https://doi.org/10.5194/essd-2026-406
Preprint. Discussion started: 18 June 2026

Open Access Earth System Science Data Discussions

A benchmark deep learning dataset for the classification of supraglacial lake drainage mechanism across the central-west Greenland Ice Sheet

Joshua H. Rines[1], Ching-Yao Lai[1], Ellianna Abrahams[1,2], Michael G. Shahin[3,4], Niall B. Coffey[1], Eojin Lee[1], and Laura A. Stevens[5]

[1]Department of Geophysics, Stanford University, Stanford, CA 94305, USA
[2]SDSS Center for Computation, Stanford University, Stanford, CA 94305, USA
[3]Department of Earth and Environmental Sciences, University of Pennsylvania, Philadelphia, PA 19104, USA
[4]Center for Remote Sensing and Integrated Systems, University of Kansas, Lawrence, KS 66045, USA
[5]Department of Earth Sciences, University of Oxford, Oxford, UK

Correspondence: Joshua H. Rines (jrines@stanford.edu)

Abstract. Supraglacial lakes on the Greenland Ice Sheet drain through physically distinct pathways: hydrofracture, moulins, lateral stream routing, and crevasse-fields. Each drainage mechanism carries unique implications for ice sheet dynamics. Existing automated classifications reduce each lake's drainage behavior to a time-series of scalar values representing the observed water surface-area and classify each lake based on drainage rate (e.g., rapid vs. slow). This scalar reduction conflates physically different drainage mechanisms, which can only be determined through consideration of full spatio-temporal tracking. Here we introduce a human-benchmarked, machine learning-ready benchmark dataset that pairs full Sentinel-2 multispectral satellite imagery time series with human-expert-labels assigned for $N = 1679$ supraglacial lakes in the central-west basin of the Greenland Ice Sheet during the 2018 ($n = 679$) and 2019 ($n = 1000$) melt seasons. The dataset is formatted as per-lake CF-1.8 NetCDF files each containing: six Sentinel-2 reflectance bands at 10 meter spatial resolution and daily cadence over the 153 day melt season (1 May to 30 September); a per-pixel binary cloud mask; co-registered lake water masks (both static and dynamic); and the human-assigned drainage classification labels. We accompany the dataset with a baseline deep learning classifier, demonstrating the utility of the dataset both in deep learning workflows and in extending lake drainage classification from rate-based to mechanism-based. The dataset is released through the Stanford Digital Repository under a CC BY 4.0 license, and the accompanying open-source `sat-tile-stack` preprocessing software under an MIT license.

1 Introduction

The surface of the Greenland Ice Sheet (GrIS) experiences seasonal partial melting from approximately May to September, producing a network of supraglacial meltwater features, namely streams and supraglacial lakes (henceforth 'lakes'). This surface meltwater influences ice sheet surface mass balance through direct mass removal via runoff, surface albedo feedbacks, and dynamical feedbacks on ice flow (Chu, 2014). The last of these has drawn particular attention from the glaciological community because some lakes are observed to drain directly into the ice via hydraulic fracturing, or hydrofracture, in which the entire lake volume propagates a fracture through the full ice thickness. This fresh surface-to-bed pathway transports surface meltwater to the bed over a matter of hours to days (e.g., Das et al., 2008; Tedesco et al., 2013). Once at the bed, this meltwater interacts with the subglacial hydrological system, reducing frictional coupling between the ice and the bed, and thereby increasing transient ice flow velocities throughout the melt season (e.g., Zwally et al., 2002; Bartholomaus et al., 2008; Bartholomew et al., 2011; Hoffman et al., 2011; Stevens et al., 2016). On shorter timescales, the basal traction loss due to surface-to-bed transfer of meltwater perturbs the stress state of the overlying ice (Stevens et al., 2015, 2024; Rines et al., 2026b; Stevens et al., 2026), which is hypothesized to be responsible for the hydrofracturing of other surrounding lakes (Christoffersen et al., 2018). As such, the dynamical consequences of these events span timescales from hours to decades: extending from the immediate velocity and stress pulses following a single drainage event, to the seasonal evolution of the subglacial drainage system that lake-drainage events help shape (Nienow et al., 2017; Andrews et al., 2014).

As dynamic as it is, hydrofracture is not the only pathway by which surface meltwater reaches the bed. Moulins, which are vertical conduits that may persist across melt seasons (Catania and Neumann, 2010), are responsible for routing the majority of supraglacial meltwater to the subglacial system across large portions of the GrIS ablation zone (Smith et al., 2015; Koziol et al., 2017). Moulins and hydrofracture events are not always independent phenomena: hydrofracture is a principal mechanism by which new moulins form (Hoffman et al., 2018). Once formed, these moulins can continue to deliver meltwater to the bed for the remainder of the melt season and even reactivate in subsequent years, serving as a persistent surface-to-bed meltwater pathway (Catania and Neumann, 2010; Banwell et al., 2016; Andrews et al., 2022). Notably, Hoffman et al. (2018) estimates that the majority of moulins in their study area (which is within the central-west region of the GrIS that we consider here) could have only formed under the elevated tensile stress produced during lake hydrofracture drainage events. Thus, hydrofracture-driven lake drainage events, and the moulins they create, are a primary control on the spatial distribution of surface-to-bed connections, and therefore exert a control on the dynamical behavior and stability of the ice sheet as a whole.

Lakes also exhibit other terminal behavior through the melt season, and these other drainage mechanisms each carry their own physical significance. Lakes may drain laterally, whereby water overtops its basin rim and routes meltwater through supraglacial stream channels to downstream features (e.g., another lake, a moulin, or off the ice sheet edge). This type of drainage redistributes water within the supraglacial hydrological network and connects to the water budgets of neighboring lakes (Catania et al., 2008; Banwell et al., 2013). Lakes may also be situated in crevasse fields, many of which experience drainage into that field of pre-existing fractures, either through a dispersed drainage or sometimes a sudden distinct hydrofracture, and represent a distinct mode of meltwater transport relevant for ice sheet dynamics (Lampkin et al., 2013; Chudley et al., 2021, 2025). Finally, some lakes exist through the full melt season without experiencing a drainage of any type. Due to their catchment topography, these lakes are hydrologically isolated from the broader supraglacial network for the duration of the melt season, though they continue to receive melt inputs from surrounding ice melting. Properly distinguishing among all five of these drainage-mechanism outcomes is critical for resolving distinct physical processes with distinct consequences for ice dynamics, surface hydrology, and the distribution and formation of surface-to-bed drainage pathways.

Though accurate identification and tracking of these drainage events at scale is central to constraining GrIS hydrology and surface mass balance, classifying these events is far from a trivial task. Previous classification work has relied predominantly on observable time-window thresholds applied to coarse spatial resolution optical imagery whereby a lake in which the surface area decreases by a specified fraction within a fixed time window (typically 2 to 6 days) was classified as rapidly draining (Selmes et al., 2011, 2013; Morriss et al., 2013; Doyle et al., 2014; Williamson et al., 2018b). Across the existing literature, these drainage rate-based classifiers estimate the share of lakes that drain rapidly each year ranges from $\sim 13\%$ to $\sim 45\%$, depending on the ice-sheet region, year, and detection method, with rapidly draining lakes representing a clear minority of the lake population (Selmes et al., 2013; Williamson et al., 2018b; Cooley and Christoffersen, 2017; Dunmire et al., 2021, 2025). While these threshold-based approaches established the foundational statistics for rapid drainage events, they conflate distinct physical processes under a single behavioral label. For an ice-sheet modeler, these rapid events would most naturally be attributed to hydrofracture. Recently, Dunmire et al. (2025) extended this body of work by training time series classification models, such as long short-term memory (LSTM) recurrent neural network architectures (Hochreiter and Schmidhuber, 1997), on combined Sentinel-1 microwave backscatter and Sentinel-2 optical reflectance time series (Torres et al., 2012; Drusch et al., 2012). In doing so Dunmire et al. (2025) developed an automated classification for nearly 10,000 GrIS lakes across the 2018 and 2019 melt seasons, classifying each individual lake into one of four categories: refreezing, rapid drainage, slow drainage, and buried.

A fundamental limitation persists, however, across all current approaches: they reduce the behavior of each lake to a scalar time series of the observed surface area, and classify lake type from the resulting one-dimensional time-series sequence, henceforth called a water surface-area curve. Though this approach aligns with the rate-based classification taxonomy, this scalar projection discards important geomorphometric contextual information that is essential to distinguishing underlying drainage mechanisms (e.g., hydrofracture) from drainage rate (e.g., rapid). Two lakes with visually indistinguishable scalar time series can drain via different physical processes, leading to fundamentally different hydrological and ice dynamical outcomes with downstream effects for understanding ice sheet evolution and sea level rise contributions. For example, a hydrofracture event and a lateral drainage can both produce a sharp drop in lake area over a single day (e.g., Figure 4 c), but only the former delivers water from the surface to the bed at the location of the lake. Recently, Ryan et al. (2026) demonstrated, using high-resolution satellite imagery, that cascading rapid lake drainages may be driven by supraglacial hydrological connections (streams), further motivating the need for lake drainage classifications based on full spatio-temporal context extractable from satellite imagery time series.

The information to separate these processes is not described solely in the surface-area curve, but rather the full, transient morphologic and morphometric context of the lake and its surroundings. This necessary information is preserved in multispectral imagery time series which can be obtained at near regular cadence from public satellite datasets. Thus, resolving drainage mechanism from satellite observations therefore requires both (i) a labeled time-series dataset in which lakes are characterized by mechanism rather than drainage rate, with the underlying spatiotemporal context preserved, and (ii) a manual labeling protocol that can be operationalized into an automated, deep-learning-based pipeline capable of extending the dataset beyond what hand-labeling alone can feasibly support. In this paper, we present the framework for achieving both of these goals directly from publicly available datasets.

https://doi.org/10.5194/essd-2026-406
Preprint. Discussion started: 18 June 2026

Here, we introduce a benchmark dataset for supraglacial lake drainage mechanism classification on the GrIS. The dataset covers lakes in the central-west (CW) GrIS drainage basin (Rignot and Mouginot, 2012) during the 2018 and 2019 melt seasons, which were cooler and warmer melt seasons, respectively. Specifically, the ice sheet surface meltwater volume nearly doubled between these subsequent years, with 3.55 km^3 in 2018 compared to 6.84 km^3 in 2019 (Zheng et al., 2023; Dunmire et al., 2025). The GrIS-wide supraglacial lake population grew by 60% from 3846 to 6146 lakes, with total lake surface area more than doubling from 1242 km^2 to 2569 km^2 (Dunmire et al., 2025). The dataset thus retains the two years as distinct populations. The dataset provides per-lake labeled NetCDF files which are readily ingestible by machine learning (ML) models. Lakes were manually labeled into five drainage mechanism categories based on interpretation of the imagery time series: no drainage (ND), hydrofracture (HF), moulin drainage (MD), lateral drainage (LD), and crevasse drainage (CD). The dataset spans two contrasting melt seasons in the same drainage basin, which together enable a direct test of cross-year generalization (i.e., training on labeled lakes from one melt season and evaluating on labeled lakes from another). To our knowledge, this is the first publicly available dataset that pairs full spatiotemporal imagery with drainage mechanism labels for GrIS supraglacial lakes, and the first to frame drainage mechanism classification as a benchmark task for deep learning on multi-modal Earth observation data.

Manual labeling is tractable for the 1679 lakes presented here, but becomes quickly untenable at the scale required for GrIS-wide, multi-year inference. Dunmire et al. (2025) identified 3846 and 6146 lakes in the 2018 and 2019 melt seasons, respectively, across the full GrIS, and lake populations and distributions continue to shift from year to year as melt extent and inland expansion of lake presence evolve with the warming climate (e.g., Leeson et al., 2015; Tedstone and Machguth, 2022; Fan et al., 2025). With hand-labeling of each lake taking a human expert ~ 2 min to complete, the ~ 5000 lakes per year puts the labeling task at > 160 per-labeler human hours of labeling effort annually. We therefore accompany the dataset with a baseline spatiotemporal deep learning classification model which ingests imagery sequences and outputs a predicted drainage mechanism. This approach expands the classified dataset population with each new year, without requiring an extensive and laborious manual labeling effort. We have also developed a custom paradigm for manually labeling the satellite imagery time series (SITS) tasks, which we share in this paper accompanied by an open-source application to enable ease-of-implementation for future labelers.

Machine learning has been shown to be effective in classifying surface hydrological features in polar glacial settings. In particular, ML methods have been developed for mapping surface meltwater (e.g., lakes, slush, ponds, and supraglacial channels) using Sentinel-1, Sentinel-2, and Landsat 8 imagery across Antarctic ice shelves (e.g., Dirscherl et al., 2020, 2021; Dell et al., 2022; Corr et al., 2022). This has enabled tracking the evolution of Antarctic meltwater extent through time across the available satellite record (e.g., Dirscherl et al., 2021; Dell et al., 2024; Baumhoer et al., 2026). ML models have also been demonstrated as powerful tools for detecting and tracking the evolution of supraglacial lakes across the Greenland Ice Sheet. Notably, Wang and Sugiyama (2024) deployed a random forest algorithm to map and track the evolution of supraglacial lake sizes for Heilprin and Tracy glaciers in north west Greenland from 2014 to 2021. Dunmire et al. (2025) used a deep learning algorithm to classify lake drainage events based on a dataset of temporal lake extent evolution across Greenland in the 2018 and 2019 melt seasons. As an extension of this foundational work, we build a dataset from satellite imagery time series, which captures full spatio-

https://doi.org/10.5194/essd-2026-406
Preprint. Discussion started: 18 June 2026

temporal context of supraglacial lake drainages. This dataset enables the classification of drainage mechanism of Greenland supraglacial lakes, something that has yet to be automated in the existing literature.

The remainder of this paper is organized as follows. Section 2 describes the construction of the dataset, including the input data sources, the `sat-tile-stack` imagery preprocessing pipeline, the manual labeling protocol, and its associated graphical user interface (GUI) that we developed into an open-source application. Section 3 presents a baseline deep learning-based classifier model, which operationalizes automated extension of the dataset. Section 4 specifies the published dataset file format, access instructions, and license information. Section 6 discusses intended use cases of the dataset, results, and known limitations. We share our conclusions in section 7.

2 Dataset construction

To build this dataset, we combine Sentinel-2 multispectral imagery time series; co-registered scalar surface-area time series from Dunmire et al. (2021); static, per-lake lake extent boundary masks from Dunmire et al. (2025); time-evolving surface water masks we derive from the normalized difference water index (NDWI) (Sect. 2.3); and, drainage mechanism labels we assign manually. This dataset spans $N = 1679$ supraglacial lakes in the central west (CW) drainage basin of the GrIS across the 2018 and 2019 melt seasons (Figure 1). We represent each lake in a standardized ML-ready NetCDF file containing daily cadence imagery at 10 meter spatial resolution with six spectral bands from Sentinel-2: red (B04), green (B03), blue (B02), near-infrared (NIR; B08), and short wave infrared (SWIR; B11, B12) (European Space Agency, 2015).

This section describes the data sources and our dataset construction pipeline (Sect. 2.1); the per-lake imagery preprocessing we perform with our open-source `sat-tile-stack` software (Sect. 2.2); the dynamic lake masking pipeline we develop for this dataset (Sect. 2.3); the manual labeling protocol and inter-rater validation (Sect. 2.4); and the resulting dataset statistics (Sect. 2.5). All components as described below are combined into the final open-source dataset (Figure 2). We provide details for accessing the data in Sect. 4.

2.1 Data sources

The dataset draws on two primary data sources: multispectral imagery from Sentinel-2, and supraglacial lake outlines and scalar water surface-area time series from the Dunmire et al. (2025) inventory. All inputs are restricted to the CW drainage basin of the GrIS as defined by Rignot and Mouginot (2012), in order to match the same basin definition, melt seasons, lake IDs as in Dunmire et al. (2025). Figure 2 shows the compilation and structure of how we construct an example dataset sample, including the per-lake imagery time series, the scalar water surface-area time series, the static lake boundaries, the dynamic water masks, and the assigned drainage mechanism labels.

Sentinel-2 multispectral imagery.

