AMTD amt-2015-370 – Reply to Referee #1

Jörg Burdanowitz, Christian Klepp, and Stephan Bakan

February 10, 2016

email address: joerg.burdanowitz@mpimet.mpg.de

In the following answer we proceed as follows. Text from the Referee #1 is shown in *italic*, our answer in **bold** and changes in the manuscript are highlighted in blue.

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.*

We very much acknowledge the careful and quick review of our manuscript by referee #1. We appreciate the degree of detail the referee put into the review pointing towards inconsistencies and shortcomings. In the following we address all raised points one by one.

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.

We agree that the wording was somewhat unclear for the reader to judge the different methods used. However, we tried to avoid the terms 3P1D and 3P2D in the abstract. As suggested, we adjusted the text as follows:

"Using two independent PP distributions represents a new method that outperforms the conventional method of using only one PP distribution to statistically derive the PP."

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.

We agree on that important point. Wherever we did not test statistical significance, "remarkable", "considerable", or "distinct" are used instead, e.g.

"The new statistical automatic PP distinction method considerably speeds

up the data post-processing within OceanRAIN [...]."
or "First, the manual method consumes a considerable amount of time [...].
However, with respect to Fig. 4, Fig. 6, and Fig. 9 the information whether 2 parameter combinations differ significantly, can be seen directly in the plots (they show box-whiskers of 100 realizations). We sharpened that information in the manuscript as follows:
"The four performance scores are calculated for both 100 realizations of 50 % randomly chosen minutes of precipitation (black boxes and whiskers in Fig. 4) and for all minutes of RS sub-data (red stars). The percentiles (5th, 25th, 50th, 75th, and 95th) illustrate how strongly the RS dataset scatters and whether performance differences between the predictor variable combinations are significant ($p = 0.95$, $n = 100$)."
referring to Fig. 4: "The addition of P and rH to T leads to a statistically significant ($p = 0.95$, $n = 100$) but only slight increase in accuracy compared to T alone."
or referring to Fig. 6: "The highest accuracy of 78 % by $T\_rH\_D_{99}$ represents a statistically significant performance increase compared to predictor combinations including RR that performed equally well in 2P1D."

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.

The actual sensing dimension of the IRSS88 is a plane. Though the precipitation particles still pass a volume, we adjusted the manuscript accordingly for the IRSS88 leaving the ODM description unchanged.
"The IR-LED of the ODM470 is only activated once at least 8 particles per minute pass the active sensing area of the precipitation detector IRSS88"
and
"The IRSS88 switches off the ODM470 after one minute without any particle passing the IRSS88 active sensing area."

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?

The translation from weather codes into PPs is not always trivial.

Whenever a weather code indicates the contribution of both liquid and/or solid precipitation particles we consider the mixed-phase. However, codes ww=96 and ww=99 to our understanding include exclusively graupel or hail (WMO: 'with hail at time of observation') that are both frozen particles. In contrast, code ww=95 states explicitly 'without hail, but with rain or snow at time of observation'. Thus, we assume that if not mentioned explicitly, no rain has been observed. Would you agree on that point, or do you see any indication of rain contributing to ww=96 and ww=99? Anyhow, these codes so far have not been observed in OceanRAIN and thus, the decision how to treat them does not directly influence the results presented in this manuscript.

WMO (2016): `http://www.wmo.int/pages/prog/www/WMOCodes/WMO306_vI2/LatestVERSION/WMO306_vI2_BUFRCREX_CodeFlag_en.pdf`

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, lets 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 overestimates 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 overprediction. 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.

Thanks a lot for pointing out this obviously unclear definition of 'snow (dis-)agreement' and 'rain (dis-)agreement'. The way we meant it is indeed your second assumption, e.g., 'rain disagreement' means that the model predicts rain, which disagrees with the reference. As I'm now aware of the ambiguity, I added the following sentence to additionally clarify the definition:

"For instance, rain disagreement means that the statistical model predicts rain that disagrees with the manual PP reference data indicating snow."

