table has been removed as part of the restructure of “Retrieval algorithm” section. See also the general comments.

Page 3614, line 7 ‘assumed aerosol intensive properties’ –Remark: Before such products can really be useful for model evaluation it must be clearly stated and transparent (1) what these assumptions are (e.g., assumption of constant microphysical properties is mentioned on page 3016) and if these are applicable/representative for the studied problem; and (2) to which extent the retrieved data product is sensitive to deviations from these assumptions.

We have added a separate discussion on the underlying assumptions of the algorithm where we also discuss the possible implication of these assumptions. We try to make clear to the reader that the effect of all these assumptions on the final result are not yet fully studied.

Page 3015-3016: LIRIC description: Unfortunately the authors refer for the detains of the LIRIC algorithm to a paper by Chaikovsky et al. (2015), which has the status ‘in preparation’ according to the references section. The other reference to the algorithm provided here is Chaikovsky et al (2012) which refers to a conference proceeding. Neither are peer reviewed or easily available to the reader. Please use other references here.

Correct. We have added new references to peer revieweed publications.

Page 3016, line 25: Regarding the retrieval errors of 20% - can this number be generalized for all dust cases? Granados-Munoz investigate the volume concentration derived by LIRIC, here mass concentrations are compared – what are the assumptions for particle densities?

The study of Granados-Munoz et al. examine a small set of cases, and their results cannot be generalized. They give, however, an indication of the order of magnitude of uncertainty that we can expect. We have tried to clarify this in the text.

We use a dust bulk density of 2.6g/cm3, similar value as the most models use. In this way we tried to remove an extra reason of discrepancy from the comparison. For the comparison of volume and mass concentration profiles are equivalent.

New text: "Granados-Munoz et al.(2014) show that the retrieval is stable to the choice of these parameters but further work is needed to generalize these results; in the examples shown in that paper, the result retrieval errors remains below 20%."

Page 3619, Table 2: Table 2 needs improvement. Model domain size is not shown in Table 2, please include the relevant information. Also, how frequently are the boundary data updated? What is meant by ''data assimilation' in this context – presumably 'dust assimilation? Overall the table is not clearly arranged, please improve the formatting.

We have added the missing information. We also tried to improve the arrangement of the table to improve its clarity.

Page 3019, line 20: Why is the NRT evaluation mentioned here while this is not a focus of the evaluation presented here? In fact it would be interesting to get some insights into possibilities of NRT evaluation of dust forecasts using the LIRIC data, mabe in the Outlook section.

To improve clarity of our argument we have integrated this part in the introduction. See also the general comments.

Page 3620, line 16 – Can you find a more recent reference than d'Almeida 1987?

Added reference to Mahowald et al. 2014.

Mahowald, N., Albani, S., Kok, J. F., Engelstaeder, S., Scanza, R., Ward, D. S. and Flanner, M. G.: The size distribution of desert dust aerosols and its impact on the Earth system, Aeolian Research, 15, 53–71, 2014.

Page 3620, line 20: 'this effect is expected to be small' – please explain why this would be the case.

We added a more detail explanation in the text. We expect that fine dust will be important percentage of the total dust volume only under strict conditions (e.g. long-range transport). We expect that when considering cases both near and far away from the source, from different sources and transport paths, the total effect of these few "fine-mode" cases will not be significant. However, as we also note in the text, this is just a guess and a quantitative estimation should be given.

New text: "When using a statistical approach, including different locations and transport paths, as in the present study, these few cases are expected to have a small effect in the overall comparison. The exact amount of fine-mode transported dust is an open issue and should be further investigated."

Page 3624, line 11: It is not surprising that RSME is strongly affected by outliers. To characterize dust events they could be characterized by the 95th percentiles which would also be a basis for a statistical comparison.

Following the reviewer suggestions we try to make the RMSE more robust. However, instead of using a 95th percentile, we detect outlier based on the discrepancy of the observations. We apply this procedure to all point statistics (concentration, centrer of mass, peak value, layer thickness). Specifically for each of these datasets, we calculate the difference of model and observations and we reject points that the difference is more than 4 standard deviations way

from the mean difference value.

Methodology section: Overall remark – thickness of dust layer would be an interesting additional test parameter – appear to have the tendency to be too diffusive in the vertical, as appears to be the case for DREAM-NMME-MACC in Figure 8. This could have important consequences for estimating CCN/IN supply at high altitudes from dust model results.

Good suggestions. We added the comparison of dust layer thickness.

Page 3625, line 16 'forecasting' -> 'simulating'

Corrected.

Page 3625, Line 22: Here cases are indicated where some of the models do not predict dust events which are detected by the observations. Of course such misrepresentation would lead to a low bias for the models, since only cases are compared where dust is observed by EARLINET/Aeronet, while other cases where models would simulate dust events at a different time (possibly due to deviations in the transport athway) would not be compared.

This is a important point. Indeed our sampling strategy could introduce some biases, if it is not interpreted correctly. Ideally, a model evaluation should consist of continuous / frequent measurements, including both dust and non-dust profiles. As the reviewer correctly points out our current dataset cannot answer the question "Does a model correctly represents the total dust burden?". It can answer however the questions "When dust arrives over Europe, do models predict the concentration and distribution correctly?". This is also a useful question for many users. We have changed the text to clearly identify the scope of the intercomparison.

