Identifying Snow-Covered Areas from Unoccupied Aerial Systems (UAS) Visible Imagery:

A Comparison of Methods

By Moradi et al. (2025)

This manuscript compares several different semi-automated classifiers for mapping snow covered area from UAS RGB images across two study sites in New Hampshire, USA. UAS data were collected several times over several winters using two different sensors. The authors use this rich dataset to conduct multiple experiments. In general, the paper is well written, but would benefit by including additional work in their defined experiments. As is, I am not convinced that the methodological framework and conclusions in the current manuscript are a substantial improvement to our greater understanding of SCA derived from UAS RGB images. With that said, I believe that major revisions are necessary before consideration for publication.

Major Comments:

Overall the principle objectives are solid and worth pursuing. However, the objectives can be addressed more thoroughly with a few additions or revisions:

Throughout various sections of the paper the authors imply several relationships or causal drivers influencing model performance but never present formal statistical testing such as an analysis of variance, t-test, or regression. Without the accompanying metrics and indicating significant differences, these relationships should be reframed as qualitative observations.

We appreciate the reviewer highlighting the need for formal statistical testing to support statements regarding differences in model performance. In the original manuscript, several comparisons among classifiers were presented primarily as qualitative observations based on consistent patterns across the orthomosaics. However, we agree that formally evaluating whether these differences are statistically significant will strengthen the manuscript.

To address this comment, we are currently implementing statistical tests to evaluate differences in classifier performance across the 26 orthomosaics following Demšar (2006). Because each classification method was evaluated on the same set of orthomosaics, we treat the dataset as a repeated-measures design and apply a Friedman test to assess overall differences among classification methods for each performance metric. When significant differences are detected, we conduct pairwise Wilcoxon signed-rank tests with Holm correction to identify which classifiers differ while controlling for multiple comparisons.

These analyses are being incorporated into the revised manuscript, including the addition of statistical results in the Results section and clarification of the corresponding statements in the Discussion. Where differences are not statistically significant, the language has been revised to present these comparisons as qualitative observations rather than causal interpretations.

We believe that including these statistical evaluations will provide a more rigorous assessment of differences among classification methods while maintaining the original goal of comparing snow detection approaches across the orthomosaic dataset.

We have added the following to our methods section and will include the test results in the first comparison.

"Differences in classifier performance across the 26 orthomosaics were evaluated using a Friedman test, which accounts for the repeated-measures structure of the dataset because each classification method was applied to the same set of orthomosaics. When significant differences were detected, pairwise

Experiment 1: Baseline performance

Overall the section is sound in its presentation of model performance metrics, although can be much improved. A major limitation in this experiment, which cascades down into the subsequent is that the authors classify images into 2 classes, snow and bare ground. Admittingly, we are only interested in snow covered area, however we are also interested evaluating methods, which means that we are interested in how the models agree, but I am more interested in how, where, and why the models disagree. And to get that at those distinctions, the UAS imagery needs to be classified into more classes like shadow, snow, vegetation, bare ground. Without such classifications it is challenging to suggest with confidence which classifier might perform best under shadowy conditions etc.

Response:

We agree with the reviewer's point and acknowledge that such a detailed classification would be valuable, however, addressing this question would require manual pixel-level labeling across the imagery to distinguish classes such as shadow, vegetation, and bare ground and to attribute specific sources of disagreement among classifiers. Developing such a detailed reference dataset is beyond the scope of the present study, whose objective was to establish a baseline comparison of available methods for distinguishing snow from non-snow. For hydrological applications and snow monitoring purposes, the key variable of interest is typically the spatial extent of snow-covered area, which is fundamentally defined by the presence or absence of snow. Consequently, many operational snow monitoring systems and remote sensing algorithms are designed to distinguish snow from non-snow surfaces, rather than to classify multiple surface types. To stay consistent with our objective, we revised the manuscript and instead of using "*bare ground*" that could imply to the specific surface condition, we used "*no-snow*" throughout the manuscript.

Moreover, some of the approaches evaluated in this study, such as thresholding the blue-band reflectance, were specifically designed to classify only snow versus non-snow and cannot be extended to identify additional surface classes.

