Identifying Snow-Covered Areas from Unoccupied Aerial Systems (UAS) Visible Imagery and Lidar: A Comparison of Methods

Mahsa Moradi[1,2], Andrew G. Fleming[1,2], Adam Hunsaker[1,2], Jennifer M. Jacobs[1,2]

[1]Department of Civil and Environmental Engineering, University of New Hampshire, Durham, NH, USA

[2]Earth Systems Research Center, Institute for the Study of Earth, Oceans, and Space, University of New Hampshire, Durham, NH, USA

Correspondence to: Mahsa Moradi (mm1631@usnh.edu),

Abstract

High-resolution imagery from Unoccupied aerial systems (UAS) offers new opportunities for mapping snow cover at fine spatial scales, particularly in regions with ephemeral and variable snow conditions. This study evaluates a range of classification strategies for generating snow-covered area (SCA) maps from UAS RGB imagery and lidar-derived snow depth collected over open areas in southern New Hampshire, USA and offers practical recommendations for producing UAS-derived SCA maps. We tested machine learning (ML), including Maximum Likelihood Estimation (MLE), Random Forest (RF), and Support Vector Machine (SVM), and threshold-based approaches, exploring the influence of input features and training set composition on classification accuracy and generalization. Results show that classifiers using full red-green-blue inputs consistently yield high performance (accuracy, balanced accuracy and F1 score above 0.96). In contrast, ML and thresholding approaches relying solely on the blue band exhibited lower balanced accuracy (0.83 to 0.86). Training with fewer than 12 orthomosaics reduced the reliability and consistency of snow cover classifiers. When fewer flights are possible, UAV flights that collectively capture the full range of snow-covered area (fSCA) between 20–60%, partial melt, and refrozen surfaces, should be prioritized. We conclude that highly accurate snow mapping from routine UAS optical imagery is possible even in landscapes with variable snow cover and ephemeral, shallow snowpacks. In such regions, lidar-derived estimates of snow-covered area were found to be highly sensitive to the selected snow-depth threshold where shallow snowpacks require low thresholds that may approach measurement uncertainty and increase uncertainty in mapped snow extent.

Keywords: Unoccupied aerial systems (UAS), Snow-covered area (SCA), Optical imagery, Machine Learning, Blue band thresholding, Lidar Snow Depth, Ephemeral Snow Cover

1 Introduction

The extent, duration, and patterns of snow-covered areas (SCA) and their accurate delineation provide insight into the winter hydrologic cycle including snow accumulation, redistribution and melt as indirect understanding of soil freeze-thaw state and evaporation (Pomeroy et al. 2007; Dong, 2018; Moradi et al., 2023) and, when summarized over time, characterize a region's snow climatology (Johnston et al., 2023). The states and process are important for characterizing snow melt water runoff (Immerzeel et al., 2009; Hammond et al., 2018; Le et al., 2023) as well as the exchange of heat and moisture between the land surface and atmosphere (Walland and Simmons, 1996; Liston, 1999; Zhang, 2005; Lundquist et al., 2024) and ecosystem controls (Liston, 1999; Baptist et al., 2010; Panchard et al., 2023).

Snow cover varies at multiple scales, and the appropriate scale depends on the application (Blosch, 1999). For many applications requiring continuous spatial coverage, satellite-scale observations are suitable (Dong, 2018; Johnston et al., 2023). For other applications requiring fine (<1 m) to moderate (10s of meters) spatial resolutions (e.g., Mott et al., 2017; Sproles et al., 2018; Harder et al., 2019; Barna et al., 2020), snow presence or absence may deviate from the coarse satellite scales and require invoking various physical features to downscale the satellite observations (Clark et al., 2011; Sturm, 2015).

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). Platforms outfitted with electro-optical (EO) cameras, which are becoming increasingly available, can provide detailed spatial information on demand in cold regions and provide numerous research opportunities (Gaffey and Bhardwaj, 2020; Verfaillie et al., 2023). While most snow cover mapping strategies have used either (1) a combination of the visible and infrared, or (2) microwave, there is mounting evidence that multispectral imagery (i.e., red-green-blue (RGB) alone or RGB and NIR) can provide reasonable results. 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).

In contrast to satellite approaches which typically use a Normalized Difference Snow Index method (NDSI) (Dozier, 1989; Gascoin et al., 2019) to map snow cover using green and SWIR bands, there is no single recommended strategy for mapping SCA using UAS RGB imagery. Promising methods for differentiating between snow and non-snow in UAS RGB imagery include machine learning (ML) classifiers such as Maximum Likelihood (MLE), Random Forest (RF), Support Vector Machine (SVM), 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 (Jacques-Hamilton et al., 2025; Eker et al., 2019). Each of these methods appears to be capable of mapping SCA with different strengths and weaknesses. Multiple studies have used K-means clustering to distinguish between bare ground

and snow, as well as ice and shadows, but the method requires manual, qualitative interpretation of the classes making it challenging to apply for multiple images (Niedzielski et al., 2018; Johnston et al., 2025). Static and dynamic blue band thresholding classifiers have shown promising results for SCA mapping, but results appear to be dependent on lighting conditions, shadowing effects, and spectral features of existing objects (Eker et al., 2019; Jacques-Hamilton et al., 2025). Additionally, there is no standard to determine the optimum blue band threshold for an orthomosaic due to varying conditions. Eker et al. (2019) determined the threshold as the minimum value of sample snow-covered pixels selected randomly from different areas. Jacques-Hamilton et al. (2025) determined the threshold value by visually inspecting snow-covered pixels and adjusting the threshold until a satisfactory alignment with the underlying imagery was achieved. However, this visual-based approach is inherently subjective, relies on human judgment, and is not scalable for automated application across large numbers of orthomosaics.

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).

Despite the promise of these tools, the different approaches have yet to be compared directly. Even indirect comparisons are challenging because the previous studies report different performance metrics. A systematic evaluation is necessary to understand the trade-offs between threshold-based approaches and the improved classification performance alongside greater training complexity and reliance on labeled data, offered by supervised machine learning classifiers in the context of snow mapping. Beyond differences in underlying assumptions and structure of previously used classifiers, the choice of predictor variables, whether limited to blue-band intensity or using the full range of RGB values, require more investigation. Furthermore, because the previous studies typically focused on the application of derived SCA products, there is limited guidance on strategies for UAS data collection to support development of automated, accurate and robust SCA mapping tool.

Here, our goal is to compare the performance of promising snow cover classification approaches in relatively homogenous open areas and to identify one or more classifiers that can consistently and accurately map SCA with limited manual input. We use 26 orthomosaics from repeat-pass UAS visible imagery and 13 lidar-derived snow depth maps collected over multiple years at two homogenous open areas in Southeastern New Hampshire, USA. First, manually labeled, randomly selected pixels are used to train and evaluate the performance of three supervised classifiers (MLE, RF and SVM) using RGB imagery as well as blue-channel only methods via SVM, static and dynamic thresholding. We further assess model generalization under limited training data availability by progressively reducing the number of training orthomosaics and quantify the resulting changes in classification performance. To evaluate the sensitivity of lidar-derived snow-cover estimates to snow-depth threshold selection, we compare fractional snow-covered area (fSCA) derived from lidar snow-depth maps against fSCA estimated from the

best-performing optical classifiers. Finally, we synthesize our findings and provide recommendations aimed at addressing key operational constraints in UAS-based snow mapping. We emphasize the critical role of representative UAS data collection as a prerequisite for developing robust and generalizable SCA classifiers.

2 Methods

2.1 Site Descriptions

Two field sites, the University of New Hampshire's Thompson Farm Research Observatory (43.1089° N, 70.9485° W) and Kingman Farm Research Observatory (43.17207° N, 70.92662° W), are located in southeastern New Hampshire, United States. The sites are separated by approximately 8 km. Both farms are actively managed and operated by the University of New Hampshire. Analyses were performed on the open field portions of the sites indicated by the hatched areas in Fig. 1.

The Thompson Farm study field is approximately 7.6 ha with elevations ranging from 10 to 35 m. The site has limited topographic relief with gentle slopes where low-lying areas within the field and clayey soils result in areas of surface water collection and icing during the winter months (Jacobs et al., 2021). Field vegetation varies between low-cut pasture grass and tall (~1 m) unkempt grasses. The surrounding mixed deciduous and coniferous forest canopy is 10 to 30 m tall (Jacobs et al., 2021). Most of the field experiences two to five hours per day of shade with up to eight hours per day at the forest-field boundary in the north portion of the field (Cho et al., 2021). An access road, running south to north, is the high point between the two fields with the eastern field being lower than the western field. Soils at Thompson Farm are classified as silt loam, sandy loam, and loamy sand in the National Cooperative Soil Survey (NCSS).

The Kingman Farm site consists of two prominent field sections, the main field and a sheltered field, that were the focus of this study. The main field is approximately 17 ha with gentle rolling terrain and maximum elevation changes of ~10 m. This field section is largely exposed with a mixed forest (deciduous/coniferous) bordering the northern edge. An access road runs east to west with the terrain on the southern side of the field. Field vegetation is primarily low-cut pasture grass with a few small sections of tall (~1 m) unkempt grass along the access road. The main field soil is classified as fine sandy loam, silt loam and loamy fine sand. The sheltered field section at Kingman Farm is a smaller (~0.7 hectare) managed field surrounded by forest. The northern edge of the field is bordered by tall conifers (~30 m), while the southern edge is bordered by deciduous trees and an access road. The sheltered field is relatively flat with elevation variations of approximately 5 m with the highest elevations near the center of the field. NCSS classifies soil in this field as fine sandy loam.

