Supplement of

A nine-year record of slush on the Greenland Ice Sheet

Emily Glen et al.

Correspondence to Emily Glen (e.glen@lancaster.ac.uk)

Supplementary Methods

S1: Comparison of RF classifier to thresholding methods

To validate our RF classifier, we compared a subset of our RF-classified slush results to those using two threshold-based methods. We pre-processed all images tested in this section using the same masking approach outlined in Section 2.2.2 in the main text.

S1.1. Normalised Difference Water Index

As stated in the main text (Section 2.5), two NDWI-based indices are often used to detect meltwater features on glacier ice: NDWI standard (e.g. McFeeters, 1996; Box and Ski, 2007) and $NDWI_{ice}$ (Equation 1; main text), an adaptation optimised for ice- and firn-covered surfaces (e.g. Yang and Smith, 2012). NDWI standard is calculated from the green and near-infrared bands, as defined in Equation S1:

$$NDWI_{Standard} = \frac{B3-B8}{B3+B8} \quad \text{(S1)}$$

Slush has been detected using an $NDWI_{ice}$ threshold alone, which is generally lower than those applied to open water and lakes. For example, Williamson et al. (2018) identified lakes using a threshold of > 0.25 , whereas slush is typically detected at lower $NDWI_{ice}$ values. For the GrIS, Yang and Smith (2012) first proposed classifying shallow water and slush using a threshold range of $0.12 < NDWI_{ice} < 0.14$ for both classes, which were later adopted by Bell et al. (2017) for the AIS. More recently, Glen et al. (2025) showed that combining both types of NDWI indices with respective thresholds, applying NDWI_standard > 0.14 alongside $NDWI_{ice} > 0.15$, improved slush detection on the GrIS while reducing misclassification with other meltwater features. A similar benefit of combining NDWI indices was also reported by Corr et al. (2021) in their mapping of supraglacial lakes on the GrIS.

In this study, we compared our RF classification results to NDWI threshold-based methods following Glen et al. (2025). Specifically, slush was identified using the combined thresholds of $NDWI_{ice} > 0.15$ (Equation 1) and NDWI_standard > 0.14 (Equation S1). All evaluation metrics presented below (and in Figure S1) therefore relate to this combined-threshold method.

S1.2. Greenness Index

We also compared our RF results to the Greenness Index (G_{ind}) developed by Covi et al. (2022). This index enhances green reflectance relative to blue and red, while reducing cloud contamination through the inclusion of the shortwave infrared band. High values of G_{ind} indicate strong green reflectance compared to other visible bands, which produces a blue hue in true-colour composites and is characteristic of surface water (Covi et al., 2022). The shortwave infrared (SWIR; B12) is used to filter out clouds: since clouds are highly reflective in SWIR, high B12 values lower G_{ind}, reducing the likelihood of clouds being misclassified as slush.

G_{ind} is defined in Equation S2:

$$G_{ind} = \frac{3 \cdot B3}{B2+B3+B4} - 2 \cdot B12 \quad \text{(S2)}$$

In our study, we applied a threshold of $G_{ind} > 1$ to classify slush. This value was determined through iterative testing against Sentinel-2 imagery. We found that > 1 provided the best match to visually identifiable slush areas.

S1.3. Analysis of evaluation metrics

Agreement between our RF classifier and the NDWI and G_{ind} methods was assessed through analysis of spatial overlap and classification metrics (Tables S4 and S5). Overall, RF showed stronger correspondence with NDWI than with G_{ind}. Spatial overlap between RF and NDWI was highest in the SW (733 km²) and the CW (41 km²), while overlap between RF and G_{ind} was lower, peaking at 436 km² in the SW and just 11 km² in the CW. Classification metrics reflected similar trends (Tables S4 and S5). For example, F1-scores were generally higher for RF compared to NDWI (0.56–0.76), indicating stronger overall agreement than RF compared to G_{ind} (0.51-0.68).

