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 have added statistical tests to evaluate differences in classifier performance across the 26 orthomosaics following Demšar (2006). Because each classification method was evaluated on the same test set, 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 to identify which classifiers differ while controlling for multiple comparisons.

These analyses have been incorporated into the revised manuscript, including the addition of the statistical testing methodology to the Methods section (2.3.3) and the corresponding results to the Results section (3.1).

"Statistical differences among classifiers were first assessed using the Friedman test (Friedman, 1937; Demšar, 2006) applied to balanced accuracy and F1 score across the repeated realizations. Post-hoc pairwise comparisons were subsequently performed using paired Wilcoxon signed-rank test (Wilcoxon, 1945; Demšar, 2006) tests with false discovery rate correction to account for multiple comparisons. Because the dynamic thresholding approach does not require labeled training data and was applied independently to each orthomosaics, the repeated train/test evaluation approach used for supervised classifiers was not applicable. As a result, this method was excluded from the statistical comparison analyses."

The results of statistical tests are added to the Results section:

"*Statistical comparison using the Friedman test revealed significant differences in both balanced*

accuracy ($\chi^2 = 32.82$, $p < 0.001$) and F1 score ($\chi^2 = 31.62$, $p < 0.001$) among the evaluated classifiers. This indicates that classifier choice had a significant impact on classification performance. Post-hoc Wilcoxon signed-rank tests on both balance accuracy and F1 score (Tables S-4 and S-5) indicated that MLE, RF and SVM achieved statistically comparable performance (adjusted $p > 0.05$), whereas both blue-band-only approaches differed significantly from the higher performing classifiers (adjusted $p < 0.05$). Although the SVM blue band and static blue band classifiers showed statistically comparable F1 scores, their balanced accuracies statistically differed significantly"

***Table S-4** - Pairwise Wilcoxon tests on balanced accuracy with FDR correction*

Classifier 1	Classifier 2	Wilcoxon statistic	raw p-value	FDR-corrected p-value	Significant at alpha=0.05
RF	SVM	14.0	0.193359	0.214844	False
RF	SVM Blue	0.0	0.001953	0.003255	True
RF	MLE	19.0	0.431641	0.431641	False
RF	Static Blue	0.0	0.001953	0.003255	True
SVM	SVM Blue	0.0	0.001953	0.003255	True
SVM	MLE	9.0	0.121094	0.151367	False
SVM	Static Blue	0.0	0.001953	0.003255	True
MLE	SVM Blue	0.0	0.001953	0.003255	Tue
MLE	Static Blue	0.0	0.001953	0.003255	True
SVM Blue	Static Blue	4.0	0.013672	0.019531	False

***Table S-5** - Pairwise Wilcoxon tests on F1 score with FDR correction*

Classifier 1	Classifier 2	Wilcoxon statistic	raw p-value	FDR-corrected p-value	Significant at alpha=0.05
RF	SVM	11.0	0.105469	0.117188	False
RF	SVM Blue	0.0	0.001953	0.003255	True
RF	MLE	9.0	0.064453	0.080566	False
RF	Static Blue	0.0	0.001953	0.003255	True
SVM	SVM Blue	0.0	0.001953	0.003255	True
SVM	MLE	12.0	0.250000	0.250000	False
SVM	Static Blue	0.0	0.001953	0.003255	True
MLE	SVM Blue	0.0	0.001953	0.003255	True
MLE	Static Blue	0.0	0.001953	0.003255	True
SVM Blue	Static Blue	9.0	0.064453	0.080566	False

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. These maps can summarize the intend of maps presented at the Figure 11 in the original manuscript more effectively.

The following text has been added to the Method section (see section 2.3.3):

"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."*

The following text has been added to the Results section (see section 3.1):

"*The disagreement maps suggested that the presence of shadows leads to the highest disagreement among models (Di > 0.4). One example of an image affected by shadows was acquired on 23 December 2020 at the TF study site. The disagreement map of this date (Fig. 5), where approximately 20% of the area exhibited disagreement values greater than 0.4, indicates 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.*

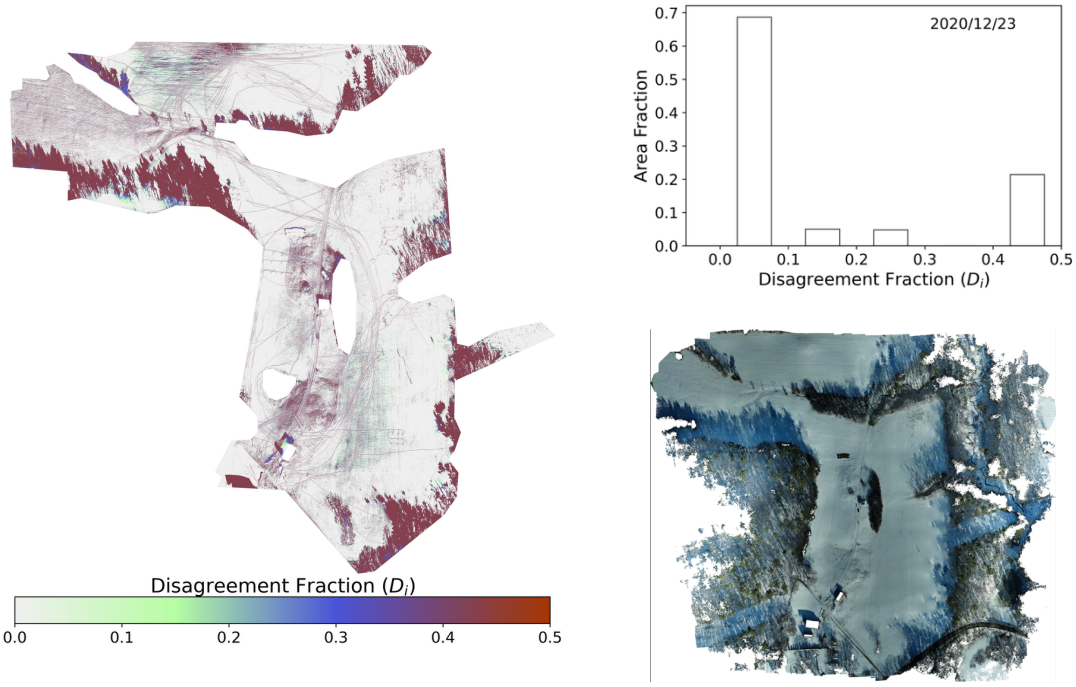


Figure 5. *Spatial distribution of disagreement fraction (Di), and corresponding histogram of disagreement fraction for the TF study site on 2020/12/23. The disagreement fraction represents the proportion of snow classification methods that disagreed with the majority-vote classification at each pixel*

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 is shown in Fig. 6. 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. Figure 7 illustrates one such example."

