



OPEN

DATA DESCRIPTOR

A Collaborative and Scalable Geospatial Data Set for Arctic Retrogressive Thaw Slumps with Data Standards

Yili Yang¹✉, Heidi Rodenhizer¹✉, Brendan M. Rogers¹✉, Jacqueline Dean¹, Ridhima Singh¹, Tiffany Windholz¹, Amanda Poston¹, Stefano Potter¹, Scott Zolkos¹, Greg Fiske¹, Jennifer Watts¹, Lingcao Huang², Chandi Witharana³, Ingmar Nitze⁴, Nina Nesterova⁴, Sophia Barth⁴, Guido Grosse⁴, Trevor Lantz⁵, Alexandra Runge⁶, Luigi Lombardo⁷, Ionut Cristi Nicu⁸, Lena Rubensdotter^{9,10}, Eirini Makopoulou¹¹ & Susan Natali¹

Arctic permafrost is undergoing rapid changes due to climate warming in high latitudes. Retrogressive thaw slumps (RTS) are one of the most abrupt and impactful thermal-denudation events that change Arctic landscapes and accelerate carbon feedbacks. Their spatial distribution remains poorly characterised due to time-intensive conventional mapping methods. While numerous RTS studies have published standalone digitisation datasets, the lack of a centralised, unified database has limited their utilisation, affecting the scale of RTS studies and the generalisation ability of deep learning models. To address this, we established the Arctic Retrogressive Thaw Slumps (ARTS) dataset containing 23,529 RTS-present and 20,434 RTS-absent digitisations from 20 standalone datasets. We also proposed a Data Curation Framework as a working standard for RTS digitisations. This dataset is designed to be comprehensive, accessible, contributable, and adaptable for various RTS-related studies. This dataset and its accompanying curation framework establish a foundation for enhanced collaboration in RTS research, facilitating standardised data sharing and comprehensive analyses across the Arctic permafrost research community.

Background & Summary

The terrestrial permafrost region in the northern hemisphere stores roughly 1600 Pg of carbon, and permafrost is warming and thawing, both through more gradual active layer deepening and through relatively abrupt thermokarst processes^{1–3}. Recent climate change has decreased permafrost extent by approximately 7%⁴, and losses of up to 70% are expected by 2100 under low mitigation scenarios⁵. Permafrost thaw can result in rapid decomposition of soil carbon, releasing greenhouse gasses (GHG, mainly CO₂ and CH₄) into the atmosphere^{2,6}. However, the minority of current Earth system models that include permafrost carbon only represent 'gradual

¹Woodwell Climate Research Center, 149 Woods Hole Road, Falmouth, MA, 02540-1644, USA. ²Institute of Space and Earth Information Science, The Chinese University of Hong Kong, Shatin, N.T., Hong Kong, China. ³Department of Natural Resources and the Environment, University of Connecticut, Storrs, CT, 06269, USA. ⁴Alfred Wegener Institute, Permafrost Research Section, Telegrafenberg A45, 14473, Potsdam, Germany. ⁵School of Environmental Studies, University of Victoria David Turpin Building, B243, Victoria, BC, Canada. ⁶Helmholtz Centre Potsdam GFZ German Research Centre for Geosciences, Telegrafenberg, 14473, Potsdam, Germany. ⁷Faculty of Geo-Information Science and Earth Observation, University of Twente, Enschede, the Netherlands. ⁸High North Department, Norwegian Institute for Cultural Heritage Research (NIKU), Fram Centre, N-9296, Tromsø, Norway. ⁹Geohazard and Earth Observation, Geological Survey of Norway (NGU), P.O. Box 6315 Torgarden, 7491, Trondheim, Norway. ¹⁰Department of Arctic Geology, The University Centre in Svalbard (UNIS), P.O. Box 156, 9171, Longyearbyen, Norway. ¹¹Geography Research Unit, University of Oulu, Oulu, 90014, Finland. ✉e-mail: yyang@woodwellclimate.org; hrodenhizer@woodwellclimate.org; brogers@woodwellclimate.org

thaw⁷, neglecting the impact of rapid terrain collapse following thaw (i.e., thermokarst) and its biogeochemical consequences. Early predictions of future permafrost carbon release estimated a significant increase in GHG emissions attributed to intensifying thermokarst activity⁶. One fundamental limitation for understanding the importance of abrupt thaw on Arctic landscapes and carbon feedback is the lack of geospatial products describing abrupt thaw distribution and changes over time.