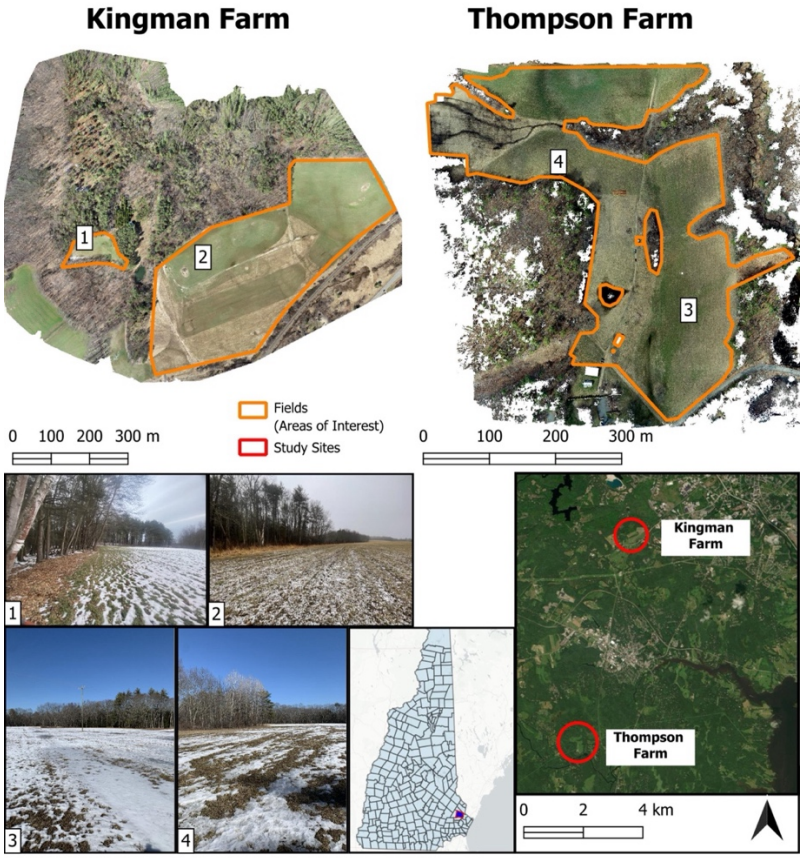


Figure 1. The two study sites, Kingman Farm and Thompson Farm, located at southern New Hampshire (basemap source: Esri, 2024).

The Thompson Farm and Kingman Farm sites are located within the ephemeral snow class, which is characterized by relatively warm temperatures, shallow snow, and a total duration of snow cover lasting for an average of 70 days per year (Sturm and Liston, 2021; Johnston et al., 2023). Based on a 20-year (2002–2021) winter climatology at the co-located U.S. Climate Reference Network sites (USCRN, Thompson site ID: 1040, Kingman Site ID: 1041), January is the coldest month with an average temperature of -4.1°C followed by February (-3.1°C) and December (-1.2°C). These are the only months having average temperatures below 0°C. From December to February, monthly mean high temperatures range from 0.6 to 3.4°C with lows ranging from -9.3

to -6.0°C. The average monthly liquid equivalent precipitation is highest in December (122 mm) and lowest in January (78 mm). February, March, and April have approximately 100 mm of precipitation on average (98 to 108 mm). The sites' combination of sub-freezing temperatures and regular precipitation results in frequent snowfall. Conversely, because above-freezing temperatures are also common, there are intermittent melting and rain-on-snow events throughout the winter. Based on fSCA derived from MODIS observations from 2002-2021 (Hall and Riggs, 2021), snow cover occurs in November (average monthly fraction snow covered area is 0.02), December (0.36), January (0.70), February (0.73), March (0.48), and April (0.04).

2.2 Data

2.2.1 UAS Optical Imagery

The DJI Phantom 4 RTK (real-time kinematic) was used throughout this study to collect RGB imagery. It has a 20-megapixel 1 inch CMOS optical sensor with a mechanical shutter. The exposure of each image was automatically adjusted using the average light metering mode with no exposure bias. In low light conditions, the Phantom 4 favors ISO and aperture adjustments over decreasing shutter speed to minimize the potential for blurring while collecting images during flight. All Phantom 4 flights were conducted at an altitude of 100 m, resulting in a ground sampling distance (GSD) of ~3 cm. Flight line spacing and flight speed were adjusted to achieve 80% side and front image overlap. Flight planning and image capture configuration were done using the DJI Pilot application built into the DJI smart controller. 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.

For all surveys, the collected UAS imagery were stitched together using Agisoft Metashape Pro photogrammetry software to produce RGB orthomosaics. The 3-band Red-Green-Blue (RGB) orthomosaics were the primary input to the snow cover detection algorithms**.** The standard photogrammetry workflow implemented in Metashape involves the generation of a surface model, leveraging image overlap and varying image perspectives, followed by the projection of each image onto the produced surface model. This process removes distortions (perspective/lens) from each image and results in a true planimetric map of the survey area. Although positional accuracy of RGB imagery was not critical for this study, all collected UAS image geotags were either post-processing kinematic (PPK) or real-time kinematic (RTK) corrected prior to photogrammetry processing (see appendix for further details).

A total of 26 UAS RGB orthomosaics were produced and used for this snow cover detection study (Table 1, Fig. S-1 and S-2). Flights were scheduled to capture a range of snow conditions. However, due to limited tie-points, most images collected during 100% snow cover could not be effectively stitched together and were excluded from this study. Of the 26 dates, 10 represented very shallow/patchy snow conditions with observed snow depths less than 4 cm, 3 were completely snow free, and the remaining 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. Figures 2 and 3 show examples of RGB orthomosaics acquired with the Phantom 4 at Kingman Farm and Thompson Farm, respectively.

Table 1. Number of UAS surveys conducted during each winter season and sensor.

Winter	UAS Optical Imagery	UAS Lidar
2020-2021	13	9
2021-2022	2	0
2023-2024	9	0
2024-2025	2	4
Total	26	13

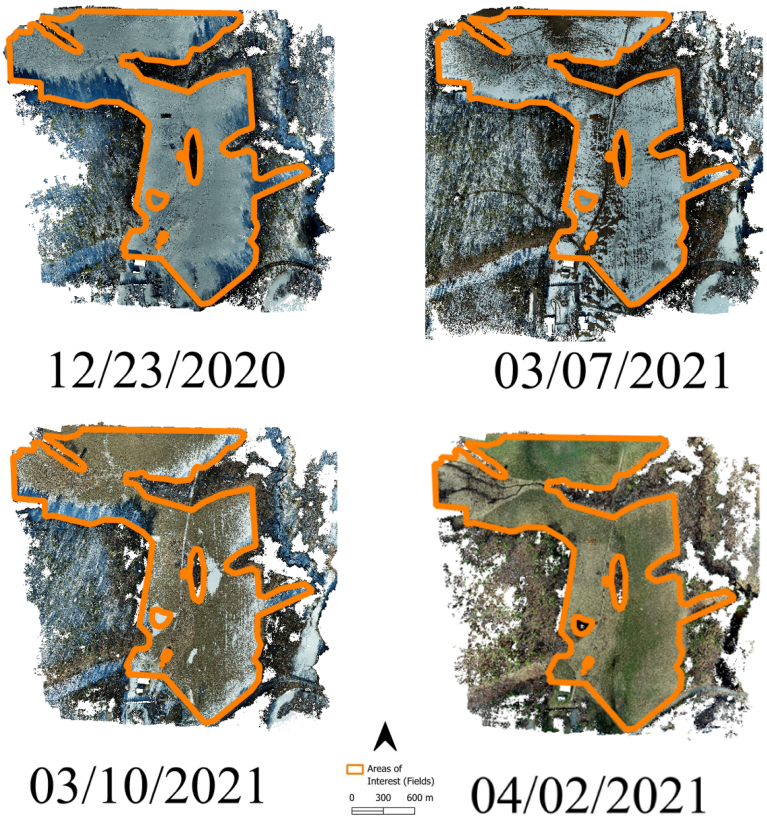


Figure 2. Example RGB orthomosaics of Thompson Farm acquired with the Phantom 4. The open-field study areas are outlined in orange.

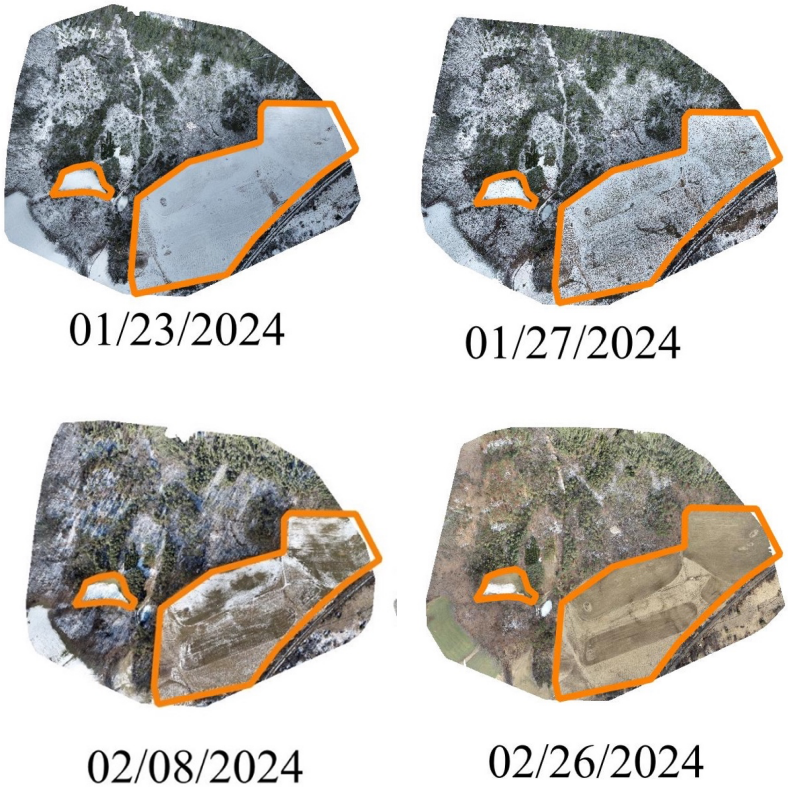


Figure 3. Example RGB orthomosaics of Kingman Farm acquired with the Phantom 4. The open-field study areas are outlined in orange.

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).

2.2.3 Ancillary Observations

Snow depth was available on dates coinciding with UAS orthomosaics. On selected dates, UAS lidar was used to produce field-scale high-resolution snow depth estimates. 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).

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. Air temperature at the start of each UAS survey, along with the mean air temperature over the six hours preceding each flight, was used as a proxy to estimate snow wetness conditions. If the 6-hour mean air temperature exceeded 0 °C, the snow was assumed to be wet at the start of the survey. Additionally, the mean observed shortwave radiation during each flight was calculated to approximate sky conditions, following the method described in the Supplementary Material (Table S-1). Summary of snow and meteorological conditions at the UAS survey dates are provided at Table S-2.

2.3 RGB-based Snow Cover Mapping

2.3.1 Ground Truth Dataset

For each orthomosaic, site-specific masks were applied to exclude surrounding forested areas, leaving only the field area for analysis. Over-saturated pixels (e.g., [255, 255, 255] or [0, 0, 0]) were also removed prior to analysis. The fields of each orthomosaic were delineated in QGIS and 50 points were randomly selected within these areas of interest. A new set of random points were generated for every orthomosaic (Fig. S-3 and S-4). For Kingman Farm, 48 points were placed in the open pasture and two in the sheltered field to represent the relative areas of the two fields. The pixel nearest to each sample point was classified manually as snow-covered or no-snow ground. The red, green, and blue band values for each of the randomly sampled pixel were extracted.

