Reply to Referee #2

We thank Referee #2 for their helpful suggestions. We replied to the comments below. The bold text refers to the referee's comments, and the text in italics are additions to the manuscript. The line numbers mentioned in the text below refer to the ACPD version of the manuscript.

I. General comments:

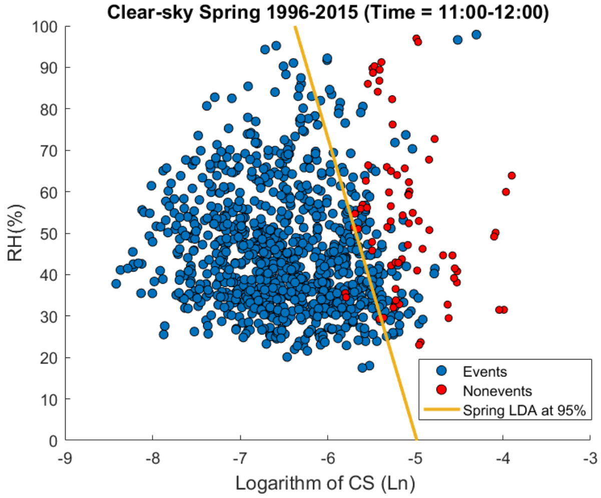
In their manuscript, the authors present an in-depth analysis of a long dataset of aerosol, meteorology, trace gas and irradiation measurements at the SMEAR station in Finland. The analysis is performed to find the key parameters that would explain new particle formation.

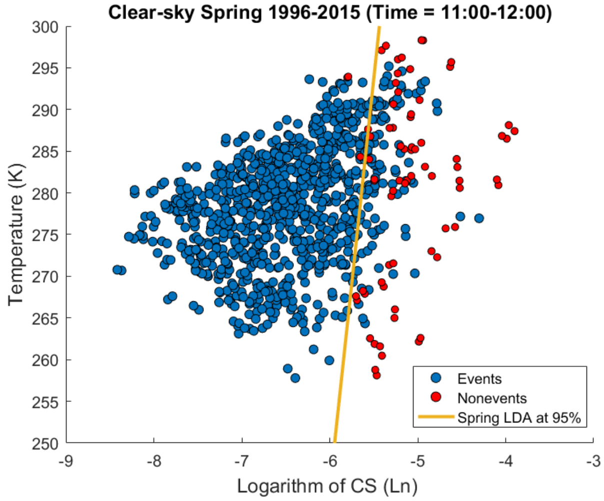
Similar analyses with the same datasets have already been performed several times, as explained by the authors. However, in this analysis the authors focus on eliminating the effect of cloudiness in the analysis, which is an interesting approach and merits publication in ACP. The data aquisition methods are described in good details, and the data analysis mostly follows the procedures described in the cited literature. Some of the specific methods for this paper could be described in more detail and the choices and justification for them should be explained in the text (see detailed comments).

1. A similar analysis without the cloudiness parameter has been performed earlier, it would be nice to see a direct comparison of the analysis of regarding the separation of events and non-events (Hyvönen et al., 2005). It should be quite straightforward to perform the same linear discriminant analysis as the Hyvönen paper for the CS and RH data (Fig 4 in the Hyvönen et al paper), and compare whether the result has changed.

Our results show that although the relative humidity seems to be a variable that influences NPF, when only clear-sky conditions are considered, the variation of RH between events and non-events does not seem to explain the occurrence of NPF. For instance, looking at figure 5c, we notice that although there might seem to be a difference in the median value of the RH when comparing event days and non-event days within each month, the percentiles seem to coincide minimizing the overall separation. However, we agree with the reviewer that it is important to compare with the suggested publication. Accordingly, we plotted clear-sky RH vs CS (below). The plots include spring clear-sky events and non-events within the time window 11:00-12:00 which has been proven shown to be the peak time of NPF Figures 8b and 9b. We also performed Linear Discriminant Analysis (LDA) according to Hyvönen *et al.* 2005 and added the 100% confidence level line to the plot. The 100% confidence limit corresponds to separating 95% of the non-events (to the right of the line in this case). We compare the corresponding spring RH vs CS plot with that of Temperature vs CS (below). The plots show that RH is as good as temperature under clear-sky conditions however it does not aid the separation (events from nonevents) further as CS sink seems to be the main controlling factor. We then conclude that during clear-sky conditions the results are somewhat different from what Hyvönen et al. 2005 who did not consider clear-sky conditions only. Based on the aforementioned results, and following the reviewer's suggestion, we add the following to the text to line 369:

Furthermore, we analyzed the effect of RH in separating the events from nonevents, similar to the study done on RH by Hyvönen et al. 2005. However, when plotting CS vs RH (data not presented), our results show no enhanced separation of events from non-events based on RH when only clear-sky conditions are considered.





2. Also, I think it should be made clear that the event probability described in Figure 13 and in section 3.3.4 is different from the equation 6, and also different from the event probability introduced in the Hyvänen et al paper. In the latter, the event probablility is computed from the LDA analysis, while in the current paper the probability seems to be directly calculated from data, and thus it is not a predictive equation. I suggest that the authors revise this part of the paper. Also, if no real propability-giving predictive equation is given, I think that aim IV in the Introduction (line 66) should be revised.

As the reviewer mentioned, the event probability presented in figure 13 and in section 3.3.4 is calculated directly using the current data set. However, introducing such results best explains the direct effect of extreme temperatures and condensation sink on classifying days as events or non-events. The aim IV in the introduction, refers to equation 6 which sets the line for variable separation during clear-sky events and non-events. First, to improve our analysis, we perform LDA analysis to our dataset, similar to the analysis presented by Hyvönen et al. 2005. That made rather equation (6) more reasonable. Accordingly, equation (6) and the corresponding figures 11 and 12 are improved and replaced. Second, to make the aims clearer, following the reviewer's comments, aim IV in line 66 is divided into two to show the independence of equation (6) accompanied by figures 11 and 12 and the figure 13 modified to "*iv) formulate an equation that predicts whether a clear-sky day with specific temperature and CS is classified as an event; v) use the clear-sky data set to calculate the NPF probability distribution based on temperature and CS*".

However, overall I think that the paper is a potentially good addition to the literature of understanding NPF, and its topic is certainly appropriate for ACP. Therefore, if the above corrections and the detailed notes given below can be considered by the authors, I would suggest publication. The corrections and revisions are, in my opinion, minor.

II. <u>Detailed notes:</u>

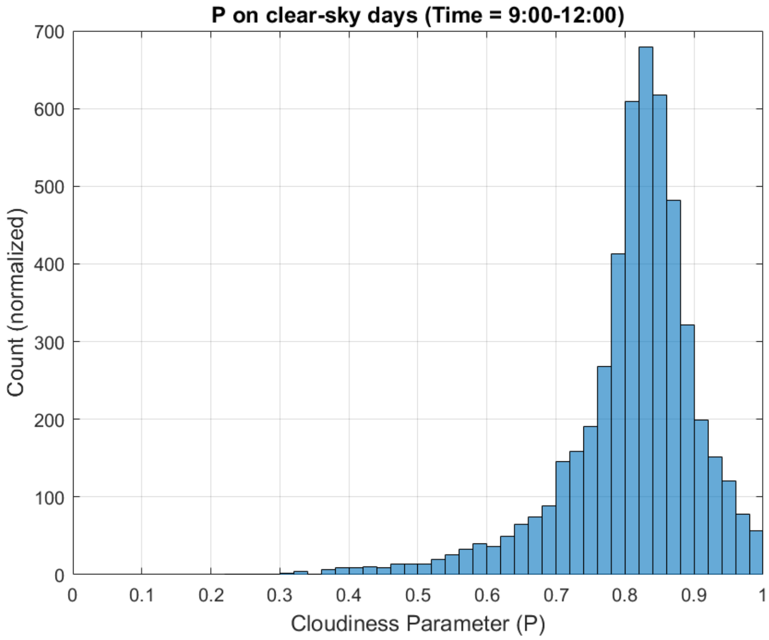
1. line 150-158 and 223-225: I do not fully understand the definition of the clear-sky day presented by the authors. Generally, it is known that particle formation occurs around noon, and that especially the mixing of the residual layer in the morning seems to have an influence. From that, I can follow that using the morning value is useful in the analysis. However, only the median P value for three hours is used. This raises the following questions: *i) Were only events that started during this three-hour window included in the analysis? *ii) Why was the median used? In this case, a time period that is for example 1 hour 29 minutes cloudy and 1 hour 31 minutes sunny gets classified as a sunny (clear-sky) day. Does the result change when the mean is used? *iii) what is the basis of using the value 0.7?