Retrogressive Thaw Slumps (RTS) are slope failures that develop in ice-rich terrain. They are one of the most rapid thermokarst features⁸ in permafrost regions. RTS typically feature a steep back wall and a low-angle bottom with displaced soil. Often initiated near water bodies, the slump deposit is rapidly removed, leaving a persistent scar. The process may continue as the active layer develops, causing back-scarp instability. RTS expansion through headwall retreat can persist until substrate or thermal conditions change, or when displaced vegetation insulates the ice-rich scarp from further thawing^{9,10}. Well-developed RTS scars or depletion areas can be easily recognised in the landscape because they commonly appear as a horseshoe-shaped depression with a different surface pattern than the surroundings, often connected at one end to a river, a thermokarst lake, or shoreline¹¹. Individual RTS typically occur over relatively small areas (<40 ha) at the landscape scale, yet can mobilise permafrost substrate and affect infrastructure, such as buildings, roads, railways, and airports^{12,13}, industrial sites¹⁴, ecosystems^{15–17}, and biogeochemical processes across watershed scales^{18–21}; at the same time they can pose a serious threat to the sustainable development and resilience of Arctic communities²² and cultural heritage sites^{23,24}. While rising air temperatures and increasing precipitation at northern high latitudes^{25–27} are expected to increase RTS abundance and activity^{28,29}, understanding of potential ecological and societal impacts across the permafrost region is hindered by challenges in broadly identifying and mapping RTS features. RTS have been studied using observational methods ranging from satellite remote sensing to airborne photogrammetry and field observations. Nevertheless, the spatial distribution and temporal progression of RTS features across large scales are not well characterised^{30,31}.

Among all the types of RTS studies, deep learning (DL) studies are the most data-intensive ones due to the state-of-the-art training paradigm of approximating a black-box function using extensive amounts of data. In the past three years, DL methods, especially convolutional neural networks (CNNs) and vision transformers (ViTs), have been increasingly used to detect RTS features in combination with satellite remote sensing^{31–35}. These recent DL model applications are largely limited to a few regions where RTS features have been sampled and manually delineated. So far no model can detect and map RTS across the Arctic. One reason is that RTS's geomorphic and land cover characteristics vary widely across permafrost regions, and non-RTS background features can closely resemble RTS features³⁴. This variability is a function of interrelated geomorphic, geological and ecological processes such as degree of RTS activity and vegetation growth^{36,37}, terrain characteristics (e.g. slope, ice content), and age since RTS initiation. Therefore, the sampling of training data needs to reflect the full spectrum of RTS appearance across the Arctic, rather than relying on RTS features from a few local regions to represent RTS across the circumpolar domain. This is particularly important because DL methods rely heavily on the amount and representativeness of the training data to achieve high accuracy and generalisation across regions. Unfortunately, RTS training data acquisition is costly in terms of time and resources, making it challenging for individual lab groups to compile a robust training dataset for broader mapping efforts. However, these challenges can largely be addressed with collaboration and thoughtful data sharing across lab groups, enabling the creation of a pan-Arctic RTS training set to build a DL model for pan-Arctic RTS detection.