2.3.2 RGB-based Snow Cover Classifiers

This study employed five classifiers for snow cover mapping, including both machine learning approaches (Maximum Likelihood Estimation (MLE), Random Forest (RF), Support Vector Machine (SVM)) and thresholding approaches (static and dynamic). The choice of predictors varied by classifier, utilizing either full RGB intensity values or exclusively the blue band reflectance. The combination resulted in seven distinct classification configurations. All classifiers were implemented in Python using scikit-learn library. When hyperparameter tuning was required, it was performed using grid search and 5-fold cross-validation applied only to the training subset. The set of parameters yielding the highest mean balanced accuracy across the cross-validation folds was selected as the optimal set. Implementation details for each classifier are provided below.

Modeled off the Maximum Likelihood Classification tool in ENVI (Environment for Visualizing Images) version 6.1, a MLE classifier was implemented in python to predict snow cover. This classification method assumes that the band values are normally distributed. Under this assumption, class statistics (i.e. mean vectors, covariance matrices, and prior probabilities) were computed from the training set for each class. The discriminant function determined the probability of each sample belonging to each class and the sample was assigned to the class with the highest probability.

The RF was developed using RGB color intensity values as predictive features. The search space for hyperparameters included variations in the number of estimators, maximum tree depth, minimum samples required to split an internal node, and the minimum number of samples required to be at a leaf node.

Two variations of the SVM classifier were applied using RGB intensity values and only blue band intensity values (SVM Blue Band). Hyperparameter tuning was performed to identify the optimal combination of the regularization parameter (C), kernel type (linear, radial basis function, polynomial, or sigmoid), and kernel coefficient (gamma).

The static blue band thresholding classifier applied a single, fixed threshold to the blue band intensity values. A pixel was classified as snow-covered if its blue band intensity exceeded this threshold. 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 as the evaluation criterion (Table S-3).

Unlike the other classifiers, the blue band dynamic thresholding does not rely on labeled ground truth data. Instead, it uses orthomosaic-specific histogram characteristics to identify a unique threshold for each orthomosaic. A reflectance histogram was constructed from the field pixels' blue band using a bin size of 5, then smoothed using a centered moving average filter with a window width of five bins to suppress noise and enhance prominent modes. Local minima in the smoothed histogram were identified and compared against a predefined threshold of either 127 or 90. Both thresholds have been previously used to distinguish snow-covered surfaces (Jacques-Hamilton et al., 2025; Salvatori et al 2011; Portenier et al 2020; Caparó Bellido and Rundquist 2021). The first local minimum exceeding this threshold was used to differentiate snow from no-snow pixels. In cases where no local minimum satisfied this condition, the predefined minimum threshold was used.

2.3.3 Performance Assessment

Baseline Performance

In this study, we first compared the performance of the classifiers trained based on data from the 26 P4 UAS orthomosaics in the reference dataset. Classifier performance was evaluated using 10 repeated random train/test realizations. For each repetition, the dataset (750 sample points from Kingman Farm and 550 sample points from Thompson Farm) was stratified into 80% training and 20% testing subsets while preserving the class distribution. To ensure a consistent and statistically comparable evaluation across classifiers, the same train/test splits were reused for all models. Model performance was evaluated on the corresponding test subset using accuracy, balanced accuracy, F1 score, and Cohen's kappa.

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.

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

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

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:

$$D_i = \min(p_i, 1 - p_i) \quad (2)$$

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

Generalization Performance Under Limited Training Data

Next, we conducted an experiment to address flight planning and understand changes in performance based on the number of flights as well as the impacts of surface condition captured during those flights used to train models. This experiment used subsets of orthomosaics to train new models, with subsets sizes ranging from 6 to 25 orthomosaics. Each subset was obtained by holding out n orthomosaics from the original set of 26, thereby containing $26 - n$ orthomosaics. When the number of possible hold-out combinations was fewer than 500 (e.g., for n = 1, 2, or 3), all were used. Otherwise, 500 combinations were randomly sampled. For each combination, a new model was trained and optimal hyperparameters were found following the methodology established in the base experiment. Model performance was assessed on the corresponding held-out orthomosaics (i.e., those excluded from training), hereafter referred to as held-out accuracy. Held-out performances were summarized and reported for the range of training flights counts. This experiment included only the top-performing classifiers from the base experiment for further evaluation due to computational time needed to conduct this experiment.

2.3.4 Performance Metrics

The agreement between the ground truth data and the model output was quantified using accuracy, balanced accuracy, F1 score, and Cohen Kappa performance metrics. In this study, a true positive (TP) is defined as both the model and the ground truth agree on the "snow" class, while a true negative (TN) indicates that both agree on the "no-snow" class. A false positive (FP) occurs when the ground truth dataset indicates “no-snow” and the model output indicates “snow”. Similarly, a false negative (FN) occurs when the ground truth dataset indicates “snow” and the model output indicates “no-snow”.

Accuracy is the number of cases where the model agrees with the ground truth data relative to the total number of instances, calculated following Eq. (3)

$$Accuracy = \frac{(TP+TN)}{(TP+FP+TN+FN)} \tag{3}$$

Balanced accuracy is the average accuracy of both the positive and negative classes. It is defined as the number of correct predictions for each class (true positives and true negatives) divided by the total number of instances within that class, with the results averaged across all classes calculated as

$$Balanced\ Accuracy = \frac{\frac{TP}{TP+FN}+\frac{TN}{TN+FN}}{2} \tag{4}$$

The F1 score is another widely used metric for evaluating classification performance, particularly for imbalanced datasets such as those having complete snow cover. The F1 score is the harmonic mean of precision and recall and is defined as

$$F1 = 2 \times \frac{Precision \times Recall}{Precision+Recall} \tag{5}$$

where precision is the ratio of correctly identified positive cases ("snow") to all predicted positives by UAS-SCA classifier

$$Precision = \frac{TP}{TP+FP} \quad (6)$$

and recall is the ratio of correctly identified positive cases to all actual positives in the ground truth

$$Recall = \frac{TP}{TP+FN} \quad (7)$$

Cohen Kappa is used to assess the reliability of a classifier while accounting for the possibility of agreement occurring by chance. It compares the observed agreement to the expected agreement under random classification and is defined as

$$K = \frac{p_0 - p_e}{1 - p_e} \quad (8)$$

where p_o, the relative observed agreement, and p_e, the expected agreement due to chance, are computed as

$$p_0 = \frac{TP+TN}{N} \quad (9)$$

$$p_e = \frac{(TP+FP)\times(TP+FN)}{N^2} + \frac{(TN+FN)\times(TN+FP)}{N^2} \quad (10)$$

where N is the total number of samples. A kappa value of 1 indicates perfect agreement, while a value of 0 suggests agreement that is no better than that from random chance.

2.4 Lidar-based Snow Cover Mapping

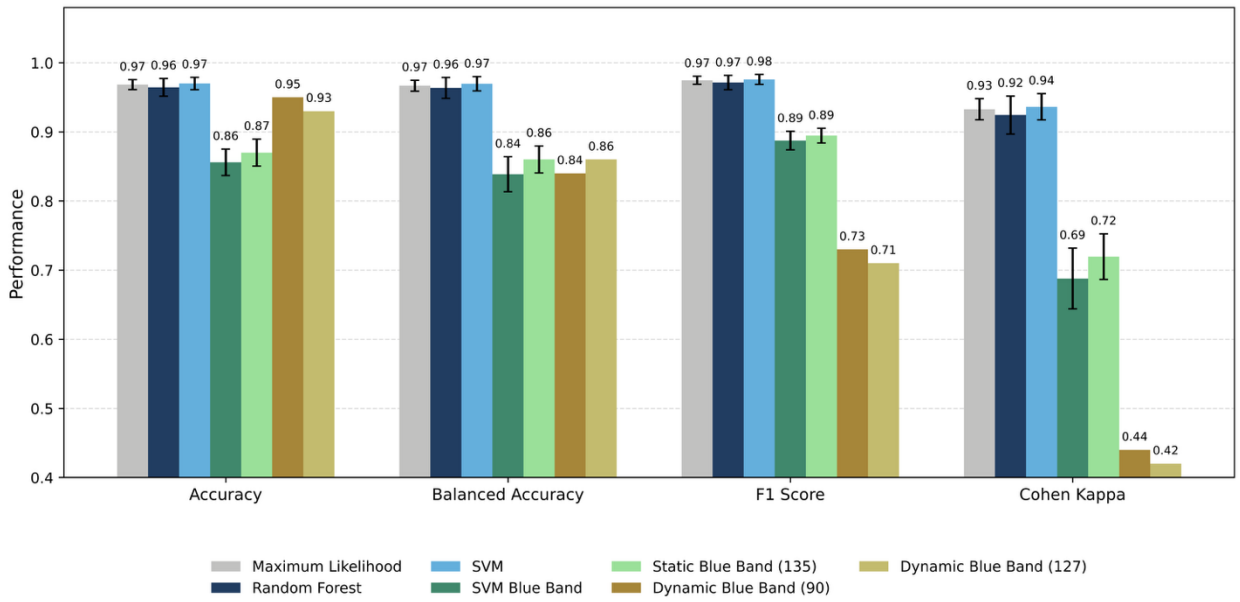
To derive binary SCA from UAS LiDAR snow depth, we applied a series of snow-depth thresholds to the 13 LiDAR-derived snow depth fields that were flown coincidentally with the optical. We tested thresholds ranging from 0 to 10 cm, in increments of 2cm, and classified pixels

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.

3 Results

3.1 Baseline Performance

The performance was evaluated for seven classification configurations for snow-covered area mapping with 26 orthomosaics using the P4 UAS platform used in the reference dataset. The results clearly separated the tested configurations into two performance groups (Fig. 4). The MLE, RF, and SVM classifiers emerged as the top performers, achieving accuracy, balanced accuracy, and F1 score values exceeding 0.95, indicating strong agreement and classification accuracy. Their high performance was further supported by Cohen kappa values exceeding 0.89. While SVM Blue Band and Static Blue Band classifiers consistently underperformed the top three classifiers, a single blue band still yields reasonable results (balanced accuracies of 0.86 and 0.84). While the Dynamic Blue Band classifier produced high accuracy values for both thresholds (0.95 for a threshold of 90 and 0.93 for 127), the classifier had the poorest performance for the other metrics. It should be noted that for the Dynamic Blue Band classifier, the reported Cohen's Kappa represents an average across individual orthomosaics. However, for orthomosaics containing only a single class (e.g., fully snow-covered or completely snow-free scenes), Cohen's Kappa becomes unreliable, as it cannot account for agreement beyond chance in the absence of class variability. Overall, all tested classification configurations have the capacity to classify SCA with the Maximum Likelihood, RF, and SVM being the most robust classifiers.