i) NPF occurs usually in the morning hours and peaks around noon (Dada et. al 2017, In Preparation). For that, favorable conditions for clustering should be available to initiate the process as well as ensure its continuation. Since the sun-cycle varies widely in Hyytiälä between seasons which affects the NPF cycle also, the time window 9:00-12:00 seems to cover all seasons equally. Also, although all events are included in this classification, the ones that occur outside our selected time window are rather very few. For consistency, the variables compared in our study are taken between the same time window 9:00-12:00. Also, we do not aim at studying exhaustively the day-by-day results, rather formulate a picture on the variables that are very different on event days in comparison to non-events. And since up to our knowledge, no clear method has been detected to identify the start time of an event, doing the data mining manually for event days of such a characteristic (start time between 9:00 and 12:00) is very time consuming and adds no heavy value for the current paper.

ii) We agree with the reviewer that it is tricky to define the day as clear/cloudy period by a median or a mean value; however, let's assume that 2 hours have P = 0.6 and the remaining 1 hour is P=0.9, this will result in a mean of 0.7, clear-sky day while this is not the reality. However, we recalculated the difference in frequency of event occurrence in case we choose the mean P value for these three hours instead of the median, the differences are acceptable demonstrated in the table below:

	Median	Mean	Percentage difference
Clear-sky events and non- events	1106	1045	5.5%

We consider this difference insignificant for our analysis. The median value is useful also because NPF is a regional-scale phenomenon, so for instance scattered clouds on an otherwise sunny day affecting the local radiation measurements (and leading to a momentarily drop in P) do not usually interrupt the regional NPF process. The histogram (Count Normalized) below also shows that within the clear-sky days, the P values calculated every half hour between 9:00 and 12:00 almost never reach a value below 0.3 (which is the threshold for complete sky cover). This result advocates the fact that if any clouds appear on a day classified as clear-sky they are mostly scattered and not thick.



iii) The value of 0.7 is found from previous studies which are mentioned in line 155. This has proven to be a value that works for different seasons, and is strict enough to exclude all totally or partly cloudy days, but not to eliminate sunny days with occasional scattered clouds passing over the station.

2. The reasoning between this central points in the methodology should be explained in much more detail, as I expect that similar analyses will be performed in the future for other sites, and therefore the method should be as robustly implemented as possible. Also, can the authors give insight on how sensitive the method is on the limit value of P chosen?

The sensitivity of the method on the limit P value is shown in Figure 1a which shows the variation of the fraction of days when using different P values. We added more details to the method section 2.2.2:

In Hyytiälä, the great majority of NPF events are initiated during the morning hours after the sunrise, yet before the noon (Dada et. al 2017, In Preparation). Since the time of the sunrise varies widely in Hyytiälä between the different seasons, the time window 9:00-12:00 seems a reasonable compromise for considering whether NPF did occur or not. We found that NPF events occurring outside our selected time window were very few. Accordingly, in this work the days were classified as cloudy or clear-sky days based on the median value of P during 9:00-12:00 each day, corresponding to the time window for new particle formation. Clear-sky days were those with a median of P > 0.7 between 9:00 and 12:00 and are the focus of this study. The median value ensures that at least half of our selected time window is clear-sky while the rest can vary between clear-sky and minor scattered clouds. The median is useful also because NPF is a regional-scale phenomenon, so for instance scattered clouds on an otherwise sunny day affecting the local radiation measurements (and leading to a momentarily drop in P) do not usually interrupt the regional NPF process. Clear-sky days were those with a median of P > 0.7 between 9:00 and 12:00 and are the focus of this study. For consistency, the variables compared in our study are taken between the same time window 9:00-12:00.