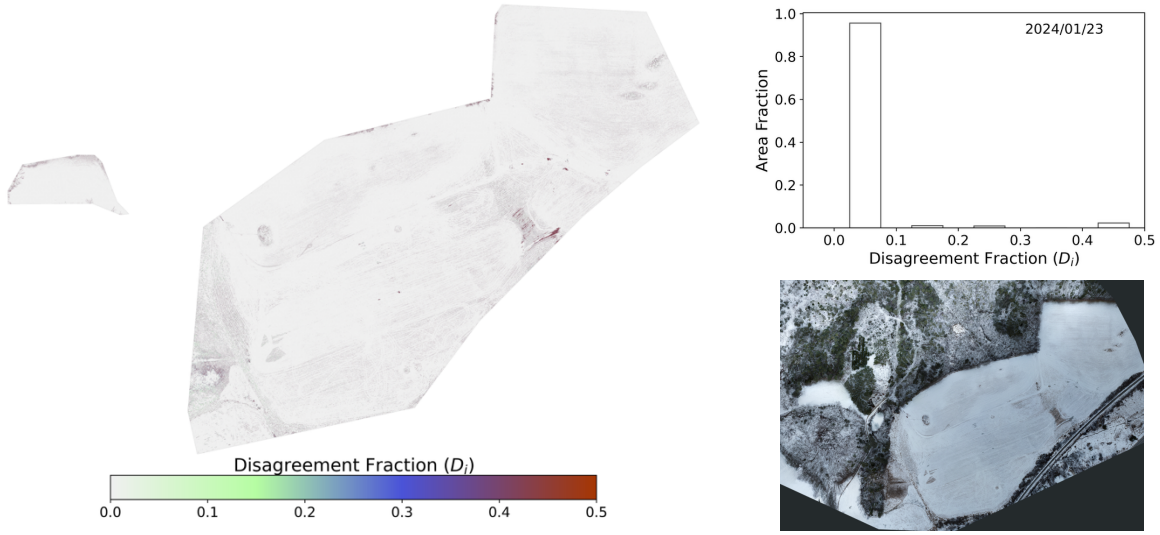


***Figure 6**. Spatial distribution of disagreement fraction (Di), and corresponding histogram of disagreement fraction for the KF study site on 2024/01/23*

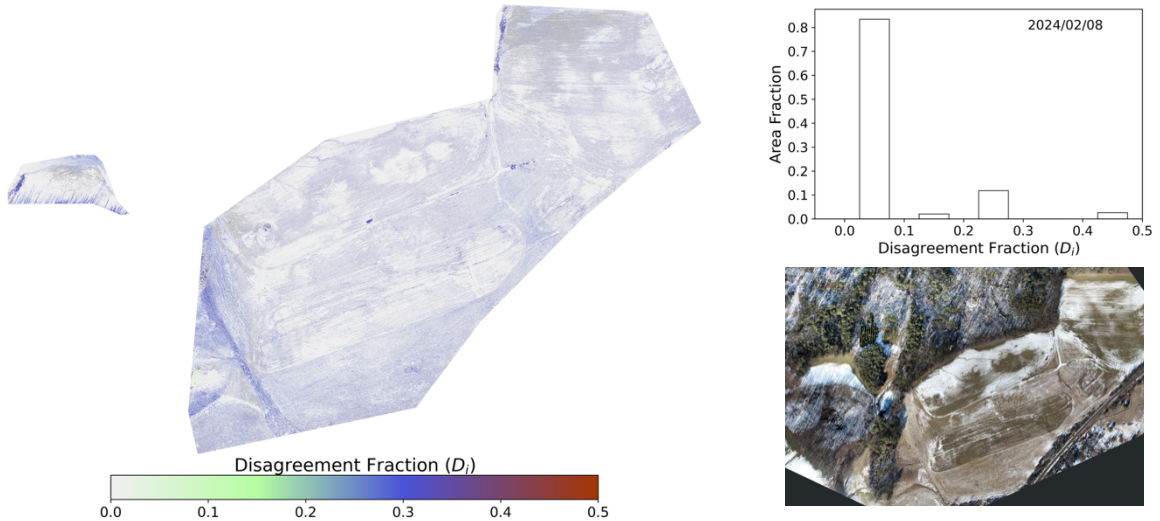


Figure 7. Spatial distribution of disagreement fraction (Di), and corresponding histogram of disagreement fraction for the KF study site on 2024/02/08

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.

		Field Site		Sensor		
Winter	Total RGB Images	TF	KF	P4	A5100	Lidar
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 (check Sections 2.2.2, 2.4, 3.3 and 4.3) 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:

The following text has been added to introduction:

"*The other widely used approach to generate SCA maps is static thresholding lidar-derived snow depth. The choice of snow-depth thresholds, however, 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 (Helbig et al., 2021, Johnston et al., 2025). Other studies adopt more conservative thresholds, commonly on the order of 3–10 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). In applications where airborne lidar snow depth products (e.g., ASO datasets) are used, a 10-15 cm threshold has frequently been applied to derive binary snow masks (Kostadinov et al., 2019; Cannistra et al., 2021; Thaler et al., 2023).*"

The following text has been added to the Method Section (Section 2.2.2 and 2.4):

"2.2.2 UAS Lidar

Lidar UAS surveys were collected using the Phoenix lidar Systems miniRANGER-3 Lite, a lightweight UAV-compatible mapping payload built around a RIEGL miniVUX-3UAV laser scanner and integrated Phoenix Air NavBox GNSS/INS navigation system. The system has a RMSEz system accuracy of 2–3.5 cm at 75 m flight altitude. The scanner operates at a 905 nm laser wavelength, supports up to five returns per pulse, and has a pulse repetition rate of 300 KHz. The integrated navigation system supports GPS, GLONASS, BeiDou, and Galileo constellations and uses a dual-frequency GNSS receiver coupled with a NovAtel OEM7 INS to estimate the sensor trajectory and orientation during flight. The INS has a reported horizontal position accuracy of 1 cm + 1 ppm RMS and IMU gyro in-run bias stability options ranging from 1.0°/hr to 0.5°/hr. Raw lidar and navigation data were processed in Phoenix SpatialExplorer version 8.1.1 to generate georeferenced point clouds, apply trajectory corrections, inspect data quality, filter noise/outliers, and produce refined survey products. Ground returns were classified using SpatialExplorer's ground-classification tools, which use a progressive TIN densification approach. The resulting ground-classified point clouds were then exported and used to generate bare-earth elevation model products for subsequent terrain analysis. Table 1 summarizes the number of UAS lidar surveys that coincided with optical observations used in this study. Nine of the 13 lidar surveys coincided with UAS optical imagery acquired using a DJI Phantom 4, while the remaining four (winter 2024-2025) coincided with optical imagery acquired using a GreenValley LiAir V70 system flown on a DJI M300 platform (see Supplementary material for more details)."

"2.4 Lidar-based Snow Cover Mapping

To derive binary SCA from UAS LiDAR snow depth, 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 2cm, and classified pixels with snow depth exceeding the threshold as snow-covered. This sensitivity analysis was designed to evaluate the influence of threshold choice on derived fSCA, particularly under shallow and patchy snow conditions where LiDAR depth uncertainty can be comparable to observed snow depths."

The following text has been added to the Result section (Section 3.3):

"*Figure 11 compares fSCA estimated from LiDAR snow depth thresholding with mean fSCA derived from optical observations and the MLE, RF and SVM classifiers. 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 Fig. 11, 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 2021/02/26, 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 2021/03/03 and 2021/03/07, 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."