Nevertheless, we agree with the reviewer that the results can be better quantified to provide additional insight into disagreement among models. To address this point, we introduced a new metric, disagreement fraction, to quantify and visualize areas where classifiers produce inconsistent predictions. The metric is defined as follows:

To quantify disagreement among snow classification methods, we computed pixel-wise disagreement metrics across all methods. The fraction of methods classifying a pixel as snow was calculated as:

$$p_i = \frac{1}{M} \sum_{m=1}^{M} x_{m,i}$$

where $x_{m,i} \in \{1,0\}$, and 1 indicates snow while 0 indicates no snow. A majority-vote classification was defined such that a pixel was considered snow-covered if $p_i \geq 0.5$. The disagreement fraction, D_i, was then computed as the proportion of methods that did not match the majority classification, which can be expressed as:

$$D_i = \min(p_i, 1 - p_i)$$

where D_i ranges from 0 (complete agreement among methods) to 0.5 (maximum disagreement). Spatial maps of D_i were produced to identify areas where snow classification results differed most strongly among the evaluated methods. These maps can summarize the intend of maps presented at Figure 11 more effectively.

The disagreement maps suggests that the presence of shadows leads to the highest disagreement among models (Di > 0.4). The figure below shows the disagreement map for the TF study site on 12/23/2020, where approximately 20% of the area exhibited disagreement values greater than 0.4, indicating that at least three of the seven tested models disagreed with the majority vote in those locations. The models contributing to this disagreement were MLE, RF, and SVM, which correctly identified snow in shaded areas, whereas the other models classified these shaded snow pixels as snow-free.

(A)   (B)   (C)
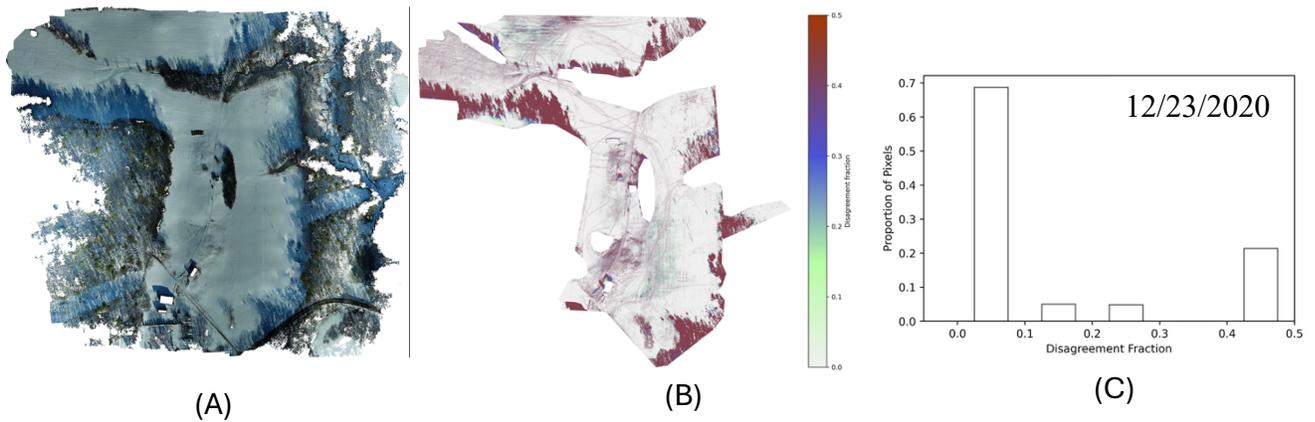
When fSCA was high (>90%) and shadow effects were minimal, the models showed strong agreement across nearly the entire study area. An example of this condition occurred at the KF site on 01/23/2024, as shown below.
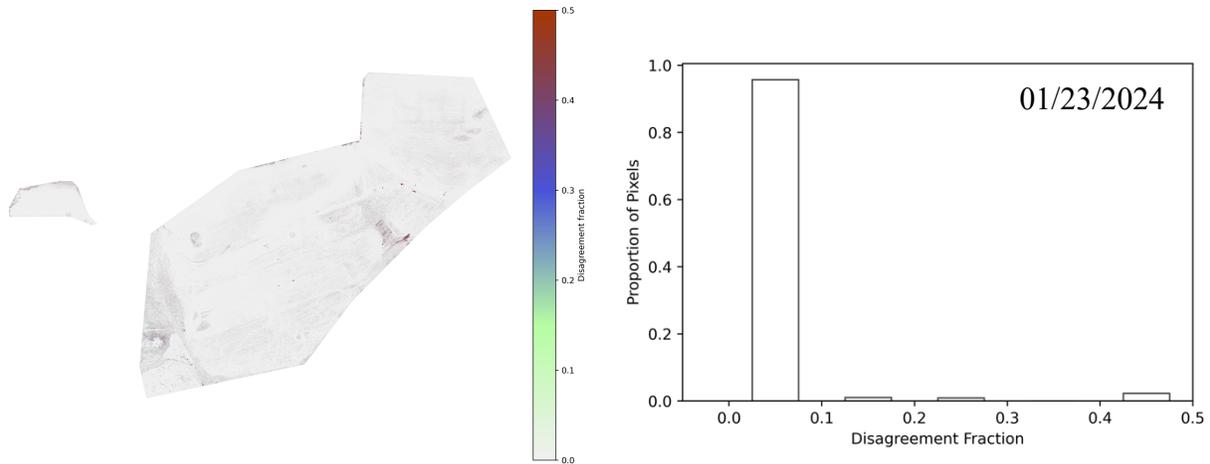
However, during melt conditions (fSCA < 0.4) and in scenes without shadows, disagreement between at least one model and the majority classification occurred more frequently across the study area compared to full snow-cover conditions. The plots below illustrate one such example:

02/08/2024
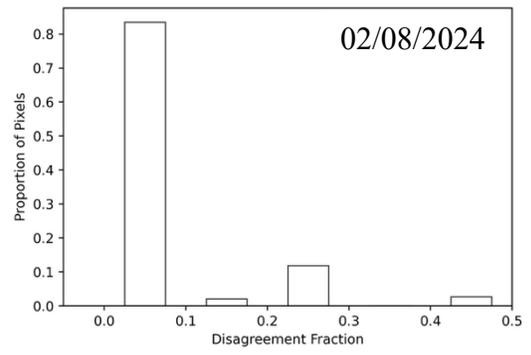
Further, this study has a valuable opportunity to incorporate lidar. Lidar can provide an independent dataset for mapping SCA. As well as help classify snow types which can further be used to aggregate model performance.

Response:
We appreciate the reviewer's suggestion to explore the use of LiDAR for mapping snow-covered area. Coincident LiDAR data with optical data were available for 13 acquisition dates, which allowed us to compare LiDAR SCA maps to the optical SCA maps. The table below summarizes the years and dates of the available dataset.

Winter	Total RGB Images	Field Site		Sensor		Lidar
		TF	KF	P4	A5100	
2020-2021	13	13	0	13	0	9
2021-2022	2	2	0	2	0	0
2023-2024	9	0	9	9	0	0
2024-2025	6	0	6	2	4	4
Total	30	17	15	26	4	13

Because the reviewer did not specify a particular approach for deriving SCA from LiDAR observations, we evaluated two methods commonly used in the literature: (1) thresholding LiDAR-derived snow depth and (2) classification based on LiDAR return intensity. The former analysis will be added to the methods and results with appropriate citations in the introduction. The latter results were inconclusive and therefore are not included in the revision. See below for more information about the two analyses.

Thresholding LiDAR-derived snow depth: We also applied the more traditional approach of deriving SCA by thresholding LiDAR-derived snow depth values to produce SCA maps. We compared the LiDAR-derived fSCA values to fSCA derived from optical images for the dates when both observations were available. The specific snow-depth thresholds used to define LiDAR-based SCA vary across studies. Some studies classify pixels as snow-covered for any positive snow depth (i.e., > 0 m), particularly when working with high-quality depth products or when focusing on continuous snowpack conditions (e.g., Helbig et al., 2021, Johnston et al., 2025). Other studies adopt more conservative thresholds, commonly on the order of 3 to 10 cm, to suppress false snow detection associated with LiDAR noise and to reflect the minimum detectable snow depth (Hopkinson, 2012; Deems et al., 2013, Cristea et al., 2017; Wang et al., 2025). In applications where airborne LiDAR snow depth products (e.g., ASO datasets) are used, a 10 to 15 cm threshold has frequently been applied to derive binary snow masks (Kostadinov et al., 2029; Cannistra et al., 2021; Thaler et al., 2023; Yang et al., 2023). Collectively, these studies highlight that LiDAR-based SCA mapping is inherently threshold-dependent, and that threshold selection should be informed by snow-depth uncertainty, spatial resolution, and the snow regime under investigation, particularly in shallow or patchy snow environments. The choice of threshold is particularly important in regions with shallow or ephemeral snowpacks, where snow depths are often comparable to measurement uncertainty and small residual errors can otherwise lead to overestimation of snow extent.

