Anonymous Referee #1
Received and published: 14 September 2011

The paper shows results of interpretation of a cluster analysis of size segregated ultrafine particle number concentration measurements performed along one year at Mace Head (Ireland). This is relevant from the point of view that it supplies information on the major air masses and processes controlling variability of ultrafine particles at an marine background site.
The cluster approach applied to hourly size distribution has been used by the team from Birmingham University. From the methodological point of view the papers is applying this methodology to a 'clean' site.
In my opinion the results are of interest for the scientific community. These are not reporting very novel findings, but give interesting data on origin and processes affecting ultrafine particles. However, also in my opinion, additional information should be provided to support the final conclusions on cluster analysis and on the interpretation of the origin of cluster groups.
Based on the above comments, I suggest publication of the paper after a moderate revision based on the comments attached below.

We thank the reviewer for considering this work relevant to ACP.

Major issues

I have the following major comments to the interpretation of results:

1) In my opinion there is not enough support of the selection of the 12 cluster result. Why not 10 or 14? The authors should give more details about the final output of clustering concerning the number of clusters.

We have now expanded this section and provide more details. The 6578 SMPS size distributions obtained at one hour resolution were then subsequently normalised by their vector-length and cluster analysed (Beddows et al., 2009). The use of cluster analysis was justified in this work using a Cluster Tendency test, providing a calculated a Hopkins Index of 0.20 and implying the presence of structures in the form of cluster in a dataset (Beddows et al., 2009). The choice of k-means clustering was made from a selection of the partitional cluster packages (Beddows et al., 2009). K-means method aims to minimize the sum of squared distances between all points and the cluster centre. Using k-means clustering, the complexity of the data set is reduced allowing characterization of the data according to the temporal and spatial trends of the clusters. In order to choose the optimum number of clusters the Dunn-Index (DI) was used, which aims to identify dense and well-separated clusters. DI is defined as the ratio between the minimal inter-cluster distance to maximal intra-cluster distance. Since internal criterion seek clusters with high intra-cluster similarity and low inter-cluster similarity, algorithms that produce clusters with high DI are more desirable. In other words, for Dunn's index we wanted to find the clustering which maximizes this index. The Dunn-Index for the results of the k-means cluster analysis for different cluster numbers showed a clear maximum for 12 clusters, some of which belonged only to specific times of the day, specific mechanisms as well as specific seasons.

2) Also the basis of the grouping of 3 clusters in each group should be explained in detail from the beginning.

We now explain the four categories from the beginning

3) The same applies to the interpretation of the origin of the groups 'open ocean' and 'coastal nucleation' and 'background clean marine'

Expanded

4) You should clarify 'back ground clean marine' if this refers to NE Europe marine background or that this represent the cleanest marine
background at Mace Head. The name is a bit confusing.

Background clean marine represents the cleanest marine background at Mace Head, which we take as NE Atlantic marine background

5) Text gives the impression that the paper have been written very fast. The reader would appreciate a bit of consistency (reporting characteristics for most parameters evaluated in all groups) in the interpretation-description of cluster and group characteristics. Especially applicable in chapter 4.1.

We revise and expand section 4.1. We also added a new concluding table (current table 4).

Specific comments

Abstract: '. . .. . . as systematically occurring and these 12 Clusters could . . .. . .' by '. . .. . . as systematically occurring. These may . . .. . .'

ok

Abstract: '. . ..more mono-modal. . .' by '. . .. . .more mono-modal (accumulation)

ok

Abstract N.E. by NE (twice)

ok

Abstract '. . ... of new aerosol particles in N. E. Atlantic Air' by '. . ... of new nano aerosol particles in NE Atlantic air'

ok

Introduction: 21680, row 6, IPCC 2001 was updated by IPCC 2007

ok

Introduction: 21680, row 25, ' a fine mode' by an Aitken mode'

ok

Pages 21685 to 2687: Apply general comment 1 here. The classification of clusters is given but no info on why 12 types, and not less or more were obtained.

Expanded and explained

Page 21688. The same applies here for the grouping of 3 clusters in each group.

Based on what grouping criteria?

Expanded and explained. It is important to note that the 12 clusters are merged in four categories: coastal nucleation, open ocean nucleation, anthropogenic and background clean marine. This was based on common physical and chemical properties among category. However, some aerosol size distribution clusters presented unique features within the same category, therefore each individual cluster is presented in this section, whereas a discussion on each category in presented in section 4 (discussion).