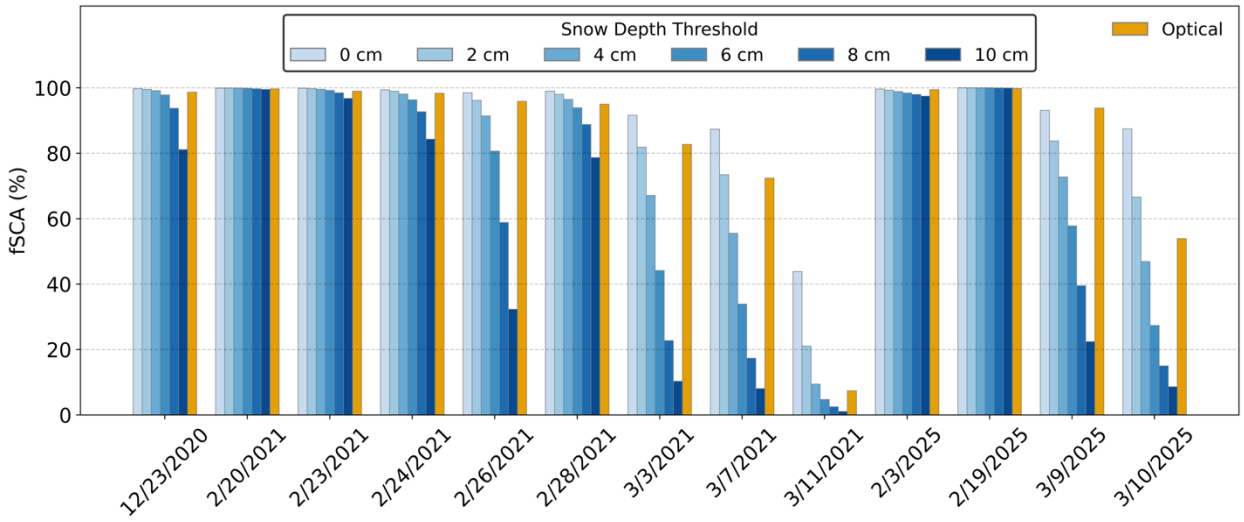


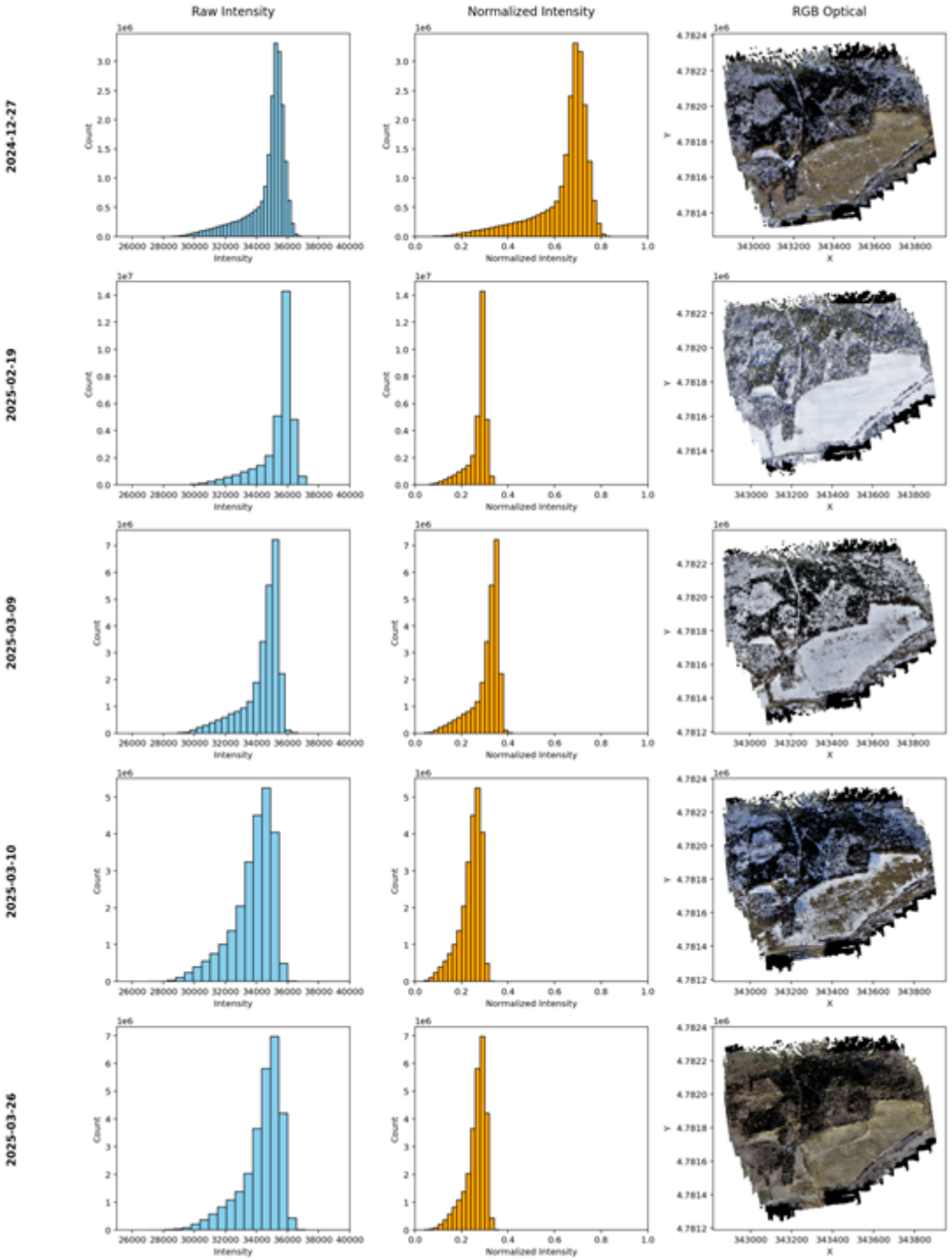
Figure 11. Fractional snow-covered area (fSCA) derived from optical imagery and lidar snow-depth maps using snow-depth thresholds of 0–10 cm.

The following text has been added to the Discussion section (Section 4.3):
"*Our analysis of SCA mapping using lidar data highlighted that in ephemeral and transitional snow regimes, where shallow but spatially continuous snow cover frequently occurs, small changes in the minimum depth threshold can therefore produce large differences in estimated snow-covered area under such conditions. This sensitivity is especially important when snow depths approach the measurement uncertainty of the lidar product (~2 cm in our case (Harder et al., 2020; Jacobs et al., 2021)), as residual errors may be incorrectly classified as snow and lead to overestimation of SCA. Consequently, threshold selection should be guided by the uncertainty characteristics of the snow-depth product, spatial resolution, and the prevailing snow regime, with particular care in environments characterized by shallow, discontinuous, or rapidly evolving snow cover*."

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/m^2]	Stdev SW_{in} [W/m^2]	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 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 previous Figure 11 were removed as the disagreement maps aggregate the findings from those maps.

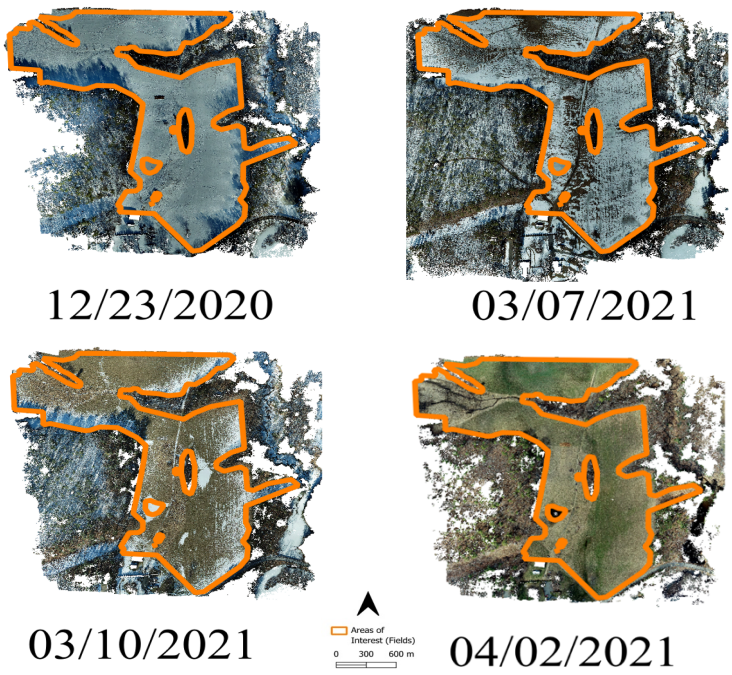


Figure 2. Example RGB orthomosaics of Thompson Farm acquired with the Phantom 4. The open-field study areas are outlined in orange