New text: "The observational dataset was selected to include only dust cases and the results should be interpreted accordingly. For example, if models represent the correct amount of dust but predict its arrival at different time, this would appear as a model negative bias in our comparison. For evaluating the simulated aerosol burden, an observational strategy with systematic measurements should be followed."

For comparing individual cases, model results not only at the location but also in neighbouring model grid cells should be taken into account. For a statistical comparison it would be a fairer approach to compare the LIRIC profiles with maximum dust cases (e.g. 95th percentile values, clear sky conditions) from the models , to compare actual dust events. Of course the number of available profiles may be too low to allow for statistical significance.

These are good points. Our comparison takes into account neighbouring grid cells in space and time by using bilinear interpolation to the measurement location to calculate the model profiles. However, to better represent the variability of each case we performed the comparison with two extra profiles at -3h and +3h of the measuring time. The results from these extra profiles have been added as errorbars in figures 5 and 6 (old figure number are 4, 5, 6, 7). In this way the variability of each case can be readily evaluated. Using the 95th percentile is a valid idea but, as the reviewer remarks, the dataset we have is too small to make such a comparison meaningful. In addition our dataset includes also "weak" dust transport events, and the 95th percentile for models could lead to overestimation.

Page 3626, line 3: Not volume concentration, but mass loads are shown in Figure 5.

Corrected.

Page 3626, line 20: Can the model low bias be due to a misinterpretation of other aerosol types as dust in the LIRIC algorithm? Since the models only provide dust this would automatically lead to too low model dust.

Indeed, there is always a chance that we misinterpret some aerosol signal as dust, but we think that this possibility is small. The LIRIC algorithm separates the contribution of fine and coarse particles, and given depolarization lidar measurements it can also separate the coarse spherical and non-spherical components. So, when no depolarization measurements are available, our "dust" profiles could include marine particles, but this contribution should be mainly in the PBL, which we excluded from our study. When depolarization measurement are available, the only non-spherical dust aerosol type that could affect the measurements are volcanic aerosols, and as far as we know our cases are not affected by such type of events.

Also, in addition to the concentration values it would be educational and straightforward to also compare modeled optical thickness values with the Aeronet data. The models will likely match the AOD values much better, which may hint to misrepresentations of particle size distributions in the models. Also, where available comparison with extinction profiles would be useful additional information.

This is a good suggestion. However we think that using other measurements to evaluate the model is beyond the scope of this paper. Our aim is to try to explore the possibilities offered from new volume retrieval algorithms and not to perform a full evaluation of the models. The later, as the reviewer correctly suggests, should certainly include many more complementary measurements. This will be the topic of our future research. We add an appropriate discussion in the conclusion section.

New text: "Understanding the causes of model and observation discrepancies should be the topic of future evaluation studies including a variety of sensors, e.g. AERONET photometer, satellite AOD measurements, and in situ measurements, to explore different aspects of dust modelling systems."

Page 3628, line 27: I do not understand 'averaged at 1km altitude ranges', please explain.

We added a more detailed description of the procedure.

New text: "The data of the models and measurements were averaged at 1km thick altitude bins (from 1 to 2km, from 2 to 3km etc.) before calculating the statistics, to give an overview of the model performance at different altitudes."

Page 3629, line 23: The slightly worse performance of the 'Western' cluster may also be due to the dust production mechanism, which may be related to wet convective events near the Atlas mountain. Regional models have notorious problems to reproduce such dust cases. However, the discrepancies in the performance between the western and eastern clusters appear to be quite small.

This is a good point. We have incorporated your suggestion in the text.

New text: "Misrepresentation of wet convection events in the region of Atlas mountains, a known problem of regional dust models, can also contribute to this discrepancy (Reinfried et al., 2009; Solomos et al., 2012)."

Page 3230, line 9: '.. clear indication that the dust vertical structure by dust models needs to be further explored ' – this is not really shown by this paper. Instead the comparisons can be interpreted that the vertical structure is actually quite well represented in the model, while the concentrations are all lower than the LIRIC retrievals. So it appears to be more a problem of the concentration than the vertical structure?

Indeed, this part was not clear. We have changed the text to clarify this point.

New text: "This first comparison indicated that dust models correctly represent in average the dust structure, but their performance for simulating individual event structure and the exact amount of dust should be further explored."

Page 3230, line 12: Studying the reasons for differences between the models would to first order be a task for model intercomparisons, observation data can only be of limited use here. Comparison of model results and observations could help to pick out the 'best' model. It will not explain the reasons for differences – which may well lie in the simulated dust emissions, which should always be intercompared first.

Correct. We have changed the text to clarify this point.
New text: "Understanding the causes of model and observation discrepancies should be the topic of future evaluation studies including a variety of sensors, e.g. AERONET photometer, satellite AOD measurements, and in situ measurements, to explore different aspects of dust modeling systems."