Page 21668, 20-25: More information on how do you attribute one type of nucleation to coastal and the other to open ocean. Also bimodal (Aitken and accumulation) monomodal (accumulation)

Ok. The name of each of the four category is derived from the clear differences seen among them: a dominant nucleation mode at sizes less that 10 nm for the coastal nucleation category (new particle formation in the nearby coastal areas), a dominant Aitken mode between 15 nm and 50 nm for the open ocean nucleation category (new particle formation occurring in far from coast open ocean areas and detected at coarser aerosol sizes when transported at Mace Head), a clear bimodality in the size distribution for the background clean marine (bimodal and coarse) and a generally more monomodal one for the continentally-influenced size distributions.

Page 21689, 14 (give number concentration in brackets as done in row 15 for the other 3 clusters.

ok

Page 21689, 16. 'largest' what this means here? Highest or coarsest?

Coarsest, modified

Page 21689, 16: 9 nm?, in table 1 you report 10 nm

Modified, 10.2

Page 21689, 18: Why low RH and coastal origin? Is the RH differences significant?

RH was found lower for coastal nucleation relative to other aerosol categories. We added a reference, Relative humidity (RH) has been observed to be anticorrelated with continental new particle formation and that this is likely due to low OH concentrations at high RH (Hamed et al. 2011).

Page 21689, 24 to 25: Rewording of the sentence required for a better understanding of the meaning.

Ok

Page 21690, 23: clusters 4, 5 and 9; or 4, 5 and 8??

8, modified

Page 21691: By interpreting 4.1.3. as anthropogenic type; do you mean anthropogenic species are completely irrelevant in all the other cluster groups?

We believe the other three categories are much less affected by anthropogenic sources.

Page 21691: 4.1.4. See general comment 4 above.

Ok

Page 21692, 5-6: If you mean the cleanest marine background at Mace Head? Why highest scattering and PM2.5?? Sea salt? Give a bit more of info

Yes, this is due to sea salt as reported previously by Dall'Osto et al (2010), section expanded.

Page 21693, 8: 'precursor gases' is very general, which ones?

Iodine related compound (as reported in O'Dowd 2002a,b), expanded

Top of page 21694: See major comment 4 above.

Ok

Page 21694, 10-29: In my opinion this is repeating finding described in the page 21693 but in a different way. Summarise and merge both sections.

Ok we removed table 4

Section 5: Please take into account major comments 1 to 4.

ok

Heading of Table 1: I do not see N.D. in table, you do not need to define.

Ok

Table 1 and Figure 1 repeat results

ok

A. Asmi (Referee)
ari.asmi@helsinki.fi
Received and published: 14 September 2011

Referee number 2

Referee comments for Dall'Osto et al, Statistical analysis..., ACPD,11,21677-21711,2011
Ari Asmi
University of Helsinki

The article uses relatively new methodology to find out 12 representative number size distributions from Mace Head coastal site, interpreting the results as a function of air mass origin and other parameters. In general I find the paper fitting from the subject matter to ACP. The results are not particularly surprising, but valid and I have no complains about the overall methodology.

The presentation is acceptable, but improvements on the text and many small corrections are needed, as the document seems to have been submitted in a hurry.

We thank the reviewer for considering this work relevant to ACP.

General comments:

My main problem from the methodology side is the overall shortness of methodology, especially on clustering process. e.g. it is mentioned that the study uses K-means clustering. It would be useful to mention that the reason it was used, as Beddows et al (2009, EST) showed that such method gives better separation and more uniform clusters than other commonly used methods. Another thing unexplained is the meaning of Hopkins number. The number of selected clusters would also require explanation, especially as the interpretations mostly concentrate on the cluster types (3 cluster per type).

As mention to reviewer 1, section expanded.

I would prefer a bit more introduction to the terminology and/or the method, so that a reader not familiar with clustering does not have to search Beddows et al (2009) or some textbook to figure out what are the main advantages and disadvantages. There should be more discussion on what the clusters actually represent, in addition to the text in the Introduction.

Expanded

As a general comment, I think the article would benefit from instead of a long explanation of different cluster properties, some way to combine the properties into either one (large) picture or table, so that the reader do not have to either go through the 7 tables or read the list of cluster properties. This could also be done in cluster type level (e.g. Coastal N, Op. Ocean N. etc), to show which are the similarities of the different types in context of size distribution shape, airmass origin, times observed and key other parameters (high PM2.5 etc). This would work very nicely as a concluding table, in addition of making the overall results more approachable.

This work presents 12 aerosol size distribution clusters, which we then merge into 4 aerosol categories. It is important to note for some of the categories (for example the background clean marine) we have SMPS clusters with marked differences. We therefore decided to present individually the 12 clusters and then describe the categories. However, we now added a concluding table (current table 4) and we hope the overall results are more approachable.