425 **Figure 4.** Comparison of classifier performance for SCA mapping across four evaluation metrics: accuracy, balanced accuracy, F1 score, and Cohen's kappa. Bars represent the mean performance across 10 repeated realizations, and error bars indicate ±1 standard deviation. Numerical labels above bars denote mean metric values.

430 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 435 (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.

The disagreement maps suggested that the presence of shadows leads to the highest 440 disagreement among models ($D_i > 0.4$). 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.

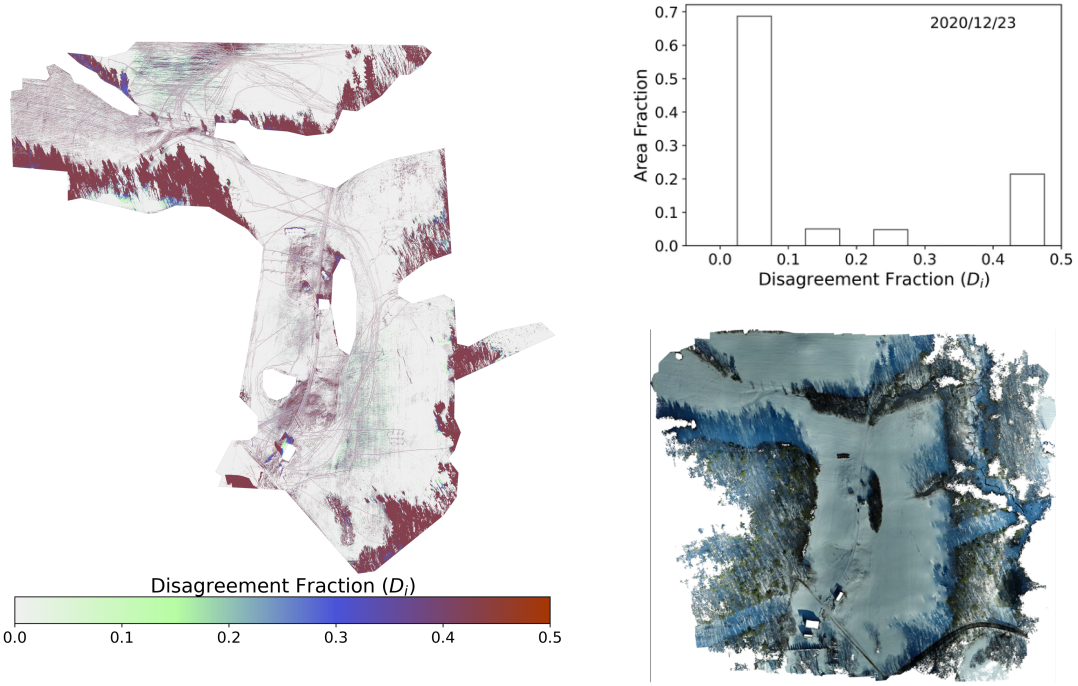


Figure 5. Spatial distribution of disagreement fraction (D_i), and corresponding histogram of disagreement fraction for the TF study site on 2020/12/23. The disagreement fraction represents the proportion of snow classification methods that disagreed with the majority-vote classification at each pixel

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.

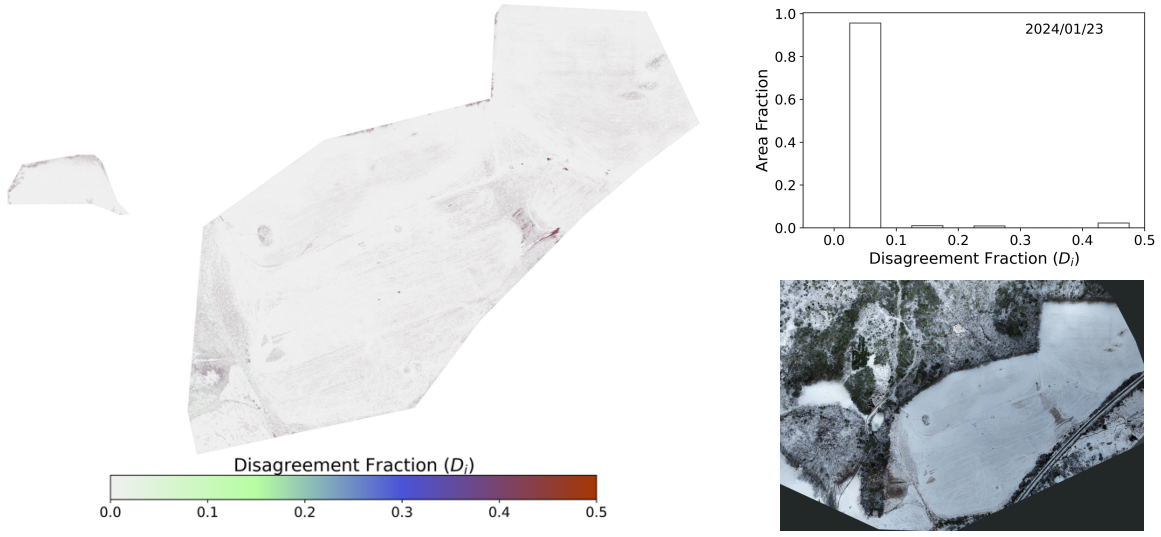


Figure 6. Spatial distribution of disagreement fraction (D_i), and corresponding histogram of disagreement fraction for the KF study site on 2024/01/23

460

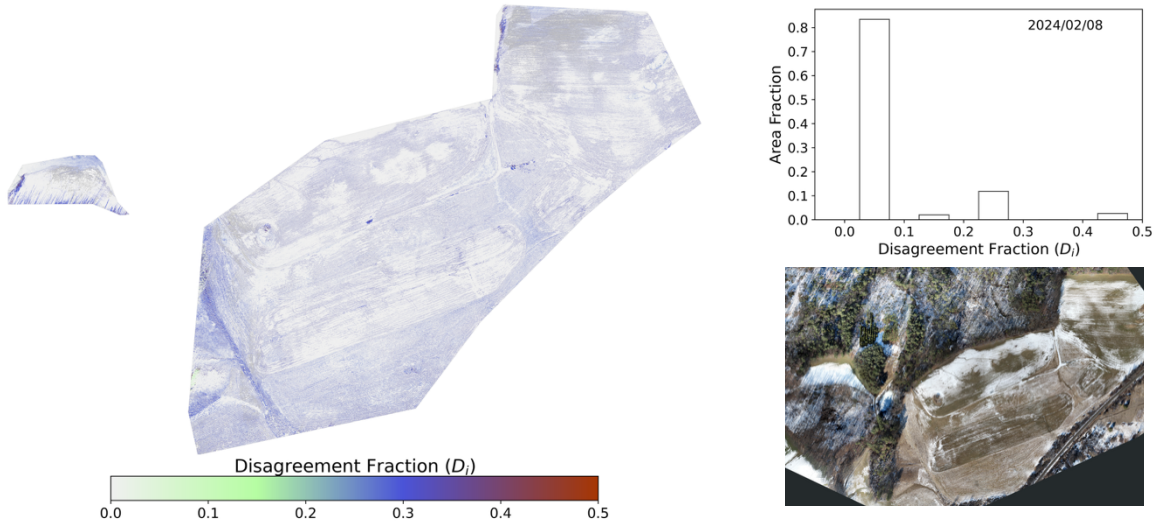


Figure 7. Spatial distribution of disagreement fraction (D_i), and corresponding histogram of disagreement fraction for the KF study site on 2024/02/08

465

3.2 Generalization Performance Under Limited Training Data

This experiment used subsets of 6 to 25 orthomosaics from the complete set of 26 orthomosaics to train and evaluate new MLE, SVM, and RF models (maximum 500 combinations). The performance was calculated for the individual orthomosaics that were "held-out" from the dataset, and an overall average of these held-out orthomosaics is presented here. Figure 8 shows that reducing the number of orthomosaics used for training only modestly reduces the median combination's performance with mean accuracies remained consistently high (at or above 0.9) for all three classifiers. However, reducing the number of orthomosaics does lead to increased variability in performance among the combinations.

All three classifiers performed comparably when trained on larger datasets (> 20 orthomosaics) and had reduced ability on certain orthomosaics when trained on small datasets (< 12 orthomosaics). RF demonstrated the weakest generalization performance on mid-sized datasets. Even when most combinations performed well, certain combinations of training and testing datasets led to reduced performance. For models trained with 20 to 12 orthomosaics, fewer than 0.5% of MLE and SVM models had held-out accuracies below 90%, whereas approximately 4% of RF models fell below the 90% accuracy.

Even when the average of all held-out orthomosaics is high, some individual orthomosaics may have reduced performance (Fig. S-5). Among the three classifiers, the RF classifier is slightly more sensitive to training data availability. RF models' median held-out accuracies declined from 0.97 to 0.90 as the number of training orthomosaics decreased, whereas MLE and SVM only decreased from 0.97 to 0.95 and 0.98 to 0.94, respectively. The RF models also had individual orthomosaic accuracies below 50% even when 18 orthomosaics were used for training (Fig. S-5). When the SVM and MLE training set included more than 14 orthomosaics, accuracy for nearly all

excluded orthomosaics exceeded 80% (Fig. S-5), though the SVM models had more low accuracy orthomosaics than MLE models. When the MLE and SVM training data sets contained about 10 or fewer training orthomosaics, some of the held-out orthomosaics had accuracies below 50%, with a few extreme cases where accuracies were as low as 0%.

