Dear Reviewer,

We thank you for doing this review and for your thoughtful and strict suggestions that helped to improve our manuscript. Below, please find your original comments in blue and our responses in black. When referencing page and line numbers, we are always referring to the new version of the manuscript.

The manuscript investigates the microphysical properties of aerosols based on a large data set at a remote site. They have used an unsupervised algorithm to classify the properties and investigated further, based on the corresponding air mass history also. The study presents a valuable data set for a long duration and follows a novel technique. However, the organization/focus of the manuscript is confusing with respect to the title, along with some other concerns. The results and discussion section could have been made more comprehensive and precise, say, the LAC results are disturbing the continuity between the particle NSD and CCN discussion unnecessarily. Sections 4 and 5 could have merged to form the summary and conclusions. One major concern is regarding the estimation of the effective hygroscopicity parameter and their further interpretations. The paper is worth publishing in the journal of Atmospheric Chemistry and Physics after considering the following aspects.

We merged original sections 4 and 5 as " Section 4 Summary and conclusions". In the new version, section 5 is Outlook.

We extended the discussion of LAC results, more details in the following responses.

The estimation of the effective hygroscopicity parameter and their further interpretations are discussed in the following response.

General Comments

- The effective hygroscopicity parameter (Petters and Kreidenweis, 2007) represent the hygroscopicity for that dry diameter. When the critical diameter (obtained from the back-integration of the NSD) is considered as the dry diameter, the corresponding K value should indicate the minimum hygroscopicity of the aerosol system at that supersaturation, since all the particles above that critical diameter should activate as CCN at lower SS itself. So how is the claim in Line 1 on Page 8 valid?

The sentence in question is "The derived κ at a given supersaturation represents the particle averaged hygroscopicity around the corresponding d_{crit}."

As said, κ is representative of particles with roughly the size of d_{crit}. It is correct that it does not give any information about particles of clearly differing sizes. But we also do not make this claim.

Also, as stated in the first version of the manuscript already, in the paragraph below Eq. 4, “In this method, aerosol particles are assumed to be internally-mixed”. This is a generally used assumption when deriving κ like this.

Therefore, within the described restrictions, κ does represent the averaged hygroscopicity, and, if it is used vice versa together with number size distributions, will enable to derive realistic CCN number concentrations. While κ does not reflect the full variability of the atmospheric aerosol, it does a good job in the described manner when used how and what it was designed for.

- Also, how well does the estimated K represent a multi-modal aerosol system, say having a distinct nucleation and accumulation mode as mentioned in the study for the moderate dust periods? Based on these discussions, what is the relevance of the claim in L9-10 in P17?

The sentence in question here is “Therefore, the CCN-derived hygroscopicity for dust and marine periods in the size range between 40 and 140 nm shows no significant difference. This is a surprising but also important result, as this may suggest the use of generalized values for κ for related cases.”

As stated above, an internal mixture is assumed, so indeed, different modes would not be resolved by this one κ value at one given d_{crit}. However, one could have assumed that an aerosol which is predominantly related to clean marine air masses will have particles in the discussed size range (40 to 140 nm) which have, even when assumed to be internally mixed, a different hygroscopicity than those in a period characterized by large desert dust contributions, due to the predominance of particles modes from different origins in the two different cases.

However, this was not what we found. Instead, even for the vastly differing cases of marine and dust related air masses, “κ during the marine period generally agreed with that of the dust period within uncertainty”. And it should be stressed that this result is valid for the mentioned size range, which is a crucial one in which particles either do or do not act as CCN, hence determining the overall CCN concentrations. So summarizing, our results can be an important contribution to possibly simplify respective modeling efforts. Your remark here does not question our claim.

What is the ‘overall average K value’ (L2, P9), an average of the K values for all the supersaturations? If so, how it can represent the overall hygroscopicity of the aerosol system?

The average κ values included all obtained values for the size between ~30 to 160 nm (critical diameter range) particles. We added the following:

“As this value was derived for different supersaturations and hence different particle sizes, it has limited use, only. However, in Sect. 3.3, we will discuss κ values in more detail related to most extremely differing air masses. κ values did not vary much between these air masses nor between different particle sizes, so that the here given average value may be of some use to characterize the aerosol at least roughly.”

- There is confusion with the data availability. Each parameter seems to have different periods of availability. It is mentioned in the Introduction (not even in the Experiment and Methods section), and so hard to follow during the Results and Discussion. How much period does Fig. 1 represent? It will be better if the measured parameters along with their observation period are presented as a table.