Question: Could one, based on the clusters presented here, use any measured size distribution in Mace Head (e.g. from earlier years) and based on some distance parameter of the clusters, select a most likely origin of the clusters. If so, could this method then be used as a Mace Head "airmass origin" filter if one wants to e.g. study only airmasses originating from open ocean?

We really appreciate this question – yes, indeed we are using aerosol size distributions as "filters" and we are currently queering large dataset like AMS, HTDMA and CCN counters to describe different air mass origins

Minor comments:

Overall, I would suggest the authors to proof-read the article one more time. I will not go through small (but numerous) small typos here.

ok

I did not see any use for the nephelometer data in table 3. Is it used somewhere?

Yes we expanded this section and we use nephelometer data to describe both background clean marine and anthropogenic aerosol size distributions.

Table 1 has no indication on a) how are the modes defined in this context, and most importantly: what are the numbers? Diameters? Statistical parameters? What are the units? Reader is left confused.

Expanded and explained Table 1. A curve-fitting programme was used to disaggregate the size distributions of each cluster into a number of lognormal distributions (nucleation, Aitcken and accumulation) whose average aerosol size diameters are reported in table 1.

In the text, there are many cases where properties are given with value+error estimate, eg. "..was found to be mP (42 ± 15%)". What are the error estimates, standard deviations? Or are they ranges? Same with figures 1, 4, 5

ok, 1sigma.

Wind direction averages are probably calculated as vector averages. Useful to mention, as otherwise quite southern-oriented averages could be though averaging.

Yes, mentioned.

Fig 2: EUCAARI

Ok

On figure 4, I would draw a coloured line on e.g 1E3 to highlight the size distribution function differences between clusters.

Ok

Figure 5, please increase the axis scale text size. What are the lower limits of the "error bars"? As in other cases, are the error bars range or deviation? A box-whisker plot would fit the subject matter much better. Overall, could you please put the subfigure markers (a, b, etc) on the top left of each figure? Subfigure e) y-axis label is in Bars, I would guess it is (according to caption) mB (or, preferably in hPa)?, Subfigure f) legend is not explaining too well what is show there with just "%". Please use e.g. "fraction nightime". Also, I would guess the solar radiation (is this global radiation measured in

10m?) is in W/m2, not in J/m2.

ok

Correct all "aitken" with "Aitken"

ok

One specific point: Why is "Cluster" written with capitalized C? It is at least consistent through the manuscript, but I would really write it as "cluster".

ok

N3 and N10 could be more in line with the notation of some earlier paper, but ND>10 are ok.

ok

Anonymous Referee #3
Received and published: 26 September 2011

In this paper, the authors use a relatively new statistical technique to categorise aerosol number-size distributions from 2008 collected at Mace Head at the GAW site. Size distributions with similar characteristics are binned into 1 of 12 types known as clusters. These clusters are then grouped into 4 more general classifications, namely coastal nucleation, open ocean nucleation, anthropogenic and background clean marine. Although the results do not present any major new insights, they do provide a means for statistically analysing large datasets in a routine way and attempt to relate these to other parameters such as air mass history.

With some re-working the paper will be acceptable for ACP.

We thank the reviewer for considering this work relevant to ACP.

Major Concerns.

The authors switch between describing individual clusters and the 4 average categories and I feel there needs to be some additions to bring all the analysis together.

We now report the individual clusters and the 4 categories and we bring the final analysis all together with a summarising table.

For example, the back trajectories starting with 'm' produce or contain all 12 clusters, concluding that air originating from those sectors are less sensitive to long range effects rather local or regional processes. The coastal nucleation event clusters are slightly mis-leading as they are presented as a separate class when in fact they are size distributions which could be background clean marine or open ocean nucleation which have had ultra-fine particles added to them. It would be interesting to know or to speculate what cluster they would fall under if only data from Dp>50nm, say, was used.

We agree both coastal and open ocean nucleation particles add on existing modes. Potentially we could remove D>50nm particles and try to better apportion accumulation modes with continental air masses. However, our approach was to consider all the aerosol size distributions and not only a part of it. Future studies will aim to separate modes by disaggregating SMPS size distributions. Specifically, we will try to apply positive matrix factorization which identified different factors. By doing this, we may be able to separate different regional modes (for example accumulation modes from continental air masses) from local ones (for example coastal nucleation events) and see how for example ultra-fine particles add on existing regional background accumulation modes.