We use Level-2A bottom-of-atmosphere reflectance from the European Space Agency's (ESA) Sentinel-2 mission (Drusch et al., 2012), accessed via Microsoft Planetary Computer's SpatioTemporal Asset Catalog (STAC) interface (Source et al.,

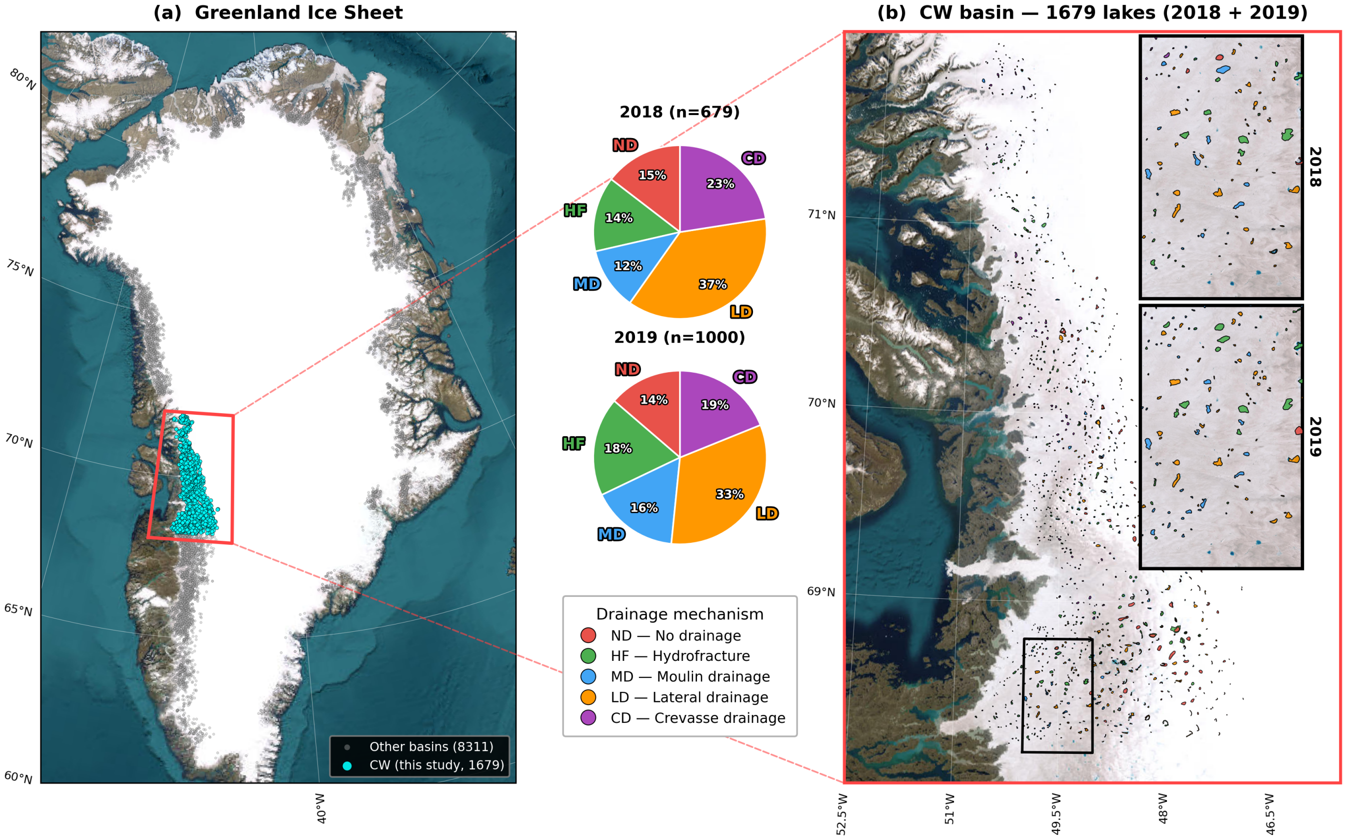
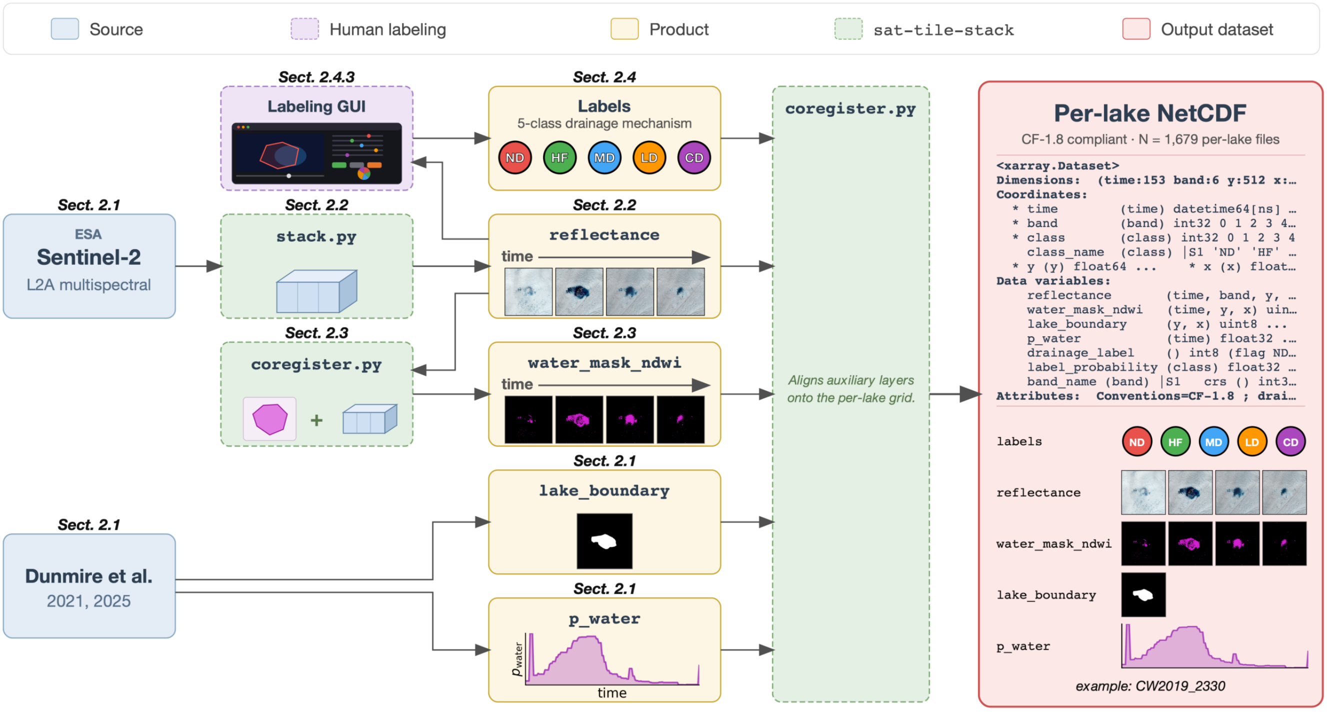


Figure 1. Study area and lake classification distributions. **(a)** All supraglacial lakes identified by Dunmire et al. (2021) across the 2018 and 2019 melt seasons within the central-west region of the Greenland Ice Sheet. The red box delineates the spatial extent of panel (b). **(b)** Zoomed-in view of the central-west region, showing the $N = 1679$ labeled lakes, plotted as their static lake boundary polygons from Dunmire et al. (2021) and colored by the human-assigned drainage mechanism class from this study: no drainage (ND, red), hydrofracture (HF, green), moulin drainage (MD, blue), lateral drainage (LD, orange), and crevasse drainage (CD, purple). Between panels are pie charts showing the per-class labeled population distributions for the 2018 (top) and 2019 (bottom) melt seasons showing clearly the class imbalance (differences) between the two melt seasons. Basemap is from ESRI World IMagery (Maxar, Airbus) | Powered by Esri.

2022). For each lake, we retrieve all available Sentinel-2 acquisitions intersecting a lake-centered window between 1 May and 30 September of the melt season in which the lake was identified, spanning the full melt-season window during which lakes form, evolve, and drain for central west Greenland. We include six of the thirteen Sentinel-2 spectral bands, selected to span the visible and infrared regions of the spectrum relevant to lake water detection and cloud discrimination: B02 (blue, 490 nm central wavelength, 10 m native resolution), B03 (green, 560 nm, 10 m), B04 (red, 665 nm, 10 m), B08 (near-infrared, 842 nm, 10 m), and B11 (short-wave infrared, 1610 nm, 20 m native, resampled to 10 m). The visible and NIR bands provide the radiometric basis for lake water detection and morphometric tracking via indices such as NDWI (McFeeters, 1996), while

https://doi.org/10.5194/essd-2026-406
Preprint. Discussion started: 18 June 2026



```
<xarray.Dataset>
Dimensions:  (time:153 band:6 y:512 x:…
Coordinates:
  * time         (time) datetime64[ns] …
  * band         (band) int32 0 1 2 3 4…
  * class        (class) int32 0 1 2 3 4
    class_name   (class) |S1 'ND' 'HF' …
  * y (y) float64 ...    * x (x) float…
Data variables:
    reflectance       (time, band, y, …
    water_mask_ndwi   (time, y, x) uin…
    lake_boundary     (y, x) uint8 ...
    p_water           (time) float32 .…
    drainage_label    () int8 (flag ND…
    label_probability (class) float32 …
    band_name (band) |S1   crs () int3…
Attributes:  Conventions=CF-1.8 ; drai…
```

Figure 2. Dataset construction workflow. External data sources (blue, ESA Sentinel-2 L2A multispectral imagery accessed through Microsoft Planetary Computer STAC catalog and the static lake boundary polygons and p_{water} water surface-area time series from Dunmire et al. (2021, 2025)) are processed by our open-source `sat-tile-stack` package (green) into four co-registered per-lake data products (yellow): the time-resolved multi-band Sentinel-2 `reflectance` tile stack (Sect. 2.2), the dynamic NDWI-derived water mask `water_mask_ndwi` (Sect. 2.3), the static `lake_boundary` polygon raster, and the scalar fractional water surface-area time series `p_water` (Sect. 2.1). The expert human labeling process via the graphical user interface (GUI; pink, Sect. 2.4.3) draws on the per-lake NetCDF files and produces labels in a CSV file. All five streams are then co-registered onto a common UTM grid and time axis via the `coregister.py` routines and embeddeed as top-level CF data variables into a single CF-1.8 compliant NetCDF per lake (red), yielding the 1679 self-contained per-lake files of the published dataset. The right-hand panel (red) illustrates the contents of a representative published file (e.t., `CW2019_2330.nc`): the on-disk `xarray` schema (top) and visualizations of each of the five embedded data products (bottom).

the B11 band supports cloud and snow discrimination. In addition to these six reflectance bands, we use the Sentinel-2 Level-2A Scene Classification Layer (SCL), a per-pixel categorical classification generated by the Sen2Cor processor (Louis et al., 2016; Main-Knorn et al., 2017) that assigns each pixel to one of twelve land-cover and atmospheric classes (including cloud shadow, cloud medium probability, cloud high probability, thin cirrus, snow/ice, water, vegetation, and bare soil). We use this during stacking to derive the binary cloud mask (Sect. 2.2)

Lake outlines and water area time series from Dunmire et al. (2025).

The lake population analyzed in this work is a subset of the GrIS-wide supraglacial lake inventory published by Dunmire et al. (2025). In particular, we limit the scope of the dataset to lakes located within the central-west drainage basin as this area represents one of the most well-studied regions in terms of dynamic supraglacial hydrology (e.g., Das et al., 2008; Stevens et al., 2015; Smith et al., 2015; Christoffersen et al., 2018; Lai et al., 2021; Stevens et al., 2024). Dunmire et al. (2025) identified 3846 lakes across the full GrIS in 2018 and 6146 in 2019; of these, 679 in 2018 and 1000 in 2019 fall within the central-west basin and constitute our dataset population. For each lake we use two products from the published Dunmire et al. (2025) dataset: the static lake boundary polygon, derived from the maximum lake extents in the 2018 and 2019 melt season and originally published in Dunmire et al. (2021), and the daily water surface-area (percentage) scalar time series (p_{water}), derived by Dunmire et al. (2025) from masked Sentinel-2 reflectance over the melt season. We note that Dunmire et al. (2025) incorporates scalar time series from other sources such as Sentinel-1 satellite aperture radar (SAR) backscatter to encode spatio-temporal behavior that can distinguish between drainage modes such as buried and frozen. For consistency with our Sentinel-2 imagery use, we use only the scalar time series from the Sentinel-2 imagery derived by Dunmire et al. (2025). The baseline classifier (Sect. 3) consumes only the scalar time series and the RGB imagery.

NDWI-derived dynamic lake masks

We additionally include a set of dynamic, time-step-dependent surface water masks, a product not included in the dataset from Dunmire et al. (2025). Whereas the static lake boundary masks represent the maximum spatial extent of each lake per melt season, the dynamic surface water masks evolve through the melt season and capture the time-varying water boundary as the lake fills and/or drains and/or refills. These dynamic masks enable a more physically faithful representation of lake water evolution through the season and may be of particular value to downstream applications that require sub-seasonal resolution of water and lake shoreline evolution. The dynamic masks are included in the published dataset. A detailed description of the dynamic masking pipeline is provided in Sect. 2.3.

2.2 Per-lake imagery preprocessing with `sat-tile-stack`

We construct per-lake Sentinel-2 imagery time series tile stacks using `sat-tile-stack`, an open-source Python package that we developed for this work. The `sat-tile-stack` package is built to support reproducible, ML-ready preprocessing of satellite image time series (SITS) for any dynamic geographic target. The package handles STAC-based scene discovery, geographic windowing, re-projection, temporal aggregation, data normalization, and serialization to CF-compliant NetCDF file format. The full `sat-tile-stack` codebase is publicly available (see Sect. 4).

The `sat-tile-stack` software is a general-purpose tool for building per-target SITS from STAC catalogs and exposes several user-configurable parameters that determine the structure of the output. In particular, users may select (i) any STAC collection with appropriate band or asset structure, including Sentinel-2 L2A, Sentinel-1 GRD, and Landsat Collection 2 L2; (ii) any subset of bands or assets from those collections; (iii) the spatial footprint of the per-target window, as a square of arbitrary pixel extent at arbitrary spatial resolution (with a warning issued if the requested resolution is finer than the sensor's native resolution); (iv) the temporal cadence of the output, as any offset alias (e.g., daily, 2-daily, weekly, monthly); (v) the aggregation rule for combining multiple observations within a cadence window (e.g., pixel-wise mean, nearest-in-time, first, or last); (vi) whether to apply robust per-image median and interquartile-range (IQR) normalization of reflectance values; (vii) which per-pixel cloud masking method to apply, with currently supported options being the Sen2Cor Scene Classification Layer (SCL) and the SWIR1 threshold method of (Williamson et al., 2018b), or user-supplied custom callable; and (viii) whether to rasterize a user-specified polygon mask from a GeoJSON source onto the output grid. For the dataset presented here, we use a specific configuration suited to supraglacial lake drainage mechanism classification from Sentinel-2 imagery, described step-by-step in the remainder of this subsection.

For each lake in the labeling population, the preprocessing pipeline executes the following steps:

1. *Lake-centered window definition.* Given the lake centroid, the pipeline computes a square window at a user-specified pixel extent and spatial resolution, centered on the lake centroid. For this dataset, we set the pixel extent to 512×512 pixels and the spatial resolution to the native Sentinel-2 10 m (parameters (iii), above), yielding a 5.12×5.12 km field of view per lake. This window is sized to capture the lake itself together with its immediate spatial contexts such as outlet streams, nearby lakes, surrounding crevasse fields, and moulins, all of which are essential for a human labeler (and subsequently a ML classifier) to distinguish the drainage mechanism. The bounding box passed to the STAC search is oversized by a factor of 1.1 relative to the final window to ensure full coverage after reprojection to the output grid in step 3, below.

2. *STAC scene discovery.* The pipeline then queries Microsoft Planetary Computer's STAC API for all Sentinel-2 Level-2A scenes intersecting the oversized lake window between a user-specified time window, which we set to the melt season bounds of 1 May to 30 September in 2018 and 2019. No cloud cover pre-filter is applied at the scene level (i.e., the STAC query filter `eo:cloud_cover < 100` returns all available scenes), so that all acquisitions are retained regardless of scene-level cloud cover. We do this because Greenland is cloudy and we still expect some meaningful visual information to exist even if the scene is partially cloud-obscured.

3. *Re-projection, temporal aggregation, and same-day scene combination.* The set of returned STAC items is passed to the `stackstac` library (Joseph, 2024), which lazily constructs a four-dimensional data arrays in `xarray` of shape [time, band, y, x] from the STAC asset URLs. Reprojection to a common spatial grid, which is necessary for ML approaches, is performed by `stackstac` using the EPSG code of the first STAC item returned by the search, which for Sentinel-2 Level-2A products corresponds to the local Universal Transverse Mercator (UTM) zone of the Sentinel-2 tile covering the lake. All requested bands are resampled to a uniform, 10 m spatial resolution on this common UTM grid. Pixels for which every requested band is simultaneously zero (e.g., the Sentinel-2 nodata convention at scene boundaries) are converted to NaN to prevent any biasing in downstream aggregation.

   The raw stack is then resampled to daily cadence by grouping all scenes acquired on the same calendar day and taking the pixel-wise mean across them, which serves as the same-day multi-scene combination step. In particular, when a lake window straddles a Sentinel-2 tile boundary or is covered by overlapping orbits, the multiple scenes on a given day are combined into a single seamless daily composite. The pixel-wise daily mean naturally handles mosaic boundaries (e.g., overlapping pixels are averaged, non-overlapping pixels retain their single-scene values) and propagates NaNs. The resulting daily stack is re-indexed onto a regular, 153-day calendar grid spanning 1 May to 30 September in each year. Tiles from calendar days lacking Sentinel-2 observations are filled with NaNs to preserve a fixed temporal shape for downstream ML models. For this dataset, we choose daily cadence and pixel-wise mean aggregation, both of which are user-configurable parameters (parameter v). Alternative cadence and aggregation rules are available for applications with different temporal sampling or scene-blending requirements. Per-scene cloud cover from the STAC `eo:cloud_cover` property is aggregated to the same daily grid.

4. *Radiometric normalization.* The `sat-tile-stack` package offers an optional per-frame, per-band robust rescaling step in which each daily 512×512 tile is centered on its median, divided by its interquartile range ($P_{75} - P_{25}$), and min-max rescaled to $[0,1]$, with `NaN` no-data pixels preserved. For the dataset presented here, this normalization is intentionally disabled. That is, the pixel values written to the `reflectance` data variable are the raw Sentinel-2 L2A digital numbers (DN) as delivered by Microsoft Planetary Computer, cast to `float32`. These values have not been divided by the quantification value $Q = 10000$ nor shifted by the ESA-applied additive offset b_{offset}. We follow Sentinel-2 community convention in naming the variable `reflectance` and the variable's `long_name` attribute documents that the stored values are "raw surface-reflectance digital numbers". Surface reflectance values are recoverable via $\rho = (\text{DN} + b_{\text{offset}})/Q$ using the per-timestep `boa_add_offset` coordinate and the dataset-level `s2_quantification_value` attribute. Leaving normalization off when building this dataset enables downstream users to decide their preferred normalization scheme without regenerating the imagery stacks. The derived binary `cloud_mask`, being a categorical flag rather than a reflectance measurement, is unaffected by any choice of reflectance normalization.