While developing robust training datasets is essential for improving DL model performance, an often overlooked yet crucial aspect is the proper representation of landscapes without RTS features. For DL models to accurately identify RTS, they must learn from both RTS features (positive data) and verified RTS-absent areas (negative data). This comprehensive approach serves two purposes: enabling systematic geographical comparisons of RTS distribution and providing balanced training data for DL models. A comprehensive training dataset must capture the complete data distribution, encompassing both positive samples that represent the full spectrum of RTS morphological characteristics and negative samples that effectively characterise the diverse non-RTS landscapes across the permafrost region. This balanced representation is particularly critical because RTS features occupy only a small fraction of the Arctic landscape, creating a significant class imbalance challenge. When trained solely on positive data, models incorrectly learn that RTS are common features, leading to high false-positive rates during prediction³⁸. While computational solutions exist, such as adjusting sample weights and loss functions, the fundamental solution lies in systematically sampling and verifying negative data during the training data development phase. Previous studies^{31,34} have demonstrated this approach regionally but pan-Arctic negative data is still needed for a pan-Arctic RTS model. Beyond DL applications, negative data also plays a vital role in RTS susceptibility studies, as demonstrated by Makopoulou *et al.*³⁹ and Luo *et al.*⁴⁰ in their machine learning-based assessments of RTS susceptibility across the northern hemisphere.

To solve the challenges discussed above, we developed a comprehensive data set - the Arctic Retrogressive Thaw Slumps (ARTS) - including 20 data sets (Table 1) for RTS digitisation polygons as 'positive data' or verified RTS-absent regions as 'negative data'. Each source data set has a few hundred to several thousand manually digitised RTS instances. These data sets are manually digitised and verified by permafrost domain experts and collectively cover a wide spectrum of Arctic RTS environments. The regions included the most-studied Arctic RTS hotspots such as Siberia, the Canadian Arctic Archipelago, the Svalbard Archipelago, and the Yukon Territory and Northwest Territories (NWT). While the ARTS data set is built for all ranges of RTS studies, we set the standard of the ARTS to fulfil the highest demand of large-scale DL modelling, including computer vision and time-series forecasting. The ARTS dataset is designed around three core principles: scalability, interoperability, and informativeness. Scalability ensures the database can efficiently update fast-changing RTS time series, incorporate new entries, and manage data growth, allowing it to evolve with ongoing research. Interoperability is achieved through unified metadata and data format standards, implementing an indexing system using a Unique Identifier (UID) generation algorithm, and features that facilitate seamless collaboration and easy access

Source	Verified RTS feature	Verified non-RTS feature	Type	Regions
Nitze <i>et al.</i> ³¹ (Ver.2)	3,579	1,300	polygon	25 locations in Russia, Canada and Alaska
Yang <i>et al.</i> ³⁴	855	3,218	polygon	Yamal-Gydan, Banks Island, Herschel Island, Horton Delta, Kolguev Island, Lena River, Tuktoyaktuk Pen.
Huang <i>et al.</i> ³²	621		polygon	Willow River, Hot Weather Creek, Banks Island
Witharana <i>et al.</i> ³³	356		polygon	Eureka Sound Lowlands
Lantz <i>et al.</i> ⁴⁸	669		polygon	Eastern Banks Island
Huang <i>et al.</i> ⁴⁹	2,494		bounding box	Pan-Arctic
Bernhard <i>et al.</i> ⁵⁰	1,832		point	Peel, Banks, Tuktoyaktuk, Ellesmere, Noatak, Yamal, Gydan, Taymyr Chukotka
van der Sluijs <i>et al.</i> ⁴¹	2,660		polygon	Peel Plateau and Anderson Plain, Tuktoyaktuk Coastlands, Northwest Territories
Bernhard <i>et al.</i> ⁵¹	1,487		polygon	Northern Taymyr Peninsula in Siberia, Russia
Lin <i>et al.</i> ⁵²	365		polygon	Banks Island and Victoria Island, Canada
Ramage <i>et al.</i> ⁵³	286		polygon	Yukon Coast, Canada
Nicu <i>et al.</i> ¹⁰	562		polygon	Nordenskiöld Land, Svalbard, Norway
Elia <i>et al.</i> , in prep.	690		polygon	Andrée and Dickson Land areas, northern Svalbard, Norway
Leibman <i>et al.</i> ⁴³	97		polygon	Yamal-Gydan, Russia
Barth <i>et al.</i> ⁵⁴	3,461		polygon	Novaya Zemlya Archipelago, Kolguev Island, Bol'shoy Lyakhovsky Island, and Taymyr Peninsula
Runge <i>et al.</i> ⁵⁵	1,769		polygon	Chukotka Coast, Iultinsky, Lower Lena, West Taymyr, Chokurdakh
Zwieback <i>et al.</i> ⁵⁶	165		polygon	Tuktoyaktuk, the Lena River delta area
Swanson <i>et al.</i> ⁵⁷	1309		polygon	Northern Alaska
Noerling <i>et al.</i> ⁵⁸	87		polygon	Yamal Peninsula
Makopoulou <i>et al.</i> ³⁹		15,905	polygon	Pan-Arctic
Total RTS digitisations count	23,529	20,434		