Visual comparisons between the three methods supported the quantitative results described above. In the CW region (Figure S1a), RF and NDWI were closely aligned spatially, while G_{ind} generally underestimated areas of slush and misclassified lake edges. However, in the NO (Figure S1b), NDWI overestimated areas of slush due to cloud contamination, which our RF classifier was not affected by. In the SW (Figure S1c), the RF classifier identified more extensive slush than either threshold method. G_{ind} again showed lower agreement, including some misclassification of shallow lakes compared to both RF and NDWI, while NDWI occasionally misclassified bright surfaces as melt. Overall, RF provided more spatially consistent results, particularly in noisy or complex conditions.

S2. Limitations of our RF classifier

Our results shown in Section S1 above as well as the main paper highlight the strength of RF in capturing the spectral expression of slush. That said, our approach has limitations, for example, we observe occasional overestimation in areal extents of slush (e.g., in the SW in 2019; Figure S1c), and some subjectivity in the manual clustering step (see Section 2.2). Despite cloud and terrain masking, commission errors may occur in certain conditions (e.g., cloudy scenes and regions with complex topography) and low sun angles can obscure melt features, especially in early and late summer. Even so, relative to previous studies, our RF approach offers a major step forward in reliably mapping slush across Greenland, with strong potential for continued refinement and wider application. Improving cloud masking, incorporating time-series filtering, and exploring multi-sensor approaches offer promising paths to enhance future slush detection and better constrain the full extent of meltwater saturation.