5. *Per-pixel cloud masking.* A binary cloud mask is computed for each daily image using the Sentinel-2 Scene Classification Layer (SCL). Pixels classified by Sen2Cor into any of the following four classes are flagged as cloud-contaminated: cloud shadow (SCL class 3), cloud medium probability (class 8), cloud high probability (class 9), and thin cirrus clouds (class 10). All other pixels, including snow/ice (class 11), water (class 6), bare soil (class 5), vegetation (class 4), and unclassified pixels (class 7), are marked as clear. The resulting binary mask is written into the per-lake NetCDF as a top-level `uint8` data variable named `cloud_mask`, sharing the `(time, y, x)` grid of the spectral reflectance so that downstream users can apply a cloud-aware filtering without recomputing the mask.

   The `sat-tile-stack` package currently implements two built-in cloud masking methods: the SCL-based mask described above and the SWIR1 threshold method of Williamson et al. (2018b), with an additional hook for a user-supplied callable. For this dataset, we use the SCL-based mask for its availability (SCL ships as a standard asset alongside the spectral bands in every Sentinel-2 Level-2A product) and for its use of a rich multi-class atmospheric classification. We note, however, that the SCL classification is known to occasionally confuse bright ice, snow, and cloud surfaces on high-albedo polar targets; users requiring an alternative cloud product can rebuild the per-lake stacks with `sat-tile-stack` configured for the Williamson et al. (2018b) SWIR1 threshold or a custom callable, in either case re-running the pipeline rather than modifying the published files.

6. *Serialization.* After all processing steps, the data array is cropped from the oversized $1.1\times$ working window back to the final 512×512 pixel window centered on the lake centroid, pixel values are cast to `float32`, and the per-lake stack is written to a single CF-1.8 compliant NetCDF file using the `write_netcdf` utility. The on-disk file from this stage holds two core data variables. The first is `reflectance`, which is of shape $[153, 6, 512, 512]$ over dimensions $[\mathrm{time}, \mathrm{band}, y, x]$, containing the six raw Sentinel-2 L2A digital numbers (DN) for bands B04, B03, B02, B08, B11, and B12 as `float32` (RGB-first ordering on the `band` axis). Raw DNs are converted to surface reflectance via $\rho = (\mathrm{DN} + b_{\mathrm{offset}})/Q$, where $Q = 10000$ is the Sentinel-2 quantification value (which itself is stored as a dataset-level attribute) and b_{offset} is the additive offset which the ESA changes per processing baseline. Both the per-timestep offset and the originating processing baseline are preserved as time-indexed coordinates. The second core variable is the `cloud_mask`, of shape $[153, 512, 512]$, which is a `uint8` CF flag variable derived from the SCL by flagging classes 3, 8, 9, and 10 as cloud-contaminated (1) and all others as clear (0). Per-timestep scalar coordinates on the `time` dimension include the aggregated scene-level cloud cover (`eo_cloud_cover`), the window-level NaN fraction (`pct_nans`), the ESA processing baseline (`processing_baseline`), and the additive offset (`boa_add_offset`). Additional coordinate variables record the numeric band index together with an auxiliary `band_name` short-name coordinate (e.g., `"B04"`), the spatial grid in UTM coordinates, the coordinate reference system (CRS).

   In a second co-registration pass, the lake-extent products from Dunmire et al. (2021, 2025), the dynamic water masks (Sect. 2.3), and the manual human-assigned labels (Sect. 2.4) are joined onto the same file as additional top-level data variables: `p_water` $(\mathrm{time},)$, a `float32` time series carrying the Sentinel-2-derived daily fractional water extent from Dunmire et al. (2025), aligned to the stack dates by calendar day; `lake_boundary` (y, x), a `uint8` raster of the static maximum-extent polygon from Dunmire et al. (2021), rasterized onto the lake's UTM grid; `water_mask_ndwi` (time, y, x), the `uint8` dynamic NDWI derived surface water mask (Sect. 2.3); `drainage_label`, a scalar `int8` CF flag variable holding the hard expert label with flag values `0–4` corresponding to ND, HF, MD, LD, CD; and `label_probability` $(\mathrm{class},)$, the `float32` soft probability vector (Sect. 2.4.3), accompanied by `class` and `class_name` coordinates. The free-text labeler notes and revisit flag are dataset-level attributes. Dataset-level attributes document the CRS, spatial transform, resolution, STAC product metadata, and the processing history of the multi stage build. The result is a single self-describing per-lake CF-1.8 compliant `NetCDF` file for each lake instance.

2.3 Dynamic lake masking

To complement the static lake boundary masks (Sect. 2.1), we additionally compute a dynamic, per-timestep water mask for every lake directly from the reflectance provided by `sat-tile-stack`. A pixel is flagged as water where the blue-to-red normalized difference water index (NDWI) exceeds a threshold:

$$\mathrm{NDWI} = \frac{(\rho_{\mathrm{B02}} - \rho_{\mathrm{B04}})}{(\rho_{\mathrm{B02}} + \rho_{\mathrm{B04}})} > 0.3\,, \tag{1}$$

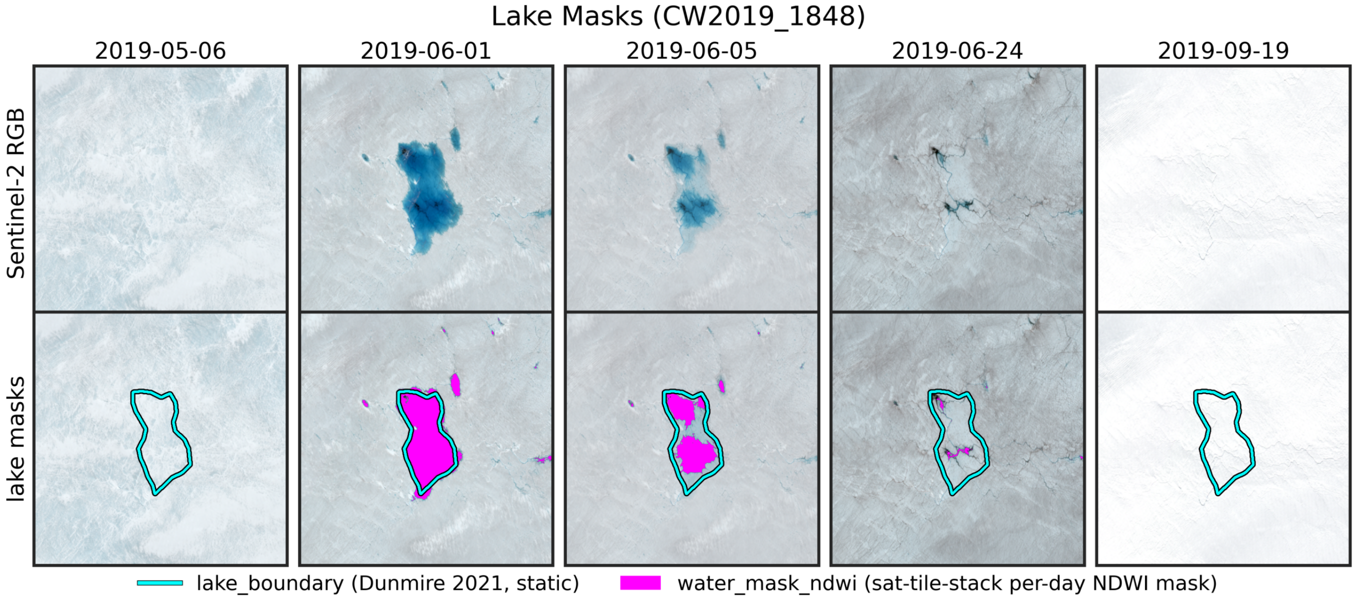


Figure 3. Static lake boundary versus dynamic NDWI-derived water mask for an example supraglacial lake: CW2019_1848 (49.59°W, 68.56°N, Fig. 4). **(Top)** five Sentinel-2 snapshots spanning the 2019 melt season. **(Bottom)** the same imagery with two co-registered water mask overlays: the static lake boundary polygon from Dunmire et al. (2021) (cyan) and the per-pixel per-day `water_mask_ndwi` computed from the file's own Sentinel-2 reflectance via NDWI masking (magenta, Sect. 2.3).

where ρ is reflectance (Kang Yang and Smith, 2013; Dunmire et al., 2025). The threshold of 0.3 was chosen as a middle ground between threshold choices that have been made in existing literature, specifically 0.5 (Dunmire et al., 2025) and 0.15 (Williamson et al., 2018a). We found this choice effective in not over-including slushy areas that are not part of a lake. Pixels in tiles with missing data are encoded as per-pixel NaNs. The dynamic mask is joined into the per-lake NetCDF as a time-varying (time, y, x) surface-water variable in the same co-registration pass that attaches the static boundary and p_{water} series, and is fully reproducible from the published reflectance via the single exposed threshold parameter (Fig. 3).

2.4 Labeling and validation

We assign each of the $N = 1679$ lakes in the dataset a drainage mechanism, which we classify by human visual interpretation of the full Sentinel-2 optical imagery time series as produced by the `sat-tile-stack` pipeline (Sect. 2.2) in the previous section. Here we describe the five-class taxonomy of lake drainage mechanisms (Sect. 2.4.1), our labeling protocol (Sect. 2.4.2), our soft probability paradigm, and the custom graphical labeling interface that we built for implementing this paradigm. (Sect. 2.4.3). We present the resulting labeled set and quality control workflow (Sect. 2.4.4), and, finally, the inter-rater agreement study used to quantify human labeling uncertainty (Sect. 2.4.4).

https://doi.org/10.5194/essd-2026-406
Preprint. Discussion started: 18 June 2026

2.4.1 Drainage mechanism taxonomy

We construct a first-of-its-kind five-class taxonomy of supraglacial lake drainage mechanisms, designed to granularly distinguish among the physically distinct pathways by which a lake's water can be, or alternatively is not, removed from its basin during a melt season (Figure 4). This differs from previous taxonomies (e.g., Dunmire et al., 2021) in that it categorizes lakes based on their drainage mode, as opposed to drainage speed. Our taxonomy follows:

0. **No drainage (ND):** The lake retains its water throughout the entire melt season without losing it through any outlet. The lake may continue to receive meltwater inputs from its immediate surroundings, but, with respect to outflow, remains hydrologically isolated from the broader supraglacial network.

1. **Hydrofracture (HF):** The lake drains rapidly via the propagation of a water-filled fracture the length of the entire ice thickness. Water is routed from the surface to the bed and disappears from view over a timescale of hours to days. Hydrofracture events typically drain water from within the lake basin itself, and are often accompanied by visible fracture scarps, remnant ice fragment blocks, and sometimes a persistent draining stream into the fracture feature (i.e., a newly generated moulin) in the post-drainage imagery. Often the lake disappears quickly and leaves behind a baby blue-colored empty lake basin in the days closely following the drainage. If a new moulin is generated, this is still classified as HF.

2. **Moulin drainage (MD):** The lake drains directly into a moulin, which is a vertical surface-to-bed conduit which often has a circular (or oblong) hole-like surface expression. The moulin must be visible in the tile (i.e., 512×512 pixels) window for the purposes of this study. Out-of-tile moulins can occur but are not considered here, as detection of farther-range moulins is beyond the scope of this work. Moulin drainages typically (but not always) proceed more slowly than hydrofractures, and the moulin itself is often, but not always, located outside the lake's water extent boundary, fed by a connecting stream from the lake. Moulin drainages are not accompanied by the formation of an obvious fracture scarp.

3. **Lateral drainage (LD):** The lake drains via one or more outlet pathways (e.g., a stream) that route water laterally across the ice sheet surface, either into another lake basin or out of the visible field of view of the imagery tile. Lateral drainage redistributes water within the supraglacial hydrological network and does not, by itself, deliver meltwater from the surface to the bed. In some cases, a lake may drain laterally via a stream to another lake basin, which then drains englacially (HF or MD). In these cases, the initial lake in question is still classified as a lateral drainage.

4. **Crevasse field drainage (CD):** The lake drains into a pre-existing field of surface crevasses. The crevasse field may disperse the water laterally, vertically, or a mix of both, and may do so rapidly or slowly. Lakes that appear to experience a distinct hydrofracture event (evidenced by a rapid drainage and a fresh scarp among the existing crevasses) are still always considered crevasse drainage events. Lakes situated within or immediately adjacent to crevasse fields whose water disappears into the field, even via an outlet stream, also fall into this class.

In our taxonomy, we consider these five classes to be mutually exclusive at the level of their dominant drainage mechanism. We add this consideration by design to physically ground the taxonomy, in that each class corresponds to a distinct drainage pathway with distinct implications for ice-sheet hydrology, surface mass balance, and ice dynamics. The taxonomy is also designed to be operationally distinguishable from Sentinel-2 time series imagery by an expert labeler with sufficient experience and surrounding spatio-temporal context. This does not, however, mean that there are not lakes that might display characteristics of multiple drainage types making it difficult for even a human expert to properly distinguish their class. The labeling process is further confounded by, among other factors, the presence of cloudy days and no-data days, both of which limit observations of the full seasonal context. In some cases, the distinguishing spatial features (e.g., HF scarp) may be small and transient, only visible on the day of or after a drainage, which clouds and no-data days may obscure. Thus, we introduce probabilistic labeling and inter-rater agreement metrics to reflect this unavoidable uncertainty (Sect. 2.4.3, 2.4.4).

2.4.2 Labeling protocol

To label each lake, a human expert reviews the full 153-frame Sentinel-2 imagery time series spanning 1 May to 30 September of the relevant melt season, and assigns a probability distribution over the five classes with confidence granularity intervals of 25% (Sect. 2.4.3). The per-labeler class assignment is then determined as the label with the maximum value in the probability vector. Labelers are discouraged from assigning a tie, but in the case of an unavoidable tie, the label is determined to be from the class with a lower class ID number. For example, a tie between HF (1) and MD (2) defers to HF. This is by design to generally defer the label of difficult-to-distinguish cases to the more dynamically consequential class. Class definitions and edge-case handling rules are documented in a written labeling protocol that is distributed to all human labelers and accessible from within the the labeling GUI (Sec. 2.4.3). The full protocol is reproduced in Appendix A. A human labeler always has the option to flag any instance for revisit, as often happens when encountering edge-case lakes. We find the most consequential edge-cases, and our chosen guidelines to be:

- *HF vs MD:* When a lake exhibits features of both hydrofracture and moulin drainage, for example when a HF event evolves into a persistent moulin or when the imagery is consistent with either mechanism, the lake is labeled as HF. The convention assigns ties to the more dynamically consequential class (e.g., HF marks the formation of a new surface-to-bed pathway) and ensures that HF events are not systematically under-counted. There are several reasons that we expect HF events could be systematically under-counted in Sentinel-2 imagery. HF scarps are sub-pixel scale in their across-crack direction and may not be directly visible for this or other reasons (e.g., a frozen ice lid, cloudy days directly following the drainage). As such, a visible HF scarp is not strictly required for the lake to be labeled HF. To account for these uncertainties without under-counting HF instances, if a lake drains rapidly interior to its lake basin, and the labeler can not demonstrate it is not a HF, in this paper's protocol, it is labeled HF.

- *Lake-to-lake drainage is always LD:* When a lake drains via an outlet stream into the basin of another (e.g., downstream) lake, the source lake is labeled LD regardless of how the downstream lake itself eventually drains. This prevents propagation of a downstream lake's drainage classification to an upstream lake's classification and preserves the locality of the mechanism label.

- *Outlet stream into a moulin is MD:* When a lake drains via an outlet stream that visibly terminates in a moulin within the 512×512 window, the lake is labeled MD rather than LD, on the grounds that the locally observable fate of the water is delivery to the bed rather than lateral redistribution. When the outlet stream meanders out of frame, the lake is labeled LD, since the downstream fate of the water cannot be confirmed from the visible scene available to the labeler (and subsequently to the ML classifier).

- *Crevasse-field-resident lakes are CD by default:* A lake located within a crevasse field is labeled CD, including lakes that drain via hydrofracture within the field. This choice reflects our belief that crevasse field drainages should be studied more in depth, as they can exhibit a range of behaviors and are relatively under-studied. The one exception to this chosen guideline is a lake that sits within a crevasse field, yet drains via a visible outlet stream that routes water out of the crevasse field or the tile bounds, to be labeled as LD under this taxonomy. This distinction is made on the grounds that when the downstream fate of the water leaves the scene, the labeler cannot reliably confirm that the water enters the crevasse field. Thus, by this logic, lakes whose outlet streams remain in-frame and are observed to enter the surrounding crevasse field, however, are CD and not LD.

- *Inlet versus outlet stream disambiguation.* Inlet streams (which deliver water to the lake) and outlet streams (which carry water away from the lake) can be visually similar in single frames. Labelers are instructed to use the temporal evolution of the lake's water level to make this distinction (e.g., rising and stable water levels generally indicate stream inflow, while falling water levels indicate stream outflow). Labelers are also instructed to apply the rule of thumb that streams typically taper from wide to narrow in the direction of flow.

2.4.3 Soft-probability labeling and graphical interface

To support efficient and reproducible labeling, we developed a custom, browser-based GUI for the labeler to use during the labeling task. This GUI is publicly available as part of the `sat-tile-stack` package (Fig. 5). The interface presents the labeler with the lake-centered 512×512 imagery at the currently selected timestep, together with a frame slider that scrubs through all 153 daily timesteps in the melt season. The labeler can review the full temporal evolution of the lake using the mouse scroll wheel, the keyboard arrow keys, or by dragging the slider directly. For each lake, the labeler assigns a discrete probability to each of the five drainage classes using a row of sliders on a right-hand panel, with the constraint that the five probabilities (discretized to 25% intervals across classes) must sum to unity. The final class label in the dataset presented in this study is the maximum of this probability vector across all labelers. But, we preserve the full soft-probability vector alongside the hard label to enable downstream users to train against soft targets, to filter the dataset by per-lake label confidence, or to define their own decision rules.

