HESS-2024-78 Leveraging a Disdrometer Network to Develop a Probabilistic Precipitation Phase Model in Eastern Canada

Response to anonymous referee 2

Many thanks to the reviewer for the very insightful comments, which helped us improve the quality of the article. Please note that additions to the article are shown in bold. The lines in this document refer to the previous version of the manuscript and may be subject to change in the revised version.

1 General comments

This is a very nice paper, and it is written well. It includes a lot of detailed analyses and discussions that make the paper very informative. I think it fits the journal well. I recommend major revision as I have several minor comments that need to be addressed before I can accept the paper for publication. Otherwise, the paper is in good shape. Also, as you revise the paper, please make sure to clarify how the automatic measurements enable mixed phase classification/ partitioning?

Thank you for the kind words. Regarding how automatic measurements enable mixed phase classification/partitioning, we explain in lines 112-136 how automatic phase measurements allow mixed-phase partitioning. We propose adding the following sentence clarify the reasoning:

Lines 122-123: *High-frequency automatic measurements do not suffer from limitations caused by mixed-phase precipitation (Froidurot et al., 2014; Harpold et al., 2017)**, as the precipitation phase can be coupled with a concurrent precipitation amount. When both phase identification and precipitation gauge measurements are made at a high frequency, phase-separated precipitation can be compiled for hourly or more timesteps, thus allowing for mixed-phase partitioning.***

Also, I suggest adding “radar-based” in the title, to clarify this is not for laser-based dendrometer.

This comment is in line with Reviewer #1 (see comment 2.1). We agree that this distinction with laser-based disdrometers is important. We suggest the new title:

Lines 1-3: *Leveraging a **radar-based** disdrometer network to develop a probabilistic precipitation phase model in eastern Canada*

2 Specific comments

2.1 Lines 21, list the atmospheric variables used.

This was also mentioned by Reviewer #1. We agree with the detailing of the variables used. Please refer to the response under their comment 2.2 regarding the proposed changes.

2.2 Abstract: partitioning is not an obvious term. I think by partitioning you mean the amount of each phase versus the type (classification). It would be helpful to define partitioning upfront.

This is indeed the intended difference between the two terms, and we acknowledge that it is not a well-established distinction. We propose adding a sentence that makes a clearer distinction between the two:

Lines 16-17: *Single-phase precipitation was also found to occur more frequently than mixed-phase precipitation.* ***This outlines the need to classify the precipitation phase, as well partitioning correctly between the solid and liquid precipitation amounts in the case of the mixed phase.***

2.3 Line 65, you may want to add one or both of these also to further support your point: https://doi.org/10.1016/j.jhydrol.2022.127884; https://doi.org/10.1029/2019GL084221

This is greatly appreciated input. The proposed references will be added to the following part of the manuscript:

Lines 63-66: *Indeed, solid precipitation is much more sensitive to undercatch (underestimation due to the wind moving hydrometeors away from the gauge) than liquid precipitation (Rasmussen et al., 2012). Consequently, an inaccurate measurement of the phase necessarily translates into an erroneous estimation of the precipitation quantity.* ***Ehsani and Behrangi (2022) showed that undercatch for solid precipitation introduced a significant bias in gridded precipitation products at both the seasonal and annual scales at higher latitudes. This highlights the need to account for the precipitation phase at the synoptic scale, especially when using precipitation products to bias-correct satellite precipitation estimates (Behrangi et al., 2019; Ehsani and Behrangi, 2022).***

2.4 Line 145, I thought the use of “However” instead of “Additional” might fit the sentence better.

We agree that it is a better fit for the sentence, it will be changed.

2.5 Line 153, I thought it would be useful to discuss the difference between phase classification and partitioning here as this may not be obvious for readers.

This is a good point, as it would further reinforce the distinction between the two terms. We propose adding the following sentence:

Lines 152-154: *The precipitation phase is classified before partitioning to accurately replicate its intricate behavior and to take advantage of the significant amount of validation data available through such a network.* ***As such, the models classify the precipitation as either solid, liquid, or mixed phase. The predicted phase then dictates the partitioning into solid and liquid fractions.***

2.6 185-186, so my understanding is that disdrometer does not "observe" phase. As you said in line 190 precipitation phase is identified according to the hydrometeor diameter-fall velocity relationships for water droplets and solid particles. So maybe you should replace "observation" with "estimation" or something similar.