6. p. 13659, l. 16: you state that Fig. 4 includes significances of performance differences. In what way? This seems unclear.
As mentioned in 2., 100 realizations go into the box-whisker plot. Accordingly the reader can easily spot with his/her own eyes whether (or not) differences in parameter performances are statistically significant ($p = 0.95$, $n = 100$), which means that the median of one parameter combination lies without the 5th to 95th percentile range of another variable combination ($p = 0.95$). Hence, the underlying probability distributions would differ by about 2σ.
"The percentiles (5th, 25th, 50th, 75th, and 95th) illustrate how strongly the RS dataset scatters and whether differences among predictor variable combinations are significant ($p = 0.95$, $n = 100$)."

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...
I see the point that dropping T in some parts of the manuscript (e.g. in Figs. 4, 6, and 9) while displaying it in others (Table 4) is misleading. The sentence that you marked as being misleading intended to clarify that T is included in all parameter combinations. This inconsistency between text and Figures I resolve by adding T to all x-axis labels (new Figures attached). The sentence now reads more general:
"Combining T with up to two other relevant predictor variables (connected by underscores) aids to assess their value in determining the PP."

8. p. 13661, l. 3: 'accordingly' is misleading. You state that 'rH' tends to increase PM (which is a negative feature). Next, you state that the combination of 'rH' with either 'D99' or 'RR' decreases 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.
First, I totally agree on the usage of 'accordingly'. Second, my

intention to use it seems reasonable but the way I related PU to PM was not explained properly (referring to your next comment). For that reason I split the argumentation for PU and PM at this point and rewrote most of the paragraph (see below). There are two points to make here: (1) the T_rH_D99 and T_rH_RR perform better than T_D99, T_RR, and T_RR_D99 with respect to accuracy and PU. (2) PM mainly decreases because of using diameter-related predictors but is slightly increased by rH. This means (as you stated) that the positive effect of RR and D99 is stronger than the negative effect of rH.
"Besides being accurate and unbiased, a small PP transition region of low PP certainty (low PU) combined with a low fraction of highly certain but misclassified PP cases (low PM) characterize a useful predictor variable combination. The PU mainly scales with the accuracy. Consequently, predictor variable combinations including rH and either D_{99} or RR reach the lowest PU of about 36 %. This low PU and thus fairly narrow PP distribution causes a slight increase in PM for T_rH_RR and $T\_rH\_D_{99}$ (1.5 %) compared to T_D_{99}, T_RR, and $T\_RR\_D_{99}$ (1.3 %). However, the positive effect of using RR or D_{99} outweighs the slightly negative influence of rH on PM. Consequently, the physical related predictor variables confirm their good performance in predicting the PP."

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 anticorrelation 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).
This is a truly interesting and vital point for the analysis that we did not address sufficiently. To understand how PU and PM are related to each other we need to consider 2P1D and 3P2D separately (3P1D gives no reasonable PU). Defining the remainder of the sum of PM and PU as X gives
$100\,\% - PM - PU = X$.
For 2P1D, X represents simply the sum of all cases that $p(rain) > 0.95$ **(same as** $p(snow) < 0.05$**) being rain and all cases that** $p(rain) <$

0.05 **(same as** $p(snow) > 0.95$**) being snow (wrt the manual reference data). This means, X represents only the certain and correctly classified cases indicating that the sum of PM and PU should ideally be small. However, this also makes clear that (and why) PU is not perfectly anti-correlated with PM. Consequently, the higher PM of 4 % in KS98 is not a direct consequence but related to the low PU of 24 % via the fitting of the PP distribution.**

For 3P2D, the definition of PM stays the same but the definition of PU changes: Instead of using the sum of $0.05 < p(PP) < 0.95$ **for all PPs we use only those cases, where** $0.05 < p(PP) < 0.95$ **is fulfilled for _all_ PPs. This is what I mean by (logical) AND 'operator' (see your comment 13). Visually speaking (Fig. 8), we use only the overlap of** $0.05 < p(PP) < 0.95$ **for all three PPs. Otherwise, PU included cases where** $p(PP) < 0.05$ **for one or two of the PPs that is by definition not uncertain (e.g. a case like** $p(snow) = 0.03$**;** $p(mix) = 0.03$**;** $p(rain) = 0.94$**). Accordingly, these latter cases influence the relation between PU and PM in addition to the certain cases both of which adding to X.**

For clarification, we added the following text when comparing the KS98-derived coefficients against the OceanRAIN-derived coefficients:

Consequently, the coefficients from KS98 better predict most uncertain cases using T_rH but miss more extreme cases such as freezing rain. For the OceanRAIN dataset, the PP prediction using the RS-fitted coefficients better reflects the OceanRAIN PP distribution compared to the KS98-fitted coefficients as indicated by the accuracy.