To me, the bigger picture analysis shows that in the cold winter months, the classical bimodal distributions of the background marine can dominate from the clean sector due to the meteorology. As the weather warms, then contributions from open ocean nucleation can add to the background and/or as the air approaches the coast then ultrafine particles can contribute to the size distribution depending on the solar radiation and tides. This is all perturbed if the wind in anthropogenically influenced by the local wind direction (clusters 6,9 and 10). This is touched upon in parts, but not really summarised. Also, how will this help the global models discussed in the introduction? How would a modeller use this analysis?

We thank for the suggestion, we expanded the text following this advice and we also added a summarising table.

I think the manuscript would benefit from a short section, maybe a very small appendix, on the cluster analysis itself. The paragraph in the text does not really explain the technique and relies on the reader to research the Beddows et al reference. Some details on the basic process and defining all the terms, how the clusters are grouped, why there are 3 clusters in each of the 4 general categories (or is that just luck?) and a plot or table of the diagnostics used to validate the results of the analysis. This provides a reference for other people wanting to use this technique whilst not clogging up the body of the text for those just interested in the results.

Ok, expanded – added summarising table

There are a lot of grammatical errors and the manuscript needs a final proof read. For example bi-modality and bimodality are both used in the abstract.

Revised

There are errors in the results section when referring to numbers in tables. For example, Page 21686, Ln 22 states cluster 7 has the lowest Nd>3 of 892. Clusters 11 and 12 have ND>3 of 764 and 773. This happens quite often and made the review much harder. The authors need to check all numbers in the text against those in the figures/tables and make sure these errors do not compromise the interpretation of the results.

Ok

Specific concerns:

Why is 'cluster' capitalised in the text?

Ok

In the introduction, the author states that 60% of the air arriving at the station comes from the clean sector and that clean air is defined as at BC concentration less than 50 ng m-3. Yet in table 3, there is not a single cluster with a BC loading less than 50ng m-3. Can the authors clarify this please? Was 2008 a dirty year?

In this study, we consider all the data available. For 2008, the BC average concentration was 210±150ng m-3, overall higher than previous years. Moreover, we acknowledge the fact that when considering all data, some local plumes (ie cars or local pollution) may affect few minutes of measurements, yet giving higher values of BC concentrations.

Please refer to tables as 'Table 3' for example,and not table 3e. If the authors feel they need to identify a column specifically, state it.

Ok

21679 Line 25 onwards. Please be consistent with units. nm in one line, then switching to um in another for the same aerosol mode.

Ok

21680 Ln 27-29, "Some examples of particle size distributions Cluster analysis for substantial SMPS datasets can be found in the literature. Similar approaches have previously been used:" Please can you clarify these sentences. Are the similar approaches the same as the analysis for the substantial SMPS datasets? If not, please give examples. Should the sentence read: "....found in the literature, where similar approaches etc"?

ok - expanded

21682 The section on the instrument description needs a little bit of tidying up. Firstly, please be consistent with the instrument name. In this section the 3025A is referred to as TSI 3025, 3025A CPC3025. Which one is it? Both a 3025 and 3025A exist as model numbers. Also, technically, 3025's are Ultrafine Condensation Particle Counters (UCPC). The description also first introduces the CPC's then the SMPS and back to C9311 the CPCs. The section would read much better if all the CPC details were in the first paragraph and then the SMPS' described.

ok - reorganised

Finally, please clarify how you can have 88% data coverage but some hours are not available so there is only 75% data coverage?

Ok - rewritten

Has the data been filtered for contamination or below detection concentrations?

Yes we did consider only validated data (by comparing data with CPC/SMPS as described in previous Mace Head studies – see Yoon 2005, 2006)

TSI Inc no longer stands for Thermo-Systems Inc. It is just TSI Inc

Ok

21683 ln 15. Define WS (first use of) Ln 19 WD define. Ln 15, Please explain what a meteorological discontinuum is. Do you mean the pressure was on average lower in August then July/Sept? This is poorly written for a journal "....somehow creating...."

ok

21684 Ln 11, suggest replacing "spring was associated with" by "spring experienced more". Ln20/21 the author is referencing figure 4 twice. It is repartition.

Ok

21685 line 5. Difference is the second largest, not the largest.