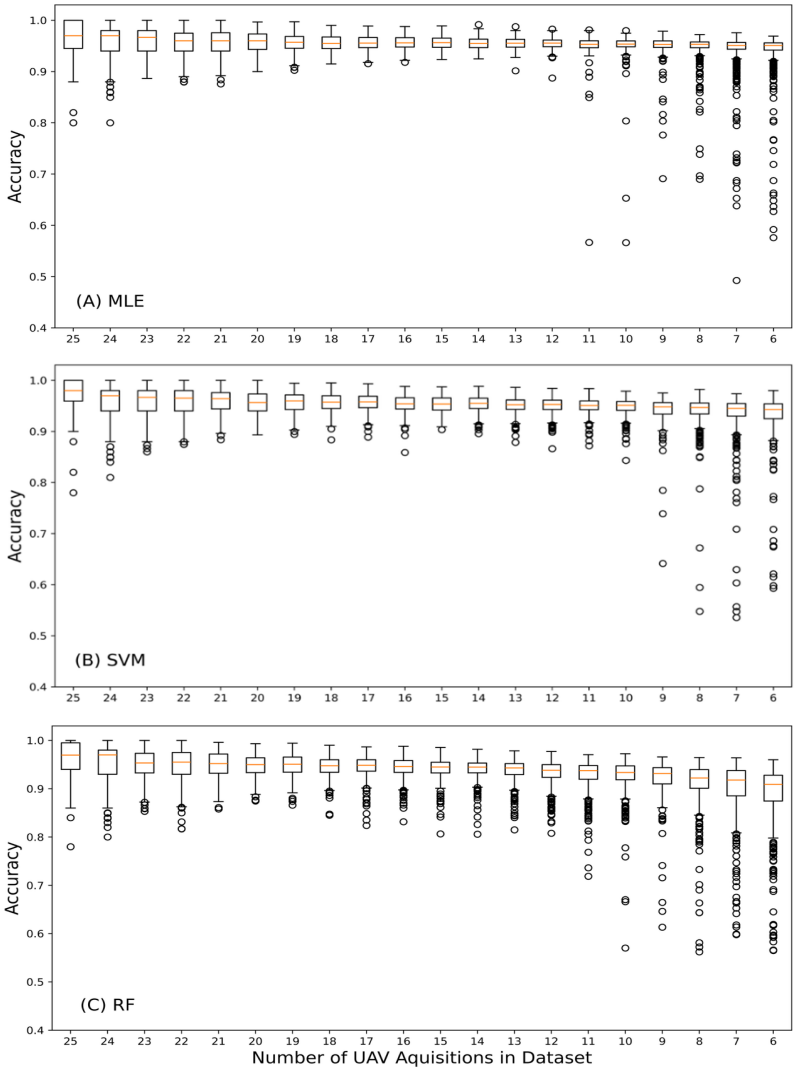


Figure 8. Boxplots of the mean held-out accuracy from each combination of held-out orthomosaics (maximum 500 combinations) by the number of orthomosaics included in the dataset for a) MLE, b) the SVM, and c) RF classifiers.

Individual orthomosaics' performance were examined. Figure 9 shows each individual orthomosaic's classification consistency when it was excluded from the training dataset plotted as

the proportion of time that the classification accuracy was less than 80% by orthomosaic. While most scenes were reliably predicted with accuracies greater than 80% for more than 95% of the combinations, some scenes were consistently more difficult to classify. For the six lowest accuracy orthomosaics, the MLE models performed better than either the SVM or the RF models. RF models have lower accuracies than either the MLE or SVM models except for the January 13th, 2025 orthomosaic from Kingman Farm. For example, the RF models' accuracies for the February 12th, 2024 orthomosaic of Kingman Farm were below 80% for nearly 40% of the combinations when that orthomosaic was held out from the training dataset, but for the same orthomosaic, MLE and SVM accuracies were seldom (10%) below 80%.

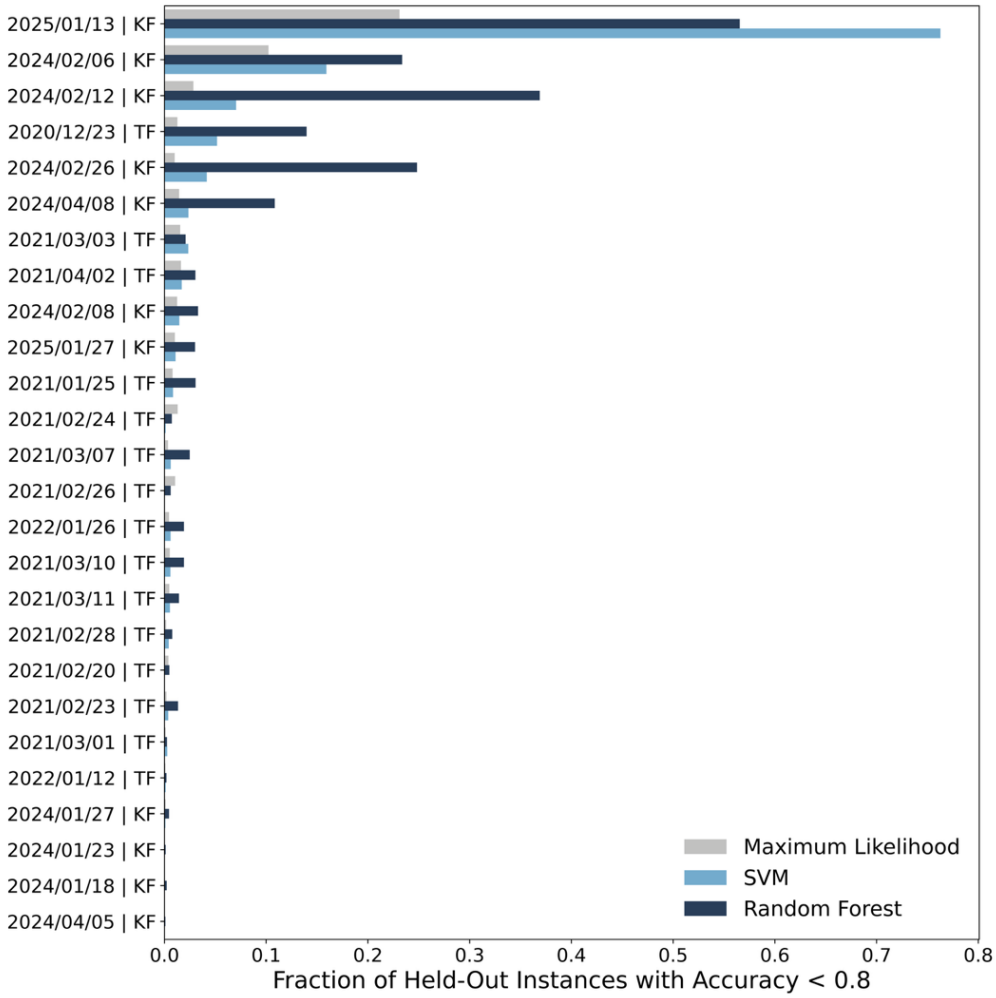


Figure 9. Fraction of held-out instances with classification accuracy below 80% for each date-site combination, shown separately for the three classifiers: Maximum Likelihood, Random Forest, and Support Vector Machine. Each bar represents the proportion of times a given orthomosaic was held-out and subsequently predicted with low accuracy.

To better understand which conditions led to reduced performance, we compared fractional snow-covered area (fSCA), snow wetness and cloud cover conditions for the ten highest- and lowest-performing models for training datasets ranging from six to 12 orthomosaics. fSCA for each orthomosaics was estimated using the models from the base experiment (i.e., models trained on the full dataset; see Fig. S-6). For each condition (i.e., fSCA, wetness, and cloud cover), we then compared its mean values from the dates included in the training sets with those of the

corresponding held-out orthomosaics for the ten highest performing and up to 10 lowest-performing when less 80% accuracy models.

For fSCA, the results for the SVM classifier are presented in Fig. 10, corresponding analyses for MLE and RF are shown in the supplementary materials (Fig. S-7 and S-8). For datasets having ten to 12 orthomosaics, average fSCA ranges from about 0.5 to 0.8 for both the training and testing datasets. Note that this size training dataset has no models with less 80% accuracy. As the dataset size decreases below ten, the highest-performing models still have fSCA values from 0.5 to 0.8, but the models having less 80% accuracy used training datasets with average fSCA values higher than 0.8. In addition, the held-out orthomosaics had lower fSCA values, sometimes dropping below 50%. This reveals that the poorest performing models used training data are dominated by snow-covered orthomosaics (often >70% fSCA), then used to predict conditions with lower or more variable snow coverage, Thus, there is a mismatch between the mean fSCA values of their training (included) and test (held-out) orthomosaics. We did not find any evidence that differences in cloud cover (Fig. S-9, S-10, S-11) or snow wetness (Fig. S-12, S-13, S-14) between training and held-out orthomosaics influenced performance.

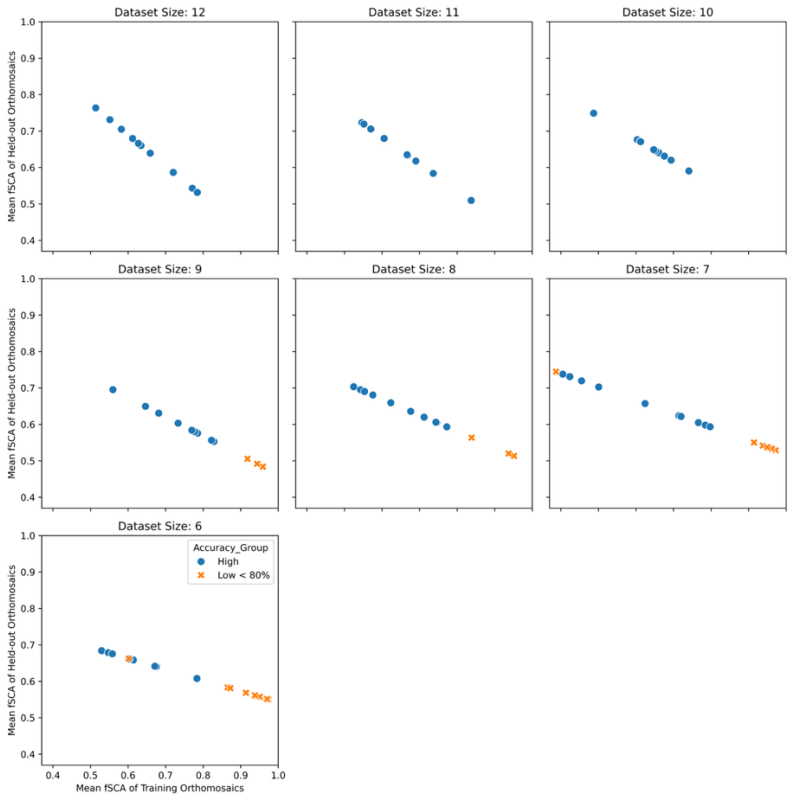


Figure 10. Mean fSCA of orthomosaics within training vs. Mean fSCA of held-out orthomosaics for the 20 combinations (10 highest and 10 lowest (<80%) accuracy) for the SVM classifier. Blue dots represent high-accuracy combinations; orange dots represent low-accuracy combinations. The individual plots differ by the number of orthomosaics in the training dataset. Training dataset sizes of 10, 11, and 12 did not have any held-out datasets with accuracy values < 80%.

3.3 Lidar-Based Snow Cover Mapping

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.