To derive binary SCA for UAS LiDAR snow depth pixels, we applied a series of snow-depth thresholds to the 13 LiDAR-derived snow depth fields that were flown coincidentally with the optical. We tested

thresholds ranging from 0 to 10 cm, in increments of 2 cm, and classified pixels with snow depth exceeding the threshold as snow-covered, then estimated the fSCA values for the study area. The results are summarized as follows:
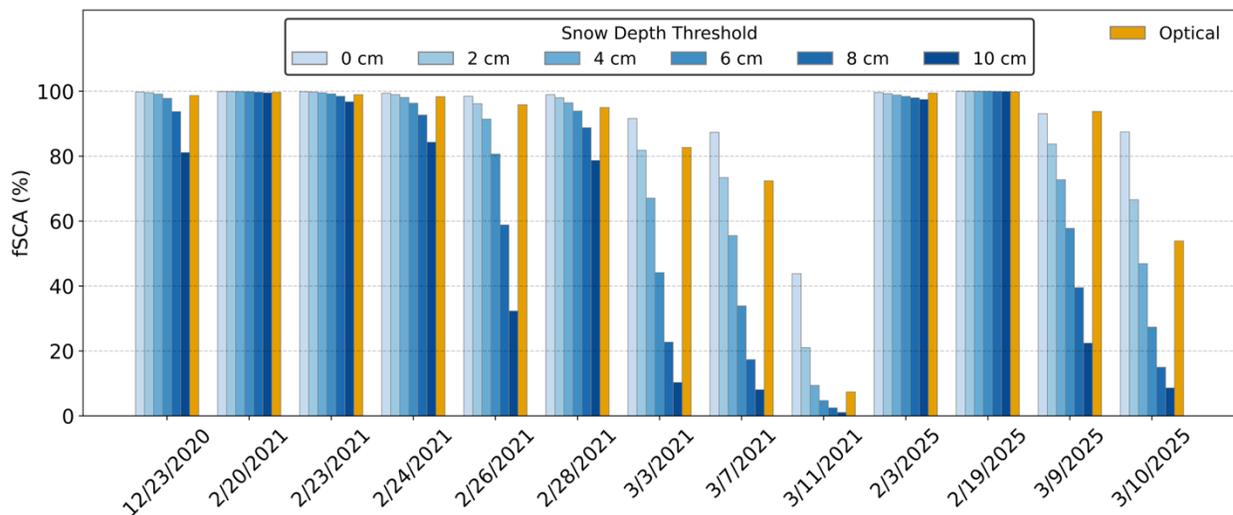
This analysis indicates that during periods of deeper snowpack, when average snow depth exceeded 10 cm (Table S-2), fSCA estimates derived from different snow depth thresholds closely matched the optical snow cover estimates, with differences among thresholds remaining small (<5%). As shown in Figure X, threshold-based estimates converge during periods of near-complete snow cover. In contrast, divergence among thresholds emerges during shallow or patchy snow conditions. For example, on 2/26/2021, the 0 cm threshold indicated nearly complete coverage (98%), while progressively higher thresholds reduced estimated fSCA to ~80% (6 cm) and ~30% (10 cm). The optical-derived fSCA (95%), used here as the reference estimate, confirms the near-complete spatial coverage of a shallow snowpack on this date.

Similar threshold sensitivity occurred on 3/3/2021 and 3/7/2021, when increasing the depth threshold from 0 to 10 cm reduced estimated fSCA by more than 60 percentage points. The optical reference fSCA on these dates was 82% and 72%, respectively, indicating discontinuous snow cover during the melt stage. During these periods, increasing the minimum snow depth threshold systematically reduced estimated fSCA, producing a pronounced gradient among thresholds. These results suggest that a substantial fraction of the landscape was covered by shallow snow detectable in optical imagery, whereas higher depth thresholds underestimated fSCA.

The sensitivity of these estimates is further influenced by the ~2 cm uncertainty associated with LiDAR-derived snow depth measurements (Harder et al., 2020; Jacobs et al., 2021). In ephemeral and transitional snow regimes, shallow but spatially continuous snow cover frequently occurs, and small changes in the minimum depth threshold can therefore produce large differences in estimated snow-covered area under such conditions.