We followed your suggestion to include D99 in Fig. 5 (see Sect. Updated Figures). This better illustrates the following explanations where we added a reference to Fig. 5. The caption of Fig. 5 has been adjusted accordingly.

"For $T\_rH\_D_{99}$, the rain/snow transition shifts with T depending on D_{99} (Fig. 5). While $D_{99} = 1$ mm shifts the rain/snow transition to even lower temperatures by about 0.5°C, $D_{99} = 5$ mm shifts it towards higher temperatures by about 2°C, both compared to T_rH derived from OceanRAIN RS sub-data."

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.*

Thank you for pointing towards the lack in clarity. The idea was to relate the variable combination that performs best in 3P1D (T_rH_D99) to that one that performs best in 2P1D (T_rH_RR) with respect to accuracy. However, in 2P1D the difference between T_rH_RR and T_rH_D99 is relatively small (0.12 %) and not statistically significant in the 2σ range which is why we wrote 'that performed equally well'. Obviously this expression seems largely unclear and we made the following modifications.

"The highest accuracy of 78 % by $T\_rH\_D_{99}$ represents a statistically significant performance increase to the remaining variable combinations in 3P1D, which contrasts to 2P1D where T_rH_RR does not perform significantly better than $T\_rH\_D_{99}$."

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?

I referred to the changed relationship of accuracy being ideally high and the PM being ideally low. In 2P1D both where anticorrelated but for 3P1D both seem to be slightly correlated. However it remains unclear whether the higher PM of T compared to T_rH_D99 is only/mostly associated with falsely classified PPs in the manual reference dataset. However, this could at least explain half of the difference as explained in p. 13663 l.27f. We found the manuscript to be clearer when dropping the whole sentence.

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 curve (=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.*

Indeed, this is a vital point in understanding the concept of the PP distributions. For the sake of consistency, we stick with the same thresholds for PM_mix as we did for PM_rain and PM_snow, i.e. p(PP)>0.95, which has to disagree with the PP from the PP reference data. However, this also means that PM_mix has basically no contribution to the (overall) PM, e.g. 5 cases for T_rH_RR. The definition of PM is already given in the manuscript (p. 13664 l.26). However, Fig. 8 seems to contradict this threshold of 0.95 for PM_mix because the PP distribution for mixed-phase (dotted line) does not exceed 0.75. Please keep in mind that the PP distribution shown in Fig. 8 represents just one realization of a fixed rH among many others. Furthermore, it remains mathematically impossible to have PM_mix>0 and PU>0 at the same time because if p(rain) and p(snow) do only overlap at p<0.05 then PU_rain and PU_snow cannot overlap anymore. We added a remark in the caption of Fig. 8

"We set $PM_{mix} > 0$ because otherwise we could not display it in the same PP distribution (rH kept constant) with $PU > 0$."

and in the text to avoid confusion.

"PM represents the percentage of all certain cases ($p(PP) > 0.95$; hatched area in Fig. 8) in which either one of the PPs disagrees with the manual PP reference data. PU as the percentage of uncertain cases ($0.05 < p(PP) < 0.95$; shaded area) represents only those cases where all PPs are uncertain after definition. We introduce this limitation because if for at least one PP $p(PP) < 0.05$ then we would not consider the PP uncertain anymore. Note that for mathematical reasons we cannot display $PM_{mix} > 0$ and $PU > 0$ in the same figure which is why we set $PM_{mix} > 0$."

13. p. 13665, l. 2: please explain what is meant by operator in this context.

By 'operator' I referred to the logical AND operator. This means as explained in the previous comment that at a certain temperature T all PU – PU_rain, PU_snow, and PU_mix – need to be greater than 0. The manuscript has been modified as we do no longer use the term 'operator', please refer to the previous comment to see the changes in the manuscript.

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.