The soft-probability paradigm reflects a deliberate design decision: drainage mechanism is not always unambiguously identifiable from Sentinel-2 imagery alone. Forcing labelers to employ a hard, one-hot class on every lake discards genuine epistemic uncertainty that is intrinsic to the labeling task and informative for model training. For example, a lake labeled $(HD = 0.0, HF = 1.0, MD = 0.0, LD = 0.0, CD = 0.0)$ represents an unambiguous hydrofracture event. In contrast, a lake labeled $(HD = 0.0, HF = 0.75, MD = 0.25, LD = 0.0, CD = 0.0)$ represents a drainage in which the labeler is reasonably confident in assigning an englacial drainage mechanism, but is not fully confident whether this occurs through MD, HF, or potentially both. Both labels are scientifically meaningful and are preserved in the published dataset.

In addition to class probability sliders, the GUI provides the following: a notes field for free-text annotations on individual lakes (preserved in the dataset), a “Flag for revisit” button for marking lakes that may require a second pass (often used for ambiguous lake instances), a “Skip” button for deferring lakes to a later time in the labeling session, and real-time labeling progress and class distribution charts that update as labels are submitted, providing the labeler with continuous feedback on their progress and the evolving statistical summary of the dataset composition. Navigation through the sample of lakes is supported through a sample list panel on the bottom right showing labeled, flagged, and unlabeled lakes. Compared to ad-hoc labeling workflows in which the labeler must manually load and inspect imagery for each lake outside of a structured interface, we find that the GUI reduces labeling time per lake by a factor of ~ 2. The GUI is launched with simple command line interface (CLI) calls, which point to the directory containing NetCDF (.nc) files built with `sat-tile-stack` and comma-separated value (.csv) files into which the labeler is populating with the labels, via the GUI.

2.4.4 Label set and inter-rater agreement

All $N = 1679$ lakes in the dataset (679 from the 2018 melt season and 1000 from the 2019 melt season) were labeled by the lead author (Labeler 1, J. H. R.) following the protocol described in Sect. 2.4.2. The lead author’s labels constitute the primary labels. The inter-rater agreement statistics reported here provide a quantitative characterization of label uncertainty against which the primary labels can be calibrated. In particular, to quantify the reproducibility and epistemic reliability of the drainage mechanism labels, we conduct a three-labeler inter-rater agreement study on a stratified random subset of the labeled lakes (Figure 6). From the full $N = 1679$ labeled lakes, we drew a subset of 200 lakes, with 100 from the 2018 melt season and 100 from the 2019 melt season. These were selected by stratified random sampling such that the class distribution within each year’s subset was matched, within sampling variability, to the class distribution of the corresponding full year. This stratification ensures that the inter-rater agreement statistics are not dominated by the most common class and provide informative agreement estimates for each of the five classes individually on an annual basis.

Two additional labelers (N. B. C. and E. L.), both co-authors of this work and following the same labeling protocol (Sect. 2.4.2), independently labeled each of the 200 lakes in the inter-rater subset. The three labels from the three labelers have an overall agreement of 72.5% with pairwise agreements of 80.0% (between Labelers 1 and 2), 83.0% (between Labelers 1 and 3), and 79.5% (between Labelers 2 and 3). The 3-labeler comparison yields a Fleiss’s κ (Fleiss, 1971) of 0.754, with pairwise Cohen’s κ values of $\kappa_{1,2} = 0.744$, $\kappa_{1,3} = 0.781$, and $\kappa_{2,3} = 0.738$ (Cohen, 1960). Notably, the three labelers diverged most strongly for their classifications of the MD class, likely reflecting the difficulty of detecting a moulin drainage from Sentinel-2 satellite imagery. The inter-rater subset is identified in the published dataset so that future users can isolate the multiply-labeled lakes for their own analyses, and the labels from all three labelers are preserved alongside the primary lead-author labels for full transparency. We emphasize that the lead-author labels remain the primary labels for dataset use;

Table 1. Class distribution of the central west (CW) basin label set across the 2018 and 2019 melt seasons.

Class	2018	2019	Total
ND	99 (14.6%)	137 (13.7%)	236 (14.1%)
HF	95 (14.0%)	184 (18.4%)	279 (16.6%)
MD	79 (11.6%)	163 (16.3%)	242 (14.4%)
LD	253 (37.3%)	327 (32.7%)	580 (34.5%)
CD	153 (22.5%)	189 (18.9%)	342 (20.4%)
Total	679	1,000	1,679

the inter-rater statistics provide a quantitative bound on the reproducibility (uncertainty) of the primary labels, along with an empirical estimate of label noise that downstream users should consider when estimating model performance.

2.5 Dataset statistics and partitioning

The class distribution across the full $N = 1679$ lake dataset is summarized in Table 1 and is, as expected for a physically grounded mechanism-based taxonomy, imbalanced. We describe the dataset statistics here, including both the class imbalance, distribution shift between years, and splitting the dataset into distinct subsets for model training. The largest class (LD) is roughly three times the size of the smallest class (MD) in 2018, but only approximately double the size of the smallest class (MD) in 2019. These shifts in the class balance reflect variable lake presence and meltwater dynamics between different melt seasons. We partition the dataset into training, validation, and test splits along two complementary axes. These splits are distributed as JSON lake-ID lists alongside the dataset (see Sect. 4). Our baseline training pipeline reads these ID lists directly for reproducibility. Below are the two axes along which the dataset is partitioned for two different training folds.

- *Combined split.* All $N = 1679$ labeled lakes (679 from CW2018 and 1000 from CW2019) are randomly pooled and partitioned into training (70%, $n = 1175$), validation (20%, $n = 336$), and test (10%, $n = 168$) subsets by class-stratified random sampling. The class distribution within each subset matches the parent pool distribution to within 0.23 percentage points across all five classes, and the train and test subsets themselves agree on every class proportion to within 0.03–0.25 percentage points (largest gap: ND; smallest: LD).

- *Cross-year split.* The full CW2018 dataset (679 lakes) is held out as the test set, so that the model, trained on the 2019 season alone, has no exposure to observations from the 2018 season during training and validation. The CW2019 pool (1000 lakes) is partitioned into training ($n = 800$) and validation ($n = 200$) subsets by class-stratified random sampling, with class distributions matching the CW2019 pool. This leaves us with a total split of training ($48\%, n = 800$), validation ($12\%, n = 200$), and testing ($40\%, n = 679$) data. This cross-year protocol is the primary generalization benchmark for the dataset and underlies the cross-year experiments reported in Sect. 6. The protocol directly evaluates whether a model trained on one melt season's lake population can predict drainage mechanisms in another season, and it operationalizes the intended use of the dataset as a living dataset for ice-sheet-wide, multi-year extension of the labeled

https://doi.org/10.5194/essd-2026-406
Preprint. Discussion started: 18 June 2026

population. Notably, the CW2018 test set exhibits an inter-annual distribution shift relative to the CW2019 training pool. For example, the most imbalanced class (MD) has 4.7% lower representation in CW2018 as compared to CW2019. This train/test split thus probes inter-annual generalization in the case of slight distribution shift.

The two splits differ in their use of an in-distribution test set: the first split reserves 10% of the dataset as an in-distribution test set indiscriminate of inherent changes in the observations, like observation year. By contrast, the second split withholds the entire CW2018 population from the model fit as the test set, and therefore partitions only the CW2019 pool into training and validation. For the data-scaling experiments reported in Sect. 6, the combined-split training set is pre-ordered using a class-balanced algorithm. At each step, the algorithm selects the next lake from the class with the lowest current count. This way, the first N identifiers in any set form a class-stratified subset for every $N \leq 1175$. Nested training subsets at $N \in \{200, 400, 600, 800, 1000\}$ are then taken as the first N entries. This guarantees monotone inclusion ($N = 200 \subset N = 400 \subset \cdots$) while preserving the parent class distribution at every subset size. The split assignments are released as JSON lake-ID lists alongside the dataset (Sect. 4) so that all dataset users can reproduce the exact training, validation, and test partitions used in this work.

3 Baseline deep learning classifier

Above we demonstrated how a human-benchmarked dataset can be composed for two melt seasons in the central west Greenland region. However, this type of manual labeling quickly becomes intractable across decades and continental scales. For example, we estimate that the total human hours spent on the labeling effort for the dataset presented here exceeds 150 hours. This estimate does not include the start-up time costs of building the dataset generation and labeling (e.g., GUI) workflows. ML approaches can provide a compelling framework for automatically extracting drainage signature from multi-temporal satellite observations. Dunmire et al. (2025) demonstrated the utility of an automated approach by applying deep-learning enabled classifiers to one-dimensional lake features for lake fate classification. However, this approach is limited to predicting on scalar water surface-area time series, and does not include sufficient external geospatial information from the overall scene to be compatible with our mechanistic taxonomy. We extend the Dunmire et al. (2025) machine learning model to incorporate spatially resolved time series, rather than only scalar summaries, by implementing a model architecture capable of jointly modeling spatial structure and temporal evolution. The convolutional long short-term memory (ConvLSTM) network, originally introduced by Shi et al. (2015) for precipitation nowcasting from radar imagery, provides exactly this capability and has been used previously for Sentinel-2-based crop and land cover classification (e.g., Rußwurm and Körner, 2018). We build directly upon this lineage in our baseline classifier model, employing a ConvLSTM-based architecture to classify drainage mechanism from both the surface-area curves like Dunmire et al. (2025), and additionally from the fully resolved imagery time series. In doing so, we extend the value of this dataset to the glaciological community, by demonstrating a successful workflow for classification of lake drainage by mechanism.

The model, training code, split files, and trained weights will be released a part of a companion paper. In this paper, we demonstrate that—even with a fixed architecture, a single set of hyperparameters chosen a priori, and no hyperparameter or architecture search—this approach provides a better baseline than training on water surface-area time series alone. This baseline classifier, by design, resolves drainage mechanism, distinguishing among the five designated classes: no drainage, hydrofracture, moulin drainage, lateral drainage, and crevasse drainage.

3.1 Architecture

The classifier, `LakeDrainageClassifier`, is a multi-stream temporal recurrent neural network with two parallel branches that process two co-registered input streams contained directly in the prepared dataset (imagery and scalar water surface-area time series), concatenate their encoded features, and pass the result to a classification head (Figure 7). This classification head outputs a probability distribution over the five drainage mechanism classes. The model has on the order of $\mathcal{O}(100,000)$ trainable parameters in this baseline configuration. The two input streams from the NetCDF-based dataset are:

- *Imagery input data stream:* The imagery branch consumes a four-channel image sequence of shape [Time, Channel, Height, Width] $= [153, 4, 512, 512]$, where the four channels are three of the six available reflectance bands (red, green, blue) plus the static lake boundary mask from Dunmire et al. (2021), broadcast across the time axis. The lake boundary mask is loaded as a fourth channel of the input tensor and broadcast across the time axis, however the baseline model reported here does not consume the mask in its convolutional encoder. The sequence is passed frame-by-frame through the network, which learns spatiotemporal features end-to-end, rather than reducing the imagery to a scalar summary at each timestep. Global average pooling (GAP) over the spatial dimension of the final hidden state yields a 32-dimensional imagery feature vector as a latent representation of each lake's spatiotemporal characteristics.

- *Scalar water presence input data stream:* The daily cadence scalar water surface-area time series p_{water} from Dunmire et al. (2025) (Sect. 2.1) is the most direct observation of drainage available to the model. Namely, a drainage event is, by definition, a decrease in lake water area. We provide the time series as a scalar sequence rather than expecting the model to derive it implicitly from the imagery since it is a direct physical descriptor of each lake's hydrological activity. The water area sequence is normalized to the $[0, 1]$ range per-lake.

The resulting feature vectors are processed and the final output is a five-dimensional vector of logits, one per drainage mechanism class. Class probabilities are obtained by softmax at inference time. A weighted cross-entropy loss is used at training time, and operates directly on the logits. More complete model architecture details are described in Appendix B.

3.2 Model training

All training hyperparameters are fixed a priori and listed in Table 2. Below, we detail a few key components of the modeling effort: normalization, class weighting, training data, and reproducibility.

Normalization. The imagery is stored on disk as raw Sentinel-2 L2A digital numbers in `float32` (Sect. 2.2). At training time, the reflectance channels of the imagery tensor are divided by the Sentinel-2 quantification value $Q = 10{,}000$ and clipped to $[0, 1]$, yielding surface-reflectance values bounded to $[0, 1]$. We omit the per-timestep additive offset b_{offset} from the conversion (Sect. 2.2): all Sentinel-2 scenes in the CW 2018 and 2019 melt seasons covered by this dataset use processing baselines for which $b_{\text{offset}} = 0$, so DN/Q is exact for this dataset. The lake mask channel is already binary and needs no further preprocessing. The water area scalar sequence p_{water} is normalized per-lake to $[0,1]$ within each melt season to remove differences of scale between lakes (Dunmire et al., 2025). All NaN values in the imagery (e.g., days without Sentinel-2 coverage) are replaced with zeros at batch construction time. The NaN values in the p_{water} time series are filled by forward-fill.

Class weighting. To mitigate the class imbalance in the training set (Sect. 2.5), the cross-entropy loss is weighted by inverse class frequency within each training fold. For example, for class c with training set count class n_c and total training population count N, the loss weight is given by $w_c = \frac{N}{(Kn_c)}$, where $K = 5$ is the number of classes. This normalization places higher loss weight on the underrepresented classes (e.g., HF, MD) and lower weight on the overrepresented classes (e.g., LD).

Training data. For the combined-year baseline, training uses all 1175 lakes in the training set, validation uses the 336 lakes in the validation set, and the test set consists of 168 lakes (Sect. 2.5). For the cross-year baseline, training and validation use the 800 lake and the 200 lake CW2019 splits, respectively ($\sim 32\%$ smaller than the training set size for the combined experiment). The test set is the full 679 lakes in the CW2018 set. The same fixed split-ID files are used on every run so that the results are reproducible and comparable across configurations.

Training metrics. The model is evaluated based on the F_1 score metric,

$$F_1 = \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} = \frac{2\,\text{TP}}{2\,\text{TP} + \text{FP} + \text{FN}}, \tag{2}$$

where TP, FP, and FN are true positives, false positives, and false negatives, respectively. This measures the harmonic mean of precision and recall, jointly penalizing both false positives and false negatives. We evaluate the baseline classifier under two training configurations (the combined split and the cross-year split, Sect. 2.5), all using the same fixed architecture (Sect. 3.1) and trained for up to 400 epochs. Training- and validation-set metrics are reported at the best-validation-macro-F_1 epoch.

Table 2. Baseline classifier training configuration.

Parameter	Value
Input imagery shape	[Time, Channel, Height, Width]= $[153, 4, 512, 512]$
Optimizer	Adam (Kingma and Ba, 2017)
Dropout	$p = 0.3$ (only employed in the classification head)
Loss function	Inverse-frequency weighted cross-entropy

4 Data availability

The benchmark dataset is publicly available through the Stanford Digital Repository at https://doi.org/10.25740/sf350xp4038 (Rines et al., 2026a), released under a Creative Commons Attribution 4.0 International (CC BY 4.0) license. The deposit contains the following.

- *Per-lake NetCDF files* for all $N = 1679$ labeled lakes (679 from CW2018 and 1000 from CW2019), organized as `CW_2018/` and `CW_2019/` subdirectories with one file per lake named `CW{year}_{lake_id}.nc`. Each file is self-contained and follows the CF-1.8 conventions. As described in Sect. 2.2, each file contains: a `reflectance` variable of shape $(153, 6, 512, 512)$ carrying the six Sentinel-2 spectral bands (B04, B03, B02, B08, B11, B12) as raw `float32` L2A digital numbers, convertible to surface reflectance via the per-timestep `boa_add_offset` coordinate and the dataset-level `s2_quantification_value` attribute; the derived binary `cloud_mask` variable of shape $(153, 512, 512)$; the daily fractional water surface area series `p_water` of length 153 from Dunmire et al. (2025); the rasterized maximum-extent lake polygon `lake_boundary` from Dunmire et al. (2021); the dynamic NDWI-derived water mask `water_mask_ndwi` (Sect. 2.3); the expert-assigned drainage label as a scalar `int8` CF flag variable `drainage_label` together with the associated 5-class soft probability vector `label_probability` over (ND, HF, MD, LD, CD); per-timestep scalar coordinates `eo_cloud_cover`, `pct_nans`, `processing_baseline`, and `boa_add_offset`; spatial coordinates and CRS metadata; and dataset-level attributes documenting free-text labeler notes, the revisit flag, and provenance information (spatial transform, STAC source, processing history).

- *Label CSVs* mirroring the labels embedded in the NetCDF attributes, released as `labels_CW_2018.csv` and `labels_CW_2019.csv`. Each row gives a lake's canonical hard label, its soft-probability vector over (ND, HF, MD, LD, CD), labeler notes, and revisit flag. The CSVs are the direct outputs of the labeling GUI (Sect. 2.4.3) and the canonical label source consumed by the training pipeline. The embedded NetCDF attributes provide the same information in a self-describing per-file form for users who consume the dataset one lake at a time. Separate CSV files provides the three-labeler annotations for the 200-lake inter-rater agreement subset (Sect. 2.4.4).

- *JSON split-ID files* specifying the combined (70/20/10) and cross-year (train CW2019, test CW2018) partitions under `splits/essd_CW/` and `splits/essd_CW_crossyear/`, the nested stratified training-subset ordering used by the data-scaling experiments ($N \in \{200, 400, 600, 800, 1000, 1175\}$), and the inter-rater agreement subset. All splits are deterministic under random seed 42 and the nested ordering guarantees that each smaller training subset is a strict subset of every larger one.