Supplementary Figures

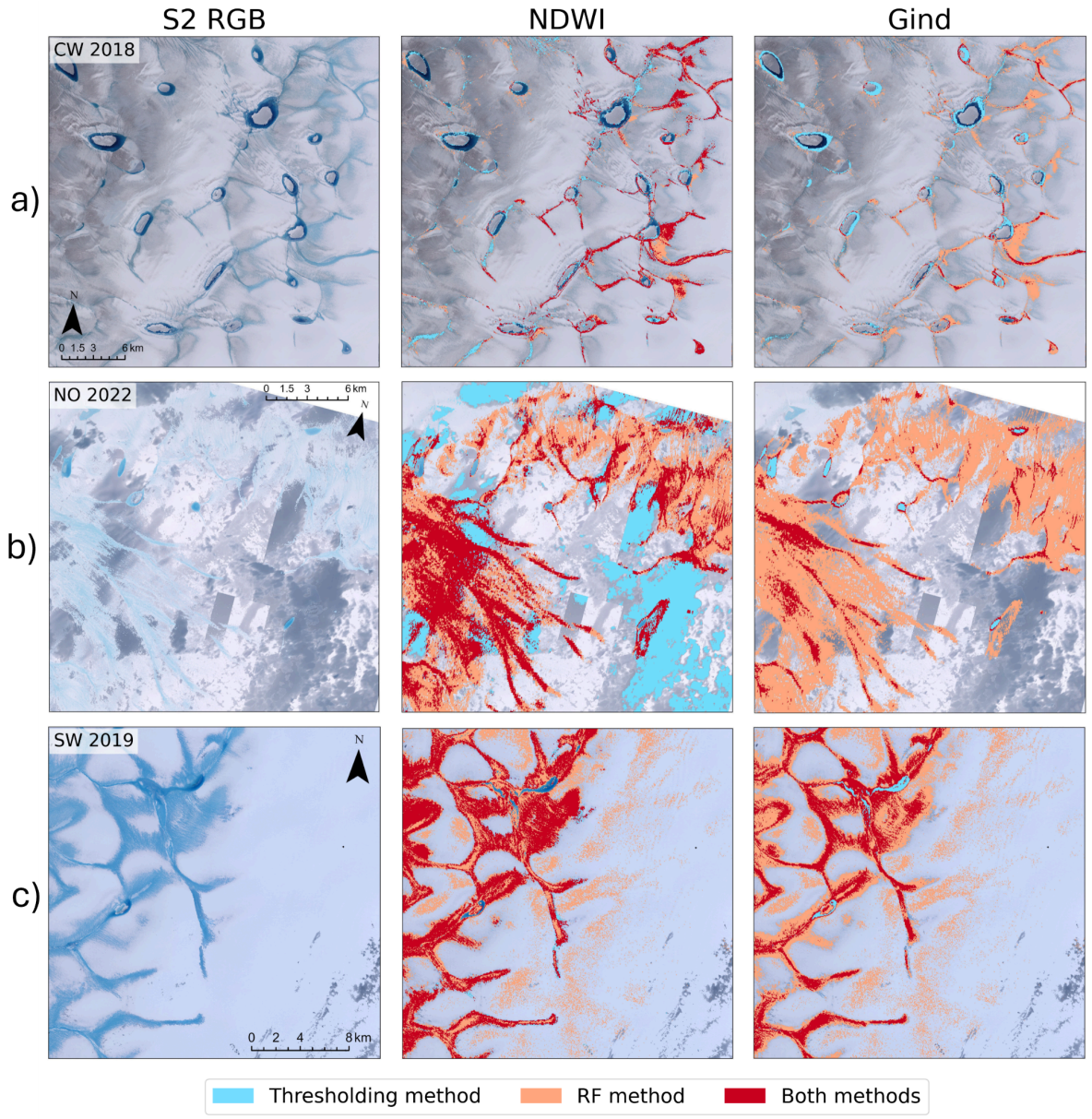

Figure S1. *Visual comparison of our RF method with both NDWI (NDWI standard and $NDWI_{ice}$ combined method; Glen et al., 2025) and G_{ind} (Covi et al., 2022) thresholding methods for slush detection, applied to S2 optical imagery across three GrIS* basins *and years: a) CW in 2018, b) NO in 2022, c) SW in 2019 (Figure 2; main text). The left column displays the original S2 RGB images for reference. The middle column shows NDWI-classified slush compared to RF-classified slush, and the right column shows G_{ind}-classified slush vs RF classified slush. Red indicates slush detected by both RF and the respective thresholding method, orange represents slush identified only by RF, and blue indicates slush identified only by the thresholding method. Of note, the NO 2022 image was intentionally selected as a complex case due to significant cloud cover and artifacts such as visible seams and spectral discontinuities between adjacent scenes resulting from the mosaicking process.*