This interpretation is correct, and it would indeed be more accurate to use terms such as "estimation" or "identification". We propose using the term "identification" in the paragraph from lines 185-197 containing the concerned sentence. However, for simplicity, we propose referring to the phase identifications as observations for the rest of the manuscript and adding at the end the following disclaimer.

Line 195: ***For simplicity, the phase identifications derived from the diameter-fall velocity relationships are referred as observations in this study.***

2.7 Lines 280-285, it is not clear to me how the aggregation of the mix of snow and rain/drizzle with rain is performed? How did you decide to convert fractions to solid or liquid phase?

The official WMO description of the mix of snow and rain/drizzle is "Rain or drizzle and snow, moderate or heavy", and under the broader rainfall category. This fact, and the verifications that are shown in Appendix B, suggest that this type of precipitation behaves more like rainfall than snowfall. Thus, for the purposes of this study, we assume that most of the precipitation is therefore rain accompanied by near-melting snow, resulting in a fully liquid precipitation. Additionally, the wording of the phase name was chosen purely for readability. As such, we also propose changing the name to "mix of rain or drizzle and snow" to better put an emphasis on the rain/drizzle part of the precipitation phase.

We acknowledge that, given the wording of the precipitation phase, the partitioning could be assumed to be 50/50 for the solid and liquid precipitation fractions, as seen for example in Wayand et al. (2016). This assumption was tested with this the data from this study and yielded unexpected results, such as a 2-m air temperature critical threshold for solid precipitation of about 3°C. This seemed highly unlikely given the literature on this subject (e.g. Jennings et al., 2018). Additionally, the discussion comparing the data used in other studies (see section 5.3) lends credibility to this study's aggregation step, as the results are similar.

We propose indicating this study's assumptions made about the fraction of solid and liquid precipitation in the mix of snow and rain/drizzle in a clearer way:

Lines 262-264: *Therefore, the disdrometer identifications of freezing rain and of mix of snow and rain/drizzle were aggregated,* ***respectively,*** *with snow and rain events.* ***For example, if an hourly precipitation has a fraction of rain and a mix of rain/drizzle and snow, it would be considered completely liquid after the aggregation.***

Furthermore, we propose adding the following statement to the section discussing instrument accuracy.

Lines 675-677: *The phase aggregation step considered the behavior of the snowpack following the different phases detected by the disdrometers.* ***Following a mix of rain/drizzle and snow, the SWE and snow height tended to decrease, a snowpack response similar to that following a rain event. It can be inferred that this type of precipitation is likely to be dominated by rainfall given the warm temperature at which it occurs and the ensuing effects on the snowpack. However, it is probable that this interpretation is specific to the disdrometers used in this study, unless evidence to the contrary emerges.*** *Phase identification errors also have the potential to introduce uncertainties in the results****, notably in the case of mixed-phase precipitation****.*

2.8 Add reference for the performance metrics used, on the other hand POD, FAR, HSS, CSI, etc. are also very popular. I also like BIAS as it shows over or under detection. How can you say you over or underpredicted?

The performance metrics used in this study are common in the machine learning field, a reference from Rokach et al. (2023) will be added. Recall is the term for POD in machine learning applications, thus a short acknowledgment of this fact will be added to the paper. As such, the precision was a logical choice for a second performance metric, for its inverse relationship with recall. The precision is also related to FAR, as it indicates the ratio of accurate predictions rather than false predictions. HSS and CSI were also calculated for this study but were not presented as they did not give significantly different results than the presented F1 score.