3. Line 198: ". . . radiation is essential for NPF as these events occur mainly during daylight hours." If radiation was essential, no NPF could be observed during nighttime. In the literature, several examples of NPF during nighttime can be found. Please rephrase.

We modified to *"radiation seems essential for NPF at this site, as the events occur almost solely during daylight hours."*

4. line 200: is SA really the main component of freshly formed particles? If heteromolecular nucleation is the prevailing mechanism, the the organic compound is as important. Both are still likely to be formed photochemically, so I think that this sentence can be fixed by just by rewording (e.g. '..because the main components of freshly formed particles are likely formed photochemically. . .')

As suggested by the reviewer, we did the change.

5. line 235-245: Please clarify also in the text and in the caption of Figure 4 that these results refer to clear-sky events only.

Line 235 is modified to *"The springtime medians are percentiles of air-mass trajectories arriving at Hyytiälä during clear-sky NPF events and non-events…."*

6. Line 251-254: As the CS is highest for event days, but not so for non-event days, does the presented conclusion that the CS is the reason for the minimum in events in summer really follow? It seems to me that in summer, events may occur despite high CS, and the actual reason for non-events is not the inhibiting effect of CS. If the authors disagree, this could be clarified.

After modifying figure 10 and the accompanying text based on both reviewers' suggestions (See comment 9 below), it appears clearer that in summer, the calculated formation rates are high also during nonevent days, yet an event is not happening. This might be explained by higher temperatures in summer which leaves the freshly formed clusters rather unstable.

7. Line 270: with monthly I think that the authors mean yearly

The whole section is rearranged to fit both reviewers' suggestions.

8. Line 280-281, '. . .low or almost no correlation. . .' something seems to be missing in this sentence.

Line 280 is modified to "*However, during non-event days, a positive correlation appears between RH and each of CO, SO_2 and NOx while the correlation between those seems to be absent during event days.*"

9. lines 331-350: I don't really understand what is shown in figure 10, and therefore also don't follow the explanation in this paragraph. What is meant by diurnal cycle here? By definition it means a repeating pattern that occurs every 24 hours, and I don't see how this could result in Figure 10. Please clarify and rewrite, or replace with the correct figure.

Figure 10 is replaced with median diurnal cycles of J_3 and CS during different seasons. The diurnal time frame is limited to 5:00-20:00 due to the incapability of calculating SA concentrations in the absence of UVB, therefore no $J_{3,C}$ values are calculated outside this time window.

10. Line 357: The procedure of finding the separating line in Fig 11-12 is described very poorly. Is this done by linear discriminant analysis (such as e.g. in the cited Hyvönen et al., (2005) paper or some other method? The authors should describe this in more detail. I'm especially concerned about the sentence "the data points have been estimated by taking the non-events with the lowest possible CS which still fit the linear ˘ separation"; was some kind of data selection applied to produce the figure?

See comment 2 in the General comments section

11. Figures: Several figures have the sentence "The lines extending 1.5 times from the central box represent the remaining of the data yet still within the relevant statistical limit. " Please clarify what this means: firstly, what is 1.5 times from the central box (the lines seem to have different lengths, eg. in fig. 5. Also, clarify what is meant by relevant statistical limit.

The extended lines are equal to 1.5 x interquartile range, and the points beyond whiskers are outliers (> 1.5 x interquartile range). Outliers are the individual stars. Maximum length of the extended line, is defined as q3 + w × (q3 – q1) while the minimum is q1 – w × (q3 – q1) where w, q1 and q3 are the mean, the 25th and 75th percentiles of the sample data, respectively. The statistical limit is defined by default as the 99.3% coverage. Based on the reviewer's comment we modify the text corresponding to the box plots for clarity to the following:

The length of the whiskers represent 1.5 x interquartile range which includes 99.3% of the data. Data outside the whiskers are considered outliers and are marked with red crosses.

References

L. Dada, R. Chellapermal, S. Buenrostro Mazon, V.M. Kerminen, P. Paasonen And M. Kulmala (2017). Method for identifying NPF event start and end times as well as NPF types (ion-initiated, particle initiated, transported..) using characteristic nucleation-mode particles and air ions. *In Preparation.*