Classification based on LiDAR return intensity: Recent work has suggested that LiDAR return intensity may outperform traditional optical imagery for SCA mapping (Ackroyd et al., 2024). Following the methodology proposed by Ackroyd et al. (2024), we investigated the use of normalized LiDAR intensity for snow classification. In theory, near-infrared LiDAR return intensity should be greater over snow-covered surfaces than over bare ground due to the high reflectivity of snow at these wavelengths (Antonarakis, Richards, and Brasington, 2008; Song et al., 2002). Ackroyd et al. (2024) showed that, after correcting for range and incidence angle, normalized intensity values exhibited a bimodal distribution that

allowed threshold-based delineation of SCA. Based on these findings, we worked closely with the Ackroyd et al. (2024) authors, Chelsea Ackroyd and Mckenzie Skiles, to apply their method to our datasets.

In our analysis, normalized intensity values did not exhibit a clear bimodal distribution that was evident in Ackroyd et al. (2024) (see the figure below). Visual comparison with coincident RGB imagery also did not reveal a consistent separation between snow-covered and snow-free surfaces in either raw or corrected intensity values. This suggests that the reflectance contrast between snow and background surfaces was not preserved in our dataset. One possible explanation is the difference in snowpack characteristics between study regions. Unlike the deep, dry snowpacks typical of many western mountainous environments, the snow in our study area was often shallow, wet, and spatially heterogeneous. During partial snow cover conditions (e.g., March 9–10), snow depths were frequently limited to only a few centimeters. Although LiDAR at 1064 nm is generally insensitive to subsurface contributions, extremely shallow snow over darker substrates may reduce return intensity relative to optically thick snow. While snow is typically more reflective than surrounding surfaces, shallow, wet, or debris-influenced snow may reduce reflectance contrast sufficiently to disrupt the bimodal intensity distribution observed by Ackroyd et al. (2024). Without a clear bimodal separation, threshold-based classification becomes unreliable, limiting the effectiveness of intensity-only SCA mapping under these conditions.
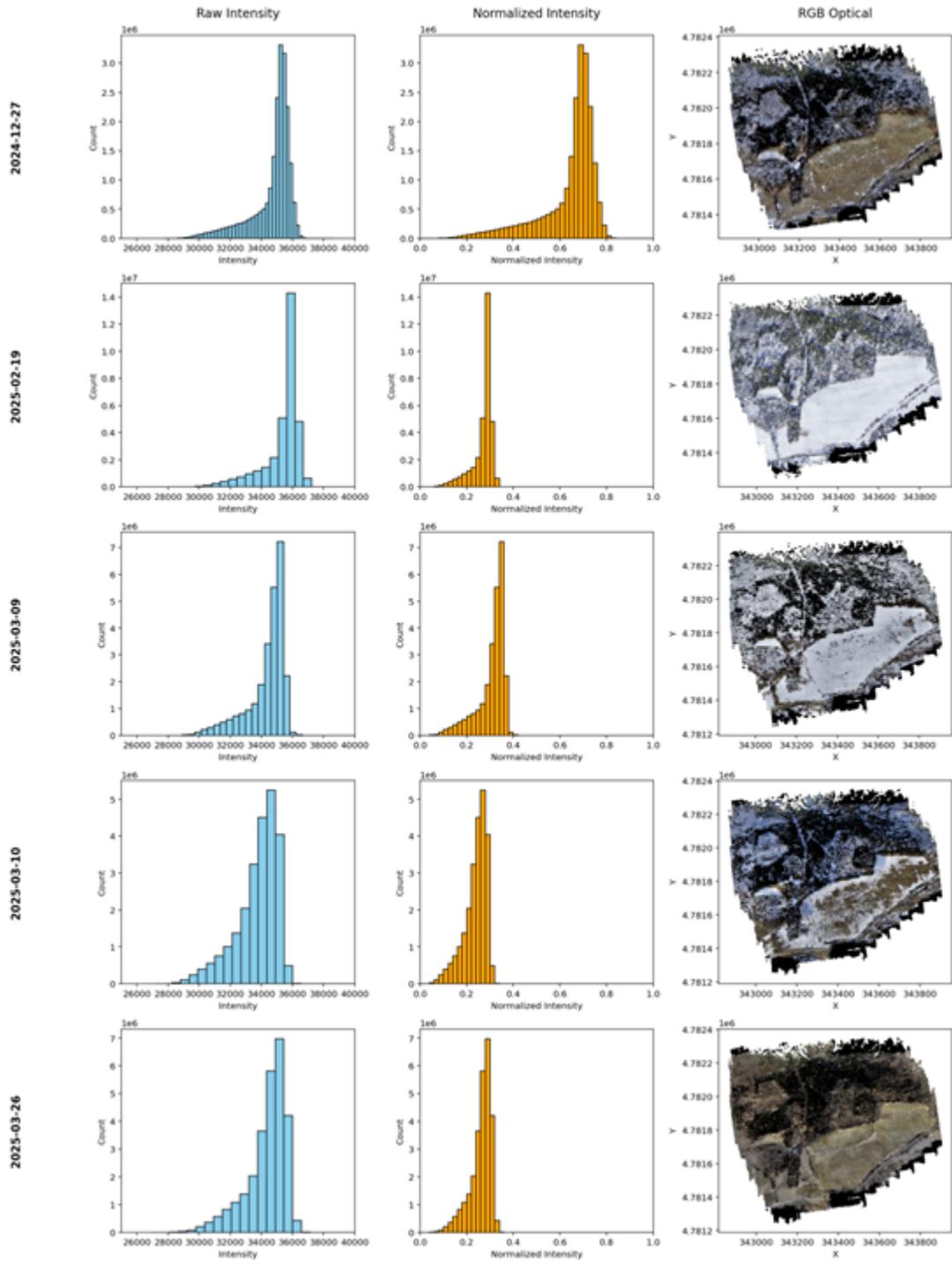
The following table (Table S-2) has been added to the supplementary material and provides details on weather conditions and snow depth on the dates of UAS acquisition.