The assessment of over and underprediction can be achieved with precision and recall. Low precision and high recall scores indicate an overprediction (i.e. the mixed phase prediction of the linear transition and psychrometric balance models). On the contrary, high precision and low recall indicates underprediction. This interpretation will be added to the description of the metrics:

Lines 295-299: *The combination of precision and recall is commonly used to evaluate model classification performance, as the metrics indicate different information. Precision indicates the proportion of correct predictions for a given phase, while recall indicates the* ***probability of detection*** *for a given phase.* ***By definition****, model precision and recall are inversely proportional.* ***The assessment of both metrics informs if a model over or underpredicts a given class. For instance, low precision and high recall indicate a class overprediction, while high precision and low recall indicate a class underprediction.*** *Therefore, a model that achieves good performance in both metrics is desirable.*

2.9 Line 316, don't you have an extra "model" in the sentence?

This is indeed the case; it will be corrected.

2.10 L323-324: I am not sure if I understand why RMSE is the same for liquid and solid phase. Explain.

In the case of this study, the RMSE on either the solid or liquid precipitation is equivalent to the root mean squared misclassified precipitation amounts. As such, if the error of predicted solid precipitation is of X mm, the error on liquid precipitation is -X mm. Of course, because of the squaring of the error, both RMSE are equal. Thus, we propose replacing the following lines to better present this reasoning:

Lines 323-325: ***Because of the partitioning between solid or liquid precipitation, the RMSE is equal to the root mean squared of the misclassified precipitation. Therefore, the RMSE is the same for both solid and liquid precipitation, and a single score is presented.***

2.11 Section 3.3. it is not clear. Did you use the same number of solid and liquid phase for training? How about testing? For example, you say 60% solid and 26% liquid were there. Did you reduce the number of solid in training to match the number of liquid phase samples?

We acknowledge that the method used was not clearly explained. The stratified K-fold method was used to maintain the precipitation phase proportions of the dataset for both the training and validation sets, to ensure that there were enough liquid and mixed phase samples in the subsets. We propose the following additions to the manuscript:

Lines 351-352: *The data were then split using an 80/20 ratio between the training and validation sets respectively, resulting in 13,339 data points for training and 3,335 for validation.* ***To account for the prevalence of solid precipitation samples****, the training and validation sets were stratified to maintain the* ***aforementioned*** *phase proportions between the two subsets* ***(60% solid, 26% liquid, 14% mixed)****.*

2.12 Section 3.5. line 390, in the PB method, is the hydromet temperature similar to Wet bulb temperature. If not, what's the difference between PB and wet bulb temperature?

While the wet bulb temperature is based on the vapor deficit of an air parcel, the hydrometeor temperature is based on the mass and energy balance of a sublimating ice sphere. In practice, they are most likely similar, since the hydrometeor's surface is considered saturated with water vapor for mass transfer calculations. However, as the conditions are mostly near saturation in this study, both the hydrometeor and the wet bulb temperature are close to the dry bulb temperature. We propose adding more context to the manuscript:

Lines 389-391: *Finally, the psychrometric energy balance (PB) model is used, which is a phase partitioning method* ***based*** *on* ***the mass and energy balance of a sublimating ice sphere*** *that integrates the relative humidity to estimate the hydrometeor temperature (Harder and Pomeroy, 2013).*

2.13 Line 395, not sure what you mean the simple threshold method also gives binary and there is no mixed phase. The probability approach can be converted to binary rain/snow. I am not sure if I understand your point.

The simple threshold model was included solely because it is still used in hydrological applications and is the simplest method available. Probabilistic models (e.g. Behrangi et al., 2018; Jennings et al., 2018) were not included as benchmark models because mixed-phase precipitation was omitted from these studies, resulting from the limitations of the direct phase observations. Direct phase observations being a qualitative measure, there is no way to properly partition the precipitation phase. Other studies using probabilistic models do include mixed-phase precipitation (e.g. Ding et al., 2014; Shin et al., 2022), however the same partitioning issue remains. We propose modifying the following parts to better explain the reasoning:

Lines 396-398: ***Previous studies that employed probabilistic models based on direct phase observations (e.g., Behrangi et al., 2018; Jennings et al., 2018) were not included as benchmark models. Mixed-phase precipitation is typically excluded from such studies, as there is no effective method to accurately partition the precipitation due to the categorical nature of direct phase observations. The above considerations make such models difficult to compare with the PGP models presented in this study.***

2.14 Line 421, from Fig 6 I don't see how median RH is around 97%. Should be lower.

Please note that the vertical axis in Figure 6b) is on a logarithmic scale, therefore most of the data points are near saturation. To provide further context, roughly 65% of the data points are above 95% relative humidity, and 21% are at saturation. The use of a logarithmic scale was deemed necessary to improve readability of the figure, as seen in figure R1 which displays Figure 6b) without the logarithmic scale.