- *Dataset size and storage.* The full benchmark dataset totals approximately 221 GB on disk: 85 GB for the 679 CW2018 lakes and 136 GB for the 1000 CW2019 lakes, averaging ~132 MB per per-lake NetCDF file. All raster data variables (`reflectance`, `cloud_mask`, `water_mask_ndwi`, `lake_boundary`) are written with zlib compression, and HDF5-chunked along their respective axes. The uncompressed per-lake total is ~1.04 GB (dominated by the $153 \times 6 \times 512 \times 512$ `float32` `reflectance` array at ~963 MB), or ~1.75 TB across the full $N = 1679$ lake dataset, giving roughly an 8× compression ratio. The baseline classifier (Sect. 3) reads a four-channel subset of `reflectance` (red, green, blue, plus the static lake boundary mask broadcast across the time axis) at training time, yielding a $[T, C, H, W] = [153, 4, 512, 512]$ `float32` input tensor of ~642 MB per lake; at the baseline batch size of 8 the corresponding training-batch tensor is ~5.1 GB.

The following external datasets were used in the construction of the benchmark dataset and are cited in Sect. 2.1, and reproduced, below.

https://doi.org/10.5194/essd-2026-406
Preprint. Discussion started: 18 June 2026

- Sentinel-2 Level-2A bottom-of-atmosphere reflectance imagery, accessed through the Microsoft Planetary Computer STAC API (https://planetarycomputer.microsoft.com).
- Supraglacial lake outlines and daily water-area-fraction time series from Dunmire et al. (2021) and Dunmire et al. (2025), available at https://zenodo.org/records/14587026.

5 Software availability

An open-source software package was developed for this work, released under the MIT license, and detailed below. The `sat-tile-stack` package (https://github.com/jharlanr/sat-tile-stack): the satellite image time series preprocessing package described in Sect. 2.2. Supports STAC-based scene discovery, per-target geographic windowing, reprojection, resampling to a common grid, user-configurable temporal aggregation, per-pixel cloud masking, optional polygon rasterization, and serialization to CF-1.8-compliant NetCDF file format. The package has been tested against Sentinel-2 L2A, Sentinel-1 GRD, and Landsat Collection 2 L2 collections. It also includes the browser-based soft-probability labeling GUI (Sect. 2.4.3) and the labeling protocol document (Appendix A) used to produce the canonical labeled dataset. The software repository `sat-tile-stack` includes a demonstration directory (https://github.com/jharlanr/sat-tile-stack/demos), which includes an example NetCDF file for lake 2019CW_1907 as well as two interactive notebooks which serve as tutorials for how to use both the `sat-tile-stack` software as well as the lake data in the dataset (Rines et al., 2026a).

6 Results and discussion

The novelty and utility of the dataset prepared here becomes clear when comparing drainage-class predictions from models that consume the full spatio-temporal context available in this dataset like the ConvLSTM approach used in Section 3, to predictions from models that are fit only on scalar representations of singular phenomena like lake area. In order to position the 5-class mechanism task within the context of existing literature, we compare the cross-year model directly to the 4-class classifier from the pioneering work of (Dunmire et al., 2025). In particular, Dunmire et al. (2025) used scalar representations of lake dynamics, derived from Sentinel-1 and Sentinel-2 satellite imagery, to categorize lakes (including the same CW 2018 and CW 2019 lakes we use here) into one of four categories: *refreezing*, *rapid*, *slow*, and *buried*. These categories describe observables that aren't grounded in the underlying physical cause that is necessary for informing hydrological models in the region. Informed by the needs of modelers, we moved our schema instead to describe the cause, rather than the rate-based fate, of lake drainage. Because the two taxonomies are not identical, both schemes are projected onto a common 3-class physically driven scheme: (*no drainage*, *englacial*, *lateral*; Fig. 8a). The (HF, MD, CD) classes in our scheme are englacial drainages, meaning that the water disappears, at least in part, vertically into the ice body below. As many lakes draining via hydrofracture are rapid, mechanistic interpretations of the *rapid* and *slow* classes from Dunmire et al. (2025) are *englacial* and *lateral*, respectively, whereas *refreezing* and *buried* directly map to *no drainage* (Figure 8a).

Under this projection (Figure 8b,c), our classifier achieves a 3-class macro F_1 score of 0.67 on val and 0.59 on test, exceeding the corresponding 0.56 and 0.43 obtained from projecting the predictions of Dunmire et al. (2025) onto the same scheme. We do not interpret this margin as evidence that the model architecture and approach is preferable to that of Dunmire et al. (2025). Rather, we emphasize that the two are trained on different taxonomies and target different scientific questions. We interpret this difference as motivation for further optimization of the 5-class classification task, as this would create a tool for the scientific community to ask a distinct question, namely what was the physical pathway (mechanism) by which water was drained from the lake. The principal value of the dataset lies in the rich, 5-class taxonomy itself, which preserves distinctions among physically distinct drainage pathways that the 3-class projection collapses, and whose disambiguation depends on the spatio-temporal context that deep learning architectures are well suited to explore.

Extending our analysis of the classifier, we focus on a cross-year experimental setup that quantifies how well our dataset supports generalization across melt seasons within a single drainage basin. The baseline classifier is trained on the primary CW 2019 training data, with the representative model selected based on the macro (all classes) F_1 score on the CW 2019 validation set ($N = 200$). The model is evaluated on the full CW 2018 labeled set ($N = 679$) as a held-out cross-year test (Fig. 9). The model achieves a 5-class macro F_1 of 0.61 on val and 0.45 on test, both above the random baseline of 0.20. The drop in F_1 score from the 2019 validation set to the 2018 test set likely reflects the cross-year class distribution shift between the two melt seasons, which experienced substantially different melt forcing (Dunmire et al., 2021; Zheng et al., 2023).

The per-class performance breakdown is strongly stratified (Figs. 9,10). Crevasse field drainage (CD) is the most learnable class on both splits (per-class F_1 of 0.84 on val and 0.71 on test) and the most visually unique in imagery (Fig. 4). Moulin drainage (MD) is the least learnable, with a particularly large val→test drop (F_1 of $0.41 \rightarrow 0.11$); the remaining classes (ND, HF, LD) cluster in the 0.44–0.68 range across both splits. The CD and LD lakes carry visually distinctive context (e.g., crevasse fields, outlet streams) that the joint imagery and water surface area-time series inputs capture well. By contrast, HF and MD lakes are distinguished by small, transient spatial features (e.g., fresh hydrofracture scarp, persistent moulin, ice blocks) that are sometimes smaller than the 10 meter resolution of our imagery. On a similar note, MD versus LD lakes are visually difficult to distinguish, even for an expert human labeler, as both often include an outlet stream and varied temporal drainage signatures. This difficulty is reflected most strongly by the worst model performance occurring when distinguishing LD from MD (see, e.g., Fig. 9).

We report here the results for the combined experiment, broken down spatially (Figure 10). In this experiment, we trained on lakes from 2018 and 2019 and evaluated on validation and test sets, as described in Sect. 2.5. The 5-class macro F_1 is 0.55 on the 2019 lakes and 0.54 on the 2018 lakes, in contrast to the $0.61 \rightarrow 0.45$ drop of the cross year model, which was trained on 2019 and evaluated on 2018 lakes. The per-class stratification described in Sect. 6 persists and is essentially year-invariant: crevasse field drainage (CD) remains the most learnable class (F_1 of 0.80 and 0.86 on the 2019 and 2018 lakes, respectively), while moulin drainage (MD) and hydrofracture (HF) remain the weakest. For example, in 2019, MD and HF predictions had F_1 scores of 0.36 and 0.47, respectively, and in 2018 the scores were 0.22 and 0.29, respectively.

These results suggest that closing the gap on these two smallest and most visually confounded minority classes may benefit from training set expansion (i.e., more labeled data), but expansion alone will likely not suffice. Closing the gap will also require careful joint consideration of model architecture, complexity, hyperparameter configuration and tuning, additional input data streams (e.g., dynamic masks of Sect. 2.3, NIR/SWIR), and explicit treatment of year-to-year distribution shift informed by external factors like weather and surface temperature (and, eventually, basin-to-basin distribution shift). These methodological, temporal, and spatial extensions are in progress and will be released as a future companion paper.

7 Conclusions

We have introduced a benchmark dataset for the classification of supraglacial lake drainage mechanism on the Greenland Ice Sheet, covering $N = 1679$ lakes across the central-west drainage basin during the 2018 and 2019 melt seasons. The dataset is packaged as standardized ML-ready CF-1.8 NetCDF files containing six Sentinel-2 reflectance bands at 10 meter resolution and daily cadence between 1 May to 30 September. Each lake file includes co-registered static and dynamic lake masks, paired water area percentage scalar sequences, and human-assigned drainage mechanism labels.

The five-class taxonomy (ND, HF, MD, LD, CD) (Sect. 2.4.1) distinguishes among physically distinct drainage pathways while remaining operationally distinguishable from Sentinel-2 time series by trained human labelers under the documented labeling protocol (Sect. 2.4.1). We demonstrate the ML-ready capabilities of this dataset in a baseline spatio-temporal classifier task trained on the CW 2019 sets, and tested on the CW 2018 sets, demonstrating that the dataset supports a non-trivial 5-class classifier whose 3-class projection performs at least as well as the projected predictions from existing approaches in the literature on the same lakes (Figs. 9, 8). The per-class breakdown, the data scaling curve (Fig. B2), and the model performance on validation and test sets indicate that model performance gains are to be expected from expansion of the labeled training set, careful consideration of distribution shift, and model optimization (e.g., tuning, architecture enhancements).

The results above identify several important considerations. The class distribution is intrinsically imbalanced (Sect. 2.5), and the per-class performance in the baseline ML classification task further suggests that labeled-set expansion may help improve model performance. The geographic scope of the dataset is currently limited to the CW drainage basin, and therefore it should not be assumed that models which are trained on this region will transfer out-of-the-box to other basins, particularly those with substantially different lake morphologies and class distributions. Temporal coverage has gaps due to cloud cover and Sentinel-2 revisit constraints, which affects transient signals that are particularly important for the HF and MD classes. As such, inclusion of a time series of binary flags indicating the usefulness of a tile at a given timestep may help the model's awareness to these complexities. The inter-rater reliability study (Sect. 2.4.4) indicates that even expert human labelers do not always agree on the mechanism class. The published soft-probability vectors allow users to filter on labeler confidence or train a network explicitly against soft targets.

We release the full dataset and software as described in Sect. 4 and Sect. 5, respectively. We intend this work to serve as a benchmark for spatio-temporal supraglacial lake classification on the GrIS, but also note that this approach may be extended to supraglacial lakes in the Antarctic and other spatio-temporal satellite imagery-informed classification tasks in which geomorphometric information is as important as reduced morphological descriptions of observed features (e.g., agriculture monitoring, civil infrastructure monitoring).

https://doi.org/10.5194/essd-2026-406
Preprint. Discussion started: 18 June 2026

Appendix A: Labeling guidelines

The following text reproduces the labeling guidelines distributed to all human labelers and embedded in the in-app help panel of the labeling GUI (Sect. 2.4.3):

Assign probability estimates for each drainage class. Probabilities must sum to 1.0. The label is automatically set to the class with the highest probability (argmax). Use probability splits (e.g., 0.50/0.50) when uncertain. Flag ambiguous cases for revisit.

A0.1 ND, No drainage.

- Lake does not drain its water in any way (laterally or englacially).
- If there is *any* outlet stream, even a small one, it is **not** ND; that would be LD.
- A lake can still be ND even if streams are visible, as long as those streams flow **into** the lake (inlets), not out of it.
- To distinguish inlet vs. outlet: streams generally taper from wide to narrow in the direction of flow, but this can be tricky, so check whether the lake level drops over time (outlet/drainage) or stays stable/rises (inlet only, likely ND).

A0.2 HF, Hydrofracture (rapid in-basin drainage).

- Rapid drainage (typically < 4 days) where water leaves **mostly within the lake extent**, not via a long outlet stream.
- *Supporting cues*: fracture scarps, ice chunks (calved blocks), sudden area decrease.
- **A visible scarp is not required.** Sentinel-2 is 10 m/pixel, but a hydrofracture itself may be only meters wide in its across-crack direction, so the crack is often sub-pixel.
- **Default to HF** when drainage is rapid and mostly in-basin, even without a clean scarp: the imagery already shows evidence of a rapid englacial event, and you cannot demonstrate it *was not* a hydrofracture.
- **Moulin tie-breaker**: a moulin observed in or near the basin during or after rapid in-basin drainage was possibly spawned by the HF event (HF events often open new moulins). Visible moulin + rapid in-basin drainage → HF, not MD.
- If truly unclear HF vs. MD → label as HF and flag for revisit.

A0.3 MD, Moulin drainage.

- Drainage into a moulin (vertical channel into the ice).
- Tends to be **slower** than hydrofracture.
- Moulin is often **outside** the lake extent; do not confuse with lateral drainage.

- Moulin must be visible within the 512×512 frame (water disappears into ice in-frame).
- Moulins sometimes become more obvious later in the season after snow falls.
- If a lake drains via an outlet stream that terminates in a moulin → label as MD.
- If a lake drains via stream AND separately into a moulin → label as MD.

A0.4 LD, Lateral drainage.

- Lake drains via an outlet stream or slush field.
- If the outlet stream meanders out of frame (even if it might end in a moulin) → LD.
- Lake-to-lake drainage is always LD: if a lake drains into another lake's basin, even if that neighboring lake drains via HF or moulin, the source lake is LD.
- If a lake sits in a crevasse field but clearly drains via an outlet stream → LD, not CD.

A0.5 CD, Crevasse drainage.

- Lake drains into a crevasse field.
- Catastrophic HF-like drainage in a crevasse field → still CD.
- Lake just outside/touching a crevasse field that drains → CD.
- Lake drains via a stream **into** a crevasse field → CD.
- Lake in a crevasse field but clearly drains via an outlet stream out of frame → LD, not CD.

Appendix B: Baseline classifier architecture and training details

The baseline classifier (`LakeDrainageClassifier`) is implemented in PyTorch and totals approximately 140,000 trainable parameters. The architecture is summarized below; the full source code is released alongside the dataset (Sect. 5).

B0.1 Imagery branch.

The 4-channel imagery sequence ($[T,C,H,W] = [153,4,512,512]$) (3 RGB plus lake boundary mask) is loaded into the model. Only the 3 RGB channels are processed frame by frame through a four-block convolutional encoder (FrontCNN). Each block applies `Conv2d` (3×3, padding 1) → `LeakyReLU` → `MaxPool2d` (2×2), with a base channel count of 8 that doubles at each block ($4 \rightarrow 8 \rightarrow 16 \rightarrow 32 \rightarrow 64$), reducing spatial resolution from 512×512 to 32×32 through the convolutional blocks, then resampled to 64×64 by an adaptive max-pool to the configured `frontcnn_out_hw`. The resulting per-frame feature

Table B1. Macro-averaged metrics for the baseline classifier under the combined and cross-year training protocols (Sect. 2.5), evaluated at the best-validation-macro-F_1 epoch, scored on the native 5-class taxonomy (ND/HF/MD/LD/CD). loss: cross-entropy loss; A: accuracy; P: macro-averaged precision; R:macro-averaged recall, and $F_1 \equiv 2\frac{P\times R}{P+R}$. Train rows are overfitting context, not performance claims.

Experiment	Split	n	loss	A	P	R	F_1
Combined	Train	1175	0.93	0.57	0.58	0.61	0.57
Combined	Val	336	1.03	0.60	0.59	0.61	0.59
Combined	Test	168	1.15	0.49	0.50	0.53	0.49
Cross-year	Train	800	0.70	0.72	0.72	0.74	0.72
Cross-year	Val	200	1.16	0.59	0.61	0.62	0.61
Cross-year	Test	679	1.44	0.48	0.48	0.48	0.45

maps are passed through a single-layer convolutional LSTM (ConvLSTM; Shi et al., 2015) with hidden dimension 32 and a 3×3 kernel, which integrates spatial features across the 153-step time axis. Spatial global average pooling over the final hidden state yields a 32-dimensional imagery feature vector per lake.

B0.2 Scalar branch.

The per-lake min–max-normalized water-area time-series p_{water} is processed by a single-layer LSTM with input size 1 and hidden dimension 16. The hidden state at the final timestep yields a 16-dimensional scalar feature vector per lake.

B0.3 Classification head.

The imagery and scalar feature vectors are concatenated into a 48-dimensional representation and passed through a single-hidden-layer multi-layer perceptron (`Linear` 48 → 64 with `ReLU` and dropout $p = 0.3$, followed by `Linear` 64 → 5), producing five-class logits. Class probabilities at inference are obtained by softmax over these logits. Training uses inverse-frequency weighted cross-entropy applied directly to the logits (Sect. 3.2).

These hyperparameters are held fixed throughout this work as a reproducible baseline. Systematic exploration of architecture, capacity, and additional input streams is deferred to subsequent work. The baseline classifier consumes three reflectance bands (red, green, blue) of the imagery tensor together with the water surface-area time series, but the published dataset preserves several additional streams that are not consumed by the baseline. These include the near-infrared (B08) and shortwave-infrared (B11, B12) reflectance bands, the derived per-pixel cloud mask, the static lake boundary mask (Dunmire et al., 2021), the dynamic lake-area masks (Sect. 2.3), the per-timestep `eo_cloud_cover` metadata, and the per-timestep `pct_nans` coordinate. Their exclusion from the baseline reflects a deliberate choice to keep the reference classifier minimal and interpretable, rather than a statement about their utility for future model variants.

Author contributions. Conceptualization: J.H.R. Data curation: J.H.R. and M.G.S. Formal analysis: J.H.R. and E.A. Funding acquisition: C.-Y.L. and L.A.S. Investigation: J.H.R., E.A., N.B.C., and E.L. Methodology: J.H.R., E.A., M.G.S., N.B.C., and E.L. Software: J.H.R., E.A., and M.G.S. Supervision: C.-Y.L. and L.A.S. Validation: J.H.R., N.B.C., and E.L. Visualization: J.H.R. Writing – original draft: J.H.R. Writing – review and editing: all authors.