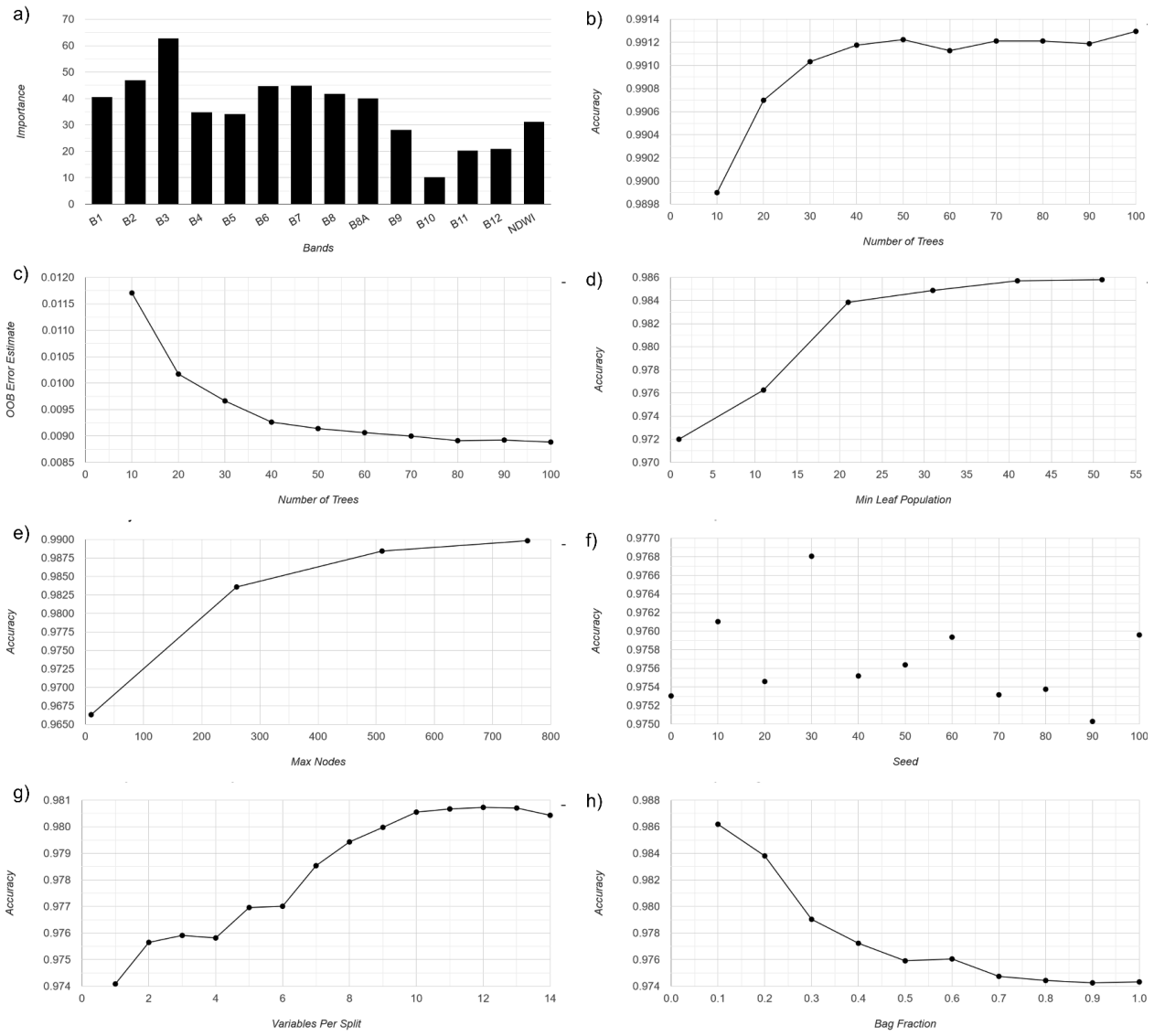

Figure S2. *Hyperparameter tuning results for the accuracy of the RF classifier in our study. a) The feature importance for the selected bands used in classification. b) The effect of the number of trees on model accuracy. c) The OOB error estimate as a function of the number of trees. d) The accuracy in relation to the minimum leaf population. e) The accuracy versus the maximum number of nodes per tree; f) The effect of different random seed values on accuracy. g) The accuracy as it relates to the number of variables considered per split. h) The accuracy versus the bag fraction.*