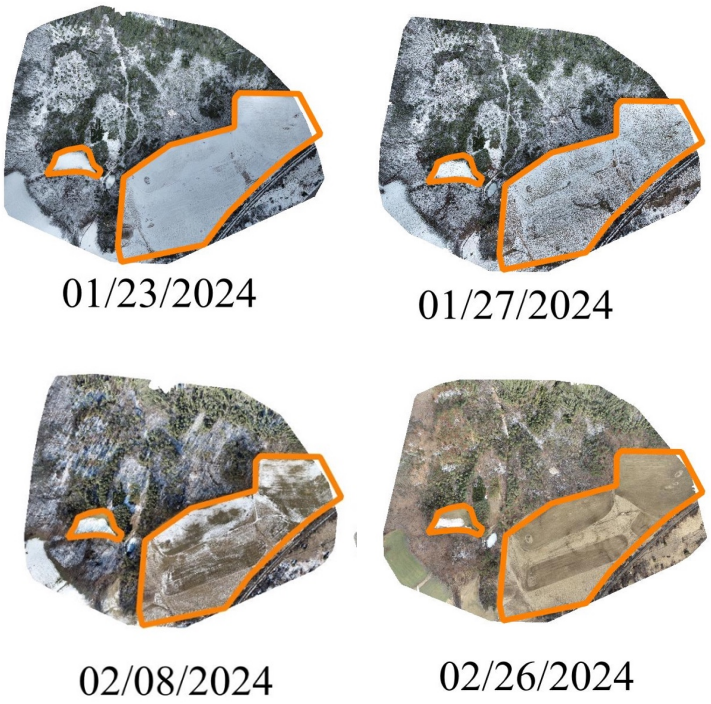


Figure 3. Example RGB orthomosaics of Kingman Farm acquired with the Phantom 4. The open-field study areas are outlined in orange.

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 S-4 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.

Review of “Identifying Snow-Covered Areas from Unoccupied Aerial Systems (UAS)

Visible Imagery: A Comparison of Methods” (Moradi et al. 2025)

General Comments

In their study, the authors present a method of mapping snow cover using drones with different sensor systems and land surface classification algorithms. The study's primary objective is to conduct an inter-comparison experiment to evaluate the performance of various snow classification methods in relation to different acquiring sensors, study sites, and training data availability. This is an interesting setup because it addresses common challenges associated with drone surveys and aims to provide recommendations on the most suitable methods for specific acquisition conditions or user applications. The manuscript is well written and mostly well structured. The scientific objects are within the scope of The Cryosphere.

Thank you for the positive words.

However, I think the analysis of the differences between the methods is a bit shallow. On the one hand, the experimental setup is relatively simplistic (i.e. binary classification of “snow” vs. “no- snow”), but on the other, it may be flawed by some uncompensated differences in data acquisition (image dates), as outlined in my general comments below. Therefore, I believe that major revisions are necessary before this study can be published. Nevertheless, expanding the experimental setup could provide a much deeper understanding of the differences between the methods and their underlying causes.

Overall Response

We appreciate the constructive feedback from both reviewers. Reviewer 1 and Reviewer 2 were generally supportive of the study concept but noted that the experimental design and data acquisition were not originally intended to support robust cross-sensor or cross-site comparisons. Both reviewers also pointed out opportunities to extend the analysis, including examining performance in forested environments (Reviewer 2) and incorporating lidar-based estimates of snow-covered area (SCA).

We agree with these assessments and have revised the manuscript accordingly. First, we removed the cross-sensor and cross-site comparisons from the manuscript to ensure that the scope of the analysis aligns with the original experimental design. Second, we added a new analysis in the Results section that uses UAS lidar observations to estimate SCA, providing an additional and independent perspective on snow detection. Finally, we expanded the Discussion to address SCA estimation in forested environments. This section explores both the potential transferability of the RF approach developed in open areas and the limitations of optical imagery in forested terrain.

We believe these revisions better align the manuscript with the strengths of the dataset while also acknowledging the challenges and opportunities associated with extending these methods to more complex environments. Our intent is to maintain the original goal of presenting results under controlled and straightforward conditions while clearly articulating limitations, challenges, and next steps for applications in more complex forested landscapes.

• Experiment images: The dataset used is certainly interesting, covering a wide range of snow cover conditions. However, a significant limitation of the cross-sensor and cross-site comparison is that the images were not acquired on the same day. It is likely that there is some kind of date-related bias between the images because the properties of the snow, the illumination and the shadows are likely to be different. Ideally, images from different sensors at the same site would have been acquired on the same day, immediately after each other. As far as I can see, these potential sources of difference are not currently accounted for in the experiments. Were the images acquired at the same time of day? What about weather conditions (e.g. cloudiness)? To what extent are the observed cross-sensor and cross-site differences actually related to acquisition dates?

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.

To provide additional context for the readers, a table summarizing the weather conditions during image acquisition has been added to the supplementary material

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/m^2]	Stdev SW_{in} [W/m^2]	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 sites: Similarly to the selection of images, I am not entirely convinced by the study area to which the image mosaics are masked (Figure 1). It looks like you have removed all areas except for the almost flat fields with sparse vegetation. While this certainly improves the separation of snow and bare ground, it means that you cannot draw any conclusions about the performance of the applied algorithms in more complex terrain, such as slopes and forests. Conversely, I suspect that the overall good performance of your classifiers (almost all approaches yield reasonable results) is due to the exclusion of challenging areas. This is made even more apparent by the fact that, in your manuscript, you mention that satellite snow cover struggles in complex terrain or forest areas compared to drone surveys (Line 59), yet your own dataset is restricted to areas where a satellite image would also most likely produce good classification results (at a coarser resolution of course). I think this is a missed opportunity, as your analysis could be much more informative if you assessed the accuracy of different methods in different landscapes.

Response: The reviewer's point is well taken; we agree that evaluating SCA mapping methods in more complex environments such as forests and sloped terrain would provide valuable insight. The overarching goal of this work was to start with the most straightforward conditions (fields, flat terrain with sparse vegetation) and establish a baseline comparison of commonly used SCA mapping approaches under controlled conditions at shallow and intermittent snow regimes and then conduct future work to move more challenging conditions. This would allow us to isolate differences among the tested classification methods without the additional confounding effects introduced by complex terrain, forest canopy, or steep slopes. We advocate for looking at the simple case first because even under these relatively homogenous surface conditions, not all tested methods produced high accuracy or provided similar performances. Evaluating the methods in a controlled setting therefore provides a necessary baseline for understanding their behavior before extending the analysis to more complex environments.

At the same time, we recognize that the dataset provides an opportunity to examine UAS-based SCA detection in forested environments, and we agree that offering some insight into these conditions would add value for readers. To address the reviewer's comment in more detail, we have expanded the Discussion to include results related to SCA mapping in forested areas. These additions highlight that extending the analysis to forested and more complex terrain represents an important direction for future work and outline the challenges associated with relying solely on optical imagery in such environments.

*"**4.4 SCA Mapping in Forest Gaps via UAS Imagery***

Although this study focuses on open-field environments to enable a controlled comparison of SCA mapping approaches, extending the analysis to forested environments remains an important direction for future work. To explore this possibility, we masked the forest canopy and applied the RF, SVM, and MLE models developed in the open-field analysis to canopy gaps within the forested portions of the Kingman Farm study site orthomosaics (see Supplementary Material for methodological details). When applied to canopy gaps, the models' accuracy, balanced accuracy, and F1 scores were comparable to those obtained in the open areas, while Cohen's kappa values decreased modestly from ~0.90 in open terrain to ~0.80 within forest openings (Figure 12).

Although the models perform well in canopy gaps, reliable SCA mapping in forested environments depends on the availability of a consistent and accurate canopy masking approach. This is particularly challenging in landscapes dominated by deciduous forests. Using a k-medoids clustering algorithm, we generated canopy masks when the forest floor was snow-covered; however, we encountered several limitations that are likely common to most unsupervised classification approaches. In particular, these methods cannot reliably distinguish branches lying on the ground from branches that remain elevated within the canopy. In deciduous forests, branch structures are sufficiently fine that modest shifts in canopy position due to wind or snow loading, as well as minor alignment differences between UAS

orthomosaics, prevent masks generated for one date from being reliably applied to other dates. These issues highlight an important limitation of relying solely on optical imagery— the focus of this study—for SCA mapping in forested environments.