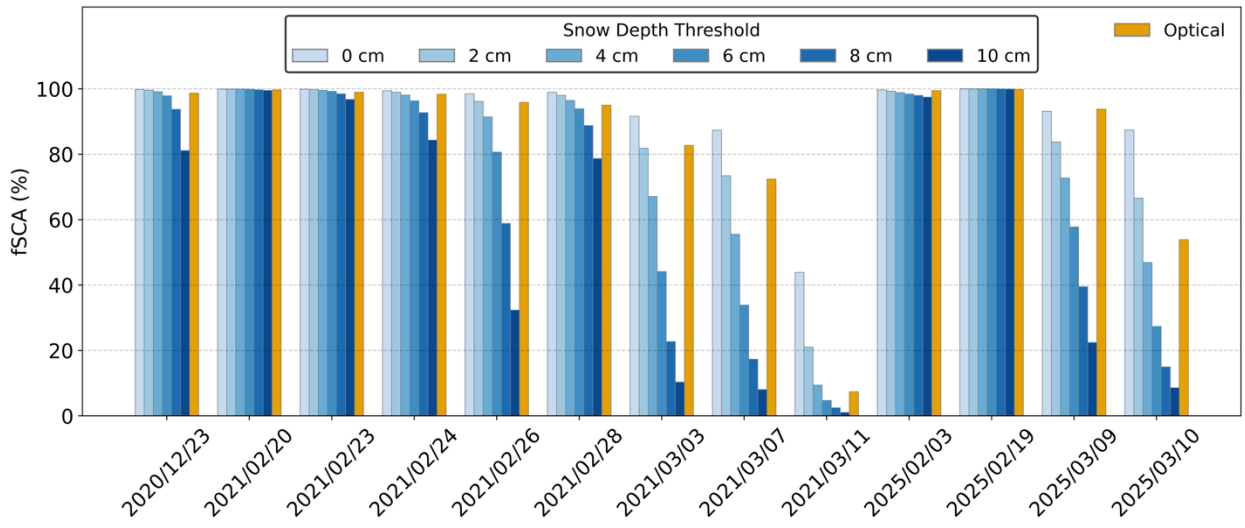


Figure 11. Fractional snow-covered area (fSCA) derived from optical imagery and lidar snow-depth maps using snow-depth thresholds of 0–10 cm.

4 Discussion

4.1 Baseline Performance

Our primary goal was to identify one or more classifiers that can consistently and accurately map SCA in open areas. The baseline assessment of tested classifiers established that the three classifiers (MLE, RF, and SVM) using full RGB inputs to be more effective classifiers for mapping SCA than classifiers relying solely on the blue band. Their ability to differentiate between snow-on and snow-free conditions for 3 cm pixels exceeded 95% for accuracy, balanced accuracy, and F1 score performance metrics. In comparison to previous studies in open areas, this study's performance matches the performance reported by Eker et al.'s (2019) overall accuracy of 95% and exceeds Belmonte et al.'s (2021) accuracy of 90.2% using a RF classifier, Cannistra et al.'s (2021) balanced accuracy of 0.82 using a convolutional neural network-based method and Johnston et al.'s (2025) 80 to 88% accuracy via k-means clustering. Taken as a whole, the UAS performance appears to be similar to satellite snow cover products' performance which typically range from about 85 to 95% (Ault et al., 2006; Hall and Riggs 2007; Gao et al., 2010; Nolin, 2010; Frei et al., 2012; Bair et al., 2016; Coll and Li, 2018) with the UAS identifying SCA at a much finer resolution (cm scale).

Classifiers relying solely on the blue band, such as SVM Blue Band, Static Blue Band Thresholding, and Dynamic Blue Band Thresholding, still achieved good results, but were consistently less accurate than their RGB-based counterparts. Of the blue band classifiers, Static Blue Band Thresholding performed the best based on its F1 Score of 0.90, balanced accuracy of 0.86 and accuracy of 0.87. While the Dynamic Blue Band Thresholding had an accuracy (0.95) comparable to the MLE, RF, and SVM, this is largely due to its strength in capturing snow when

fSCA is high, its balanced accuracy of 0.78 reflects its diminished ability to capture mixed conditions. The dynamic thresholding approach generally underestimates snow-covered pixels, as indicated by a higher rate of incorrectly classifying of snow free pixels than snow covered pixels (Tables S-6 and S-7). While this classifier is designed to adapt to varying illumination conditions unique to each orthomosaic, its effectiveness diminishes when no local minimum is detected in the histogram (e.g., fully snow-covered days) making its performance highly dependent on the pre-defined minimum threshold. Figures S-15 and S-16 clearly demonstrate how the value of the pre-defined minimum threshold can influence classifier performance when the intensity histogram lacks a clear local minimum. Moreover, the dynamic thresholding approach can be memory-intensive when constructing histograms using all pixels from high-resolution orthomosaics for large study areas. As an alternative, random sampling of the orthomosaic pixels may be used to reduce memory requirements but would likely further reduce classification performance due to less representative histogram estimation. Despite its simplicity, the Static Blue Band classifier is preferable to the dynamic approach. Where reduced performance is an acceptable trade-off for increased ease of use and users are knowledgeable about conditions that may be problematic, the blue band threshold classifier appears to offer a reasonable compromise that merits additional vetting.

The clear difference between using classifiers with three bands as compared to the single blue-channel is somewhat unexpected because the blue-channel is a widely used classifier. As reviewed by Webster & Jonas (2018), the use of the single blue-channel is the favored method for RGB imagery from field cameras because blue light is reduced in the canopy view fraction (e.g. Reid and Essery, 2013; Frazer et al., 1999) and saturates over most snow-covered scenes causing loss of sun-lit/shadow contrast (Wolter et al., 2012; Dozier, 1989). Using UAS RGB imagery,

Jacques-Hamilton et al. (2025) reporting satisfactory results by thresholding the blue-channel values following Salvatori et al.'s (2011) approach for mapping snow cover using field camera images. Their blue pixel intensity threshold value of 90 was determined using a qualitative approach in which they adjusted the threshold until pixels classified as snow cover agreed with what they saw in their photos. This is considerably lower than the threshold value of 127 used for field cameras (e.g. Salvatori et al 2011, Portenier et al 2020, Caparó Bellido and Rundquist 2021). Our results suggest that while the blue band is highly sensitive to snow due to its strong reflectance properties (Nolin, 2010), combining it with the red and green bands provides more reliable snow classification, particularly when mapping shadowed snow, dirty snow, and wet snow. Figure S-17 shows examples of SCA maps on a partially snow-covered day and a fully snow-covered. It appears that the reduced blue band reflectance caused by shadows, impurities, or melting led to misclassification of snow-covered surfaces leading to a reduction in the field's fSCA values by 15 to 25% in comparison to fSCA that was properly captured by MLE, RF, and SVM.

4.2 Recommendation for methods, transferability, and UAS data collection strategies

While MLE, SVM and RF classifiers produced comparable baseline performance, the assessment of their generalization performance revealed that RF models may be more sensitive to some surface conditions, possibly due to its reliance on multiple decision trees that are impacted by the absence of key features in outlier orthomosaics. Conversely, SVM and MLE demonstrated relatively more stable behavior, indicating that they may be less prone to overfitting and could offer better generalization for snow detection when the training set is not representative of a new orthomosaic's conditions. MLE, while simple, seems to perform consistently in the majority of cases.

Our generalization experiment offered insights into how both the number and characteristics of orthomosaics influence the development of a robust model, highlighting the challenges posed by data availability and diverse snow cover conditions when mapping snow using UAS RGB imagery and machine learning classifiers. For dynamic winter conditions similar to those in our study area (characterized by an ephemeral snowpack) at least 12 orthomosaics spanning a range of snow cover conditions was enough to result a reasonable classification accuracy (exceeding 80%) when employing robust classifiers such as MLE, SVM, and RF. However, beyond just the number of orthomosaics, such a dataset should be representative of a diverse range of snow cover conditions that pose challenges for snow mapping. Key examples of challenging conditions include:

Thin snow: The two orthomosaics with the highest frequency of accuracies below 80% (January 13th, 2025, *KF* and February 6th, 2024, *KF*) in the generalization experiment also exhibited the lowest median accuracy across all models (Fig. S-18). Although both orthomosaics were predominantly snow-covered, with fSCA exceeding 60%, the observed snow depth was relatively shallow (3 to 4 cm). This consistently poor generalization performance, regardless of classifier, underscores the inherent difficulty of detecting shallow snow cover using only optical imagery. These findings are not surprising for these optically thin snowpacks. In the visible wavelengths, where there is little absorption and scattering is dominant, light penetrates snow by 10's of centimeters beneath the surface (deeper for blue than red). If snow is shallower than this then almost certainly the observed signal is being darkened by the ground below. In contrast, methods that use the NIR, where ice absorption/grain size dominates, would be expected to perform better because light does not penetrate below 1-2 cm of snow. That said, our experiment

used numerous other orthomosaics taken when snowpacks are equally thin. In most cases, these thin snowpacks were successfully classified, suggesting that even if the ground is making the snow appear darker than it actually is, it still is brighter than most other surface types in this mapping area. Special cases such as shallow and wet snow or snow that is shallow and shadowed require additional study (personnel communication S. McKenzie Skiles, March 19, 2025).

Patchy snow: The results of the generalization assessment suggest that poor generalization performance may be rooted in an fSCA mismatch between training and test sets. When classifiers are not exposed to a diverse range of snow cover conditions, their ability to identify snow accurately in less homogeneous scenes is significantly impaired. This vulnerability is particularly evident for smaller training sizes, where the representation of snow variability decreases. Patchy snow scenes often contain a mix of exposed ground, water-stained snow, and potentially dirty snow with soil or organic debris. These features cause snow RGB values to deviate from values for fresher and cleaner snow typically represented in training datasets with higher mean fSCA. These findings align with previous studies in snow classification and remote sensing, which have shown that snow detection algorithms, particularly those relying solely on RGB or optical inputs, are sensitive to scene complexity and snow spatial heterogeneity (Painter et al., 2009; Dietz et al., 2012). Our results further show that this sensitivity is not limited to spectral thresholds or empirical rules but extends to machine learning classifiers as well. Both RF and SVM rely on feature-space separation, and if the features of aged or dirty snow overlap too much with non-snow classes, the models may misclassify them unless those cases were explicitly represented during training. MLE, in turn, assumes underlying distributional properties for each class. When snow pixels in patchy conditions no longer follow the same statistical profile learned from clean snow scenes,

misclassification becomes inevitable. Previous studies have also found that snow reflectance varies with solar illumination, becoming brighter and more neutral under direct sunlight and darker and bluer in shadow or diffuse light (Dozier, 1989; Painter et al., 2003). However, our results did not find any evidence that variability in weather conditions at the time of UAS acquisition limited generalization performance.

No snow: Three additional orthomosaics with a high fraction of low-accuracy outcomes in the generalization experiment (*February 12th, 2024, KF*, *February 26t, KF, 2024*, and *April 8th, 2024 KF*) further illustrate the limitations of relying solely on RGB data for snow mapping. These cases, all from the Kingman Farm site, feature no snow condition with a whitish appearance that can be easily misclassified as snow (Fig. S-19). This situation is likely not a problem if the same research team collects the imagery and processes it because the lack of snow is self-evident. If a different team processes imagery, this issue can be addressed by a manually scanning the orthomosaics prior to classifying snow.