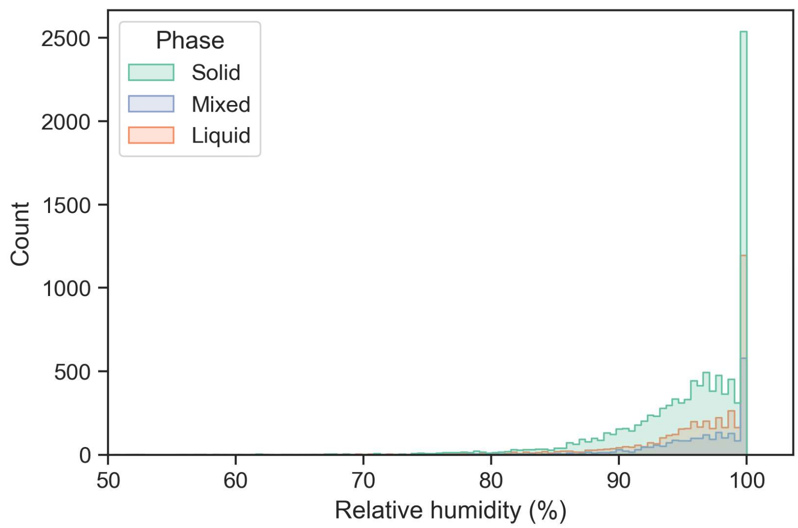

Figure R1: Relative humidity counts separated by precipitation phase.

2.15 Lines 445-455: I don't see any evidence in supporting your performance evaluation. Do you have a graphic or statistics in support of what you stated here about the performance evaluation? If so, refer to them in your text.

This is a very good point, as it is true that this statement has no supporting evidence at this point in the manuscript. We propose rearranging and merging the paragraphs from lines 445-455 and from lines 461-472.

Lines 445-455 and 461-472:

Figure 7 shows the phase density distribution of the benchmark models and the PGP models. *The* ***corresponding weighted*** *classification scores of the models* ***are presented*** *in Table 3.* ***The phase density distributions show the limitations of the benchmark phase partitioning models, namely that the mixed phase is absent or overrepresented compared to the observations. However,*** *ST performs well in all three metrics due to the low likelihood of mixed phase occurrence. When evaluating the overall classification performance using the F1 score, LT follows ST because of a disparity between precision and recall that affects its F1 score. The lower recall score for LT can be attributed to its overprediction of the less frequent mixed phase, which, in turn, negatively affects the recall of other phases. This enhances the model's weighted precision by decreasing the number of false positives in non-mixed-phase prediction. The same reasoning can be more extensively applied to PB's weighted scores. The mixed phase's overlap with other phases significantly decreases the model's overall recall.* ***Due to the relationships used to create the benchmark models, the overlap between all three phases is not accurately represented. By including relative humidity, PB can model the phase overlap, but this does not improve the modelled phase distribution densities with respect to the observations.***

The weighted F1 score for the PGP models shows that they have a more robust performance, as they yield high weighted precision and recall scores, while having a small disparity between

both scores. *The PGP models reproduce the observed phase overlap well, but slightly overpredict the mixed phase, affecting both the solid and liquid-phase predictions. PGP_basic overpredicts the most the mixed phase, while the difference between PGP_hydromet and PGP_full is marginal. This result suggests possible improvements to PGP models, particularly for mixed-phase precipitation.*

2.16 Can you show the observation reference in Fig. 7?

This is a good suggestion to improve the ease of comparison with the observations. Here is the modified figure that will be added to the manuscript.

