GENERAL COMMENTS:

This discussion paper introduces a new automatic precipitation phase distinction algorithm for optical disdrometer data over the global ocean.
The introduction highlights current short-comings of in-situ precipitation measurements, which eventually serve as reference to microwave precipitation records. The authors present the benefits of ODMs related to extreme meteorological conditions and poor data density and highlight the need of time-saving post-processing procedures. In this respect, convincing arguments regarding automatic precipitation phase (PP) distinction algorithms of different complexity are brought up, as these not only accelerate data analysis, but minimize subjectivity inherent to manual observations.
After a brief instrument description and a mathematical background, the manual reference data set as well as the OceanRAIN data basis is introduced. It becomes clear that the high fraction of snow and mixed-phase precipitation within OceanRAIN under extreme meteorological conditions, next to high measurement frequencies, is beneficial to creating the PP distinction algorithms.
The following three subsections focus on the different PP distinction algorithms, including elaborations on their (dis-)advantages. The main findings are well illustrated in form of several demonstrative figures.
A short discussion towards the end relates the findings to recent publications. Comparing statistical scores derived in this work with those of former studies suggests that the new phase distinction algorithm considerably improves the model performance.
This discussion paper is valuable for the scientific precipitation community, as the proposed automatic distinction algorithm is transferable to other particle size distribution sampling instruments. It will thus help to further characterize in-situ surface precipitation uncertainties and in consequence support the creation of higher-quality surface precipitation reference data sets. I recommend publishing the paper in AMT once the suggested minor revisions (see below) have been considered.

SPECIFIC COMMENTS:

1. p. 13646, ll. 17 ff: the wording is somewhat confusing regarding 3P1D and 3P2D. You may want to first point out that two different approaches exist for RMS. In this context, state that 3P2D represents a new approach, which outperforms the 'conventional' 3P1D version.
2. You frequently mention (in-)significance of differences, e.g. on p. 13659, 13660, 13662, 13665, 13666, 13670. Be careful with this term, as using it implies that the (in-)significance has been statistically confirmed. You may want to either refer to a different term or include statistical evidence.
3. p. 13651, l. 1 ff: the description of the ODM470 setup is misleading, as you use 'sensitive optical volume' for two different things: once for the IRSS88 (l. 5), once for the disdrometer (l. 7). Is this done on purpose? In the latter case it seems obvious you are referring to what is seen on the left-hand side of Fig. 1, whereas the IRSS88 is shown on the right-hand side of Fig. 1.
4. p. 13654, l. 16 ff: you assign a snow flag to combinations of graupel and hail (ww = 96, 99). Can you explain why you exclude the mixed-phase flag for these weather codes?
5. p. 13658, l. 22 ff: the bias score definition is misleading. As to my understanding of l. 12, 'rain disagreement' implies that the manual PP is 'rain', whereas the model PP is 'snow'. Keeping this in mind, let's briefly focus on the following example: Rain agree = 100, snow agree = 100, rain disagree = 20, and snow disagree = 30. Following your definition of bias score (l. 22 ff), this results in b = (100+20)/(100+30) ≈ 0.92, i.e. the model *over*estimates snow fall. This is contradictory to the example data, as (to my understanding) the model predicts 'snow' in 20 cases, whereas the manual distinction gives 'rain'. Vice versa, the model predicts 'rain' in 30 cases, when the manual distinction gives 'snow'. To sum up, the model predicts *more* rain than snow (with respect to the manual reference), i.e. rain *over*prediction. However, your definition of bias score (l. 22 ff) seems correct, if one assumed 'rain disagreement' to *imply model = 'rain' and observed = 'snow'* (in contrast to my translation of 'rain disagreement' further above). If this was the correct definition, its repetition on p. 13660, l. 18 f is also logical. Summing up, please clarify what 'rain disagreement' implies for both model and observation.
6. p. 13659, l. 16: you state that Fig. 4 includes significances of performance differences. In what way? This seems unclear.
7. p. 13659, l. 17 ff: the wording is misleading. Taking a look at Fig. 4, it is not obvious that *all* of the shown predictor combinations include the air temperature T (this cannot be reproduced). One may think that the accuracy only exceeds 88% if a 'T' is explicitly quoted. Furthermore, the connection by underscores directly implies that *three*

predictor variables are considered. This, for example, does not account for merely *two* predictor variables, such as 'P' (i.e. 'P'). You may want to include something like 'combining T with *two* other relevant predictor variables'…