Thanks for spotting that mistake, we switched 'rain' and 'snow' curve accordingly in the text. We also updated the Fig. 11 as suggested using underscores for the predictor combinations and 2P1D instead of 2P, displayed in Sect. Updated Figures.

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).

Yes, this comparison is constrained to using the KS98 coefficients and the number was previously mentioned in p13661, l.14. However, we see the lack in clarity and thus added the following in that sentence:

"Schmid and Mathis (2004) find an overprediction of snow cases (bias 0.82), very similar to the OceanRAIN RS snow overprediction (bias 0.8) using the same KS98 derived 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.

Yes, we first refer to the coefficients derived over Finland (Fig. 5, red curve) that show a narrower rain/snow transition compared to OceanRAIN. However, instead of 'narrower' it should be 'wider' PP distribution, which we corrected in the text. We further improved the comparison between OceanRAIN and the other datasets by considering T range and width of rain/snow transition separately.

"The OceanRAIN data using 3P2D reveals a wider rain/snow transition zone compared to data derived over Finland (Koistinen and Saltikoff, 1998). The rain/snow transition in OceanRAIN occurs at slightly lower temperatures compared to the data from Finland as well as NCEP DS464.0 global ocean ship data (Dai, 2008). The difference in the rain/snow transition zone likely originates from heterogeneous spatial and seasonal sampling in OceanRAIN that is likely to decrease with an increasing OceanRAIN time series."

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.

Yes, we refer to the same definition (common on the precipitation community, e.g.) where we mentioned the POD for mixed-phase (p.13668, l.6). However, in the conclusions I rounded up the POD of 0.25 to 0.3, which I see now is confusing. Therefore, 0.3 is replaced by 0.25.

"Mixed-phase precipitation carries the largest uncertainty of the three PPs and is most challenging to detect for the new algorithm with a probability of detection of about 0.25 using the predictor variable combination $T\_rH\_D_{99}$ and 3P2D."

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?

As you stated, many satellite retrievals cannot distinguish between rain, snow, and mixed-phase. With the help of the new PP distinction algorithm and OceanRAIN precipitation rates one could quantify satellite errors with respect to PP. This helps in particular to identify systematic retrieval errors with respect to PP. To clarify that aspect we adapted the manuscript as follows.

"The PP probability further allows error characterizing other precipitation datasets such as satellite data using OceanRAIN precipitation rates to unveil systematic errors with respect to PP."

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°C) and very high T (> 8°C), as the same is valid for Table 2 (compare p. 13656, l. 9 f). Please explain the difference.

The 268340 cases listed in Table 2 include all cases with a precipitation rate > 0. Excluding air temperatures below -6 and above 8°C, particles counts of < 20 and latitudes between -45°N and 706°N leaves 165632 cases. The difference to 165915 results from erroneously counted missing values when calculating the pdf. The value on Fig. 3 has been corrected. However, 4 cases exist where

$rh < 40\,\%$, which is why the sum of the numbers shown in the new Fig. 3 is by 4 smaller.
Concerning Table 2, the 164994 excluded also air temperatures of -6°C and of 8°C. Including these air temperatures results in 165632, which we corrected in Table 2. For RS, the sample size increases to 149635 cases while all rounded no-rain fractions remain unchanged.

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.
Good point, we added that information and a reference in the caption.
"RSM and RS include only those minutes with at least 20 particles of precipitation falling at mid- or high latitudes at air temperatures around the freezing point (see Sect. 2.3)."

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?
Increasing the data density for a comprehensive statistical analysis of precipitation, with emphasis on global oceans and high-latitudes, remains one of the key motivations of the OceanRAIN project. The overall goal is to provide a data base suitable for satellite retrieval validation, ship radar calibration and in turn analysis of scale dependencies of the precipitation parameter. All six ODMs are owned by the OceanRAIN project and measurements are continuously ongoing. The price of the instrument strongly depends on units built and would significantly drop with increased production. The OceanRAIN data collection effort is secured until the end of 2017. By that time, the time series of Polarstern will be at least twice as long (2010-2017). A total of 5 ships are long-term equipped with the OceanRAIN instrumen-

tation, namely the German RVs Polarstern (since 2010; entire Atlantic), Meteor (since 2014; tropical/subtropical Atlantic), the new Sonne (since 2015; tropical/subtropical Eastern Pacific), the Russian ship Akademik Ioffe (since 2010; entire Atlantic) and the Australian RV Investigator (since 2016; Indian/Pacific/Southern Oceans). We constantly seek for new long-term ship installation co-operations that offer support for instrument operation. Applications for the prolongation of OceanRAIN including further instrumentation are planned beyond 2017 to establish OceanRAIN as a truly long-term data set to the research community.