  Figure 1 represents data from April 2008 to December 2017.

  We added a new table (Table 1 in the new version of the manuscript) to summarize all data information, including measured parameters, instrumentation, time resolution, and sampling periods.

- The absorption coefficient is corrected using a theoretically (Mie) derived scattering coefficient assuming a 'less absorbing' marine aerosol system. However, the same study highlights the seasonal presence of dust aerosols. In that case, how relevant is the scattering correction applied to the reported absorption coefficient values?

  While the correction applied here is needed, as described in the text, it is not perfect. To make this clear, the authors have inserted the following sentences on page 5, line 29:

  "However, due to the assumption of a refractive index and the assumption of spherical particles, the quality of calculated scattering coefficients are not sufficiently good, e.g., for the use in radiative transfer calculations. Therefore, scattering data are not presented. The calculated scattering coefficients serve as the best estimate for minimizing artifacts in the absorption measurements."

  And at the end of the paragraph, one sentence further:

  "It can be seen later in the study (Sect. 3.2.1.), that further improvements for the scattering corrections are of importance, especially for high single-scattering albedos since Eq. 2 is merely a first order correction. A deeper investigation requires instruments not affected by scattering artifacts, e.g. photo-acoustic photometers."

- The introduction needs a thorough revision. The authors should clearly specify the objectives and relevance of this paper systematically. Why the unsupervised Ml is preferred in this study as mentioned in L3, P3? The data strength and location details can be moved to the later (Experiment and Methods) section.

  Our impression from both reviews was, that stronger reasoning for our choice of the machine learning method was needed upfront in the text. Therefore, we added more information on this topic in the introduction section to explain the difference between supervised and unsupervised machine learning (page 3, line 12).

  In this study, we did not have preassigned clusters before we performed the machine learning algorithm. Therefore, unsupervised machine learning algorithms were the better solution. We know that particles in different size ranges have different formation routes, sources, and behaviors. For example, we can assume that aerosols with high Aitken and low coarse mode are different from aerosols with low Aitken and high coarse mode. The high or low concentrations of Aitken and coarse mode particles can be reified into a distance in space. This is similar logic to the k-means clustering algorithm, which uses the Euclidean distance to measure the similarities between objects. Therefore, we used the k-means in this study and it successfully classified the aerosol particle into different types.

We wonder if the reviewer can accept the introduction as it is now. We feel that it follows a central theme which we did not want to cap unless the reviewer will still express a strong discontent.

- Another concern is the lack of appropriate references which might have enriched the discussions more. A few examples are;

  Section 2.4: studies like Furutani et al., 2008; Jayachandran et al., 2017; 2021, etc has followed this approach at other parts of the globe

  Studies like Nair et al., (2020) have investigated the CCN characteristics during the mixing of distinct air masses based on the clustering of aerosol NSD, which are not cited or discussed.

First of all, let us be clear that we looked at the correct publications by citing the doi of these publications here:

Furutani et al. (2008): https://doi.org/10.1016/j.atmosenv.2007.09.024

Jayachandran et al. (2017): https://doi.org/10.1016/j.atmosenv.2017.06.012

Jayachandran et al. (2021): https://doi.org/10.1016/j.atmosres.2021.105976

Nair et al. (2020): https://doi.org/10.5194/acp-20-3135-2020

If these are the publications you meant, we are sorry, but we do not understand this comment.

Section 2.4. introduces the derivation of particle hygroscopicity, originally suggested by Petters & Kreidenweis (2007). Since then, this method has been used in a lot of studies by a lot of different groups, and also by us. We do not see how adding a few of these studies, e.g., the ones you suggested, or others, would improve this section. We feel that this does not belong to the method introduction.

Moreover, the studies you mention are neither connected to our study area nor to the focus of our study (examining a long-term data-set which covers several years). We did not see an easy and straightforward way to include these publications without omitting others, as our focus is not about reviewing results on CCN but on proposing machine learning for discovering patterns in long-term data.

We did, however, include Jayachandran et al. (2022) (if we found the correct study, this is actually only printed just now, so the year is 2022). We did this towards the end of Sect. 3.2.2:

"A somewhat comparable result of similar κ for differing aerosol was obtained recently by Jayachandran et al. (2022), examining aerosol during the Indian summer monsoon: Their κ values scattered over a wide range. But when comparing mean and median values for dry and wet periods they were similar, although total particle number concentrations and N_{CCN} varied strongly.