Page 3230, line 17: I agree with the usefulness of the ensemble model approach – however, in the presented cases an ensemble approach would not have helped the fact that all models have lower concentrations than the LIRIC retrieval.

Indeed, the ensemble model would not fix all the differences between models and observations. But it would combine the good structure representation of two models with the better concentration of the other two and, in average, would bring the simulation nearer to the observations. We have changed the text to clarify this point.

New text: "In total, the study hints that an ensemble dust models products would better simulate the dust observations, even if some discrepancies would remain."

In fact, well tested models would have been evaluated not only with optical thickness measurements but also with surface concentration measurements and if possible particle size distribution. That all models (which are at least in part well established) have lower concentrations that the LIRIC retrievals is rather curious. Can you explain this? Please note that in most dust models the emission flux is 'tuned', i.e. scaled for best match with a set of observations. A useful dust model evaluation should therefore not be limited to one set of observations (but should at include e.g. both concentration and optical thickness observations)

We fully agree that a complete dust model evaluation would require a number of different observations. Our aim here is just to introduce a new element in the spectrum of different

data that can be used in model evaluation. Concerning the bias, one reason for the discrepancy could be the sampling strategy, as the reviewer noted before. Our datasets includes only measurements that dust was actually observed, but this could be balance-out by cases where the model predicted dust when it was not observed.

However, there are other studies that indicate a negative bias on the model side. Similar results were presented also at Mona et al. 2014, where model data were compared with lidar extinction measurements. A negative model bias is also found in the real-time evaluation against AERONET and MODIS (total) AOD at the SDS-WAS (http://sds-was.aemet.es/forecast-products/forecast-evaluation). So, while the dataset we are using has some weaknesses, we think that it points to the same direction as other studies.

Page 3230, line 21: 'automated retrieval algorithms' – that would indeed be useful in particular for NRT retrievals or even dust assimilation.

Indeed, we are working on it and hope to have results soon.

Anonymous Referee #2

Page 3614, line 18 and following lines: the forward scattering computations are typically based on Mie theory. It is an old problem that Mie theory still is the choice of light scattering model, even if dust properties are modeled. If Mie theory is used in deriving (some of) the modeling results the authors need to provide an uncertainty analysis. In how far does Mie theory create uncertainties that make it virtually impossible to compare your results to the LIRIC results? Do you find good agreement for the wrong reasons? AERONET retrievals are based on a different light-scattering model.

The point we tried to make in this paragraph was not clear. In our comparison we did not use Mie code to convert volume concentration to extinction values. This is what was done in previous studies and, as the reviewer correctly points out, it could lead to significant errors. Instead, LIRIC uses the photometer inversion to interpret the observed aerosol and directly compare the retrieved volume concentration profiles to model results. We have rewritten the paragraph to clarify the meaning of our argument.

It would be very helpful, if you showed error bars for each of your data points, e.g. Figure 5 (and many other figures); another option would be if you showed characteristic error bars for some data points. You should also include information on the extreme (maximum) uncertainties.

We have added error bars to both point measures and profiles. As we mention in the general part, the full characterization of uncertainties for the LIRIC algorithm is still an open issue. We now explain this in the text and carefully specify under each plot what is represented by the error bars. See also our general comments.

What was the quality of the lidar profiles in general? The LIRIC algorithm certainly requires a reasonable signal-to-noise ratio in order to derive meaningful profiles of the data products used in this study. What optical depths did you find for these cases? Did these optical depths fulfill thresholds required for the AERONET retrievals? Did you check if the aerosol conditions remained constant (homogeneous in time and/or space) between the AERONET observations and the lidar

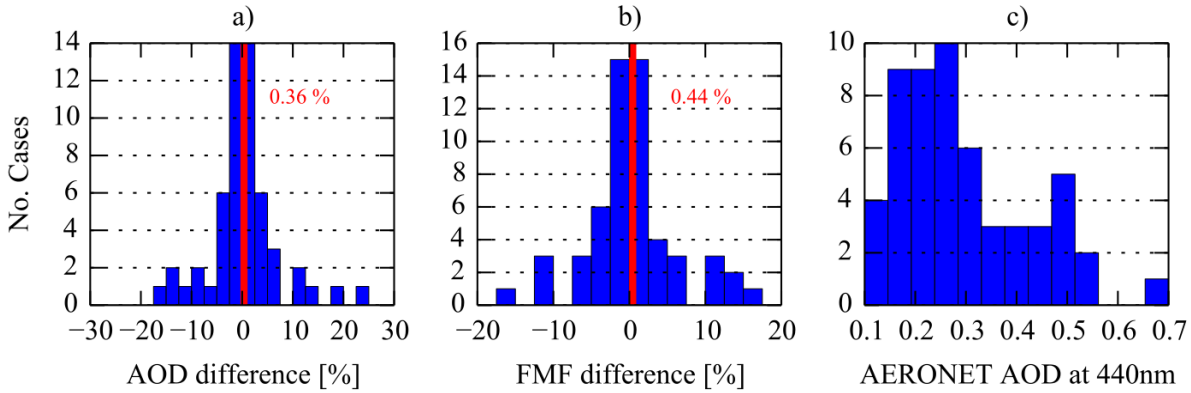
measurements. Did you check time series of optical depths, time series of backscatter profiles (or at least the range-corrected backscatter signals)? Did you check temporal/spatial variations of Angstrom exponents (backscatter related and/or extinction related if multiwavelength Raman lidar data were available) during the time between AERONET and lidar observations? Your manuscript does not show that you considered any of (respectively all) these potential quality assurance tests (there are more tests available like checking time series of depolarization ratio profiles and curtain plots) before you derived the dust profiles with LIRIC. Please provide a table in which you list the optical depths for these cases. Also list information on variability of optical depths, Angstrom exponents, backscatter coefficients and illustrate by how much these parameters changed between lidar and sun photometer observations. One or two concise scatter plots might be another option. It would be helpful if you also show comparisons of modeling results of extinction profiles (can they be provided by the models?) to Raman lidar measurements (many of the stations can provide you with extinction profiles, even if they are not used in LIRIC). This comparison would provide additional proof of the quality of the modeling results.