Additional challenges arise from branches and litter present on the snow surface, which can reduce classification reliability and complicate identification of true canopy gaps where snow accumulates on the ground. These effects are particularly pronounced in deciduous forests under leaf-off conditions. Furthermore, snow intercepted and retained on branches cannot be easily distinguished from snow on the ground using optical imagery alone. Together, these limitations highlight the challenges of relying on optical data for snow mapping in forested landscapes and motivate the exploration of existing satellite fSCA and NDSI methods which use canopy cover information and integration of complementary observations, such as LiDAR, which can provide additional information about sub-canopy snow distribution."

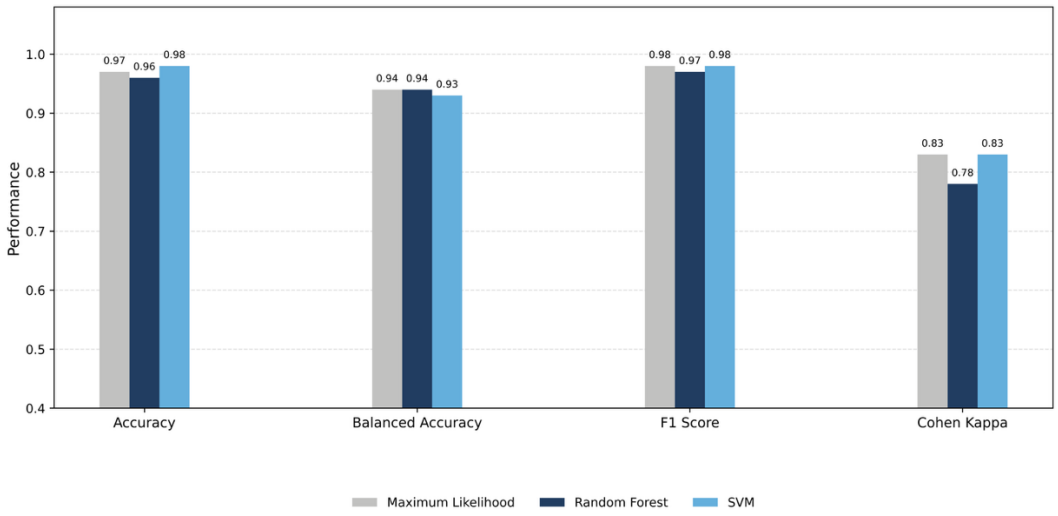
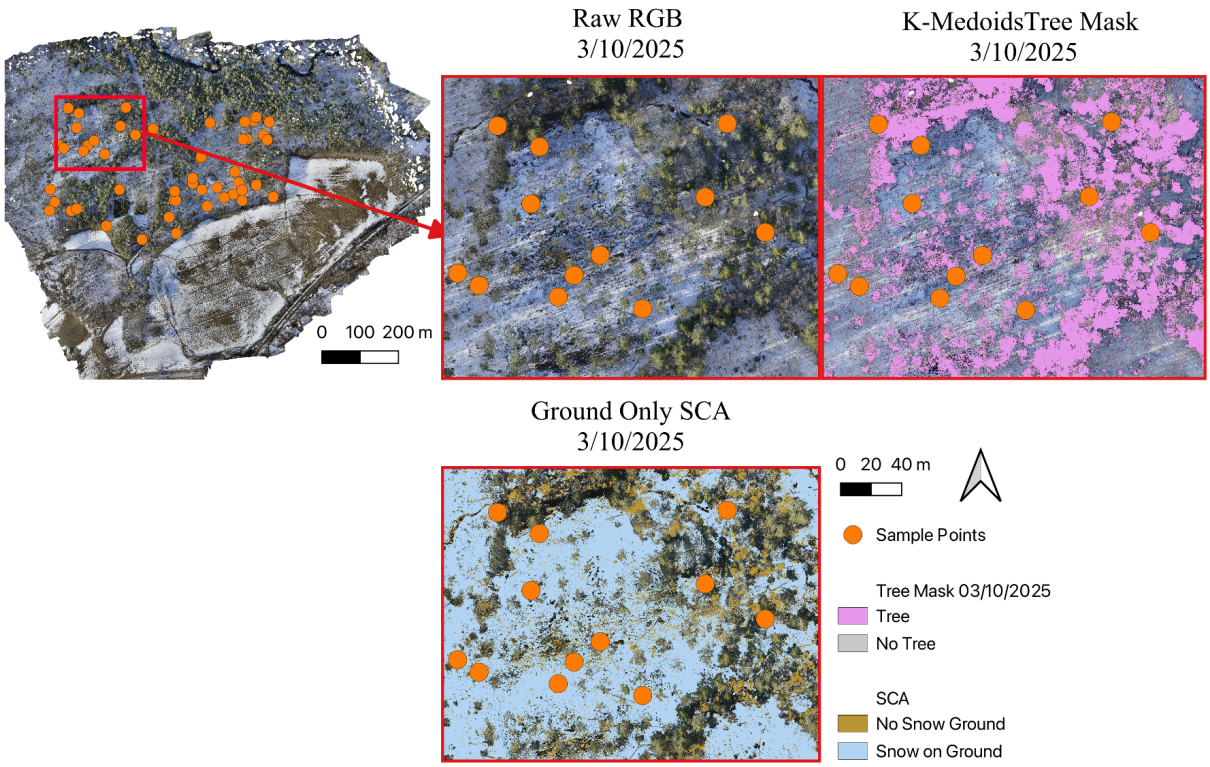


Figure 12 *- Performance metrics of the three best-performing models from the reference SCA analysis, evaluated in forested area canopy gaps using 11 orthomosaics collected at the Kingman Farm study site.*

Also, additional details describing our workflow for estimating SCA in forest gaps have been added to the supplementary material.

"In this study, we mapped snow cover in forested environments by first identifying and masking tree canopies. Tree canopy masks were generated using an unsupervised clustering approach applied to the orthomosaic imagery. A representative subset of pixels was randomly sampled from each image to capture the range of surface conditions present in the scene. The sampled pixel values were transformed into the CIELAB color space and reduced using principal component analysis (PCA) to improve feature separability and reduce dimensionality. K-medoids clustering was then applied to classify pixels into multiple clusters representing different surface types. The cluster corresponding to tree canopy was identified through visual inspection and used to generate a binary mask that removed forested areas from each orthomosaic. After masking the canopy, 50 validation points were randomly sampled from the remaining gaps between tree canopies and labeled as snow or no snow. The models trained in the open-field experiments presented in this manuscript were then applied to estimate snow cover at these validation points, and their performance was evaluated. An example of the classification of tree canopy and subsequent SCA classification of the ground is shown in the following figure."



The reviewer is also correct that an earlier version of the manuscript included discussion in the Introduction related to forest environments that was not directly supported by the analyses presented. Because the present study does not explicitly target forested terrain, we have revised the Introduction to remove discussion of forest- and terrain-specific challenges associated with satellite SCA products. Instead, the Introduction now focuses on the conditions directly addressed in this study, namely shallow and spatially discontinuous snow cover.

"Additionally, UAS-based snow mapping may be less affected than coarse-resolution satellite snow products in shallow or patchy snow conditions because centimeter-scale imagery reduces mixed-pixel effects and allows better detection of fine-scale snow heterogeneity, whereas satellite products can exhibit increased omission errors when snow cover is thin or spatially discontinuous"

Data availability: Your data availability statement only refers to meteorological data which is available elsewhere. What about your survey data? The raw images and your intermediate classification results should also be made available via an open data repository.

Response:
Thank you for the comment. This is an extensive dataset that we would like to make available to the community. We have requested approval from our funding agency (DoD) to make the dataset publicly available. Unfortunately, the release of the dataset must remain pending until this approval is granted.