22. General comment: your results indicate that the bias scores is exclusively below 1, which indicates an overestimation 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?
This is indeed a very interesting question. So far, we cannot say whether we reached a limit of predictable accuracy by the logistic regression model or whether we can still significantly reduce the bias with the help of additional ancillary data. For sure, the statistical PP distinction model also strongly depends on spatial and temporal sampling of the calibration data set.
A future opportunity to decrease the snow-bias might be additional vertical information. Recently, an ODM470 has been deployed on the Australian RV Investigator that is also equipped with a vertically pointing Micro Rain Radar as well as a scanning C-band dual-pol radar. With the help of these instruments one can infer additional information about the vertical distribution of precipitation using the bright band. We are confident to identify potential systematic errors in the current PP distinction algorithm due to e.g. insufficient sampling.
However, we added one sentence to the discussion as a kind of outlook to state that this is a point worth investigating in the future when it comes to PP separation.
"Hence, OceanRAIN is likely to face the same problems underpredicting rain when supercooled raindrops fall under prevailing temperature inversions. Further work is needed in order to clarify whether we need additional

ancillary data to reduce the bias or whether the logistic regression model is unable in providing a less biased PP prediction."

Technical Corrections:

1.p. 13646, ll. 16: grammar. Better: 'An accuracy of 81.2 % is reached for...'.

Text modified as suggested.

"Including mixed-phase ($> 165,000$ min), an accuracy of 81.2 % is reached for two independent PP distributions with a slight snow overprediction bias of 0.93."

2. p. 13650, l. 4: one PP distribution distinguishes between two PPs.

Corrected.

3. p. 13650, l. 18: to develop a robust PP...

Corrected.

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.

Yes, we mean the difference between theoretical rain rate and theoretical snow rate. We modified the sentence to clarify this and replaced 'reveal' by 'identify'. "Large differences between theoretical rain and snow rate can help to identify a plausible PP."

5. p. 13655, l. 28: obviously, the remaining rain fraction is 0.43. This part of the sentence can be left out.

I agree and dropped this part of the sentence.

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.

We removed the sentence completely as the next sentence (first sentence of next Section) starts with the same information.

7. p. 13657, l. 5: ..., we later apply....

Corrected.

8. p. 13660, l. 28: this is probably a typing error. Please check whether T_T2h should be replaced by T_rH.
This is correct, thanks for spotting that typo! However, this sentence has been already revised with respect to specific comment #8.

9. P. 13665, l. 17: you repeat T_RR_D99 twice. Please replace one of them by T_D99.
Corrected.

10. P. 13666, l. 13f: structure. You may want to change this to e.g. By discriminating three PPs, 3P1D and 3P2D enable....
Corrected.

11. P. 13669, l. 9: remove ( ).
Citation corrected.

12. P. 13670: l. 10: your summary is written in past tense. Chance test to tested.
Tense corrected.

13. P. 13683, caption: serve as instead of serve es.
Typo corrected.

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).
We replaced all inconsistent labels in Figures 5 and 11 as well as incomplete labels ('$T\_$ ...') in Figures 4, 6, and 9. Please find these updated Figures in a separate Section.

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.
We updated Fig. 11 (Sect. Updated Figures) as suggested. $1 - p(snow)$ denotes the snow distribution for 3P2D but shown as a rain distribution to be better comparable to the other PP distributions.

16. generally: stick to one version of setting commas when listing > 2 items (sometimes you set a comma, sometimes you don't).
We set all missing commas before 'and' when listing more than 2 items.