Competing interests. The authors declare no competing interests.

Acknowledgements. J.H.R, C.-Y.L., M.G.S., N.B.C., and E.L. acknowledge funding from NSF's Office of Polar Programs through OPP-2344690. Support to L.A.S. was provided by the UK Natural Environmental Research Council (NE/Y002369/1) and the Oxford University Press. We thank Leigh Stearns (University of Pennsylvania) for the helpful discussions throughout the development of this work. We also thank SDSS Center for Computation for the access to and support with computational resources for this research, especially Brian Tempero and Bob Clapp. The authors acknowledge the use of Claude Code in assisting in the organization of the data and software repositories but note end results have been human-validated.

https://doi.org/10.5194/essd-2026-406
Preprint. Discussion started: 18 June 2026

References

Andrews, L. C., Catania, G. A., Hoffman, M. J., Gulley, J. D., Lüthi, M. P., Ryser, C., Hawley, R. L., and Neumann, T. A.: Direct observations of evolving subglacial drainage beneath the Greenland Ice Sheet, Nature, 514, 80–83, https://doi.org/10.1038/nature13796, 2014.

Andrews, L. C., Poinar, K., and Trunz, C.: Controls on Greenland moulin geometry and evolution from the Moulin Shape model, The Cryosphere, 16, 2421–2448, https://doi.org/10.5194/tc-16-2421-2022, 2022.

Banwell, A., Hewitt, I., Willis, I., and Arnold, N.: Moulin density controls drainage development beneath the Greenland ice sheet, Journal of Geophysical Research: Earth Surface, 121, 2248–2269, https://doi.org/10.1002/2015JF003801, 2016.

Banwell, A. F., Willis, I. C., and Arnold, N. S.: Modeling subglacial water routing at Paakitsoq, W Greenland, Journal of Geophysical Research: Earth Surface, 118, 1282–1295, https://doi.org/10.1002/jgrf.20093, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/jgrf.20093, 2013.

Bartholomaus, T. C., Anderson, R. S., and Anderson, S. P.: Response of glacier basal motion to transient water storage, Nature Geoscience, 1, 33–37, https://doi.org/10.1038/ngeo.2007.52, 2008.

Bartholomew, I. D., Nienow, P., Sole, A., Mair, D., Cowton, T., King, M. A., and Palmer, S.: Seasonal variations in Greenland Ice Sheet motion: Inland extent and behaviour at higher elevations, Earth and Planetary Science Letters, 307, 271–278, https://doi.org/10.1016/j.epsl.2011.04.014, 2011.

Baumhoer, C. A., Koehler, J., Wouters, B., Lhermitte, S., Dietz, A. J., and Kuenzer, C.: PolarLakes: A Bi-Weekly Dataset of Supraglacial Lakes on Antarctic Ice Shelves from Multi-Sensor Satellite Observations (2015–2024), https://doi.org/10.5194/essd-2025-569, 2026.

Catania, G. A. and Neumann, T. A.: Persistent englacial drainage features in the Greenland Ice Sheet, Geophysical Research Letters, 37, 2009GL041 108, https://doi.org/10.1029/2009GL041108, 2010.

Catania, G. A., Neumann, T. A., and Price, S. F.: Characterizing englacial drainage in the ablation zone of the Greenland ice sheet, Journal of Glaciology, 54, 567–578, https://doi.org/10.3189/002214308786570854, 2008.

Christoffersen, P., Bougamont, M., Hubbard, A., Doyle, S. H., Grigsby, S., and Pettersson, R.: Cascading lake drainage on the Greenland Ice Sheet triggered by tensile shock and fracture, Nature Communications, 9, 1064, https://doi.org/10.1038/s41467-018-03420-8, 2018.

Chu, V. W.: Greenland ice sheet hydrology: A review, Progress in Physical Geography: Earth and Environment, 38, 19–54, https://doi.org/10.1177/0309133313507075, 2014.

Chudley, T. R., Christoffersen, P., Doyle, S. H., Dowling, T. P. F., Law, R., Schoonman, C. M., Bougamont, M., and Hubbard, B.: Controls on Water Storage and Drainage in Crevasses on the Greenland Ice Sheet, Journal of Geophysical Research: Earth Surface, 126, e2021JF006 287, https://doi.org/10.1029/2021JF006287, _eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2021JF006287, 2021.

Chudley, T. R., Howat, I. M., King, M. D., and MacKie, E. J.: Increased crevassing across accelerating Greenland Ice Sheet margins, Nature Geoscience, 18, 148–153, https://doi.org/10.1038/s41561-024-01636-6, 2025.

Cohen, J.: A Coefficient of Agreement for Nominal Scales, Educational and Psychological Measurement, 20, 37–46, https://doi.org/10.1177/001316446002000104, 1960.

Cooley, S. W. and Christoffersen, P.: Observation Bias Correction Reveals More Rapidly Draining Lakes on the Greenland Ice Sheet, Journal of Geophysical Research: Earth Surface, 122, 1867–1881, https://doi.org/10.1002/2017JF004255, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/2017JF004255, 2017.

Corr, D., Leeson, A., McMillan, M., Zhang, C., and Barnes, T.: An inventory of supraglacial lakes and channels across the West Antarctic Ice Sheet, Earth System Science Data, 14, 209–228, https://doi.org/10.5194/essd-14-209-2022, 2022.

Das, S. B., Joughin, I., Behn, M. D., Howat, I. M., King, M. A., Lizarralde, D., and Bhatia, M. P.: Fracture Propagation to the Base of the Greenland Ice Sheet During Supraglacial Lake Drainage, Science, 320, 778–781, https://doi.org/10.1126/science.1153360, 2008.

Dell, R. L., Banwell, A. F., Willis, I. C., Arnold, N. S., Halberstadt, A. R. W., Chudley, T. R., and Pritchard, H. D.: Supervised classification of slush and ponded water on Antarctic ice shelves using Landsat 8 imagery, Journal of Glaciology, 68, 401–414, https://doi.org/10.1017/jog.2021.114, 2022.

Dell, R. L., Willis, I. C., Arnold, N. S., Banwell, A. F., and de Roda Husman, S.: Substantial contribution of slush to meltwater area across Antarctic ice shelves, Nature Geoscience, 17, 624–630, https://doi.org/10.1038/s41561-024-01466-6, 2024.

Dirscherl, M., Dietz, A. J., Kneisel, C., and Kuenzer, C.: Automated Mapping of Antarctic Supraglacial Lakes Using a Machine Learning Approach, Remote Sensing, 12, 1203, https://doi.org/10.3390/rs12071203, 2020.

Dirscherl, M. C., Dietz, A. J., and Kuenzer, C.: Seasonal evolution of Antarctic supraglacial lakes in 2015–2021 and links to environmental controls, The Cryosphere, 15, 5205–5226, https://doi.org/10.5194/tc-15-5205-2021, 2021.

Doyle, S. H., Hubbard, A., Fitzpatrick, A. A. W., van As, D., Mikkelsen, A. B., Pettersson, R., and Hubbard, B.: Persistent flow acceleration within the interior of the Greenland ice sheet, Geophysical Research Letters, 41, 899–905, https://doi.org/10.1002/2013GL058933, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/2013GL058933, 2014.

Drusch, M., Del Bello, U., Carlier, S., Colin, O., Fernandez, V., Gascon, F., Hoersch, B., Isola, C., Laberinti, P., Martimort, P., Meygret, A., Spoto, F., Sy, O., Marchese, F., and Bargellini, P.: Sentinel-2: ESA's Optical High-Resolution Mission for GMES Operational Services, Remote Sensing of Environment, 120, 25–36, https://doi.org/10.1016/j.rse.2011.11.026, 2012.

Dunmire, D., Banwell, A. F., Wever, N., Lenaerts, J. T. M., and Datta, R. T.: Contrasting regional variability of buried meltwater extent over 2 years across the Greenland Ice Sheet, The Cryosphere, 15, 2983–3005, https://doi.org/10.5194/tc-15-2983-2021, 2021.

Dunmire, D., Subramanian, A. C., Hossain, E., Gani, M. O., Banwell, A. F., Younas, H., and Myers, B.: Greenland Ice Sheet Wide Supraglacial Lake Evolution and Dynamics: Insights From the 2018 and 2019 Melt Seasons, Earth and Space Science, 12, e2024EA003 793, https://doi.org/10.1029/2024EA003793, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1029/2024EA003793, 2025.

European Space Agency: SENTINEL-2 User Handbook, Tech. Rep. GMES-S1OP-EOPG-TN-13-0001, Issue 1, Rev 2, European Space Agency, https://sentinels.copernicus.eu/documents/247904/685211/Sentinel-2_User_Handbook, 2015.

Fan, Y., Ke, C.-Q., Luo, L., Shen, X., Livingstone, S. J., and Lea, J. M.: Expansion of supraglacial lake area, volume and extent on the Greenland ice sheet from 1985 to 2023, Journal of Glaciology, 71, e4, https://doi.org/10.1017/jog.2024.87, 2025.

Fleiss, J. L.: Measuring Nominal Scale Agreement Among Many Raters, Psychological Bulletin, 76, 378–382, https://doi.org/10.1037/h0031619, 1971.

Hochreiter, S. and Schmidhuber, J.: Long Short-Term Memory, Neural Computation, 9, 1735–1780, https://doi.org/10.1162/neco.1997.9.8.1735, 1997.

Hoffman, M. J., Catania, G. A., Neumann, T. A., Andrews, L. C., and Rumrill, J. A.: Links between acceleration, melting, and supraglacial lake drainage of the western Greenland Ice Sheet, Journal of Geophysical Research: Earth Surface, 116, https://doi.org/10.1029/2010JF001934, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1029/2010JF001934, 2011.

Hoffman, M. J., Perego, M., Andrews, L. C., Price, S. F., Neumann, T. A., Johnson, J. V., Catania, G., and Lüthi, M. P.: Widespread Moulin Formation During Supraglacial Lake Drainages in Greenland, Geophysical Research Letters, 45, 778–788, https://doi.org/10.1002/2017GL075659, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/2017GL075659, 2018.

Joseph, G.: stackstac: Load a STAC collection into xarray with dask, https://github.com/gjoseph92/stackstac, 2024.

Kang Yang and Smith, L. C.: Supraglacial Streams on the Greenland Ice Sheet Delineated From Combined Spectral–Shape Information in High-Resolution Satellite Imagery, IEEE Geoscience and Remote Sensing Letters, 10, 801–805, https://doi.org/10.1109/LGRS.2012.2224316, 2013.

Kingma, D. P. and Ba, J.: Adam: A Method for Stochastic Optimization, https://doi.org/10.48550/arXiv.1412.6980, arXiv:1412.6980 [cs.LG], 2017.

Koziol, C., Arnold, N., Pope, A., and Colgan, W.: Quantifying supraglacial meltwater pathways in the Paakitsoq region, West Greenland, Journal of Glaciology, 63, 464–476, https://doi.org/10.1017/jog.2017.5, 2017.

Lai, C.-Y., Stevens, L. A., Chase, D. L., Creyts, T. T., Behn, M. D., Das, S. B., and Stone, H. A.: Hydraulic transmissivity inferred from ice-sheet relaxation following Greenland supraglacial lake drainages, Nature Communications, 12, 3955, https://doi.org/10.1038/s41467-021-24186-6, number: 1, 2021.

Lampkin, D. J., Amador, N., Parizek, B. R., Farness, K., and Jezek, K.: Drainage from water-filled crevasses along the margins of Jakobshavn Isbræ: A potential catalyst for catchment expansion, Journal of Geophysical Research: Earth Surface, 118, 795–813, https://doi.org/10.1002/jgrf.20039, _eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1002/jgrf.20039, 2013.

Leeson, A. A., Shepherd, A., Briggs, K., Howat, I., Fettweis, X., Morlighem, M., and Rignot, E.: Supraglacial lakes on the Greenland ice sheet advance inland under warming climate, Nature Climate Change, 5, 51–55, https://doi.org/10.1038/nclimate2463, number: 1, 2015.

Louis, J., Debaecker, V., Pflug, B., Main-Knorn, M., Bieniarz, J., Mueller-Wilm, U., Cadau, E., and Gascon, F.: Sentinel-2 Sen2Cor: L2A Processor for Users, in: Proceedings of the Living Planet Symposium 2016, ESA SP-740, pp. 1–8, ESA Publications Division, Prague, Czech Republic, 2016.

Main-Knorn, M., Pflug, B., Louis, J., Debaecker, V., Müller-Wilm, U., and Gascon, F.: Sen2Cor for Sentinel-2, p. 3, https://doi.org/10.1117/12.2278218, 2017.

McFeeters, S. K.: The use of the Normalized Difference Water Index (NDWI) in the delineation of open water features, International Journal of Remote Sensing, 17, 1425–1432, https://doi.org/10.1080/01431169608948714, 1996.

Morriss, B. F., Hawley, R. L., Chipman, J. W., Andrews, L. C., Catania, G. A., Hoffman, M. J., Lüthi, M. P., and Neumann, T. A.: A ten-year record of supraglacial lake evolution and rapid drainage in West Greenland using an automated processing algorithm for multispectral imagery, The Cryosphere, 7, 1869–1877, https://doi.org/10.5194/tc-7-1869-2013, 2013.

Nienow, P. W., Sole, A. J., Slater, D. A., and Cowton, T. R.: Recent Advances in Our Understanding of the Role of Meltwater in the Greenland Ice Sheet System, Current Climate Change Reports, 3, 330–344, https://doi.org/10.1007/s40641-017-0083-9, 2017.

Rignot, E. and Mouginot, J.: Ice flow in Greenland for the International Polar Year 2008–2009, Geophysical Research Letters, 39, https://doi.org/10.1029/2012GL051634, _eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2012GL051634, 2012.

Rines, J. H., Lai, C.-Y., Abrahams, E., Shahin, M. G., Coffey, N. B., Lee, E., and Stevens, L. A.: Central West Greenland Supraglacial Lake Drainage Classification Dataset (2018–2019), https://doi.org/10.25740/sf350xp4038, 2026a.

Rines, J. H., Lai, C.-Y., and Wang, Y.: Longitudinal stress induced by basal slippery patch, https://doi.org/10.31223/X50J2P, preprint, 2026b.

Rußwurm, M. and Körner, M.: Multi-Temporal Land Cover Classification with Sequential Recurrent Encoders, ISPRS International Journal of Geo-Information, 7, 129, https://doi.org/10.3390/ijgi7040129, arXiv:1802.02080 [cs], 2018.

https://doi.org/10.5194/essd-2026-406
Preprint. Discussion started: 18 June 2026

Ryan, J. C., Datta, R. T., and Cooley, S. W.: Mechanisms of Surface Meltwater Ponding and Drainage on the Greenland Ice Sheet Revealed Using SkySat Imagery and Deep Learning, AGU Advances, 7, e2025AV002 030, https://doi.org/10.1029/2025AV002030, _eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2025AV002030, 2026.

Selmes, N., Murray, T., and James, T. D.: Fast draining lakes on the Greenland Ice Sheet, Geophysical Research Letters, 38, https://doi.org/10.1029/2011GL047872, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1029/2011GL047872, 2011.

Selmes, N., Murray, T., and James, T. D.: Characterizing supraglacial lake drainage and freezing on the Greenland Ice Sheet, The Cryosphere Discussions, 7, 475–505, https://doi.org/10.5194/tcd-7-475-2013, 2013.

Shi, X., Chen, Z., Wang, H., Yeung, D.-Y., Wong, W.-k., and Woo, W.-c.: Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting, https://doi.org/10.48550/arXiv.1506.04214, arXiv:1506.04214 [cs], 2015.

Smith, L. C., Chu, V. W., Yang, K., Gleason, C. J., Pitcher, L. H., Rennermalm, A. K., Legleiter, C. J., Behar, A. E., Overstreet, B. T., Moustafa, S. E., Tedesco, M., Forster, R. R., LeWinter, A. L., Finnegan, D. C., Sheng, Y., and Balog, J.: Efficient meltwater drainage through supraglacial streams and rivers on the southwest Greenland ice sheet, Proceedings of the National Academy of Sciences, 112, 1001–1006, https://doi.org/10.1073/pnas.1413024112, 2015.

Source, M. O., McFarland, M., Emanuele, R., Morris, D., and Augspurger, T.: microsoft/PlanetaryComputer: October 2022, https://doi.org/10.5281/zenodo.7261897, 2022.

Stevens, L. A., Behn, M. D., McGuire, J. J., Das, S. B., Joughin, I., Herring, T., Shean, D. E., and King, M. A.: Greenland supraglacial lake drainages triggered by hydrologically induced basal slip, Nature, 522, 73–76, https://doi.org/10.1038/nature14480, number: 7554, 2015.

Stevens, L. A., Behn, M. D., Das, S. B., Joughin, I., Noël, B. P. Y., van den Broeke, M. R., and Herring, T.: Greenland Ice Sheet flow response to runoff variability, Geophysical Research Letters, 43, 11 295–11 303, https://doi.org/10.1002/2016GL070414, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/2016GL070414, 2016.

Stevens, L. A., Das, S. B., Behn, M. D., McGuire, J. J., Lai, C.-Y., Joughin, I., Larochelle, S., and Nettles, M.: Elastic Stress Coupling Between Supraglacial Lakes, Journal of Geophysical Research: Earth Surface, 129, e2023JF007 481, https://doi.org/10.1029/2023JF007481, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1029/2023JF007481, 2024.

Stevens, L. A., Nettles, M., Larochelle, S., Okal, M., Falconer, E., Turner, N., Rines, J., Lai, C.-Y., and Lu, G.: Ice-sheet hydro-fracture not advanced inland by lower-elevation lake drainages in Kalaallit Nunaat, Nature Communications, 17, 4598, https://doi.org/10.1038/s41467-026-73033-z, 2026.