From a practical standpoint, insights from this study can guide future data collection, dataset design, and model development. It is essential to represent the full spectrum of snow depth and fSCA from no snow to almost full snow cover, and especially the transitional states (e.g., fSCA between 20–60%). While not measured in this study, representativeness likely extends to a range of snow metamorphism stages. As snow ages, grain size growth and increased density reduce its reflectance, particularly in visible bands, leading to a duller appearance in RGB images (Warren, 1982; Green et al., 2006). Wet, refrozen, or contaminated snow further lowers reflectance, making it spectrally similar to bright ground, introducing noise and complexity (Kaufman et al., 2002).

These effects result in a strongly non-linear relationship between fSCA and observed snow's spectral characteristics, even in the limited RGB space.

Finally, several considerations should be taken into account when applying RGB classifiers for operational applications. First, for transfer of trained classifiers across sensors, attention should be given to differences in sensor radiometric characteristics. We note that in this study, the RGB reflectance of the P4 sensor spanned the full dynamic range (i.e., 0–255). In cases where a sensor exhibits reduced sensitivity in a spectral channel or produces compressed or biased RGB values, it is critical to apply bias correction to the intensity values prior to classification to ensure comparability and preserve model performance. 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.

4.3 Lidar-based Snow Cover Mapping

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.

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 (Fig. 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 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.

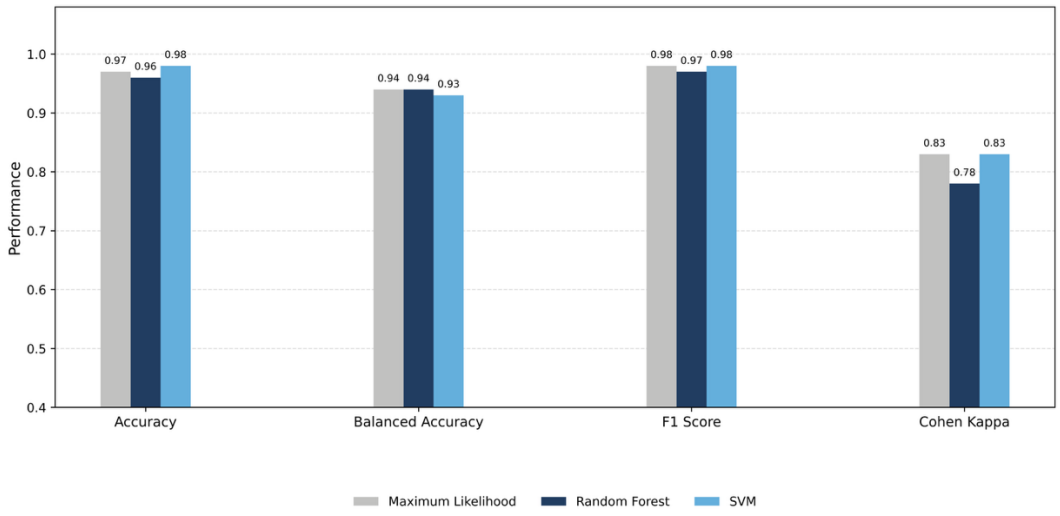


Figure 12 - Performance metrics of the three best-performing models from the reference SCA analysis, evaluated in forested area canopy gaps using 11 orthomosaics collected at the Kingman Farm study site.

5 Summary and Conclusion

Mapping snow cover persistence via UASs is useful for detailing snow spatial variability over large areas, and has direct relevance to agriculture, hydrology, mobility, and land surface modeling. This work compares seven classification configurations for producing snow cover estimates from RGB UAS's high resolution (5 cm) observations over two study sites located in southern New Hampshire, USA having open, agricultural landscapes. 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). The tested classifiers included machine learning methods (MLE, RF and SVM) and thresholding approaches (static and dynamic). Predictor selection varied by classifier, using either full RGB intensity values or the

blue band alone. Among them, MLE, RF, and SVM using RGB inputs consistently outperformed others, with accuracy, balanced accuracy, and F1 scores exceeding 0.95. In contrast, classifiers relying solely on the blue band showed lower overall performance (accuracy: 0.90; balanced accuracy: 0.85; F1 score: 0.81, on average). For snow mapping using UAS optical imagery in regions characterized by shallow, ephemeral snow and dynamic winter conditions, we found that including at least 12 orthomosaics spanning a broad range of snow conditions, especially including transitional states (e.g., fractional snow-covered area between 20–60% during melting, reaccumulation and aging phases) can yield accuracies above 80% with MLE, SVM, and RF. While MLE, SVM, and RF performed comparably when trained on large, representative datasets, MLE and SVM emerged as more reliable and consistent choices when data availability is limited. Where reduced performance is acceptable, the blue band threshold classifier appears to offer an easy-to-use method, but users are advised to carefully vet their dataset for conditions that may be problematic.

We also investigated the use of lidar-derived snow-depth observations for SCA mapping. Our analysis demonstrated that lidar-based estimates of snow-covered area are highly sensitive to the selected snow-depth threshold in ephemeral and transitional snow regimes. These conditions are characterized by relatively shallow snowpacks and therefore require lower depth thresholds to accurately represent snow extent. However, such thresholds may approach the measurement uncertainty of the lidar product, increasing uncertainty in the resulting SCA estimates.

Data Availability

Meteorological observations used in this study were obtained from United States Climate Reference Network (USCRN) stations and are publicly available from the NOAA National Centers for Environmental Information (NCEI) at https://www.ncei.noaa.gov/access/crn/.

Author Contributions

Methodology: MM, AF, AH, JMJ; Data curation: AF, AH, MM; Software: MM; Formal analysis: MM, AF; Visualization: MM, AF; Writing (original draft preparation): MM, AF, AH, JMJ; Writing (review and editing): MM, JMJ

Competing Interests

The authors declare that they have no competing interests.

Acknowledgments

The authors want to thank Jeremy Johnston and Rashik Koirala who participated in the data collection which made this work possible. We also thank Lee Friess for editing this manuscript. This material is based upon work supported by the Broad Agency Announcement Program and the Cold Regions Research and Engineering Laboratory (ERDC-CRREL) under Contracts W913E521C0006 and W913E523C0004. Distribution A: Approved for public release. Distribution is unlimited.

References

Ackroyd, C., Donahue, C.P., Menounos, B. and Skiles, S.M., 2024. Airborne lidar intensity correction for mapping snow cover extent and effective grain size in mountainous terrain. GIScience & Remote Sensing, 61(1), p.2427326.

Ault, T. W., Czajkowski, K. P., Benko, T., Coss, J., Struble, J., Spongberg, A., Templin, M., and Gross, C.: Validation of the MODIS snow product and cloud mask using student and NWS cooperative station observations in the Lower Great Lakes Region, Remote sensing of Environment, 105, 341–353, 2006.

Belmonte, A., Sankey, T., Biederman, J., Bradford, J., Goetz, S., and Kolb, T.: UAV-Based Estimate of Snow Cover Dynamics: Optimizing Semi-Arid Forest Structure for Snow Persistence, Remote Sensing, 13, 1036, https://doi.org/10.3390/rs13051036, 2021.

Blöschl, G.: Scaling issues in snow hydrology, Hydrological Processes, 13, 2149–2175, https://doi.org/10.1002/(SICI)1099-1085(199910)13:14/15<2149::AID-HYP847>3.0.CO;2-8, 1999.

Breen, C., Currier, W. R., Vuyovich, C., Miao, Z., and Prugh, L.: Snow depth extraction from time-lapse imagery using a keypoint deep learning model, Water Resources Research, 60, e2023WR036682, 2024.

Cannistra, A. F., Shean, D. E., and Cristea, N. C.: High-resolution CubeSat imagery and machine learning for detailed snow-covered area, Remote Sensing of Environment, 258, 112399, https://doi.org/10.1016/j.rse.2021.112399, 2021.

Caparó Bellido, A. and Rundquist, B. C.: Semi-automatic fractional snow cover monitoring from near-surface remote sensing in grassland, Remote Sensing, 13, 2045, 2021.

Cho, E., Hunsaker, A. G., Jacobs, J. M., Palace, M., Sullivan, F. B., and Burakowski, E. A.: Maximum entropy modeling to identify physical drivers of shallow snowpack heterogeneity using unpiloted aerial system (UAS) lidar, J. Hydrol., 602, 126722, https://doi.org/10.1016/j.jhydrol.2021.126722, 2021.

Clark, M. P., Hendrikx, J., Slater, A. G., Kavetski, D., Anderson, B., Cullen, N. J., Kerr, T., Örn Hreinsson, E., and Woods, R. A.: Representing spatial variability of snow water equivalent in hydrologic and land-surface models: A review, Water Resources Research, 47, https://doi.org/10.1029/2011WR010745, 2011.

Coll, J. and Li, X.: Comprehensive accuracy assessment of MODIS daily snow cover products and gap filling methods, ISPRS J. Photogramm. Remote Sens., 144, 435–452, https://doi.org/10.1016/j.isprsjprs.2018.07.003, 2018.

Cristea, N.C., Breckheimer, I., Raleigh, M.S., HilleRisLambers, J. and Lundquist, J.D., 2017. An evaluation of terrain-based downscaling of fractional snow covered area data sets based on LiDAR-derived snow data and orthoimagery. Water Resources Research, 53(8), pp.6802-6820.

Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. Journal of Machine Learning Research, 7, 1–30.

Dietz, A., Davis, R. E., and Dozier, J.: Wide-area surface reflectance characteristics of snow: Implications for multispectral snow-cover retrievals, Int. J. Remote Sens., 33, 7753–7776, https://doi.org/10.1080/01431161.2012.701391, 2012.

Dong, C.: Remote sensing, hydrological modeling and in situ observations in snow cover research: A review, Journal of Hydrology, 561, 573–583, 2018.

Dozier, J.: Spectral signature of alpine snow cover from the landsat thematic mapper, Remote Sensing of Environment, 28, 9–22, https://doi.org/10.1016/0034-4257(89)90101-6, 1989.

Eker, R., Bühler, Y., Schlögl, S., Stoffel, A., and Aydın, A.: Monitoring of Snow Cover Ablation Using Very High Spatial Resolution Remote Sensing Datasets, Remote Sensing, 11, 699, https://doi.org/10.3390/rs11060699, 2019.

Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. Journal of the American Statistical Association, 32(200), 675–701. https://doi.org/10.1080/01621459.1937.10503522

Frei, A., Tedesco, M., Lee, S., Foster, J., Hall, D. K., Kelly, R., and Robinson, D. A.: A review of global satellite-derived snow products, Advances in Space Research, 50, 1007–1029, https://doi.org/10.1016/j.asr.2011.12.021, 2012.

