Review of the manuscript: 'Mediterranean desert dust outbreaks and their vertical structure based on remote sensing data', submitted to ACP by Gkikas et al.

This manuscript by Gkikas et al. aims at describing the horizontal and vertical distribution of desert dust during 'intense events'. The methods followed to reach this aim include the use of satellite data from passive and active sensors. The work is largely an extension of a previous paper by Gkikas et al. (2013), and this makes it a bit lacking in originality. The manuscript could provide potentially interesting information. However, in its present form it shows some major weaknesses that substantially compromise the validity of the results reached. These weaknesses, as well as some additional comments, are detailed below. For what follows, I cannot recommend the paper for publication until these major weaknesses are properly addressed. I also recommend review of the language as the text is often confusing and difficult to follow.

Major weaknesses of the work:

1) Methodological Problems.

Although the study follows a methodology almost identical to the one already published in Gkikas et al. (2013), in my view this methodology builds on assumptions that should AT LEAST be further commented and 'tested' to make the reader understand how reliable the derived results are. In particular a key point of the work is the identification of 'intense dust events' ant their separation into 'strong dust episodes' and 'extreme dust episodes' by the so-defined 'objective and dynamic satellite algorithm'. This selection is basically fully dependent on the AOD threshold chosen, defined as 'AOD Mean' in the text. Unfortunately I could not find any definition of this 'AOD Mean' in the text other that '*the mean (Mean) and the associated standard deviation (SD) are calculated for the whole study period*' (page 27688, lines 18-20).

1.1) What does this 'Mean' mean?

You use over 10 years of AOD data, so: are you getting a 1x1-resolved 'AOD Mean' by simply averaging the 1x1-resolved AOD time series corresponding to each pixel? (This is what I understood reading Gikaks et al. (2013), but this is not clarified in this manuscript).

What's the total number of AOD data points you have in each pixel? Are all years within the record equally represented? Are all seasons (or even months) in each year equally represented?

You should show this is the case, otherwise the 'Mean' AOD value you get might be meaningless. In fact, given the inter-annual variability of AOD, and, above all, given its marked seasonal variability, an unequal data coverage of the different years/seasons could lead to a 'biased' Mean AOD.

More explicitly: if, for example, the number of data points in winter is 50-60% of those in summer (this is likely due to enhanced cloud cover in winter) you derive a summer-biased 'Mean' AOD which alters all the subsequent analysis.

Therefore, please provide a clear indication of the number of data points you are including in your statistics (for example with maps in an Appendix), and of their inter- and intra-annual distribution. If the monthly-resolved data coverage is not uniform (as I suspect), and if you still prefer to perform the whole analysis using a single, pixel-resolved 'AOD Mean' threshold value, then you should rather obtain it as an average of monthly-mean AODs.

A map of the derived AOD Mean values would also be of help to better interpret the final results.

1.2) I also see another problem in the computation of the 'Mean AOD'. Your algorithm firstly performs selection of the 'strong dust episodes' and of 'extreme dust episodes' based on the 'Mean AOD' threshold and AFTERWARDS uses the AI and FF information to exclude possible 'not-dust' cases. However, this means that if, by paradox, all the days of the time series are 'dust-affected', and assuming for simplicity a normal distribution of those AOD values around their mean, following your approach only a very limited fraction of them would be classified as 'strong' dust (the 2.2%) or 'extreme dust' (the 0.003%) (this is because in a normal distribution the 2.2% and the 0.003% of data exceed respectively the 'Mean + 2 s. d.' and 'Mean + 4 s. d.' thresholds you fixed).

This is to say that, in my view, dust events should be firstly filtered out of the record to compute a sort of 'dust-free' mean AOD; otherwise, again, you get a biased 'Mean AOD' to which to compare the 'dust-affected' record. The way to do this could be based on a combined analysis including the other parameters you consider in your analysis (Angstrom, AI and FF). Obviously, this problem mostly affect those regions with higher dust-events frequency. Can you comment on that?

2) Dataset problems.