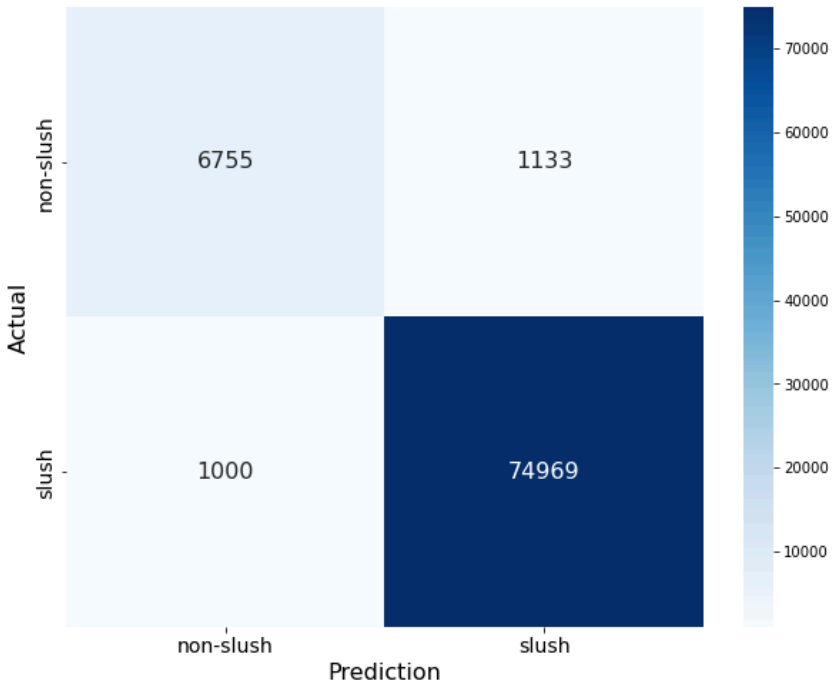

Figure S3. *Confusion matrix for the RF classifier applied to slush detection. The matrix compares predicted versus actual values for two classes: "slush" and "non-slush." The vertical axis represents the actual labels, while the horizontal axis represents the predicted labels. The colour bar on the right indicates the number of observations, with darker shades representing higher counts.*

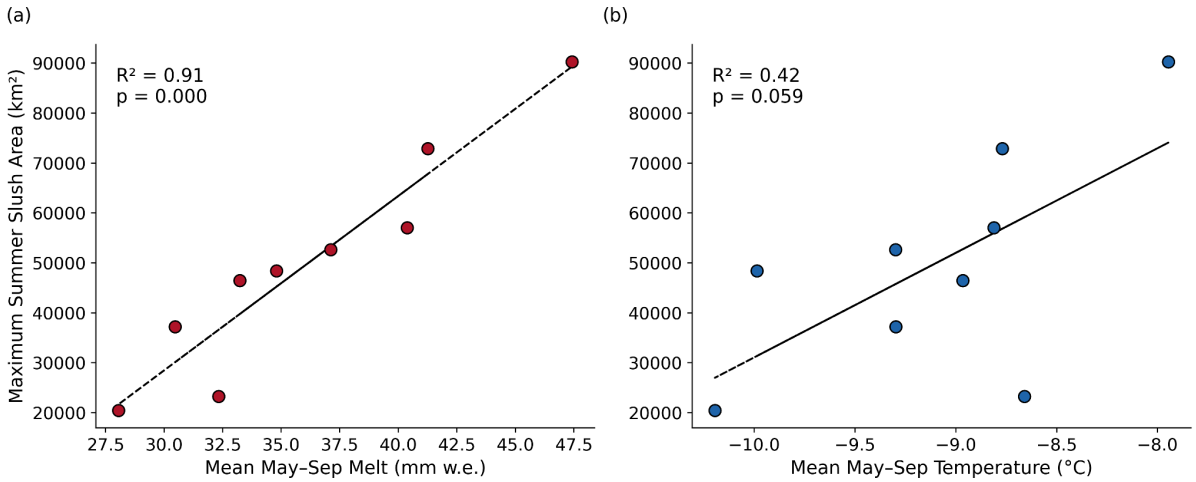

Figure S4. *For the entire GrIS, relationships between maximum summer slush area (km²) and RACMO-derived mean May–September (a) surface melt (mm w.e.) and (b) 2 m air temperature (°C) from 2016–2024. Each panel shows a linear regression with the corresponding coefficient of determination (R²) and p-value.*