Updated Figures

Modified Figures 3, 4, 5, 6, 7, 9, and 11:

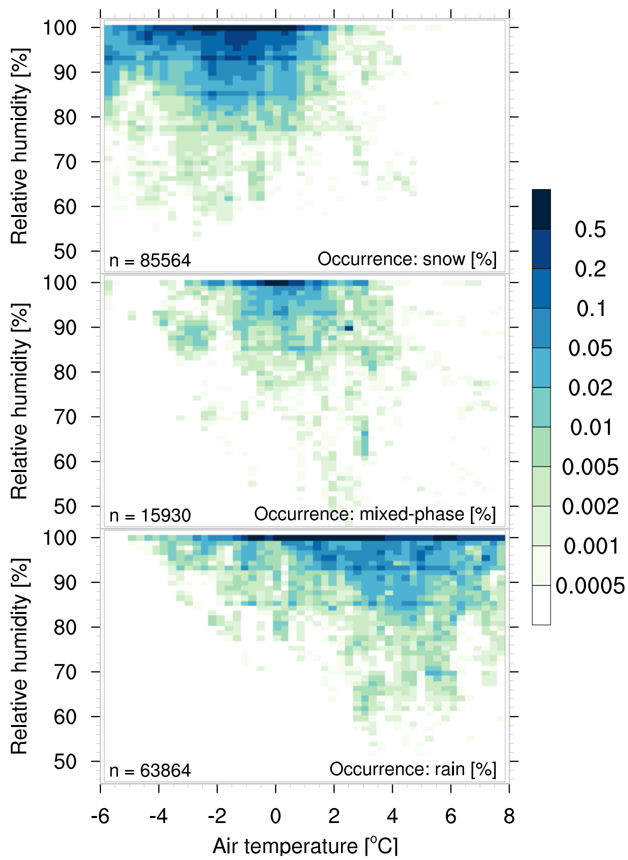

Figure 1: 2d-histogram shows relative occurrence [%] for each PP (top: snow; middle: mixed-phase; bottom: rain) after manual PP distinction from OceanRAIN RSM dataset of RV *Polarstern*. n denotes the number of minutes used per PP (165,632 in total).

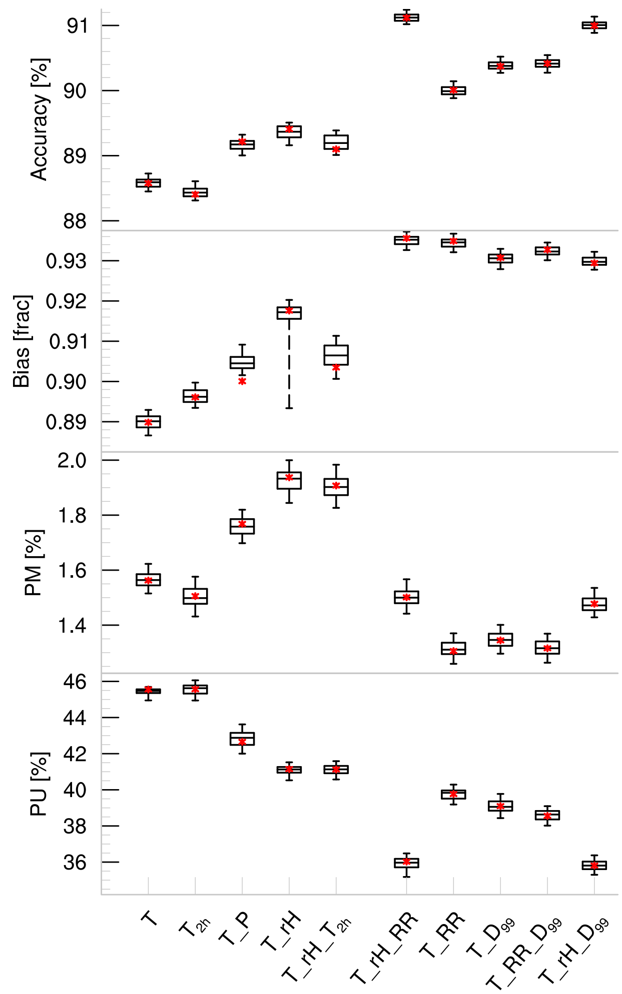

Figure 2: Box-whisker plot displays inter-quartile spread (black box: 25th, 50th, and 75th percentile) and lower (whisker: 5th percentile) as well as upper (95th percentile) extremes, calculated from 100 realizations of each 50 % randomly chosen minutes of precipitation from RS sub-data. Red stars denote the values for 100 % of RS sub-data. Accuracy [%], bias score [frac], percentage misclassified (PM: Fraction of disagreeing cases with high certainty of $p > 0.95$ in %) and percentage unclassified (PU: Fraction of uncertain cases of $0.05 < p < 0.95$ in %) serve as performance scores using the calculated coefficients in Table 2 against the manually determined PP reference data. Labels indicate variable combinations, whereby all combinations include T.