Specific Comments

Line 50-52: Unoccupied aerial systems (UAS) are emerging as a reliable, low-cost method for collecting snow data at finer spatial resolution than satellites and are not limited by orbital revisit times.“ => While it is true that drones provide much better resolution than non-commercial satellite images, the logistical effort required depends on the use case. For example, the surface reflectance products of the Landsat/Sentinel missions have a revisit time of a few days. They are distributed free of charge by the ESA and NASA, and they include infrared channels, which are great for mapping snow. Conversely, drones require an operator in the field, surveys of large areas take time, and they are limited by weather conditions. So, all in all, there are specific applications where one system is superior to the other. If you want to compare satellites and drones in the introduction (which makes sense), you should briefly summaries the advantages and disadvantages of drones and satellites, and which data is useful for which application.

Response:
Thank you for this suggestion. We agree that the relative advantages of UAS and satellite imagery depend strongly on the application and that UAS should not be presented as universally “low-effort” or superior. We revised the Introduction to provide a more clarity on specific applications of UAS as follows:

“*Satellite products (e.g., Landsat and Sentinel) provide free, repeatable multispectral coverage suitable for regional snow monitoring, but their coarser spatial resolution and susceptibility to cloud cover and mixed-pixel effects can limit their application. In contrast, unoccupied aerial systems (UAS) offer much finer spatial resolution and flexible, on-demand data acquisition, enabling improved monitoring of snow in heterogeneous environments. These capabilities make UAS particularly well suited for capturing highly dynamic snow conditions, such as ephemeral or transitional snow cover. However, their application may be constrained by weather conditions and demands from operational logistics (Sturm, 2015; Verfaillie et al., 2023).*”

Line 57-60: “… does not appear impacted to the same extent as satellite snow cover products in forested areas, complex terrain and shallow snow conditions …” => add brief explanation why there is less impact

Response:
Thank you for the suggestion. We have added a brief explanation clarifying why UAS-based snow mapping may be less impacted than coarse-resolution satellite snow products in heterogenous and shallow snow conditions. The primary reason is that satellite snow mapping errors often arise from (i) mixed pixels at coarse spatial resolution, (ii) reduced detectability when snow is thin or sparse. For example, Hall and Riggs (2007) note that omission errors are greatest when only trace amounts of snow are present and that snow/cloud confusion (including cloud-shadowed land) contributes to errors under thin/sparse snow conditions. In contrast, UAS imagery is acquired at centimeter-scale resolution and can therefore better resolve fine-scale snow heterogeneity (reducing mixed-pixel ambiguity). We revised the text accordingly to reflect these mechanisms more explicitly.

Since the present study does not target forested environments, forest-specific challenges associated with satellite SCA mapping were removed from the text.

“Additionally, UAS-based snow mapping may be less affected than coarse-resolution satellite snow products in shallow or patchy snow conditions because centimeter-scale imagery reduces mixed-pixel effects and allows better detection of fine-scale snow heterogeneity, whereas satellite products can exhibit increased omission errors when snow cover is thin or spatially discontinuous (Ault et al., 2006; Hall and Riggs, 2007; Gao et al., 2010; Frei et al., 2012).”

Line 164-192: You provide a lot of technical details for each drone/sensor system. Instead, why not add a table that summarises the technical differences between the systems, such as sensor specifications, camera settings and flight line parameters? You could then focus the text on the key differences between the systems. However, if you decide to stick with the current version, that is fine too.

Response:

Because the cross-sensor experiment was removed from the study, the associated descriptions of the UAS platforms and sensors were also removed from the Methodology section.

• Line 176 & 189: Please specify which DEM was used, and whether the same DEM was used to process each dataset

Response:
The text has been updated for clarity as follows:
"A subset of the global Shuttle Radar Topography Mission (SRTM) 1 arc-second digital elevation model (DEM) was uploaded and referenced to ensure a consistent flight altitude was maintained throughout each flight."

• Line 208-209: "… most images collected during complete snow cover could not be effectively stitched together and were excluded from this study." => I would imagine that it is not a significant issue for the purposes of your study, as separating snow from bare ground in images with continuous, thick snow cover is likely easier than in cases with thin, fractional snow cover (as you demonstrate later). Nevertheless, for an actual drone mapping campaign to monitor snow cover evolution, this would present a significant challenge, as the creation of orthomosaics would fail for certain acquisition dates. At the very least, I would mention in the discussion that your test dataset is not representative of complete snow cover applications.

Response:
We appreciate the reviewer highlighting this point. The difficulty in generating orthomosaics during periods of complete snow cover arises from the lack of stable visual features required by structure-from-motion algorithms to reliably match tie points between images. As the reviewer notes, this issue is unlikely to affect the core objective of the present study, which focuses on classification performance under shallow and spatially heterogeneous snow conditions where distinguishing snow from bare ground is more challenging.

We agree that this limitation is important to acknowledge in the broader context of drone-based snow monitoring. To address this comment, we added text in the Discussion noting that the dataset used in this study is not representative of conditions with continuous, thick snow cover and that orthomosaic generation can become unreliable in such cases due to the absence of persistent surface features with citations to the structure-from-motion snow literature. We also note that this limitation primarily affects image reconstruction rather than snow classification itself and that future drone mapping campaigns targeting continuous snow cover may require alternative acquisition strategies (e.g., inclusion of ground control features, lower flight altitudes, or complementary sensors).

"*Second, it should be noted that SCA mapping using UAS RGB imagery faces limitations under continuous, deep snow cover conditions. Under such conditions, orthomosaic generation from visible imagery can become challenging because structure-from-motion algorithms rely on identifying stable visual features to establish tie points between overlapping images. Uniform snow surfaces often lack*

sufficient texture and contrast, which can reduce reconstruction quality and, in some cases, prevent reliable orthomosaic generation. This limitation primarily affects image reconstruction rather than the snow-classification methods evaluated here. Future drone-based snow-mapping campaigns targeting continuous snow cover may therefore require alternative acquisition strategies, such as the inclusion of artificial ground control features, lower flight altitudes, or the integration of complementary sensing technologies."

• Line 209: What approximate snow depths do you mean by "uniform snowpack"?

Response:
Good point, uniform implies spatially consistent. For clarity, the manuscript is updated as follows:
"*Of the 26 dates, 10 represented very shallow/patchy snow conditions with observed snow depths less than 4 cm, 4 were completely snow free, and the remaining 22 dates had snowpacks with snow depths ranging from ~4 to ~45 cm within the open fields. For the other winters, the average snow depths were ~10 cm*".

• Line 220-222: "Field cameras were installed following the method used in NASA's 2020 SnowEx field camera campaign … " => You refer to personal communication here, but do not explain what was actually done. Could you please provide a description of the approach that was taken?

Response:

The appropriate citation is added to the manuscript along with additional details:

"*On dates when lidar was not flown, snow depth observations were acquired from either automated ultrasonic snow depth sensors (Judd Communications LLC) or from Moultrie Wingscapes Birdcam Pro Field Camera time lapse images of snow stakes (i.e., PVC poles spray-painted red and marked with 1 cm and 10 cm increments), both deployed in the open field at the respective site and sampled every 15-minutes. Field cameras were installed following the approach used in NASA's 2020 SnowEx field camera campaign in Grand Mesa, Colorado (see Proulx et al., 2022 for details).*"

The methodological details provided in the cited manuscript (Proulx et al., 2022) are as follows:
""Moultrie Wingscapes Birdcam Pro Field Cameras were used to capture images of the snowpack relative to a 1.5 meter marked PVC pole following the method used in NASA's 2020 SnowEx field camera campaign in Grand Mesa, CO (personal communication, 16 November 2020). Three cameras were used; one was in the open field, one was in the coniferous forest, and one was in the deciduous forest (Fig. 1). Each camera was mounted approximately 0.85 m above the ground and placed approximately 5.5 m from its respective PVC pole. Each camera's field of view included the entirety of the PVC pole, some of the ground surface below the pole, and some open area above the pole. Each PVC pole was spray-painted red and was marked with 1 cm and 10 cm increments. The cameras captured images of the poles every 15-minutes for the duration of the study period. Snow depth was derived by manual inspection of the photos and recorded to the nearest cm."