2.1) You make several efforts in the text to highlight the differences of this work with respect to the previous one (Gkikas et al., 2013) which, as mentioned above, is very similar to this in terms of methodology and structure. In fact, one of the differences with that paper is the extension in time and the use of the additional similar datasets from Modis-AQUA. So, one of the potentially interesting points of the study relates to the differences found between the TERRA and the AQUA-based results. Unfortunately, at the present stage, the validity of this comparison is completely jeopardized by the different time-periods used in the manuscript for the two Modis sensors. In those cases in which differences between the two outcomes are found and commented in the text, there is always the ambiguity whether those differences

are due to a 'real' aerosol (mostly desert-dust in this case) diurnal variability, or rather to the different time-period covered by the two datasets (note that this point is also connected to my point 1). It is well know that AQUA has a shorter time-coverage with respect to TERRA, but I do not see the reason not to limit the analysis to the period 2003-2012, which is common to both sensors.

In this respect, in Section 4 (page 27691) you specify that: '*In order to investigate this difference in detail we have also applied the satellite algorithm, over the period 2003–2012, i.e. that of Aqua, using MODIS-Terra retrievals as inputs. Through this analysis (results not shown here), it is evident that there is a very good agreement between the satellite algorithm's outputs, for the periods March 2000–February 2013 and 2003– 2012, revealing a constant dust episodes' regime. Therefore, the discrepancy appeared between MODIS-Terra and MODIS-Aqua spatial distributions in Fig. 2, is attributed to the diurnal variation of factors regulating the emission and transport of dust particles from the sources areas.*' Apart from the fact that you decide not to show in the text an element that would be fundamental in your analysis, in my view this sentence is by no means sufficient to justify the use of a different time-period coverage of the AQUA and TERRA Modis data in your study.

I think the same period for both sensors should be used to strengthen the results reached.

2.2) In the present form, description of the datasets used and of the way the different variables are matched is lacking. Description of the single datasets does not allow to get all the necessary information to understand their advantages and limits. I found some more details in Gkikas et al. (2013), but it is a bit annoying to always go to that paper to better understand the current one. Please provide more details to your Section 2 (e.g., 1- resolution of the AI from TOMS and OMI is a bit different, can you specify this? how do you match these values with the MODIS derived ones? 2- What's the exact meaning of 'Quality assurance-weighted' data?).

In Section 2.1.1 you only give the expected accuracy of the AOD data used. Which is the accuracy of the other MODIS-derived parameters employed in the study? How does this accuracy change above land and ocean? Is it sufficient to make this products suitable to be employed for scientific purposes?

The algorithm uses the information on Angstrom Exponent (AE), AI and FF (plus r_{eff} over sea) to select 'strong' dust and 'extreme' dust events. However, there is very little information in the text on HOW the matching between AOD, AE, FF (plus r_{eff}) and AI is operatively done at the pixel level. In particular the manuscript lacks in describing the statistics of the coincident multi-parameter dataset.

I guess you do not always have ALL the parameters available at the same time. What happens in case you do not have coincident datasets? How frequent these cases are? What's the impact of this on the final outcome of your study?

3) Presentation of Results.

3.1) Results are reported in Section 4 which however includes a large body of material intended to provide a sort of 'validation' of the method followed (comparison with AERONET and in situ PM10 data). This is a bit confusing as, in my view, the logical sequence would be to present the methodology, then check/demonstrate its validity, and only at that point present the results obtained by that methodology.

I would therefore rename the relevant sections accordingly.

Additionally, for its contents, the evaluation of the Method (Section 4.2) within Section 4 does not represent a real 'validation' but rather a 'comparison' with other datasets. Just to mention an example: the comparison of the AERONET AOD to the MODIS one (Figure 5) does not represent a validation of your 'objective and dynamic satellite algorithm' but rather a 'validation' of the Modis-AOD-retrieval algorithm.

Therefore, I would avoid using expressions as 'the performances of the satellite algorithm are evaluated', widely used throughout the text, and rather refer to this material as: 'the results of the satellite algorithm are compared to…', which is completely different.

3.2) In the same Section, I also believe the comparison with PM10 measurements has little validity in the context of this work. You want to report on the vertical structure of desert dust events, as clearly highlighted in the title of the manuscript. Your results show that several dust events do not reach the ground (see for example Figure 12), so: what do you expect to derive from the straightforward comparison of (columnar) AOD to (ground-level) PM10? I think this topic is of potential interest in general, but not in the form it is presented here. I would remove this part from the manuscript, as it does not add much to the text and rather makes it more confusing and weak.

Other general comments

- Title: I would suggest to modify it as 'Mediterranean Intense Desert Dust Outbreaks from columnar and vertically–resolved remote sensing data'

- Please define somewhere at the beginning of the manuscript the term 'intense dust events' you often refer to in the text and use the acronym IDD to refer to it. Specify clearly that, according to your classification, IDD events are divided into 'strong dust events' and 'extreme dust events' (and use respectively the acronyms SDD and EDD to indicate them throughout the text). This will improve its readability.