The reviewer is correct in that a complete quality control/assurance procedure is an open question. It is true that in the original processing we didn't apply a centralized processing. To cover these we implemented the following steps:

- **For each LIRIC retrieval we compared the AOD at 440nm during the AERONET inversion and the average AOD during the lidar measurement. If during the lidar measurements there was no direct-sun AERONET measurements, the average AOD was calculated by linearly interpolating the nearby measurements. In cases were large discrepancies were detected, the aerosol situation was checked using ancillary measurements and, if available, the lidar AOD at 532nm was also recalculated using Raman measurements; we converted the AOD from 532nm to 440nm assuming an angstrom exponent 1. If the difference of the two AODs was larger than 30% the LIRIC profile was rejected. Despite the differences in individual cases, the dataset has almost no bias (mean difference ~0.36%).**
- **The same procedure was used for the AERONET Fine mode fraction (FMF). The only difference was that for this comparison we were based only on AERONET measurements, as we couldn't derive comparable estimations of FMF from lidar measurements. If the difference of the two FMFs was larger than 30% the LIRIC profile was rejected. Despite the differences in individual cases, the dataset has almost no bias (mean difference ~0.44%).**
- **For each of the 61 LIRIC retrievals, we run HYSPLIT back-trajectories for both LIRIC and photometer measurements. The start time of the trajectories was rounded to the nearest hour. The trajectories were qualitatively inspected to find any possible significant change in the origin of the observed air masses. In case that such a difference we checked possible ancillary measurements (e.g. collocated ceilometer measurements) to justify the use of the two measurements. If this was not possible the LIRIC profile was rejected.**

In many cases the AOD at 440nm was lower than the AEROENT level 2 threshold of 0.4. We recognise that these retrievals could have larger uncertainties. However, the AERONET threshold was found too strict in other studies (e.g. Li et al. 2014), and would limit the available cases for our example dataset, especially in remote regions. This could introduce a bias against "weak" dust transport events. For this reason we kept cases with AOD at 440nm larger than 0.1 and provide to the reader a histogram detailing the available AODs.

In total, applying these quality control criteria we had to remove 6 cases, leaving our quality-checked dataset with 55 measurements. During this processes, two more cases were



reprocessed to achieve better agreement between photometer and lidar measurements.

The results will be summarized in the above figure, that will be included in the final manuscript for clarify the assumptions and weaknesses of the dataset.

Li, J., Carlson, B. E., Dubovik, O. and Lacis, A. A.: Recent trends in aerosol optical properties derived from AERONET measurements, Atmos. Chem. Phys., 14(22), 12271–12289, doi:10.5194/acp-14-12271-2014, 2014.

Page 3616, line 3: "constant microphysical properties" are assumed throughout the atmosphere. Please provide sufficient evidence in your paper that microphysics did not change significantly for the case you used in this paper. Otherwise this study is affected by a large unknown error source.

It is true that the effects of this assumption are hard to quantify. The text at this point was not clear, we have added now a more detailed description about all of LIRIC's assumptions in the new version. We would like to specify that LIRIC assumes that the micro-physical properties of each mode do not change with altitude. This would not be true if two different aerosol type are present in the same mode. For example, the assumption will be invalid if the atmosphere includes two coarse non-spherical aerosol types e.g. dust and volcanic aerosol. We argue in the text that such cases are rare, but we clearly now mention that we cannot exclude this possibility.

New text: "First, each aerosol mode is considered to have constant microphysical properties with altitude, and only vary its concentration. In case that two aerosol types are averaged in one mode, e.g. when smoke and urban particles are both present in the atmosphere, this assumption will introduce some errors. When no depolarization measurements are present, so the coarse mode is not separate to spherical and non-spherical parts, the coarse mode could include both marine and dust particles, but this will affect mainly the PBL. With depolarization measurements available, LIRIC retrieves the coarse spheroid mode, and this could incorporate more than one aerosol types if the atmosphere includes desert and volcanic dust, or even dust from two very different sources. These cases are rare and will have small effect in a statistical comparison. We cannot exclude, however, that they can become important for specific cases."