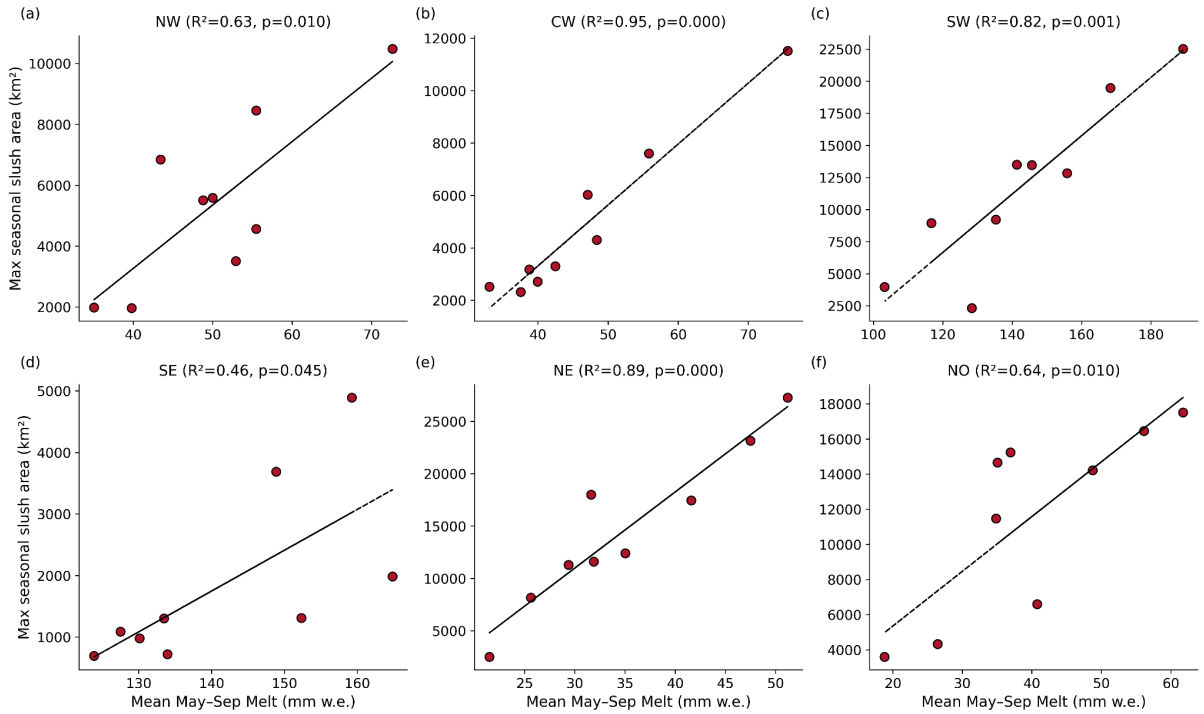

Figure S5. *For all six drainage basins of the GrIS (a-f), the relationship between maximum seasonal (i.e. May–September) slush area (km²) and mean seasonal snowmelt (mm w.e.) from RACMO for the nine melt seasons from 2016–2024. Each panel shows a linear regression with the corresponding coefficient of determination (R²) and p-value.*