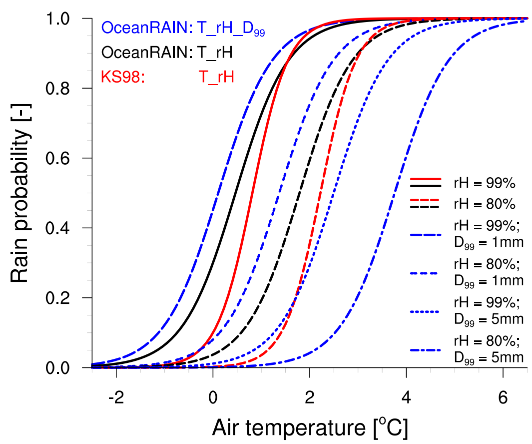

Figure 3: Rain probability using regression coefficients from Table 4 for OceanRAIN RS sub-data (2P1D) with the predictor variables T_rH (black), $T\_rH\_D_{99}$ (blue) both fitted against OceanRAIN, compared to KS98-recommended coefficients for T_rH (red). Dashed lines (black, red) indicate a PP distribution where rH is set to 80 % while for solid lines it is set to 99 %. For $T\_rH\_D_{99}$ (blue lines), D_{99} is set to either 1 or 5 mm in addition to rH.

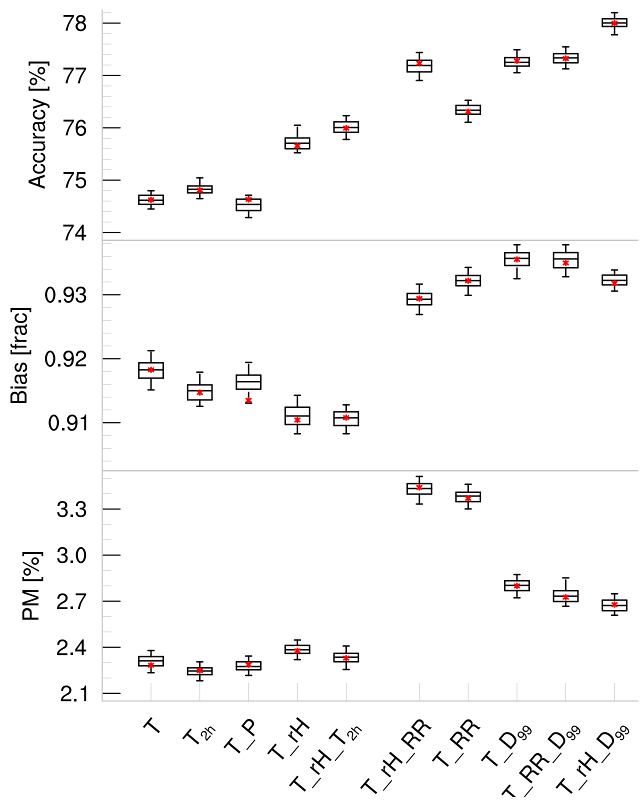

Figure 4: Performance of fit is shown for different combinations of atmospheric variables as in Fig. 4 for RSM sub-data. All variable combinations again include T.

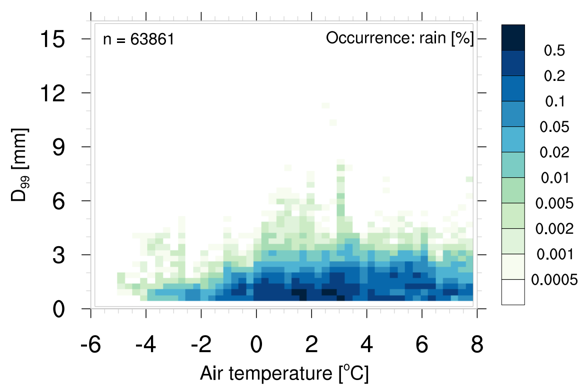

Figure 5: 2d-histogram of temperature and the 99th percentile of the particle diameter for cases classified as rain by the manual PP estimation in RSM.