Page 3610, line 6: "under certain assumptions". Please be specific about what these assumptions

are, how they affect your results, and if measurement situations of (mixed dust) were present for the cases considered here.

We have added a separate paragraph discussing in detail the major assumptions and a qualitative discussion of the introduced uncertainties.

I am also wondering if the model results naturally end up in a bias, as the models are used for dust modeling, whereas LIRIC can be used for dust, mixed-dust and non dust caes. In how far were the "dust cases" used in this study significant enough (e.g. dust optical depth). Can you compare dust modeling to LIRIC retrieval results of "dust" profiles as the plumes may have contained other aerosol components?

The benefit of the LIRIC retrieval is exactly that. It can separate the contribution of dust from that of other aerosol types, so we can compare directly dust volume concentration profiles from models and observations. The algorithm will not be able to separate dust and volcanic aerosol particles, but such cases are few and well studied.

Page 3616, line 16: you write that "LIRIC has proven to be a robust algorithm for aerosol volume concentration retrieval": I take issue with this comment. I am not aware of any study in which the LIRIC results were compared to direct measurements of volume concentrations, e.g. by carrying out in-situ measurements of particle size distributions (e.g. aircraft measurements). Please rephrase such comments and discuss the results of LIRIC. You need to make clear that LIRIC has not been validated. Explain why you have confidence in the LIRIC results. Provide references to corroborate your explanations. Show an uncertainty analysis. Please also remove comments like "actual size distribution" (line 9, page 3610), as this implies that LIRIC "measures" size distributions.

We have removed the comment and now point out at the open issue concerning LIRIC validation / uncertainty characterization.

Section 2.2, page 3613 and table 1: please be more specific how the data products in table one are related to the methodology, particularly as some data products are retrieved (LIRIC), some can be measured, some data products may be more of qualitative nature; again: LIRIC has not been validated with independent methods.

We have removed this table as part of the restructuring and condensing the paper, as described in the general comments.

Page 3615, description of LIRIC and reference to Chaikovsky et al.: please provide a short description of LIRIC. Chaikhovsky seems under review. Most of the references are not peer reviewed literature. As the LIRIC results are at the core of your study it is justified that you outline the most important parts of LIRIC in this paper. Please provide a discussion of the errors and assumptions involved in LIRIC retrievals and by how much LIRIC results might be biased. It may not necessarily be the case that only the model results are biased.

We have added new references and a more extended description of LIRIC. We also added a discussion of underlying assumptions and uncertainties, and clearly stated that quantitative estimation of LIRIC uncertainties are still missing.

Page 3619 (lines 5-27) – 3620 (lines 1-7) and also rest of section 3: This section is about the methodology, but it is kept very unspecific. Please provide hard facts, numbers, and references. A

case study would help significantly in explaining the various concepts explained in this part of the methodology section.

Indeed, we have integrated large part of this section to the introduction to make our argument more clear. Following the reviewers suggestions, we have added a new figure that gives a visual overview of LIRIC input, the retrieved product, and the LIRIC/model comparison.

Page 3624, lines 26/27 (I am mainly repeating what I already wrote before): you write that the time difference was kept small. This is not sufficient. Did you check if wind direction changed between lidar and sun photometer measurement? I do not find information on the possibility that, e.g. the center of mass (peak of the backscatter profiles can certainly be used as an approximation) changed between lidar and sun photometer observations. Such a variation would distort the profiles (from LIRIC), and I am wondering if such an effect would lead to a misinterpretation of the quality of the modeling results. Did you check if optical depth between time of lidar and SPM measurement change? That might indicate a change of (at least) extensive properties. How can you rule out that intensive properties did not change? Please provide some evidence that the lidar and sun photometer data did in fact provide optical data that describe similar aerosol conditions, regardless of the temporal differences between lidar and SPM measurements.

As described before we now implement an extended quality check examining differences in AOD and fine mode fraction, and compared the back trajectory analysis for all photometer and lidar cases. Our analysis shows that the differences in AOD and FMF remain low, and in any case have an average very near 0, as expected. This means that these differences don't introduce a bias in the total dataset. It is worth mentioning that a change in the centre of mass between photometer and lidar measurements will not affect the quality of LIRIC. The parameter that would effect would be a change in the quantity or type of aerosols in the atmosphere, and not their altitude.

Page 3625, line 14 – and following lines: you compare center of mass and peak value of mass. It might be worthwhile if you include modeling results that do not only describe the grid point of the lidar/sunphotometer location. It certainly could be the case that the model puts the center of mass at a neighboring grid point in a different height. It would be helpful to know if differences of center of mass are less (or larger?) if you use neighboring grids. In fact, such an analysis would also be helpful for the other data products.

This is correct. Indeed our model dataset does not consist of data from a single grid point but uses bilinear interpolation from 4 neighbouring point of the station location. The same procedure is performed for two model profiles (stored every 3 hours) and the final model profiles is extracted as a linear interpolation between the two. So our model data take into account both the spatial and temporal variability of the model performance. To study also in the variability of each case, we performed the same procedure for two profiles at -3h and +3h from the time of measurements. We indicate the range of results obtained in this way as errorbars in figures 5 and 6 (old figure number are 4, 5, 6, 7). In this way the variability of each case can be evaluated.