Ok

21686, Ln 7 there is no table 3d, Leave as table 3 or table 3, column d. Same for table 3e, Ln 21 why is it peculiar that a background marine number concentration should be low? Also, it is not the lowest, clusters 11 and 12 have ND>3nm <892

ok

21687 Table 3e should just be table 3. Ln 12 the TEOM loading for that cluster is 16.5. Ln 18. The ND>3 is in fact he lowest at 764. The difference is the second lowest. Ln 21. Maybe justify that statement about the coarse mode and sea salt with reference to the higher wind speeds and the O'Dowd paper 1997 (I think). Ln 26. The ND>3 at 773

is the second lowest. The TEOM concentration is 13.3

ok - expanded

21689 Ln 2, Sect. Should be section. Ln 5 replace among with during.Ln 13 they showed the highest ND>3 concentrations, not ND>10. Please be specific. Ln 15,

ok

2840 is not lower than 1905. Ln 16. The mode is not at about 9 nm, that is where a SMPS point is. The mode itself, if a mode was fitted, would be around 10-11nm. I am being pedantic, but the authors claim Aitken mode aerosol are for Dp>10nm. This really strikes at the heart of distinguishing what class of aerosol cluster 3 is and how the classes are defined. I would argue there is good cause to have cluster 3 at the open ocean nucleation cluster. Ln 20 - page 21690. This needs some re-wording. I can see the shift in the peak of the occurrences, but the explanation needs elaborating more for readers unfamiliar with coastal nucleation.

We rewrote this part. We exclude cluster 3 is due to open ocean nucleation as it presents a distinct diurnal variation (similar to cluster 1-2, figure 5).

21690 Ln 20 Cluster 8 had the 6th highest BC loading.....Ln 23, should it be 4,5 and 8?

Ok

21691 Ln 12. Table 2 not table 1. Also, I do not agree with the statements in this section. Firstly, table 2 does not show correlations. It lists occurrences. Secondly, cluster 9 cmP and cP have occurrences of 15 and 16, the two lowest. It is more likely to find cluster 9 in mP air masses. Furthermore, how does 6 fit into this picture of air mass origin? Is this basically telling the reader that the source of the air does not matter if the local wind direction is outside the clean sector? That is my conclusion. Also, the last statement needs re-wording and clarifying. The sub micro particles are dominated by the nucleation events in number. What I believe the authors to mean is that the scattering data confirms the large number of accumulation mode aerosol, as seen in figure 4c. Ln 24 cluster 1 is more westerly than cluster 11 and close to cluster 12.

We expanded the text and add a local/regional contribution and the inland wind dependency.

21692 Ln 2 wrong table referenced again. Also, the argument about the high scattering and high PM does not hold up against the data from other clusters. There are other clusters with PM loadings higher than cluster 7 but with lower scattering eg cluster 1 + 3. Cluster 9 has a lower PM loading but a similar scatter to cluster 7. Can the authors discuss this? The nepholometers will only detect particles down to about 100nm, so it is not the nucleation particles having an effect.

Expanded, indeed sea salt affects PM and scattering properties.

21693 Sect should be section. Ln 4 - 10. This needs to re-writing. Comments like Cluster 3 did not present a clear seasonality, somehow in between Cluster 1 and Cluster 3. It should be cluster 2 at the end and if there is no clear trend how can it 'somehow' be between the other two? A trend exists or it does not. Ln 23 cluster 5 peaked in September, not the summer months. Ln 29 cluster 9 does not spike, it is a clear seasonal trend. Also, where is figure 11?

Ok

21694 Ln10 - 19. This is complete repartition. There is nothing in table 4 that cannot be seen in figure 7 and has not already been stated. Remove this section. Ln 25 from Yoon et al onwards. How does this discussion contribute to the paper?

Ok - removed

21695 Ln 13. The authors results do not support the statement that cluster 11 dominates summer months. It is clear there are cluster 11 events in winter/spring.

Ok - modified

Table 2. Given that the authors have assigned clusters a type, this table would be better organised into these types, rather than in ascending order of cluster number. Eg, 1,2,3 4,5,8, 6,9,10, 7,11 and 12. Same for Table 3 and 4.

Ok – we added a summarising table (current table 4)

Fig2. Why are the dots joined up? This is a frequency plot, not a time series. Markers or bars only please.

Ok

Fig 4. Legend incorrect. I believe the last sentence should be (e) shows average. The legend also needs explaining. Why if e shows the average of the 4 cluster types are there 5 traces?

Ok – the main issue with Figure 3d is the major difference between the 2 bimodal distributions (cluster 7-11) and the unique coarser mode ones described by cluster 12. The fact during winter time we have two different scenarios for background clean marine (7 and 12) is a novel finding, which would be lost if current figure 3 had only one size distribution describing "background clean marine". We describe it in the text.

Fig 7 and table 4 are showing the same data in different formats. Suggest using just one and organising them into the cluster types.

ok