Tedesco, M., Willis, I. C., Hoffman, M. J., Banwell, A. F., Alexander, P., and Arnold, N. S.: Ice dynamic response to two modes of surface lake drainage on the Greenland ice sheet, Environmental Research Letters, 8, 034 007, https://doi.org/10.1088/1748-9326/8/3/034007, 2013.

Tedstone, A. J. and Machguth, H.: Increasing surface runoff from Greenland's firn areas, Nature Climate Change, 12, 672–676, https://doi.org/10.1038/s41558-022-01371-z, 2022.

Torres, R., Snoeij, P., Geudtner, D., Bibby, D., Davidson, M., Attema, E., Potin, P., Rommen, B., Floury, N., Brown, M., Traver, I. N., Deghaye, P., Duesmann, B., Rosich, B., Miranda, N., Bruno, C., L'Abbate, M., Croci, R., Pietropaolo, A., Huchler, M., and Rostan, F.: GMES Sentinel-1 mission, Remote Sensing of Environment, 120, 9–24, https://doi.org/10.1016/j.rse.2011.05.028, 2012.

Wang, Y. and Sugiyama, S.: Supraglacial lake evolution on Tracy and Heilprin Glaciers in northwestern Greenland from 2014 to 2021, Remote Sensing of Environment, 303, 114 006, https://doi.org/10.1016/j.rse.2024.114006, 2024.

Williamson, A. G., Banwell, A. F., Willis, I. C., and Arnold, N. S.: Dual-satellite (Sentinel-2 and Landsat 8) remote sensing of supraglacial lakes in Greenland, The Cryosphere, 12, 3045–3065, https://doi.org/10.5194/tc-12-3045-2018, 2018a.

Williamson, A. G., Willis, I. C., Arnold, N. S., and Banwell, A. F.: Controls on rapid supraglacial lake drainage in West Greenland: an Exploratory Data Analysis approach, Journal of Glaciology, 64, 208–226, https://doi.org/10.1017/jog.2018.8, 2018b.

Zheng, L., Li, L., Chen, Z., He, Y., Mo, L., Chen, D., Hu, Q., Wang, L., Liang, Q., and Cheng, X.: Multi-sensor imaging of winter buried lakes in the Greenland Ice Sheet, Remote Sensing of Environment, 295, 113 688, https://doi.org/10.1016/j.rse.2023.113688, 2023.

Zwally, H. J., Abdalati, W., Herring, T., Larson, K., Saba, J., and Steffen, K.: Surface Melt-Induced Acceleration of Greenland Ice-Sheet Flow, Science, 297, 218–222, https://doi.org/10.1126/science.1072708, 2002.

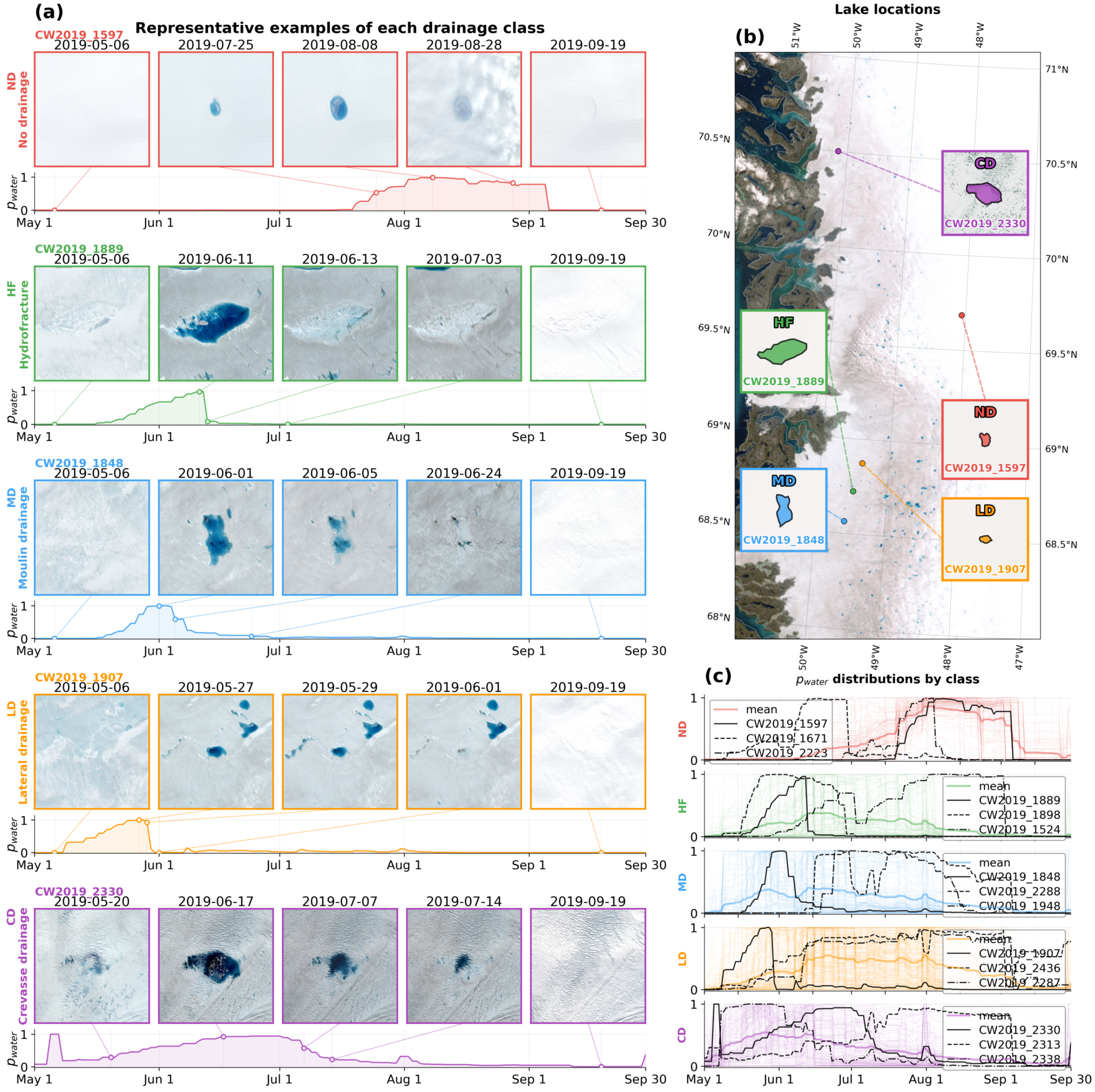
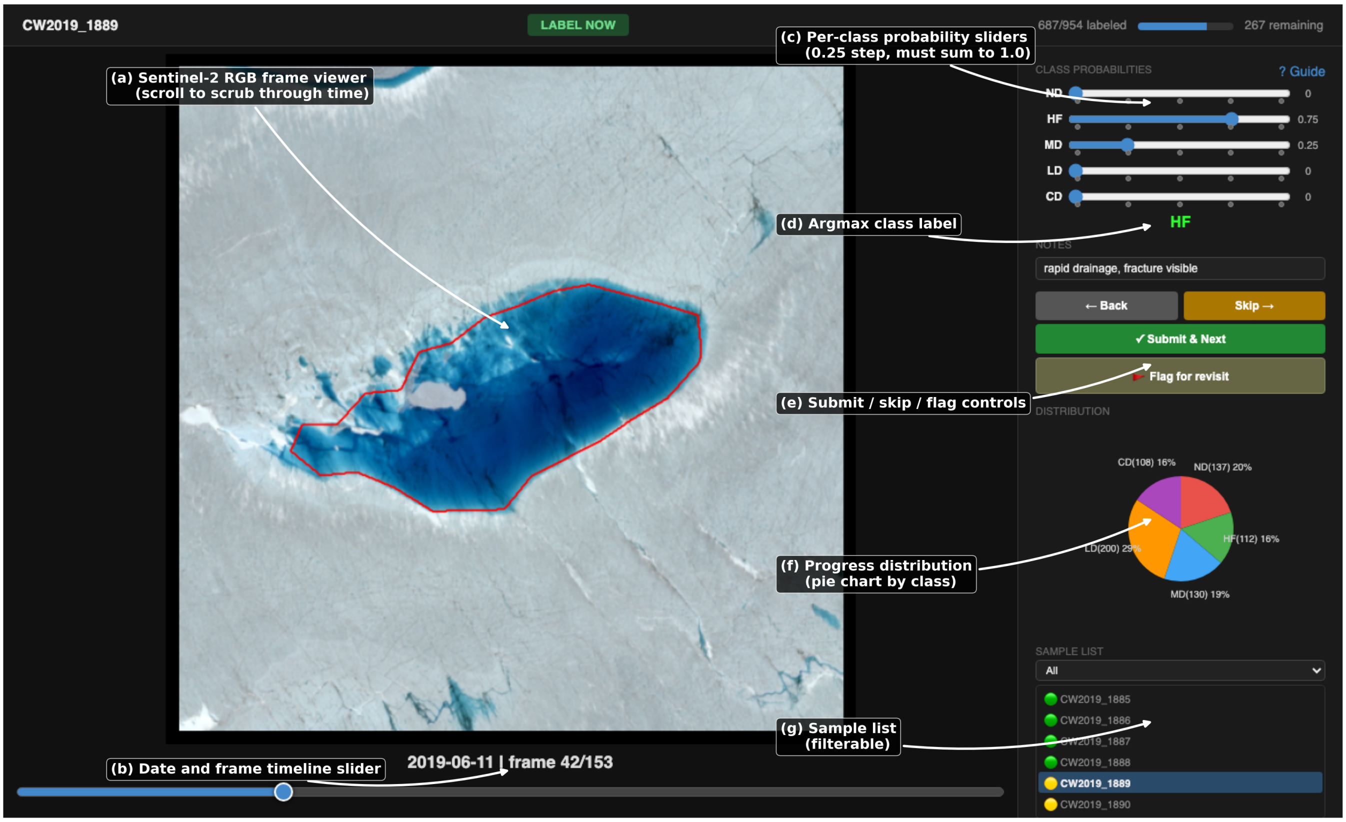


Figure 4. Representative examples of the five-class supraglacial lake drainage mechanism taxonomy from the 2019 melt season. **(a)** For each class (ND: no drainage, HF: hydrofracture, MD: moulin drainage, LD: lateral drainage, and CD: crevasse drainage), Sentinel-2 images from five key timesteps showing the example lakes' evolution, with the lake's daily fractional water surface-area extent series p_water from Dunmire et al. (2025) plotted beneath. **(b)** Geographic locations of the five example lakes within the central west Greenland region (Esri World Imagery basemap | Powered by Esri). Insets show each lake's Dunmire et al. (2021) polygon filled in the class color. **(c)** Per-class p_water time series for the full labeled population of each class (faint color), with the time series mean (bold curve) and three representative lakes per class (black curves) illustrating the variability in the scalar water surface-area time series p_water across (and within) classes.

https://doi.org/10.5194/essd-2026-406
Preprint. Discussion started: 18 June 2026



ND 0
HF 0.75
MD 0.25
LD 0
CD 0

HF

rapid drainage, fracture visible

All

- CW2019_1885
- CW2019_1886
- CW2019_1887
- CW2019_1888
- CW2019_1889
- CW2019_1890

2019-06-11 | frame 42/153

Figure 5. Labeling interface used to assign soft-probability drainage mechanism labels to each per-lake timestack (e.g., CW2019_1889 shown). **(a)** Main view: lake-centered Sentinel-2 imagery at the currently selected timestep, time-scrubbable by mouse scroll, arrow keys, or (b). **(b)** Frame timeline slider. **(c)** Per-class probability sliders, discretized at 25% intervals and constrained to sum to 1.0. **(d)** Hard label, assigned as the argmax of the soft probability vector, below which sits free-text notes field. **(e)** Submit, back, skip and “Flag for revisit” controls. **(f)** Pie chart of per-class label distribution. **(g)** Filterable sample list for access navigation across the full lake population. The GUI is implemented in `sat-tile-stack.labeling` (Sect. 2.4.3) in our workflow.

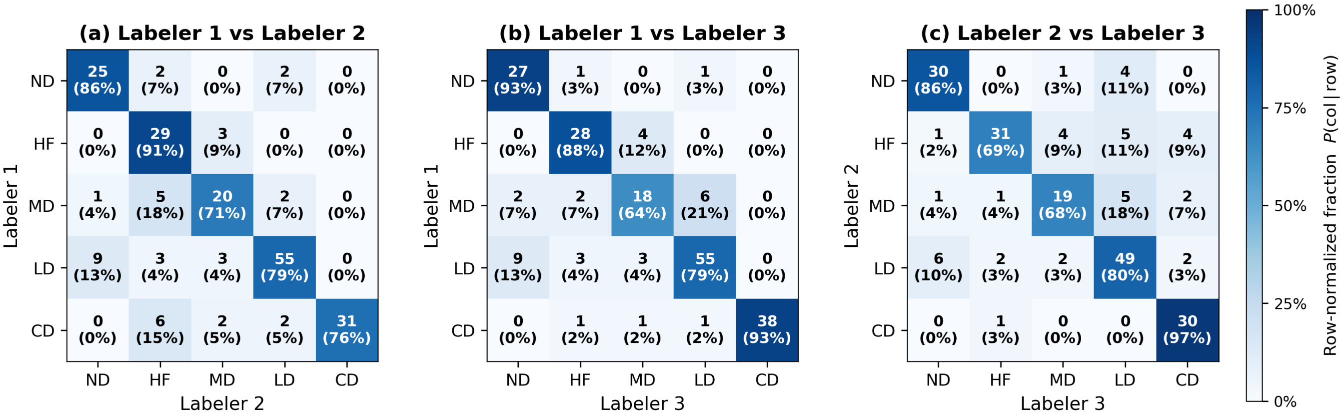


Figure 6. Comparison between three labelers' classification of a random subset of $n = 200$ lakes (100 from 2018, 100 from 2019; Sect. 2.4.4). Each panel shows a row-normalized pairwise confusion matrix between two labelers: **(a)** the lead author (Labeler 1) and Labeler 2, **(b)** Labeler 1 and Labeler 3, and **(c)** Labelers 2 and 3. Within each panel, the cell at row i and column j reports the count (top number) and row-normalized fraction (percentage, colorbar) of lakes that the row-axis labeler assigned to class i and the column-axis labeler assigned to class j, $P(\text{column label} \mid \text{row label})$. Rows sum to 100 %. Diagonal cells quantify per-class agreement while off-diagonal cells display disagreements. Overall three-labeler agreement is 72.5 %. Unsurprisingly, the largest pairwise disagreements concentrate along the MD row, reflecting the visual difficulty of identifying moulin drainage from Sentinel-2 imagery at 10 m spatial resolution.

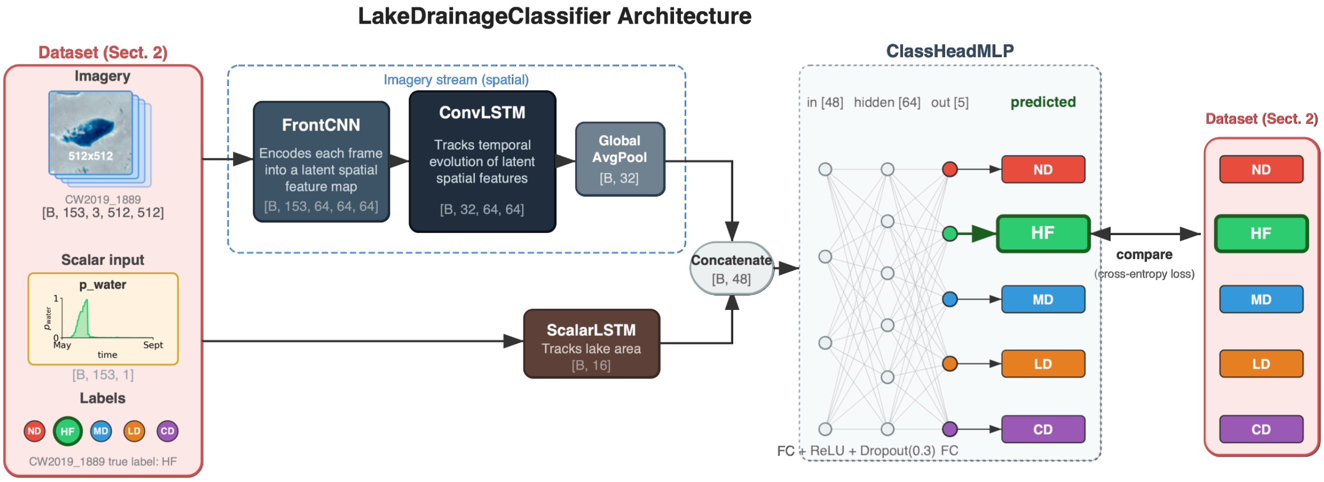


Figure 7. Architecture of the baseline `LakeDrainageClassifier`, a two-stream temporal recurrent neural network with ≈140,000 learnable parameters (Sect. 3, Appendix B) which ingests the dataset (left, Sect. 2). The imagery branch (top) takes the per-lake Sentinel-2 time series and learns temporal evolution of spatial features in each frame, while the scalar branch (bottom) takes the daily water surface-area time series p_water and learns its temporal evolution. The two multimodal datasets are joined together and passed through a fully connected (FC) classification neural network to produce a prediction for drainage mechanism class (ND, HF, MD, LD, CD) which are compared to the human-assigned hand-labels in the dataset (right). Square-bracketed tuples (e.g. $[B, 153, 3, 512, 512]$, where B is batch size) denote tensor dimensions at each stage as the data is processed by the network.