- The LAC data is mentioned and a monthly mean picture is shown. But, no more discussions on that! Any reasons?

The light absorbing carbon (LAC) data originally was not included in the classification as

a) no LAC data was available for 17.52% of the measurement period for which particle number concentration data is available, and b) we wanted to focus on the more often used and more widely needed particle number size distributions.

However, as this topic was mentioned in both reviews, we now included and discussed the LAC data sorted according to the derived clusters in Sect. 3.2.1 and a new figure (Fig. 8 in the new version of the manuscript). The main part of this discussion is at the end of this section and is as follows:

"While, as mentioned above, σ_{abs} and $\sigma_{abs,corr}$ were the lowest for the marine type cluster and the highest for the heavy dust cluster, the pattern showing up for these LAC data (Fig. 8) are not fully conclusive. It is striking that the moderate dust cluster has the second-lowest LAC values, although one could assume that the moderate and heavy dust clusters would be more similar in this respect. On the other hand, the freshly formed cluster, which resembles the marine cluster plus additional small freshly formed particles, has much higher LAC values than the marine cluster, while strong absorption by freshly formed particles is not to be expected. We assume that large uncertainties presented by the influence of light scattering are causing these issues. Therefore, this suggests that further measurements with methods that are not influenced by the light scattering coefficient are necessary. These can be photo-acoustic photometers for measuring the light absorption coefficient or Single Particle Soot Photometers (SP2) for measuring the refractory black carbon."

Specific comments

L13, P2: the 'physics, chemistry, and biology' usage seems too qualitative!

We changed this sentence and extended the information to be more precise.

"Marine aerosol particles' hygroscopicity and their ability to act as CCN can be controlled by marine ocean processes such as biological activity and wind-dependent sea spray generation (Quinn et al., 2015). For example, in the North Atlantic Ocean, O'Dowd et al. (2004) found that the organic fraction dominated the sub-micron aerosol mass and contributed 63% (45% water-insoluble and 18% water-soluble) during algae bloom periods, while this value decreased to 15% during the lowest ocean activity periods."

L5-6, P3: What is the relevance of this statement?

As machine learning becomes more and more important for the evaluation of large data-sets in our community, this sentence should inform readers about the fact that there are machine learning algorithms which are freely available and easy-to-use in many platforms, to help encourage their use. We made a slight amendment by adding one example: ", such as Scikit-learn in Python".

Figure 2 is specified as the monthly mean, but there is no such information about Figure 1. Is it hourly mean?

The LAC and PNSD data are hourly mean values. We added a new table (Table 1 in the new version of the manuscript) to summarize this information. We also added "hourly averaged" to the caption of Fig. 1.

Fig. 7: What is the significance of linear scaling apart from the logarithmic one? Not clear from the text.

It is clear to see the Aitken and accumulation mode, and Hoppel minimum in the linear y-scale plot. The super-micron or sea-spray aerosol mode can be better seen in the logarithmic y-scale plot. We show both linear and logarithmic y-scale for the convenience of readers.

L7-11, P13: Confusing. It is obvious that the nucleation mode particles contribute less to the volume as the mass distributions. As seen in the figure, the nucleation mode aerosol system will have a large total aerosol concentration. But it is mentioned that the NPF happens in marine air mass with low particle concentrations. Please justify this statement.

To make this part easier to read, we divided it into three sentences. Also, we changed "low particle number concentrations" to " low particle volume concentrations".

As an additional explanation, we amended the preceding sentence:

"While we did not measure PNSDs below 20 nm, it can still be safely assumed that such high concentrations of small particles indicate new particle formation events which must have happened recently in the respective air masses, which is known to occur e.g., in the marine free troposphere, followed by downward mixing of the particles (Korhonen et al., 2008, Merikanto et al., 2009)."

L28, P15: during 'this' previous study??

We changed "this" to "the". Now it is: "The freshly-formed and heavy dust clusters did not show up during the previous study and are described here for the first time for CVAO."

L7-9, P16: Not clear.

This refers to "Also, it is clear that the atmospheric N_{HM} (i.e., $N_{CCN,0.30\%}$), meaning also particles being able to act as CCN, is about 2.5 times higher during dust periods than during marine periods at CVAO."

We reformulated this to:

"Also, it is clear that the atmospheric N_{HM}, which best represents number concentrations to be expected for CCN in the atmosphere, is about 2.5 times higher during dust periods than during marine periods at CVAO."