Proulx, H., Jacobs, J.M., Burakowski, E.A., Cho, E., Hunsaker, A.G., Sullivan, F.B., Palace, M. and Wagner, C., 2022. Comparison of in-situ snow depth measurements and impacts on validation of unpiloted aerial system Lidar over a mixed-use temperate forest landscape. The Cryosphere Discussions, 2022, pp.1-20.

• Figure 2 & 3 & 4: The images are relatively small, and it is not always easy to tell the difference between those with similar snow conditions. I suggest moving these figures to the supplementary material. In the main text, you could show a figure with fewer, larger panels instead. For example, select one or two images with distinctly different snow conditions for each site and each camera. This would allow readers to compare the differences between sites and cameras with and without a large snow cover fraction.

Response:
Figures have been revised to display a smaller set of representative images at a larger size. The original versions of these figures have been moved to the supplementary material.

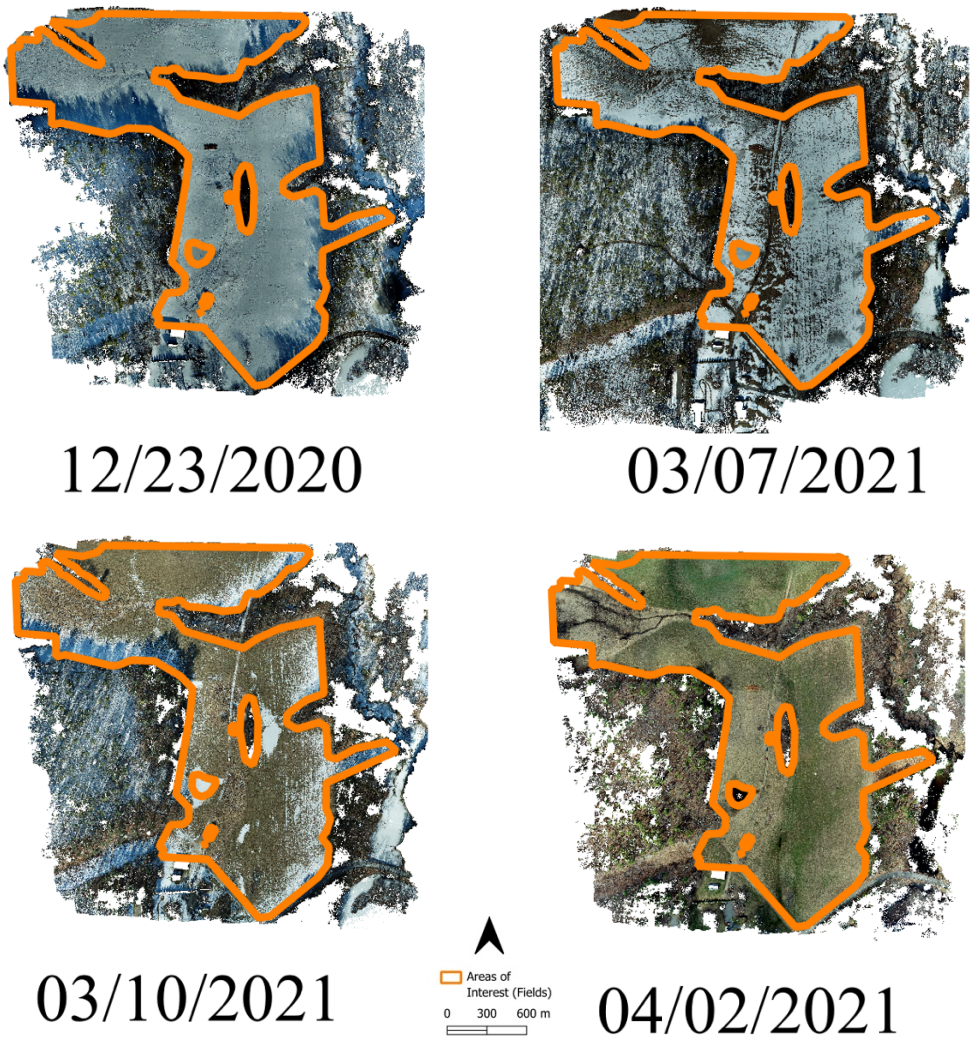


Figure 2. Example RGB orthomosaics of Thompson Farm acquired with the Phantom 4. Survey dates are shown below each orthomosaic. Areas of interest are outlined in orange.

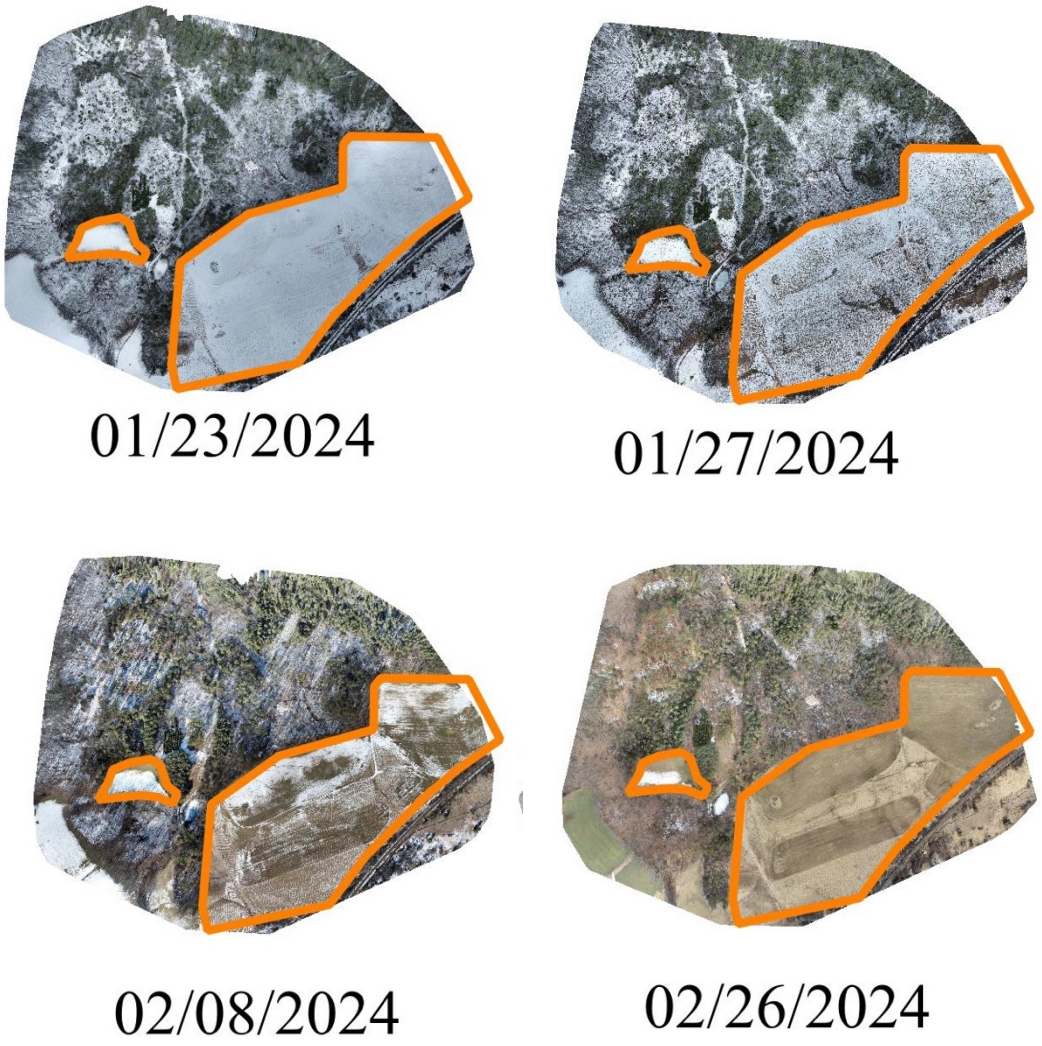


Figure 3. Example RGB orthomosaics of Kingman Farm acquired with the Phantom 4. Survey dates are shown below each orthomosaic. Areas of interest are outlined in orange.