Line 5, page 3626: please avoid giving the reader the impression that LIRIC provides the truth ("the model underestimated the total amount of dust"). There are too many model assumptions involved in LIRIC and there is no independent study that would allow us to judge for which measurement situation LIRIC is correct.

Correct. We have improved the wording.

Page 3607: dust models are validated with optical depths data?

Following also the comments of 1st reviewer, we have specified that dust models are evaluated using optical depth and in-situ measurements.

Section 3, general comment: I think a considerable part of the text belongs to the introduction. This part of the methodology section is very unspecific. I would like to see a case study in which you guide the reader through an example: data, LIRIC data products, modeling results, including uncertainty analysis. The figures show a nice overview on the analysis of many measurement examples taken from several stations. But it is nearly impossible to judge, how well LIRIC performs.

As described before, we have tried to condense the section, and add a figure to illustrate the steps the necessary steps and metrics used in the methodology.

Please combine figures in plates. This would make the paper more organized in view of the relatively short text. I suggest that you combine Figures 4 – 7 into one plate, Figure 8 and 9 into one plate, figures 10 -13 into one plate, and figures 15, 16 into one plate.

Good point. We have condensed several figures to combined plates. The old figures that are now grouped are (2, 3), (4,5,6), (10, 11, 12, 13), (15, 16). We think that combining figures 8 and 9 would be more confusing, so we left them as before.

Interactive comment from G.P. Gobbi

We would like to thank Prof. Gobbi for providing his input and comments.

I wish to thank the authors for citing (page 3609, line 26) my ACP paper of 2013. Still, I believe the most pertinent papers which should be cited in the context of this article are the ones written starting 2001 and reported below. This is because those papers addressed the same issues this article addresses (lidar estimates of dust volume/mass profiles and comparisons with dust model equivalent forecasts). Specifically:

Kishcha, P., Barnaba, F., Gobbi., P., Alpert, P., Shtivelman, A., Krichak, S.O., Joseph, J.H. (2005), Vertical distribution of Saharan dust over Rome: comparison between 3-year model predictions and lidar soundings.doi:10.1029/2004JD005480 J. Geophys. Res., 110, D06208,

Which compares 34 lidar retrievals of dust volume profiles collected in Rome over a 3-year time span, against the Tel Aviv DREAM model forecasts of dust volume profiles. In 2007 such comparison was extended to the Skiron and to the Barcelona DREAM models:

Kishcha, P., Alpert, P., Shtivelman, A., Krichak, S., Joseph, J., Kallos, G., Spyrou, C., Gobbi, G.P., Barnaba, F., Nickovic, S., Perez, C., and J.M. Baldasano. Forecast errors in dust vertical distributions over Rome (Italy): Multiple particle size representation and cloud contributions. J. Geophys. Res., 112, D15205, doi:10.1029/2006JD007427,2007.

Conversion of aerosol backscatter signals (from single-wavelength polarization lidar) to the dust volume profiles employed in these two papers was based on the (non-spherical) aerosol scattering model published in: Barnaba F. and Gobbi. G.P., " Lidar estimation of tropospheric aerosol extinction, surface area and volume: Maritime and desert-dust cases", Journal of Geophysical Research, 106-D3, 3005-3018, 2001 (correction to Table 7 and 8 in Barnaba and Gobbi,JGR, 107, D13, 10.1029/2002JD002340, 2002).

The outcomes of the above aerosol scattering model (Barnaba and Gobbi, 2001) have been validated in the following papers:

Against sunphotometer extinction measurements in: Gobbi, G. P., F. Barnaba, M. Blumthaler, G. Labow and J. R. Herman, Observed effects of particles non-sphericity on the retrieval of marine and desert dust aerosol optical depth by lidar, Atmospheric Research, 61, 1-14, 2002

Against in-situ Extinction, Volume and surface area measurements in: Gobbi, G. P.,F.Barnaba, R. Van Dingenen, J. P.Putaud, M. Mircea, and M. C. Facchini, Lidar andC1298in situ observations of continental and Saharan aerosol: closure analysis of particlesoptical and physical properties, Atmospheric Chemistry and Physics, 3, 2161-2172,2003,

Against extinction measurements at 351nm in: Barnaba, F., F. De Tomasi, G. P. Gobbi, M. R. Perrone, and A. Tafuro, Extinction versus backscatter relationships for lidar applications at 351 nm: maritime and desert aerosol simulations and comparison with observations, Atmospheric Research, 70, 229–259, 2004.

Relationships between Aerosol extinction and volume obtained by the aerosol scattering model (Barnaba and Gobbi, 2001) have been published in: Barnaba, F. and G. P.Gobbi, Aerosol seasonal variability over the Mediterranean region and relative impact of maritime, continental and Saharan dust particles over the basin from MODIS datain the year 2001. Atmospheric Chemistry and Physics, Vol. 4, 2367-2391, 2004.

Ansmann et al. (ACP, 12, 9399, 2012), showed these "mass-specific extinction coefficients" to match most of the observed/calculated ones published in the relevant literature at that date.