Date	Site	Sensor	Start Time (local time)	End Time (local time)	Mean air temp (during flight) [°C]	Stdev air temp (during flight) [°C]	Mean SW_{in} [W/m²]	Stdev SW_{in} [W/m²]	Sky Condition	Field camera snow depth (cm)	Judd mean snow depth (cm)	Lidar mean snow depth (cm)
12/23/2020	TF	P4	9:26:00	11:10:00	-1.05	0.48	304.76	38.26	Partly Cloudy	7	9.21	12.8
1/25/2021	TF	P4	10:43:00	11:40:00	-1.86	0.56	428.25	12.55	Cloudy	0	0	No data
2/20/2021	TF	P4	12:04:00	12:44:00	-1.1	0.28	479.63	132.73	Cloudy	20	17.4	24.5
2/23/2021	TF	P4	12:16:00	12:55:00	5.01	0.23	541.13	144.44	Mostly Clear	15	13.8	17.6
2/24/2021	TF	P4	10:41:00	11:20:00	7.55	0.24	590.63	18.59	Clear	10	11.5	13.3
2/26/2021	TF	P4	11:57:00	12:38:00	0.16	0.18	632.75	3.41	Cloudy	7	9.06	8.7
2/28/2021	TF	P4	12:21:00	13:01:00	5.64	0.09	261.63	16.71	Clear	7	9.56	12.5
3/1/2021	TF	P4	16:37:00	17:16:00	5.36	0.16	29.5	17.02	Cloudy	4	7.72	No data
3/3/2021	TF	P4	12:00:00	12:39:00	3.24	0.34	459.63	177.73	Partly Cloudy	3	7.25	5.4
3/7/2021	TF	P4	11:25:00	12:06:00	-2.1	0.25	679.33	2.74	Clear	0	6.38	4.5
3/10/2021	TF	P4	13:50:00	14:34:00	7.84	0.17	541.67	27.6	Clear	0	3.23	No data
3/11/2021	TF	P4	15:12:00	15:51:00	18.89	0.35	265.13	62.83	Clear	0	1.65	0
4/2/2021	TF	P4	14:19:00	14:58:00	2.49	0.21	406.5	94.09	Clear	0	0	No data
1/12/2022	TF	P4	14:47:00	15:26:00	0.43	0.09	19.5	10.07	Clear	4	4.2	No data
1/26/2022	TF	P4	12:50:00	13:32:00	-6.03	0.18	490.11	15.76	Clear	0	0	No data
1/18/2024	KF	P4	13:39:00	14:43:00	-3.25	0.1	162.31	42.14	Partly Cloudy	8	6.4	No data
1/23/2024	KF	P4	14:26:00	15:17:00	3.21	0.03	86.1	23.04	Cloudy	6	6.7	No data
1/27/2024	KF	P4	13:46:00	14:26:00	3.4	0.16	94.25	17.3	Mostly Clear	4	4.6	No data
2/6/2024	KF	P4	14:36:00	15:34:00	3.69	0.14	259.45	40.05	Clear	4	5.4	No data
2/8/2024	KF	P4	13:19:00	13:59:00	6.25	0.14	436	17.19	Clear	2	3.8	No data
2/12/2024	KF	P4	15:40:00	16:27:00	6.12	0.8	80.9	59.04	Partly Cloudy	0	0	No data
2/26/2024	KF	P4	10:45:00	11:25:00	5.34	0.88	491.11	139	Mostly Clear	0	0	No data
4/5/2024	KF	P4	15:16:00	16:02:00	5.61	0.13	95.78	22.5	Cloudy	15	11.9	No data
4/8/2024	KF	P4	14:41:00	15:27:00	16.48	0.76	246.78	96.5	Clear	0	0	No data
12/27/2024	KF	MR	13:25:00	14:10:00	3.5	0.12	165.3	34.01	Cloudy	No data	No data	No data
1/13/2025	KF	P4	15:38:00	16:31:00	0.86	0.3	0	0	Cloudy	No data	No data	No data
2/3/2025	KF	MR	12:15:00	13:00:00	2.81	0.17	356.3	8.43	Partly Cloudy	No data	No data	16.6
2/19/2025	KF	MR	12:25:00	13:15:00	-4.84	0.29	424.27	8.67	Mostly Clear	No data	No data	40.7
3/9/2025	KF	MR	14:35:00	15:20:00	4.35	0.24	192.9	27.29	Partly Cloudy	No data	No data	6.9
3/10/2025	KF	MR	16:45:00	17:25:00	12.31	0.28	196.56	44.89	Mostly Clear	No data	No data	4.3