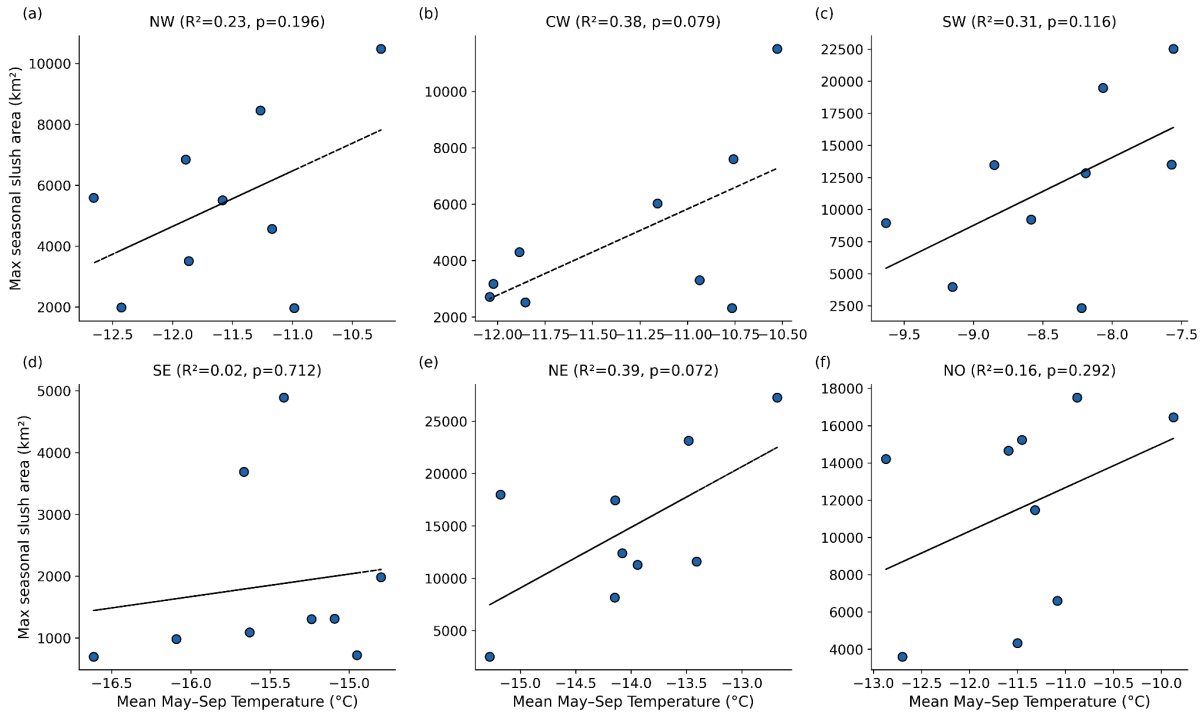

Figure S6. *For all six drainage basins of the GrIS (a-f), relationship between maximum seasonal (i.e. May–September) slush area (km²) and mean seasonal 2 m air temperature (°C) from RACMO for the nine melt seasons from 2016–2024. Each panel shows a linear regression with the corresponding coefficient of determination (R²) and p-value.*

Supplementary Tables

Table S1. *List of selected S2 images used for training and validation across basins. Image IDs correspond to GEE asset identifiers from the COPERNICUS/S2 collection (Gorelick et al., 2017).*

Image Date	***Sentinel 2 Image ID***
2016-05-03	COPERNICUS/S2/20160503T180920_20160503T231352_T20XMP
2016-05-16	COPERNICUS/S2/20160516T145922_20160516T150205_T22WEA
2017-06-16	COPERNICUS/S2/20170616T151911_20170616T152036_T22WED
2017-08-15	COPERNICUS/S2/20170815T151911_20170815T152131_T22WFB
2018-06-04	COPERNICUS/S2/20180604T143919_20180604T143920_T26XNJ
2018-09-13	COPERNICUS/S2/20180913T145911_20180913T145910_T26XNM
2019-07-13	COPERNICUS/S2/20190713T150921_20190713T150916_T26XNP
2019-07-21	COPERNICUS/S2/20190721T151809_20190721T151812_T23WM
2019-08-01	COPERNICUS/S2/20190801T144759_20190801T144759_T22WFV
2019-08-02	COPERNICUS/S2/20190802T164901_20190802T164904_T26XMP
2020-06-13	COPERNICUS/S2/20200613T161829_20200613T161831_T21XWD
2020-07-26	COPERNICUS/S2/20200726T180919_20200726T180916_T20XMN
2021-06-21	COPERNICUS/S2/20210621T144749_20210621T144900_T22WES
2021-08-05	COPERNICUS/S2/20210805T135729_20210805T135732_T25WFS
2021-08-22	COPERNICUS/S2/20210822T143931_20210822T144327_T22WFS
2021-08-23	COPERNICUS/S2/20210823T141011_20210823T141007_T24WWV
2022-07-04	COPERNICUS/S2/20220704T140739_20220704T140740_T25WDS
2022-08-16	COPERNICUS/S2/20220816T164851_20220816T164854_T25XDL
2022-08-25	COPERNICUS/S2/20220825T180919_20220825T181113_T19XEH
2023-05-01	COPERNICUS/S2/20230501T142741_20230501T142742_T23WPN
2023-07-02	COPERNICUS/S2/20230702T150801_20230702T150959_T22WEC
2023-07-03	COPERNICUS/S2/20230703T161831_20230703T161832_T21XWC
2023-07-16	COPERNICUS/S2/20230716T185919_20230716T185920_T21XWK
2023-08-15	COPERNICUS/S2/20230815T171859_20230815T171900_T20XML