Line 250: "… the field area of interest process was isolated … " => What do you mean?

Response:

The text has been revised for clarity:

"*For each orthomosaic, site-specific masks were applied to exclude surrounding forested areas, leaving only the field area for analysis*."

Line 251: "… forested areas was removed …" => see my general comment about selection of areas

Line 287-289: It is tedious for readers to have to look up the threshold value in the supplement. If the value is fixed, just mention it here and refer to the supplement for details.

Response:
A list of the tested thresholds has been added to the manuscript to improve readability:
"*The optimal threshold selection was guided by 5-fold cross-validation on a range of candidate thresholds (95, 105, 115, 127, 135, 145), using balanced accuracy (i.e., the average accuracy of both the positive and negative classes) as the evaluation criterion*"

Line 302: The results section is divided into subsections for each experiment. I suggest doing the same here, creating subsections for the 'base experiment', 'second experiment', and so on.

Response:
Subsections were added with a new naming convention (see the next comment).

Line 302: To improve readability, you could replace "base experiment", "second experiment", "third experiment", "fourth experiment" with a more (mathematically) name convention, e.g. SCA reference

Response:
Thank you for this suggestion. We changed the two experiments names to "*Baseline SCA*" and "*Generalization Performance Under Limited Training Data*" both in the methods and results sections.

Line 404: "… is rooted in differences in snow conditions." => see my general remark about image dates and conditions

Response:
To address this concern, the cross-sensor experiment has been removed from the manuscript.

Line 440: In the conclusion you state that 12 images are the minimum requirement for a good training set (based on Figure 8?). Please elaborate why you conclude that 12 images is the threshold.

Response:
Twelve images were adopted as the minimum sample size in this study (Figures 8 and S1). Although the SVM classifier was generally more robust, maintaining good performance with as few as 10 images, the performance of the MLE and RF classifiers deteriorated across a larger number of image combinations when fewer than 12 images were used. In these cases, model behavior became less stable, as evidenced by an increased number of outliers with classification accuracy below 80%.

Line 521: "… identify one or more classifiers that can rapidly …" => Remove "rapidly" because you do not analyze processing speed.

Response:

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

Line 685-688: How do your results validate “satellite snow cover products, converting satellite NDSI observations of fSCA, and downscaling satellite SCA observations”? Explain this or delete.

Response:
Previous studies have used higher-resolution satellite snow maps (e.g., Landsat-derived snow cover or fSCA) as reference data to validate coarser-resolution snow products and to diagnose mixed-pixel behavior under patchy snow conditions. For example, Salomonson and Appel (2004) used Landsat ETM+ as higher-resolution reference information to derive the relationship between MODIS NDSI and fractional snow cover within 500-m pixels. Similarly, Rittger et al. (2013) evaluated MODIS snow-cover mapping methods using Landsat-based snow maps as reference. Additional work has validated MODIS fractional snow cover retrievals using temporally and spatially coincident Landsat scenes across multiple regions.

UAS-based optical imagery provides centimeter-to-decimeter spatial resolution, substantially finer than publicly available satellite imagery (e.g., Sentinel-2 at ~10 m) and even typical commercial optical imagery (~1–3 m). As a result, UAS snow-covered area (SCA) maps can serve as a more detailed reference dataset for validation and mixed-pixel analyses, particularly in landscapes where snow cover is highly heterogeneous and prone to patchiness. To date, UAS SCA maps have been less commonly used for satellite snow-cover downscaling than higher-resolution satellite products, but examples do exist (Johnston et al., 2025). Instead, several studies have primarily used UAS-derived snow maps to support MODIS-scale (500 m) fractional snow cover (fSCA) estimation by training machine-learning models with UAS observations as reference data (Liang et al., 2017; Liu et al., 2020). Our study offers a guidance on effective approaches to map SCA using UAS observations that potentially can be used for the above-mentioned applications.

We added appropriate references for the mentioned applications.

“*It further offers guidance on effective strategies for producing UAS derived SCA maps that can later be used for a range of purposes such as validation of satellite snow cover products, converting satellite NDSI observations to fSCA, and downscaling satellite SCA observations (Salomonson and Appel, 2004; Rittger et al., 2013; Liang et al., 2017; Johnston et al., 2025)*.”

Line 705: Data availability => Drone images and classification masks should be made available to reproduce results. Technical Corrections

Response:
Thank you for the comment. This is an extensive dataset that we would like to make available to the community. We have requested approval from our funding agency (DoD) to make the dataset publicly available. Unfortunately, the release of the dataset must remain pending until this approval is granted

Line 62: Here you could add a reference to large-scale snow cover products based on NDSI (e.g. https://essd.copernicus.org/articles/11/493/2019/)

Response: We added the Gascoin et al. (2019) reference.

Gascoin, S., Grizonnet, M., Bouchet, M., Salgues, G. and Hagolle, O., 2019. Theia Snow collection: High-resolution operational snow cover maps from Sentinel-2 and Landsat-8 data. *Earth System Science Data*, *11*(2), pp.493-514.

Line 67: replace „blue-channel thresholding“ with „blue-channel“ reflectance (or similar)

Response:
The text has been revised as suggested:

"*Promising methods for differentiating between snow and non-snow in UAS RGB imagery include machine learning (ML) classifiers such as maximum likelihood, random forest, support vector machine, k-means clustering (Liang et al., 2017, Belmonte et al., 2021, Johnston et al., 2025; Niedzielski et al., 2018), and thresholding approaches such as static or dynamic thresholding of blue-channel reflectance*."

Line 159: There is no citation of MODIS data.

Response:
The appropriate reference has been added.

Line 226: missing reference NOAA's publicly available archive (ADD-REFERENCE)"

Response:
The appropriate reference has been added:

"*Subhourly air temperature and shortwave radiation data was obtained from the USCRN stations located at the study sites, through NOAA's publicly available archive (NOAA NCEI, 2025) Air temperature (°C) and incoming shortwave radiation (W/m²) from 2020 to 2025 at five-minute temporal resolution were used to calculate the mean and standard deviation during each survey*."

NOAA National Centers for Environmental Information (NCEI). U.S. Climate Reference Network (USCRN) Data. National Oceanic and Atmospheric Administration. https://www.ncei.noaa.gov/products/land-based-station/us-climate-reference-network; Accessed: April 01, 2025.

Figure 1: The panel on the bottom right appears to be a Google Earth background. I think there should be a reference to this somewhere.

Response:
The appropriate citation has been added:

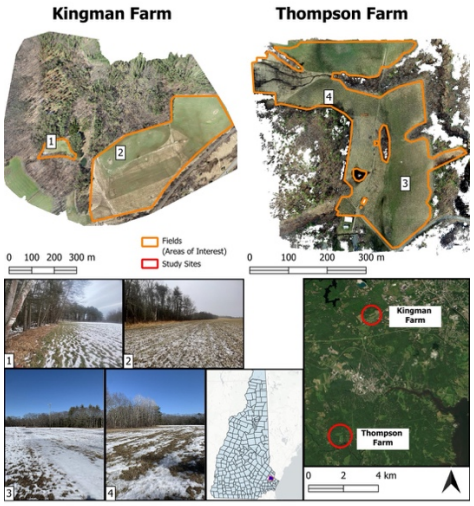


Figure 1. The two study sites, Kingman Farm and Thompson Farm, located at southern New Hampshire (basemap source: Esri, 2024).

Esri, 2024, . World Imagery [Basemap]. ArcGIS Online, https://services.arcgisonline.com/ArcGIS/rest/services/World_Imagery/MapServer

Figure 5 and other bar charts: Do not include both red and green bars in the same graph.

Response:
The figure is updated with new set of colors