Experiment 2: Cross-sensor Comparison

Though relevant, I believe as the experiment is currently setup this is not only a cross-sensor but also a cross-date comparison. In order to better isolate the differences model performance across sensors, if the model is trained using images from camera A, then it needs to be tested using images co-collected imagery from A and B. This allows us to distinguish model performance elative to camera A.

Experiment 3: Across Sites

Similar to experiment 2, I think working with UAS imagery the day and time of day of

acquisition is so important. In this experiment, it is again difficult to parse out what parts of model performance are being affected by location or day of acquisition?

Results for experiment 2 and 3 were similar in that differences in model performance were not from sensor or location but rather differences in date of acquisition. It would be ideal if both experiments had isolated the effect of acquisition date.

Reviewing these three experiments I would suggest excluding the sensor comparison and even only focus on the images acquired from the A5100. Even though that reduces the total number of acquisitions from

36 to 10. I believe the reduction in imagery is worth the ability to incorporate lidar into the study, which will allow for the introduction of an independent snow cover map as well as aggregating snow cover classifications by snow depth.

Response:
We agree that because the data collected by different sensors were not acquired simultaneously and environmental conditions differed between acquisition dates, the analysis would not fully support the intended comparison. For this reason, we decided to remove Experiments 2 and 3 from the manuscript. However, the main analysis has remained the same using the 26 images collected with the P4 platform, as they span a wider range of snow conditions across different stages of accumulation and melt periods.

Specific Comments:

The current figures are a bit repetitive and difficult to read mainly because there is little variability in the performance metrics. These data and model outputs can be presented in several interesting ways including aggregated area values as timeseries or more qualitative approaches that identify model outputs misclassed as shadow, for example.