Gaffey, C. and Bhardwaj, A.: Applications of Unmanned Aerial Vehicles in Cryosphere: Latest Advances and Prospects, Remote Sensing, 12, 948, https://doi.org/10.3390/rs12060948, 2020.

Gao, Y., Xie, H., Lu, N., Yao, T., and Liang, T.: Toward advanced daily cloud-free snow cover and snow water equivalent products from Terra–Aqua MODIS and Aqua AMSR-E measurements, Journal of Hydrology, 385, 23–35, https://doi.org/10.1016/j.jhydrol.2010.01.022, 2010.

Green, R. O., Dozier, J., Roberts, D., and Painter, T.: Spectral snow-reflectance models for grain-size and liquid-water fraction in melting snow for the solar-reflected spectrum, Ann. Glaciol., 34, 71–73, https://doi.org/10.3189/172756402781817042, 2002.

Hall, D. K. and Riggs, G. A.: Accuracy assessment of the MODIS snow products, Hydrological Processes: An International Journal, 21, 1534–1547, 2007.

Hall, D. K. & Riggs, G. A., 2021,. MODIS/Terra Snow Cover Daily L3 Global 0.05Deg CMG. (MOD10C1, Version 61). [Data Set]. Boulder, Colorado USA. NASA National Snow and Ice Data Center Distributed Active Archive Center. https://doi.org/10.5067/MODIS/MOD10C1.061.

Hammond, J. C., Saavedra, F. A., and Kampf, S. K.: How Does Snow Persistence Relate to Annual Streamflow in Mountain Watersheds of the Western U.S. With Wet Maritime and Dry Continental Climates?, Water Resources Research, 54, 2605–2623, https://doi.org/10.1002/2017WR021899, 2018.

Harder, P., Pomeroy, J.W. and Helgason, W.D., 2020. Improving sub-canopy snow depth mapping with unmanned aerial vehicles: lidar versus structure-from-motion techniques. The Cryosphere, 14(6), pp.1919-1935.

Helbig, N., Bühler, Y., Eberhard, L., Deschamps-Berger, C., Gascoin, S., Dumont, M., Revuelto, J., Deems, J.S. and Jonas, T., 2021. Fractional snow-covered area: scale-independent peak of winter parameterization. The Cryosphere, 15(2), pp.615-632.

Hopkinson, C., Collins, T., Anderson, A., Pomeroy, J. and Spooner, I., 2012. Spatial snow depth assessment using LiDAR transect samples and public GIS data layers in the Elbow River watershed, Alberta. Canadian Water Resources Journal/Revue Canadienne Des Ressources Hydriques, 37(2), pp.69-87.

Immerzeel, W. W., Droogers, P., de Jong, S. M., and Bierkens, M. F. P.: Large-scale monitoring of snow cover and runoff simulation in Himalayan river basins using remote sensing, Remote Sensing of Environment, 113, 40–49, https://doi.org/10.1016/j.rse.2008.08.010, 2009.

Jacobs, J. M., Hunsaker, A. G., Sullivan, F. B., Palace, M., Burakowski, E. A., Herrick, C., and Cho, E.: Snow depth mapping with unpiloted aerial system lidar observations: a case study in Durham, New Hampshire, United States, The Cryosphere, 15, 1485–1500, https://doi.org/10.5194/tc-15-1485-2021, 2021.

Jacques-Hamilton, R., Valcu, M., Kwon, E., Versluijs, T. S. L., and Kempenaers, B.: Measurement error in remotely sensed fractional snow cover datasets: implications for ecological research, Environmental Research: Ecology, 4, 015005, https://doi.org/10.1088/2752-664X/ada8b3, 2025.

Johnston, J., Jacobs, J. M., and Cho, E.: Global Snow Seasonality Regimes from Satellite Records of Snow Cover, Journal of Hydrometeorology, 25, 65–88, https://doi.org/10.1175/JHM-D-23-0047.1, 2023.

Johnston, J. M., Jacobs, J. M., Hunsaker, A., Wagner, C., and Vardaman, M.: Applications of Snow-Covered Areas from Unoccupied Aerial Systems (UAS) Visible Imagery: A Demonstration in Southeastern New Hampshire, Remote Sensing, 17, 1885, 2025.

Kaufman, Y. J., Kleidman, R. G., Hall, D. K., Martins, J. V., and Barton, J. S.: Remote sensing of subpixel snow cover using 0.66 and 2.1 μm channels, Geophys. Res. Lett., 29, 28-1–28-4, https://doi.org/10.1029/2002GL014825, 2002.

Kostadinov, T.S., Schumer, R., Hausner, M., Bormann, K.J., Gaffney, R., McGwire, K., Painter, T.H., Tyler, S. and Harpold, A.A., 2019. Watershed-scale mapping of fractional snow cover under conifer forest canopy using lidar. Remote Sensing of Environment, 222, pp.34-49.

Le, E., Janssen, J., Hammond, J., and Ameli, A. A.: The persistence of snow on the ground affects the shape of streamflow hydrographs over space and time: a continental-scale analysis, Frontiers in Environmental Science, 11, 2023.

Liang, H., Huang, X., Sun, Y., Wang, Y., and Liang, T.: Fractional Snow-Cover Mapping Based on MODIS and UAV Data over the Tibetan Plateau, Remote Sensing, 9, 1332, https://doi.org/10.3390/rs9121332, 2017.

Liston, G. E.: Interrelationships among Snow Distribution, Snowmelt, and Snow Cover Depletion: Implications for Atmospheric, Hydrologic, and Ecologic Modeling, Journal of Applied Meteorology (1988-2005), 38, 1474–1487, 1999.

Lundquist, J. D., Vano, J., Gutmann, E., Hogan, D., Schwat, E., Haugeneder, M., Mateo, E., Oncley, S., Roden, C., and Osenga, E.: Sublimation of snow, Bulletin of the American Meteorological Society, 105, E975–E990, 2024.

Moradi, M., Cho, E., Jacobs, J. M., and Vuyovich, C. M.: Seasonal soil freeze/thaw variability across North America via ensemble land surface modeling, Cold Regions Science and Technology, 209, 103806, 2023.

NASA Land Processes Distributed Active Archive Center (LP DAAC). MOD09GA v061: MODIS/Terra Surface Reflectance Daily Global 500 m. NASA EOSDIS.

Niedzielski, T., Spallek, W., and Witek-Kasprzak, M.: Automated Snow Extent Mapping Based on Orthophoto Images from Unmanned Aerial Vehicles, Pure and Applied Geophysics, 175, 3285–3302, https://doi.org/10.1007/s00024-018-1843-8, 2018.

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.

Nolin, A. W.: Recent advances in remote sensing of seasonal snow, J. Glaciol., 56, 1141–1150, https://doi.org/10.3189/002214311796406077, 2010.

Painter, T. H., Davis, R. E., Rittger, K., and Dozier, J.: Multiple endmember spectral mixture analysis of hyperspectral imaging spectrometer data with application to snow cover, Remote Sens. Environ., 85, 598–610, https://doi.org/10.1016/S0034-4257(03)00015-8, 2003.

Painter, T. H., Rittger, K., McKenzie, C., Slaughter, P., Davis, R. E., and Dozier, J.: Retrieval of subpixel snow-covered area, grain size, and albedo from MODIS, Remote Sens. Environ., 113, 868–879, https://doi.org/10.1016/j.rse.2009.01.001, 2009.

Panchard, T., Broennimann, O., Gravey, M., Mariethoz, G., and Guisan, A.: Snow cover persistence as a useful predictor of alpine plant distributions, Journal of Biogeography, n/a, https://doi.org/10.1111/jbi.14689, 2023.

Pettorelli, N., Laurance, W. F., O'Brien, T. G., Wegmann, M., Nagendra, H., and Turner, W.: Satellite remote sensing for applied ecologists: opportunities and challenges, J. Appl. Ecol., 51, 839–848, https://doi.org/10.1111/1365-2664.12261, 2014.

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.

Pomeroy, J., Gray, D., Brown, T., Hedstrom, N., Quinton, W., Granger, R., and Carey, S.: The cold regions hydrological model: a platform for basing process representation and model structure on physical evidence, Hydrological Processes: An International Journal, 21, 2650–2667, 2007.

Portenier, C., Hüsler, F., Härer, S., and Wunderle, S.: Towards a webcam-based snow cover monitoring network: methodology and evaluation, The Cryosphere, 14, 1409–1423, 2020.

Rittger, K., Painter, T.H. and Dozier, J., 2013. Assessment of methods for mapping snow cover from MODIS. Advances in Water Resources, 51, pp.367-380.

Salomonson, V.V. and Appel, I., 2004. Estimating fractional snow cover from MODIS using the normalized difference snow index. Remote sensing of environment, 89(3), pp.351-360.

Salvatori, R., Plini, P., Giusto, M., Valt, M., Salzano, R., Montagnoli, M., Cagnati, A., Crepaz, G., and Sigismondi, D.: Snow cover monitoring with images from digital camera systems, Ital. J. Remote Sens, 43, 137–145, 2011.

Sturm, M.: White water: Fifty years of snow research in WRR and the outlook for the future, Water Resources Research, 51, 4948–4965, 2015.

Thaler, E., Crumley, R. and Bennett, K., 2024. Estimating snow cover from high-resolution satellite imagery by thresholding blue wavelengths: Supporting Data (No. https://doi. org/10.5440/2294083; NGA312). Next Generation Ecosystems Experiment-Arctic, Oak Ridge National Laboratory (ORNL), Oak Ridge, TN (US).

Verfaillie, M., Cho, E., Dwyre, L., Khan, I., Wagner, C., Jacobs, J. M., and Hunsaker, A.: UAS remote sensing applications to abrupt cold region hazards, Frontiers in Remote Sensing, 4, 1095275, https://doi.org/10.3389/frsen.2023.1095275, 2023.

Warren, S. G.: Optical properties of snow, Rev. Geophys., 20, 67–89, https://doi.org/10.1029/RG020i001p00067, 1982.

Walland, D. J. and Simmonds, I.: Modelled atmospheric response to changes in Northern Hemisphere snow cover, Climate Dynamics, 13, 25–34, https://doi.org/10.1007/s003820050150, 1996.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. Biometrics Bulletin, 1(6), 80–83. https://doi.org/10.2307/3001968

Wolter, P., Berkley, E., Peckham, S., Singh, A., and Townsend, P.: Exploiting tree shadows on snow for estimating forest basal area using Landsat data, Remote Sensing of Environment, 121, 69–79, 2012.

Zhang, T.: Influence of the seasonal snow cover on the ground thermal regime: An overview, Reviews of Geophysics, 43, https://doi.org/10.1029/2004RG000157, 2005.