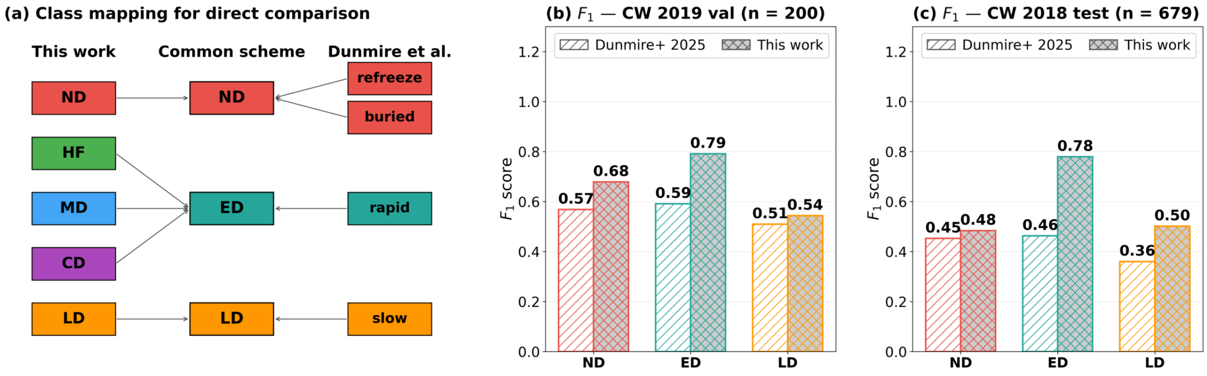


Figure 8. Three-class comparison of the cross-year baseline (this work) compared to the automated time-series based classifier of Dunmire et al. (2025) on the same lakes. **(a)** Class mapping diagram. Note that the two classification results are not directly comparable, as the classes are not exactly the same. However, to assess the impact of incorporating spatial information from imagery (this work), we attempt a preliminary comparison by mapping our 5-class taxonomy (ND, HF, MD, CD, LD) and the 4-class taxonomy (refreeze, buried, rapid, slow) from Dunmire et al. (2025) onto a common scheme: no drainage (ND), englacial drainage (ED = HF + MD + CD; refreeze + buried), and lateral drainage (LD; slow). **(b)** Per-class F_1 on the CW 2019 validation set ($n = 200$). **(c)** Per-class F_1 on the CW 2018 test set ($n = 679$). Both methods are scored against our 5-class manual labels projected to the common scheme.

https://doi.org/10.5194/essd-2026-406
Preprint. Discussion started: 18 June 2026

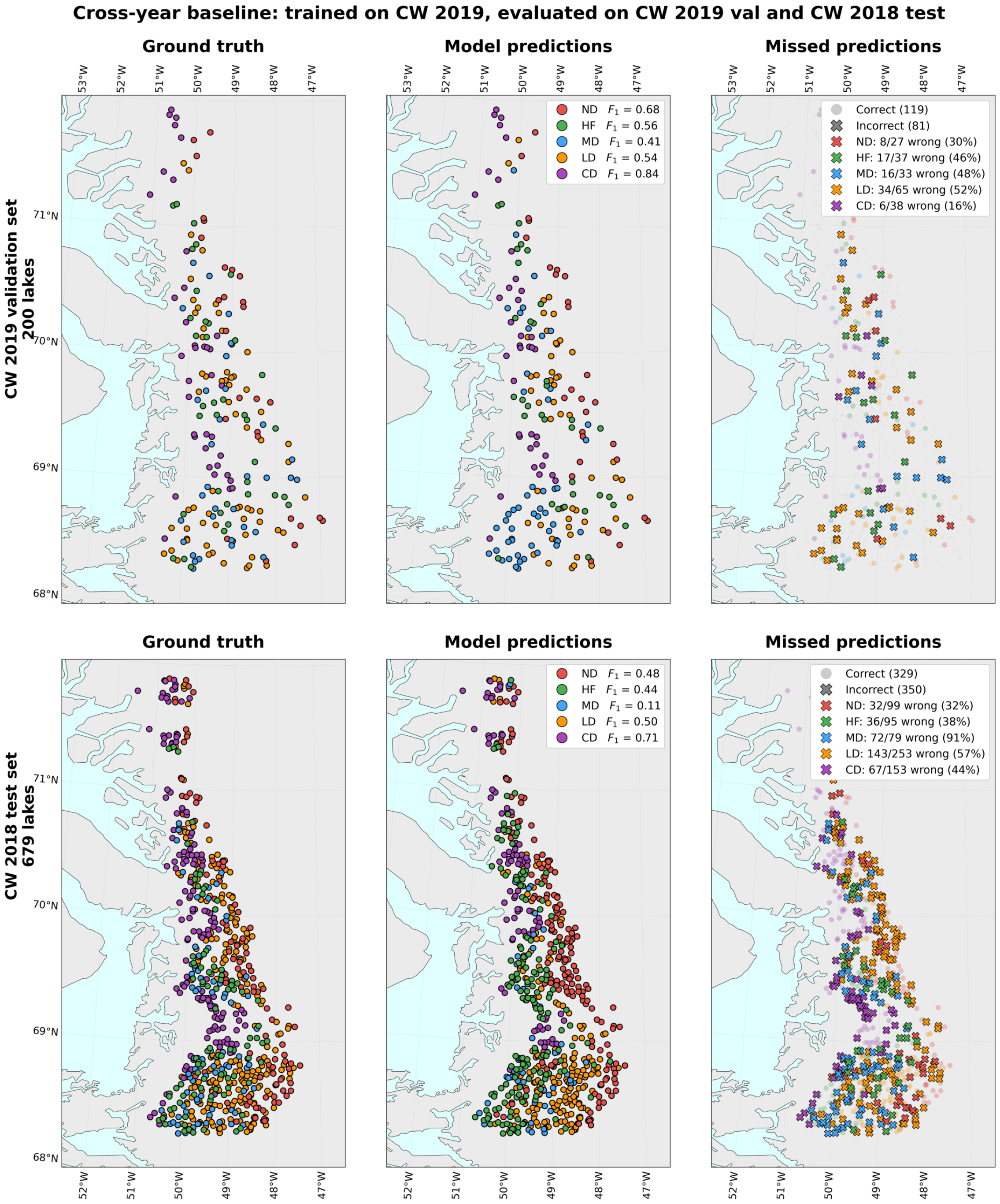


Figure 9. Cross-year baseline model results: trained on the CW 2019 training data and evaluated on the CW 2019 validation set (top row, $n = 200$) and the held-out CW 2018 test set (bottom row, $n = 679$). **Left:** ground-truth (expert) class labels at each lake's location. **Center:** model predictions with per-class F_1 in the legend. **Right:** prediction outcomes showing correct (light color) and incorrect (crosses colored by the true class).

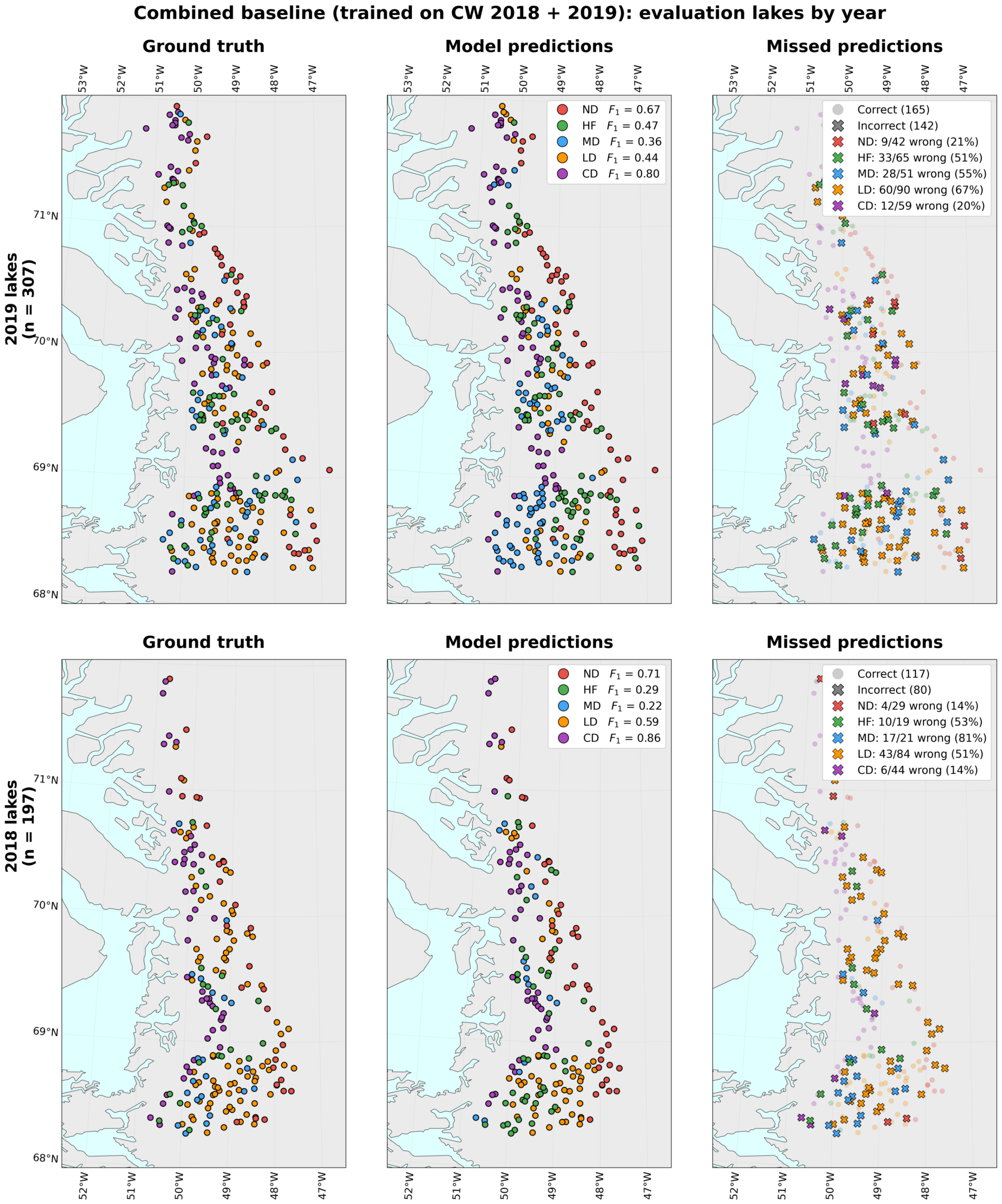
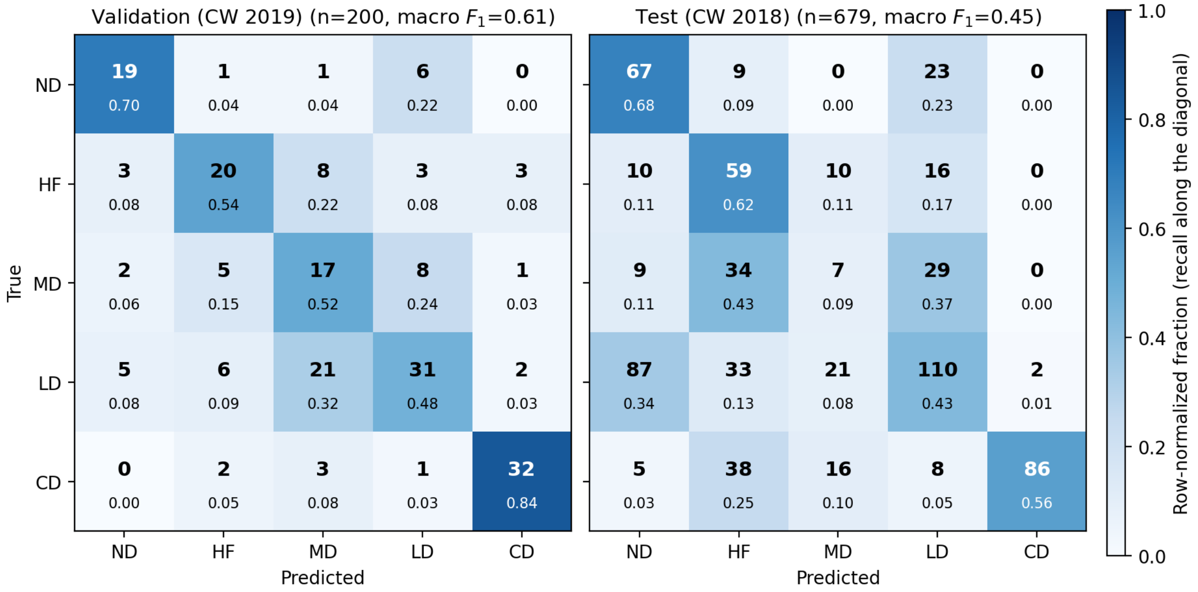


Figure 10. Combined baseline model results: trained on the combined CW 2018 + 2019 training data and evaluated on the held-out validation and test lakes, showing 2019 lakes (top row, $n = 307$) and 2018 lakes (bottom row, $n = 197$). **Left:** ground-truth (expert) class labels at each lake's location. **Center:** model predictions with per-class F_1 in the legend. **Right:** prediction outcomes showing correct (light color) and incorrect (crosses colored by the true class).



Validation (CW 2019) (n=200, macro F_1=0.61)

True \ Predicted	ND	HF	MD	LD	CD
ND	**19** 0.70	**1** 0.04	**1** 0.04	**6** 0.22	**0** 0.00
HF	**3** 0.08	**20** 0.54	**8** 0.22	**3** 0.08	**3** 0.08
MD	**2** 0.06	**5** 0.15	**17** 0.52	**8** 0.24	**1** 0.03
LD	**5** 0.08	**6** 0.09	**21** 0.32	**31** 0.48	**2** 0.03
CD	**0** 0.00	**2** 0.05	**3** 0.08	**1** 0.03	**32** 0.84

Test (CW 2018) (n=679, macro F_1=0.45)

True \ Predicted	ND	HF	MD	LD	CD
ND	**67** 0.68	**9** 0.09	**0** 0.00	**23** 0.23	**0** 0.00
HF	**10** 0.11	**59** 0.62	**10** 0.11	**16** 0.17	**0** 0.00
MD	**9** 0.11	**34** 0.43	**7** 0.09	**29** 0.37	**0** 0.00
LD	**87** 0.34	**33** 0.13	**21** 0.08	**110** 0.43	**2** 0.01
CD	**5** 0.03	**38** 0.25	**16** 0.10	**8** 0.05	**86** 0.56

Figure B1. Confusion matrices for the cross-year baseline. Left: CW2019 validation set ($n = 200$). Right: CW2018 test set ($n = 679$). Rows are true classes, columns are predicted classes, and cell shading is normalized within each row.

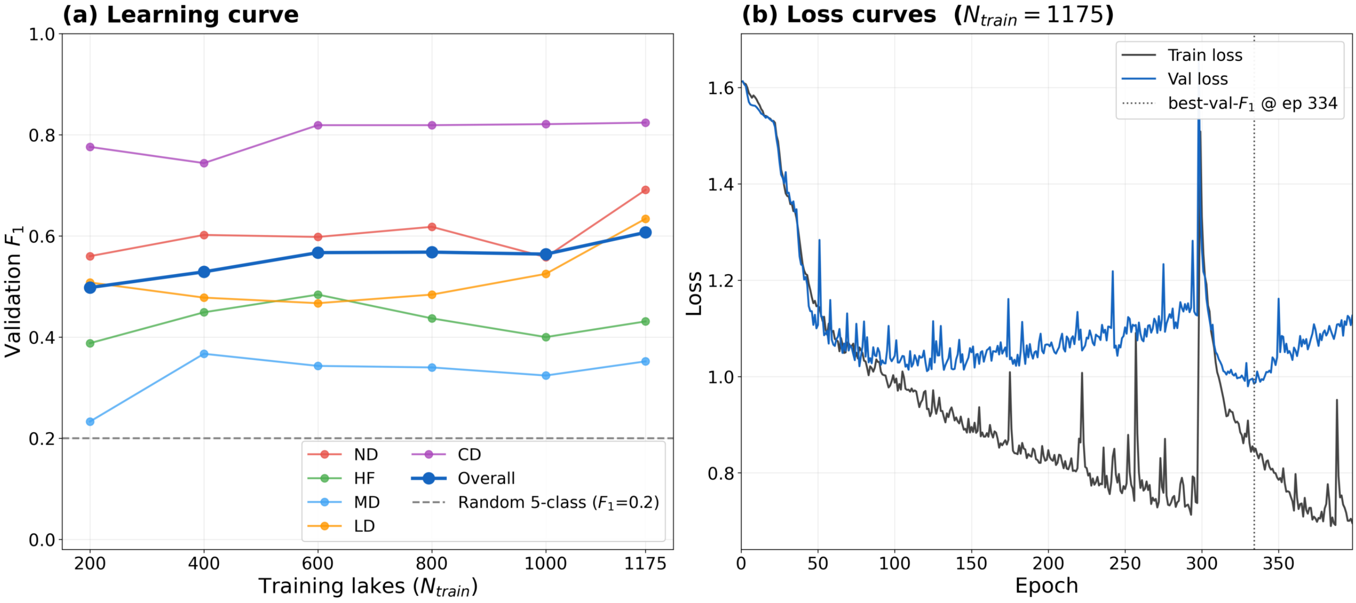


Figure B2. Data scaling curves and loss curves for the model under the combined training protocol. **(a)** Data-scaling curve: per-class and macro-averaged validation F_1 as a function of training-set size $N \in \{200, 400, 600, 800, 1000, 1175\}$, with all other hyperparameters held fixed at the values in Table 2. Each point is the peak over the run's full training extent; macro F_1 is logged every epoch but per-class F_1 only every 10, so per-class markers are sampled at the nearest 10-epoch checkpoint to the peak and need not average exactly to the overall line. The dashed gray denotes the F_1 score of random guesses ($= 0.2$) with no predictive power. **(b)** Train/validation loss for the canonical $N = 1175$ run. The dotted line marks the best validation F_1 epoch (the model-selection checkpoint), after which the widening gap between training and validation losses reflects overfitting.