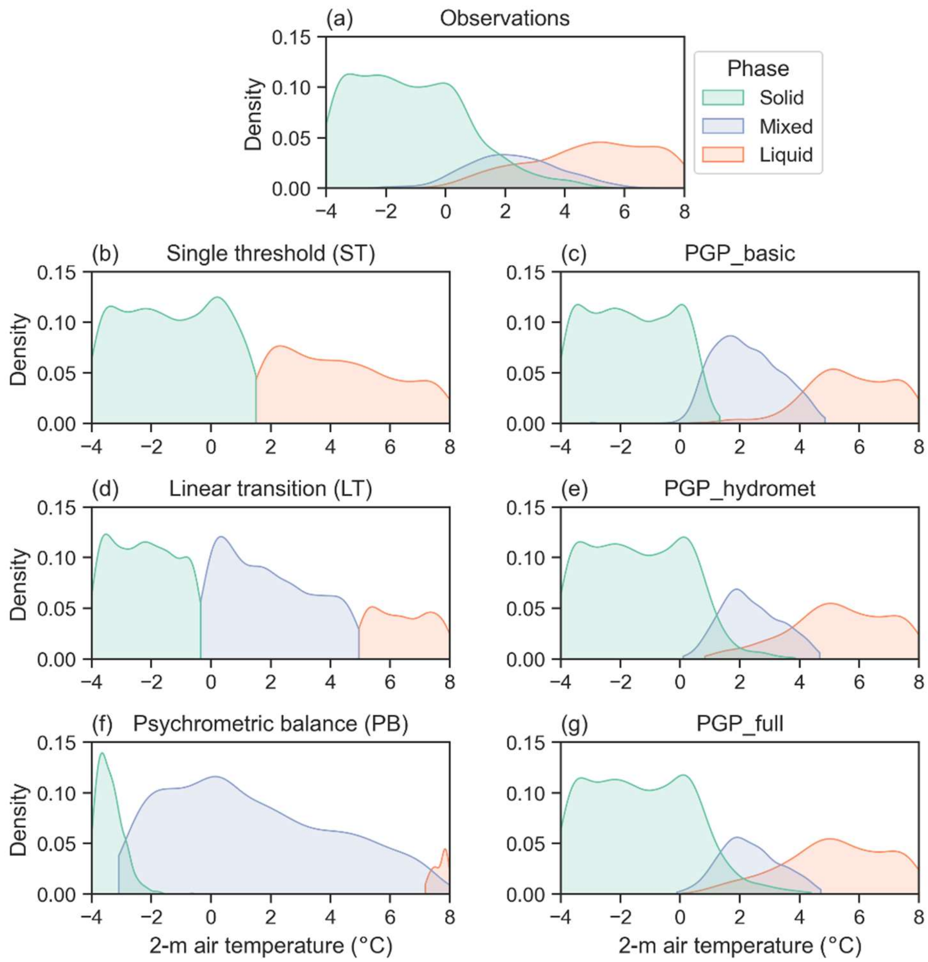

Figure 7: ***Hourly*** *phase distributions according to 2-m temperature of the (a)* ***observations, (b)*** *single threshold,* ***(c)*** *PGP_basic,* ***(d)*** *linear transition,* ***(e)*** *PGP_hydromet,* ***(f)*** *psychometric balance and* ***(g)*** *PGP_full. PGP model details are summarized in Table 2.*

2.17 Line 512, add “S” . Table 4 show”s”

Thank you, this will be corrected.

2.18 REMOVE lines 512-521. This is a repeat ! later in Lines 530-538:

Thank you, this will be corrected.