Table S2. *Hyperparameter settings for the RF classifier, detailing our chosen values for key parameters*

Hyperparameter	Value	Description
Number of Trees	50	Controls the number of trees in the forest. More trees improve stability and accuracy but increase computational demand.
Bag Fraction	0.7	Fraction of the training data used for each tree. A lower value increases variance in the predictions made by individual trees, while a higher value reduces it but may introduce bias.
Minimum Leaf Population	20	Minimum number of samples required at a leaf node to prevent overfitting by ensuring sufficient data per leaf.
Maximum Nodes	500	Limits the number of nodes per tree, balancing complexity and simplicity.
Seed	30	Ensures reproducibility by initialising the random number generator used for tree building.
Variables per Split	10	Controls the number of randomly chosen variables that are considered at each split in a tree.

Table S3. *Performance metrics derived from the confusion matrix of the RF classifier (Figure S3) used for slush detection.*

	Precision (%)	**Recall (%)**	**F1-score (%)**	**True Positives (TP)**	**False Positives (FP)**	**False Negatives (FN)**	**True Negatives (TN)**
Non-slush	*85.6*	*87.1*	*86.4*	*6,755*	*1,133*	*1,000*	*74,969*
Slush	*98.7*	*98.5*	*98.6*	*74,969*	*1,000*	*1,133*	*6,755*

Table S4. *Performance metrics (Recall, Precision, F1-Score) and spatial overlap areas of slush for the comparison of RF vs. NDWI classifications across all six GrIS* basins.

RF VS. NDWI										
Image	**Region**	**Date**	**Overall Agreement**	**Kappa**	**Precision**	**Recall**	**F1-Score**	**RF Area (km²)**	**NDWI Area (km²)**	**Overlap Area (km²)**
1	CW	Aug-18	0.96	0.52	0.79	0.73	0.76	86.04	67.82	41.33
2	SW	Aug-19	0.68	0.35	0.79	0.68	0.64	2019.85	758.16	733.17
3	NW	Aug-21	0.82	0.31	0.66	0.65	0.66	134.74	126.68	54.77
4	NO	Aug-22	0.77	0.29	0.66	0.64	0.65	353.22	299.01	142.12
5	SE	Aug-23	0.85	0.16	0.76	0.55	0.56	400.55	71.90	47.15
6	NE	Aug-24	0.9	0.25	0.71	0.59	0.62	430.4	176.25	88

Table S5. *Performance metrics (Recall, Precision, F1-Score) and spatial overlap areas of slush for the comparison of RF vs. Gind classifications across the all GrIS* basins.

RF VS. G_{ind}										
Image	**Region**	**Date**	**Overall Agreement**	**Kappa**	**Precision**	**Recall**	**F1-Score**	**RF Area (km²)**	**G_{ind} Area (km²)**	**Overlap Area (km²)**
1	CW	Aug-18	0.94	0.16	0.64	0.56	0.58	86.04	33.19	11.06
2	SW	Aug-19	0.61	0.21	0.75	0.6	0.54	2019.85	461.64	435.64
3	NW	Aug-21	0.87	0.38	0.78	0.65	0.68	134.74	66.21	44.5
4	NO	Aug-22	0.79	0.1	0.84	0.53	0.51	353.22	27.97	25.03
5	SE	Aug-23	0.85	0.1	0.77	0.53	0.52	400.55	39.26	27.14
6	NE	Aug-24	0.91	0.14	0.78	0.54	0.56	430.4	61.76	39.98