Response:
To address the issue of repetitive maps, Figures 3 and 4 have been revised to show a smaller set of representative images at a larger size, while the original versions have been moved to the supplementary material. In addition, the multiple panels in Figure 11 were removed as the disagreement maps aggregate the findings from those maps.
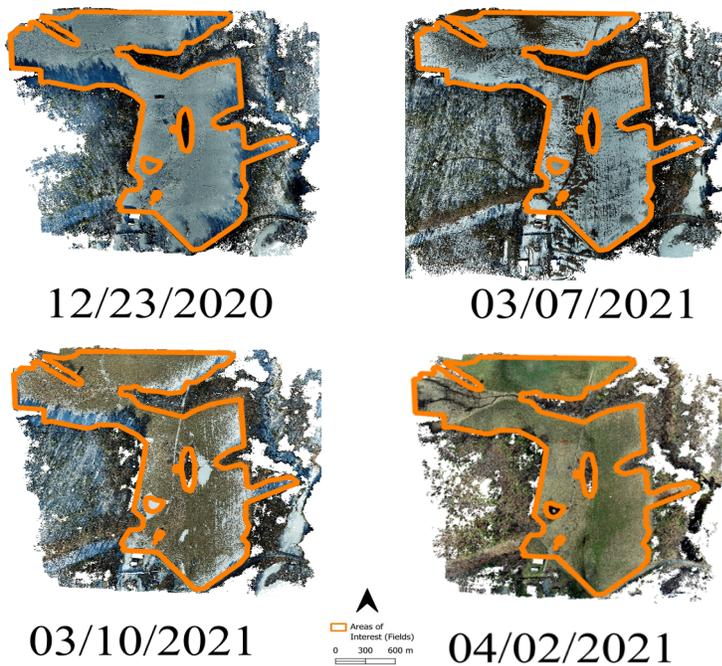

Figure 3. Example RGB orthomosaics of Thompson Farm acquired with the Phantom 4. Survey dates are shown below each orthomosaic. Areas of interest are outlined in orange.
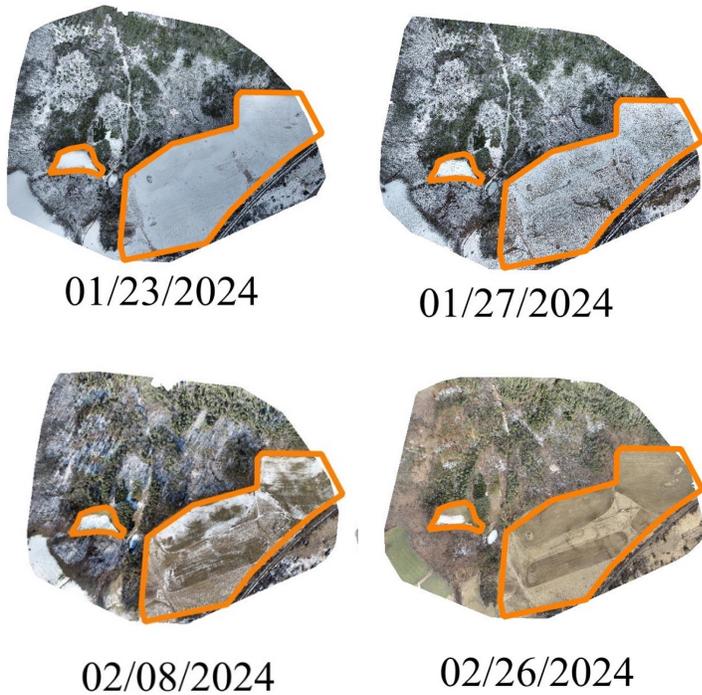
Figure 4. Example RGB orthomosaics of Kingman Farm acquired with the Phantom 4. Survey dates are shown below each orthomosaic. Areas of interest are outlined in orange.

Random forest is spelled out in several places. It should be consistently presented as 'RF.'

Response:

Thank you for pointing this out. We reviewed the manuscript and ensured that the abbreviation *RF* is used consistently throughout the text.

In the sampling approach, how were shadows excluded form the random sample? It seems it could be possible that a randomly selected pixel could be a shadow? Were classes equally distributed, 25 for snow and 25 for bare ground?

Response:

To clarify, we would like to emphasize that the shadowed areas were not excluded from the random sampling. Figure S16 shows that several randomly generated sampling points fall within shadowed regions (e.g., on 1/25/2021 and 2/20/2021). The sampling locations were generated randomly, with no constraints imposed to control their distribution across different surface characteristics.

In the discussion 4.1, it states 'Our primary goal was to identify one or more classifiers that can rapidly… ' Rapidly implies that model performance speed was tested, no metrics are presented but could be valuable.

Response:

Thank you for pointing this out. The word "*rapidly*" has been removed from the text for clarity.