It is worth mentioning that the Kishcha et al. (2007) paper was based on comparisons of 34 lidar-derived volume profiles collected at one single station, while this AMTD paper addresses 61, asynchronous profiles collected amongst ten lidar stations. Still, correlations between model and lidar retrievals are similar in both exercises. Conversely, Kishcha et al. (2007) found no particular bias in the differences between model (Skiron and BSC-DREAM) and lidar volume estimates, while an important (mean) negative bias is reported here.

Thank you for the useful suggestions. We have modified the paper to discuss and present the related works of *Kishcha et al.* that help put our work in a wider context.

Indeed the two studies seem to point to different directions but there are several differences that make them not directly comparable. The main difference is that the BSC-DREAM used at Kishsca et al. is an older version than the one used here (BSC-DREAM8bV2). In the Kischsca et al. paper the lidar profiles include total volume concentration retrievals while LIRIC utilizes multipspectral and depolarization information to separate spherical and non-spherical aerosol volume concentration. The measurements at the mentioned paper include

only cases from March-June while the present study includes some cases also in other periods. Lastly the selection of cases is different, as Kischsca et al. choose dust cases based on volume depolarization ratio and scattering ratios. In total we think that the two studies are not directly comparable.

As a support that our results are not in contrast with existing literature, we should mention that similar negative bias was found in the study of Mona et al., ACP, 2014 that compared 12 years of lidar and DREAM8b profiles. A negative bias for most of the models is also found in the model evaluation at SDS-WAS against AERONET and MODIS data (http://sds-was.aemet.es/forecast-products/forecast-evaluation).

In this respect, to help understanding the reasons of such bias it might be useful to compare lidar and model profiles at each single station, so to address the effects that both spatial variability and large variability in number of profiles available at each station (from 1 to 18, Table 4) can have on the average results presented here.

We fully agree that further work is needed to understand the source of the model / observation discrepancy. With our current dataset, however, we cannot get useful results by examining the cases by station: some stations contribute only few profiles that are not enough to get meaningful statistics. This is exactly the reason why we performed the east/west cluster analysis, that both show similar bias. But we acknowledge that this is a remaining issue. We have updated the manuscript to reflect these concerns.

In fact, "bias" is defined here with respect to a retrieval, not actual measurements of dust volume or mass. Since European Guidelines exist to evaluate the Saharan dust contribution to PM10, and PM10 measurements are widespread and easily accessible all over Europe, it might be of benefit to the paper to assess if the models employed in this work keep showing similar negative biases when comparing their dust loads at the ground to the dust content derived from PM10 ground measurements.

Indeed to fully understand the biases, we would need to perform a full model evaluation using a step of complementary measurements. However, as we pointed out also to reviewer 1, we think that using other measurements to evaluate the model is beyond the scope of this paper. Our aim is simply to present the new volume retrieval algorithms as a new part of such evaluation and not to perform an evaluation of the model. The latter, as both reviewers correctly suggest, should certainly include, many complementary measurements. We added a discussion in the conclusion section, and tried to better clarify the scope of the paper.

Interactive comment from A. Ansmann

First of all I have a minor problem with the co-author list: There are, e.g., 5 co-authors from Athens, but only 3 LIRIC profiles (probably produced by one person), and then there is Evora (co-author V. Carrasco, new EARLINET scientist?) with 18 profiles (contributing 30% to the 61 considered LIRIC profiles), but Evora has only one co-author.

Furthermore: What about the quality of the Evora data, when obviously analysed by a scientist of which the LIRIC expertise is at least unknown? As mentioned this is a minor issue, but I did not understand on which rules the co-author list is based.....

The co-authors from each observation and modelling group were given from the groups (suggestion 2-3) in a way to give proper credit to people that actually contributed to this work. However, some of the co-authors are not related to performing measurements and providing data but to data analysis and the overall preparation of the manuscript. Concerning Athens, please note that the 5 co-authors are coming from 2 independent research institutes (National Observatory of Athens and National Technical University of Athens), with distinct contributions in this manuscript. Please also note that we have 2 co-authors from Evora (V. Carrasco and Sergio Pereira). S. Pereira is a well-known member of EARLINET community and has considerable experience using LIRIC. In any case, all retrievals were centrally checked and discussed before accepting them in the study.

Introduction:

General remark: Except for the abstract, all acronyms or short names should be explained (give full explanation!) at their first appearance. . .. DREAM, BSC, NMME, MAAC, POLIPHON,

Done.

Page 3609: The first EARLINET network paper on a Saharan dust outbreak was presented by Ansmann et al. (JGR, 2003). Should be cited.

Correct. We have added missing references to better represent EARLINET work.

New text: "The vertical distribution of dust over Europe has been studied using active remote sensing instruments such as lidars (Ansmann et al. 2003, Papayannis et al. 2005, 2008)"

Algorithms and Models:

Page 3612, lines 14-17: One could also mention one of the important Chinese photo-meter networks: Ground-based aerosol climatology of China: aerosol optical depths from the China Aerosol Remote Sensing Network (CARSNET) 2002–2013 H. Che, X. Zhang, X. Xia, P. Goloub, B. Holben, H. Zhao, Y. Wang, X. Zhang, H. Wang, L. Blarel, B. Damiri, R. Zhang, X. Deng, Y. Ma, T. Wang, F. Geng, B. Qi, J. Zhu, J. Yu, Q. Chen, and G. Shi Atmos. Chem. Phys. Discuss., 15, 12715-12776, 2015

Thank you for the suggestion. We have added the references.