8. p. 13661, l. 3: 'accordingly' is misleading. You state that 'rH' tends to *in*crease PM (which is a *negative* feature). Next, you state that the combination of 'rH' with either 'D99' or 'RR' *de*creases PU (which is a *positive* feature). 'Accordingly' would only make sense, if the second statement is a logical consequence of the first statement (which it is not, especially because the diameter-related parameters seem to be most important for the PU reduction (and not rH). Please clarify.
9. p. 13661, l. 13: you state that KS98 PU is much lower (in comparison to OceanRAIN), to the expense of a much higher PM of 4%. Is this increase in PM a direct consequence of PU? Can it be explained by the fact that a lower PU implies a narrower uncertainty (0.05 < PP p < 0.95) range, more data exceeding the PP p of 0.95 and thus an increase in the chance of misclassification of certain cases (i.e. PM)? However, this seems to not always be the case, as the anti correlation between PU and PM (Fig.4, Fig. 6, Fig. 9) is not -1. Please clarify. Regarding the dependency on D99: it may be worth including it in Fig.5 (i.e. differentiation between D99 = 1 mm and D99 = 5 mm).
10. p. 13662, l. 23 ff: this statement is misleading. First, why are you comparing T_rH_D99 (3P1D) to T_rH_RR (2P1D), instead of comparing it to T_rH_D99 (2P1D)? I.e. you are comparing two different things here. Second, I agree that the accuracies of 2P1D and 3P1D have a similar behavior (apart from the accuracies of 3P1D being much lower) and that the accuracy of T_rH_D99 is highest in case of 3P1D. However, your next statement seems wrong – the predictor variable combinations including RR do *not* perform equally well in 2P1D. This only accounts for T_rH_D99 and T_rH_RR. However, the accuracies of T_RR and T_RR_D99 (both 2P1D) are considerably lower. Please clarify.
11. p. 13663, l. 13: the sentence starting with 'The correlation of… ' seems unclear - what do you want to express? Is it that the correlation coefficient of accuracy and PM do not necessarily need to be -1? It also remains unclear why PM of T_rH_D99 is above that PM of T. I agree that the inclusion of D99 is beneficial; so is the higher PM of T_rH_D99 assumed to be associated with wrong *manual* PP assignments only?
12. p. 13664, l. 26 f: Fig. 8 is very helpful in visualizing the different PM and PU regimes as a function of precipitation phase. In this context, you repeat the definition of PM. While this is trivial for the two individual PP distributions (hatched for pp < 0.05 and > 0.95), it remains unclear how the range of 'PM_mix' is derived. In case of the rain distribution, PM_rain is derived as a ratio between the limit of certainty (you set it to 0.95) and the maximum of the solid curce (=1). Does this also account for PM_mix, i.e. something like 0.67/0.72? Or is PM_mix derived graphically? Please elaborate on this.
13. p. 13665, l. 2: please explain what is meant by 'operator' in this context.
14. p. 13666, l. 15 ff: the connection seems to be incorrect. Comparing to Fig. 11, 2P1D approaches the *rain* distribution of 3P2D at lower T (i.e. the rightmost dotted curve and the blue curve become very close), whereas it approaches the *snow* distribution for higher of 3P2D at higher T.
15. p. 13667, l. 8: the comparison to OceanRAIN is misleading. Where does the (poor) OceanRAIN bias score of 0.8 come from? This is neither reflected in Fig.4, nor in Fig.6 and Fig.9. Please indicate whether this comparison is constrained to using the *KS98 coefficients* only (in contrast to the OceanRAIN fitted coefficients).
16. p. 13671, l. 10 f: are you referring to Fig. 12? If so, it shows (next to Dai Ocean data) observations from the *Swiss Alps*, not Finland! Or does 'data from Finland' refer to the derived *coefficients*, which were obtained from Finland data and applied to the Swiss Alps data? Please clarify.
17. p. 13671, l. 28 f: regarding 'probability of detection': Please explain what it means in this context. Is it equal to the definition given on p. 13668 (l. 3)? Does the value of 0.3 find expression in one of the Figures of the manuscript or where does it come from? Please elaborate on this.
18. p. 13672, l. 4 f: please indicate how the PP probability could serve as a measure of error in context of satellite measurements. Many satellite retrievals do not differentiate between different PPs to date, so in what sense can an (improved) PP discrimination be helpful?
19. p. 13682, caption: the number of minutes used (165915) differs from the number listed in Table 2 (164994). This difference cannot originate from the fact that Fig. 3 excludes very low T (< -6 degC) and very high T (< 8 degC), as the same is valid for Table 2 (compare p. 13656, l. 9 f). Please explain the difference.
20. p. 13677, caption: regarding 'non-trivial': the wording is somewhat ambiguous. You may want to mention that non-trivial implies highest PP uncertainty, that data equatorward of 45S and 70N has been omitted, and that minutes with less than 20 observed particles have been excluded. Or (to keep it short) refer to the manuscript text, where those three features are listed.
21. P. 13671, l. 7 f (e.g.), regarding the extension of the OceanRAIN data base: although the maintenance of the contributing instruments is somewhat simple, ODMs are expensive. It is beyond question that your presented