- Please define somewhere at the beginning the exact study region considered.

Minor Comments

There are several minor revisions the manuscript would need. However, as major revisions are requested and the text will probably change a lot in its revised version, a not exhaustive list of minor comments is given below.

Section 2

Section 2.1.1. The ocean and land Angstrom exponents are computed using different wavelengths. Please explain why you use the same threshold for them in your algorithm.

Page 27686, Lines 8-13. Confusing, please explain better

Section 2.2.2

Please specify better which data from AirBase and/or EUSAAR are you using (if this part will be kept in the revised version, which I discourage).

Section 3

Page 27689

> Line 22: It is not the quality of results to be improved but rather the quality of input data. Please rephrase.

> Lines 25-28: '…*in the present version of the algorithm are not implemented temporal filters, concerning the availability (masking out of AOD grid cells with less than 50% available data of the time-series) of raw AOD data, in contrast to Gkikas et al. (2009, 2013)*'. This sentence is not clear to me. Do you mean in this work L3 pixels having less than 50% of L2 data are included? If so, do you consider this as an improvement of the methodology?

> Line 29 to the next page: Rephrase as this sentence is confusing. At the beginning I understood you performed the analysis only considering CF < 0.8, but then in Section 4 you mention (Page 27692, Lines 9-12): '…*The analysis has been repeated (results not shown here) considering only AODs associated with cloud fractions lower/equal than 0.8,…*' . I think you should restrict the analysis JUST to the cases with CF < 0.8, as AOD is not reliable above this threshold.

Section 4

Page 27690, Line 21-22: '..*The obtained patterns are in a very good agreement with those presented by Gkikas et al. (2013),..*' . This is just an example of similar comments you often insert in the text

to comment your results. However my suggestion is to avoid repetition of this concept as it is rather obvious that results comparable to those presented in Gkikas et al. (2013) are obtained in this study, which follows a very similar methodology.

Page 27692, Lines 9-12: '…*The analysis has been repeated (results not shown here) considering only AODs associated with cloud fractions lower/equal than 0.8,…*' . You are commenting something which is not shown and provide no explanation of the differences found. As already commented above for the same sentence, I think your study should be limited to the cases with CF < 0.8, avoiding most of the comments at the end of Section 4.1.

Page 27693, Section 4.2: The title of this section is inappropriate for the reasons explained in my general comments 3.1 and 3.2.

Page 27694

Lines 5-7: Do you mean you '*…found at least one strong or extreme dust episode*' over the whole period considered?

Lines 5-19: This part is quite confusing. What's your aim here and how are you pursuing it? How do you define a 'clim' value? (what does it mean 'calculated from all the available retrievals'? how many data-points are used? Do these data cover the same period of your satellite dataset?).

Additionally, when you refer to AERONET data please use 'ground-based' instead of 'ground' as this latter can be confused with 'in situ'.

Section 4.2.1

Subsections of Section 4.2.1 should be numbered.

If I understand correctly, you are comparing the 1x1 degree satellite data with the AERONET data (measured in a specific site). Please at least comment on the expected spatial variability of AOD within a 1x1 cell, and therefore on the validity of such an approach.

Page 27694, Line 22: '*…346 pixel level intense DD episodes*'. To understand the relevance of this number it would be important to mention somewhere how many pixels you have in your domain, how many of these are classified as intense DD (IDD).

Section 4.3

How much do the Calipso-based result change if you use the overall calipso database and its aerosol type discrimination to investigate desert dust, not limiting only to those cases previously classified as IDD in your scheme?

Figures:

Figure 1

The scheme of the work is exactly the same of Figure 2 in Gkikas et al. (2013), which is quite inconvenient, please modify highlighting differences from that work or remove the Figure.

Figure 2

- It should be enlarged as it is not very readable at the moment.

- As commented above, it should refer to the same period for Aqua and Terra.

- There is a clear discontinuity between land and ocean, can you comment?

- Change the color scale to more clear numbers (e.g. 0-10, top, 0 – 3 bottom)

Figure 3

- It should be enlarged as it is not very readable at the moment.

- In the i-plots, I see a problem of misclassification over the Po valley in Italy. Can you comment on that?

Figure 12

- It should be enlarged as it is not very readable at the moment.

- Please specify the units of the backscatter values.