Page 3615, line 10: OPAC: How can OPAC be used to separate dust and nondust! The original OPAC data base only deals with spherical particles.

Indeed, the text in this part was not clear. In the mentioned study OPAC was used to derive the volume-to-extinction ratio for dust and non-dust particles. We have changed the text to clarify the meaning.

New text: "In a similar approach, Nemuc et al. (2013) derive the volume-to-extinction ratio of different aerosol types from the Optical Properties of Aerosols and Clouds (OPAC) database."

Page 3615, lines 14-22: I would like to see a more balanced discussion on the two methods, POLIPHON versus LIRIC. LIRIC is not just of advantage in many situations, t needs clear skies, three wavelengths, it needs assumptions on particle optical properties for the overlap region (from the lidar up to 500-1000m above the lidar), and also not too complicated vertical aerosol layering with mixtures of dust, urban haze, smoke, and marine particles. POLIPHON seems to be more robust, needs only one wavelength, has no overlap effect when looking at lofted dust layers, and does obviously not need cloudless conditions.

As described in the general comments, we have now included a more detailed discussion of LIRIC, its underlying assumptions, its limitations, its relationship relating to other algorithms, including POLIPHON.

The need for clear skies (cloudless conditions) for the application of the LIRIC method may bias the model-observation comparison because the probability is high that wet removal effects are underestimated.

The dust particles observed with LIRIC can be affect by wet removal processes throughout their transport path (e.g. over Africa and the Mediterranean). In this respect, the limitation that LIRIC poses is that clouds/wet removal were present in the very last part of the transport. However the cloud-free conditions could indeed bias the dataset towards in specific dust transport paths etc. We have added a note about this possibility in the description of the dataset.

New text: "The selection of cloud-free sky could bias our sampling to meteorological conditions and transport paths that favour such cloud-free weather."

Page 3615, line 23: How was the data set of 61 profiles produced? All groups delivered their signals, and the LIRIC products were analysed by one group, or all the groups analysed their own signals. Should be clearly mentioned. Even this will certainly introduce some variability in all the observed data and the data comparisons.

This is a good point. Each station performed its own retrievals but all results were centrally checked. In the updated version, we specified this in the text. Following reviewer 2 suggestions, we also implemented some quality checks (see also general comments)

Page 3616, lines 16-25: If I read these papers of Wagner et al. (2013) and Granados-Munoz et al. (2014) I do not have the impression that LIRIC is robust. . ., and the errors are always below 20%. A very careful preparation of the signal profiles and all the input parameters seems to be necessary before using LIRIC.

Correct. We have removed the claim that LIRIC is a robust algorithm, and now clearly state that the exact error characterization is still an open issue. We also clarify that Granados-Munoz results are specific for these cases, and results should be extrapolated with caution.

Page 3622, lines 11-12: I think this is speculation that the hydrophobic dust particles change their microphysical (size, shape) and optical properties (including depolarization features) as a function of coating and water uptake. Yes, that may be true, but how large are these effects. Is there any reference for the hypothesis?

Indeed these changes, especially in the depolarization, are party a speculation. There are however clear indication that dust after long range transport is partly hygroscopic (e.g. Perry et al., 2004 – Dust transport from Asia to America – see also Sulivan et al. 2007). Concerning depolarization, even if the more water uptake was limited to more hydrophilic fine mode dust particles, this could affect the overall depolarization of the dust population by few percent. We have changed the text to make the speculative part clear, and added a reference.

Perry, K. D., Cliff, S. S. and Jimenez-Cruz, M. P.: Evidence for hygroscopic mineral dust particles from the Intercontinental Transport and Chemical Transformation Experiment: HYGROSCOPIC MINERAL DUST PARTICLES, Journal of Geophysical Research: Atmospheres, 109, D23S28, doi:10.1029/2004JD004979, 2004.

Sullivan, R. C., Guazzotti, S. A., Sodeman, D. A. and Prather, K. A.: Direct observations of the atmospheric processing of Asian mineral dust, Atmospheric Chemistry and Physics, 7(5), 1213–1236, 2007.

Page 3625, lines 2-5: As mentioned, LIRIC needs cloud-free conditions during the lidar observations. This means that the dust cases considered here are (in the majority) not affected by any wet deposition which can remove significant amounts of dust. This point should be mentioned.

As discussed before, wet deposition could affect our dataset throughout the transport, from the source to the observation station. LIRIC needs cloud-free conditions only during observation. In the new version, we have added a note about the possible cloud-free biases that could be included in the dataset.

Figure 8: Is it possible to put variability bars (standard deviations) to the observational profile and also to the modelled ones?

Good suggestions. We have added to all average concentration profiles error bars indicating standard error of the mean value.

Figures 9 and 15: standard deviation bars?

Added.