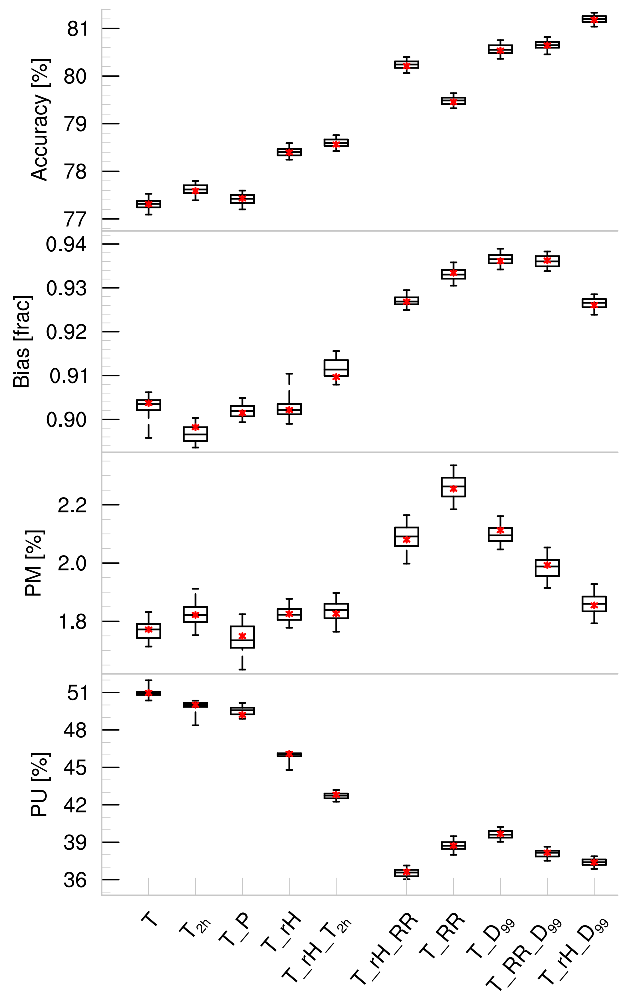

Figure 6: As Fig. 4 but for RSM including mixed-phase using two independent PP distributions (3P2D). The calculation of PM and PU differs from Fig. 4 as displayed and explained in Fig. 8.

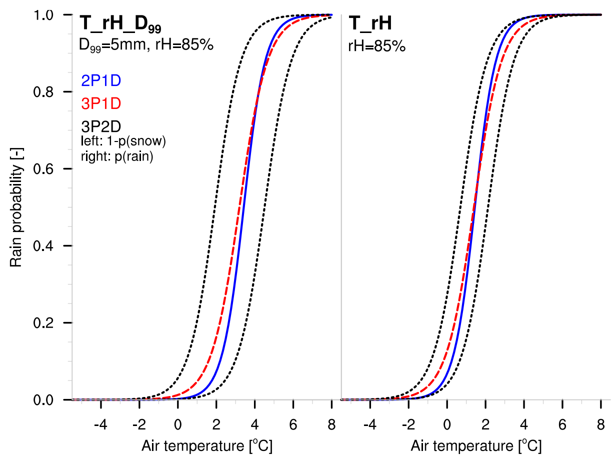

Figure 7: Air temperature shown against predicted PP by the different methods: two PPs (2P1D; solid blue), three PPs one-distribution (3P1D; dashed red), and three PPs two-distribution (3P1D; dotted black). 3P2D consists of two curves (left: snow distribution as $1 - p(snow)$; right: rain distribution as $p(rain)$) for the calculated coefficients of $T\_rH\_D_{99}$ (left panel; $rH = 85\,\%$, $D_{99} = 5$ mm) and T_rH (right panel; $rH = 85\,\%$).