2.19 Line 571, I don't understand this sentence "Improving PGP models' ability to accurately predict the mixed phase is manifold"

Reviewer #1 made a similar comment about this sentence as well. We agree that it is vague and superfluous. Here are the proposed changes:

Lines 571-575: ***The scoring scheme for permutation importance must be carefully selected according to the model and use case. In this instance,*** *the PGP models tend to overpredict the mixed phase, which also negatively impact their ability to predict the other phases. In turn, this also affects the models' partitioning error,* ***which indicates that their overall performance is reliant on accurate phase classification****. For these reasons, the chosen scoring scheme for the permutation importance is the weighted F1 score, to consider* ***the classification of*** *the imbalanced* ***phase dataset****.*

2.20 Lines 627-628: Do you have references to back up your statement here?

The statement is based on the results presented in this study, where both LT and PB greatly overpredicted mixed-phase precipitation. However, we infer that this is due to the calibration method. For instance, the mixed-phase prediction of the benchmark models could be artificially constrained to reduce overprediction and improve classification performance. Such constraints would impact the models' partitioning performance, as the benchmark model results were optimized for partitioning, hence the use of "trade-off" in the manuscript. We propose modifying the following to better reflect this reasoning:

Line 627-630: *These models were calibrated to minimize partitioning error, but in doing so, they are biased toward predicting mixed-phase precipitation.* ***The mixed-phase prediction of the benchmark models could be artificially constrained to reduce overprediction and improve classification performance. Such constraints would however increase the benchmark models' partitioning error, given that they were calibrated according to solid precipitation fraction. Therefore****, there is a trade-off between classification and partitioning error for precipitation fraction-based models such as LT and PB.*

2.21 Line 635, could be helpful if you refer to the figure or table that supports your overprediction claim

This is a good point, it will be added.

2.22 Line 785, can you remind based on which figure the "phase overlap between 1.5 and 3.5°C" was concluded?

This is a great point. To improve the overall coherence of the manuscript, it will be added. Here is the modified sentence:

Line 785-786: *It successfully reproduced the phase overlap between 1.5 and 3.5°C **seen in Figures 6 and 7**, where the probability of mixed phase was highest.*

3 References to be added to the manuscript

Behrangi, A., Singh, A., Song, Y., and Panahi, M.: Assessing Gauge Undercatch Correction in Arctic Basins in Light of GRACE Observations, Geophysical Research Letters, 46, 11358-11366, https://doi.org/10.1029/2019GL084221, 2019.

Ehsani, M. R. and Behrangi, A.: A comparison of correction factors for the systematic gauge-measurement errors to improve the global land precipitation estimate, Journal of Hydrology, 610, 127884, https://doi.org/10.1016/j.jhydrol.2022.127884, 2022.

Rokach, L., Maimon, O., and Shmueli, E.: Machine Learning for Data Science Handbook, 3, Springer Cham, https://doi.org/10.1007/978-3-031-24628-9, 2023.

4 References in this document

Behrangi, A., Yin, X., Rajagopal, S., Stampoulis, D., and Ye, H.: On distinguishing snowfall from rainfall using near-surface atmospheric information: Comparative analysis, uncertainties and hydrologic importance, Quarterly Journal of the Royal Meteorological Society, 144, 89-102, https://doi.org/10.1002/qj.3240, 2018.

Ding, B., Yang, K., Qin, J., Wang, L., Chen, Y., and He, X.: The dependence of precipitation types on surface elevation and meteorological conditions and its parameterization, Journal of Hydrology, 513, 154-163, https://doi.org/10.1016/j.jhydrol.2014.03.038, 2014.

Jennings, K. S., Winchell, T. S., Livneh, B., and Molotch, N. P.: Spatial variation of the rain-snow temperature threshold across the Northern Hemisphere, Nat Commun, 9, 1148, https://doi.org/10.1038/s41467-018-03629-7, 2018.

Shin, K., Kim, K., Song, J. J., and Lee, G.: Classification of Precipitation Types Based on Machine Learning Using Dual-Polarization Radar Measurements and Thermodynamic Fields, Remote Sensing, 14, 3820, https://doi.org/10.3390/rs14153820, 2022.

Wayand, N. E., Stimberis, J., Zagrodnik, J. P., Mass, C. F., and Lundquist, J. D.: Improving simulations of precipitation phase and snowpack at a site subject to cold air intrusions: Snoqualmie Pass, WA, Journal of Geophysical Research: Atmospheres, 121, 9929-9942, https://doi.org/10.1002/2016JD025387, 2016.