results will become even more robust once the data base grows. However, how realistic is the scenario that especially the high-latitude data density (sampled by ODMs) will continuously grow in the near future, keeping in mind the instrument costs?

22. General comment: your results indicate that the bias scores is exclusively below 1, which indicates an *over*estimation of snow events by the model. This also accounts for biases derived in the framework of other studies, which are specifically mentioned in Section 4. Some of the bias scores listed are even as low as 0.8. The question arises as to whether all of the proposed algorithms are subject to fundamental shortcomings (do they miss an important predictor variable, e.g.?) or should the bottom line be that bias scores exceeding 0.94-0.95 are as good as we can get?

TECHNICAL CORRECTIONS:

1. p. 13646, ll. 16: grammar. Better: 'An accuracy of 81.2% is reached for...'.
2. p. 13650, l. 4: one PP distribution distinguishes *between* two PPs .
3. p. 13650, l. 18: to develop *a robust* PP… .
4. p. 13655, l. 4f: '… this comparison can reveal…' - Do you mean the comparison between the calculated theoretical rain and snow rates? The wording is awkward.
5. p. 13655, l. 28: obviously, the remaining rain fraction is 0.43. This part of the sentence can be left out.
6. p. 13656, l. 28: you may want to (re-)move this last sentence, as it does not fit into the context of Section 2.3.
7. p. 13657, l. 5: '…, *we later* apply…'.
8. p. 13660, l. 28: this is probably a typing error. Please check whether 'T_T2h' should be replaced by 'T_rH'.
9. P. 13665, l. 17: you repeat T_RR_D99 twice. Please replace one of them by T_D99.
10. P. 13666, l. 13 f: structure. You may want to change this to e.g. 'By discriminating three PPs, 3P1D and 3P2D enable…'.
11. P. 13669, l. 9: remove '( )'.
12. P. 13670: l. 10: your summary is written in past tense. Chance 'test' to 'tested'.
13. P. 13683, caption: 'serve *as*' instead of 'serve *es*'.
14. P. 13684, Figure 5: be consequent with labels in the caption, i.e. replace the hyphens by underscores (as is done in the manuscript text).
15. P. 13690, Figure 11: be consequent with labels in the caption, i.e. replace '2P' by '2P1D' (as is done in the manuscript text). Also, '1-p_snow' is likely to mean '1-p_rain'. Regarding the caption: You may want to swap T_rH and T_rH_D99 in the text, as T_rH_D99 is shown *first* (Fig. 11a). Additionally, the caption would become clearer if it was split into two sentences.
16. generally: stick to one version of setting commas when listing > 2 items (sometimes you set a comma, sometimes you don't).