Table 1. Summary of RTS data sources.

for contributors and users alike. Informativeness is maintained by mandating key metadata for reproducibility, including both RTS-present and RTS-absent digitisations, and utilizing peer-reviewed data sources. These principles collectively ensure that ARTS remains a dynamic, accessible, and reliable resource for comprehensive Arctic permafrost research, capable of adapting to the evolving needs of the scientific community.

We designed a Data Curation Framework to establish a standardised, scalable, and collaborative approach to managing geospatial RTS data across the Arctic. Its primary intention is to overcome challenges associated with compiling diverse datasets from multiple sources, ensuring consistency, and facilitating ongoing contributions. The framework encompasses a unified metadata structure, a robust unique identifier system, standardised naming conventions and units, protocols for handling intersecting digitisations, and data storage and contribution guidelines. By implementing these standards, the framework aims to enhance data quality, improve interoperability, and streamline the process of integrating new information. This approach not only ensures the long-term usability and reliability of the ARTS dataset but also promotes collaboration among researchers, ultimately advancing our understanding of Arctic permafrost dynamics and their global implications. The Framework is not only a design but also a mature implementation, we provided tested codes and tutorials in Python and R for automated processing and streamlining the metadata formatting process in the ARTS GitHub repository (<https://github.com/whrc/ARTS>).

Methods

Data Curation Framework. The Data Curation Framework is our proposed standard and guideline for RTS data set creation, compilation, storage and future contribution. It also regulates the metadata requirements, formatting and indexing. The processing and compilation of the component raw data sets of the ARTS strictly followed the guidance of the Framework. Individual components of the Framework for RTS data curation are described in the following subsections.

Collating Standalone Data Sets. We compiled 20 peer-reviewed RTS digitisation datasets (Table 1), obtained through published links or directly from corresponding authors. The data collating process involved three main steps guided by the Framework. First, we cleaned the data by verifying and eliminating redundancies, while supplementing missing metadata using published information and direct communication with source laboratories. Next, we standardised the datasets by harmonising metadata field names, formats, and units. We generated UIDs for each polygon, assigning identical UIDs to multiple delineations of the same RTS instance due to temporal changes or different imagery sources. Relationships between intersecting digitisations were documented. Finally, we integrated all processed datasets into a comprehensive GeoJSON file. This rigorous approach ensures a consistent, high-quality compilation of RTS data suitable for large-scale analysis and future contributions. All raw RTS data sets that were collated for the ARTS can be downloaded from the GitHub repository (https://github.com/whrc/ARTS/tree/main/raw_data).

FieldName	Format	Required	Description
CentroidLat	Decimal Degrees	True	Polygon centroid latitude in EPSG:4326, round off to 5 decimal places
CentroidLon	Decimal Degrees	True	Polygon centroid longitude in EPSG:4326, round off to 5 decimal places
RegionName	String	True	Name of the geographical region
CreatorLab	String	True	Data creator and associated organization
BaseMapDate	String	True	Date of base map used for RTS delineation in YYYY-MM-DD for a single date, range of dates should be separated by a comma
BaseMapSource	String	True	Name of the satellite sensor used for RTS digitisation
BaseMapResolution	Number	True	Resolution of the imagery used for RTS digitisation (meters)
TrainClass	String	True	'Positive' for genuine RTS and 'Negative' for background
LabelType	String	True	Type of digitisation, e.g. 'Polygon', 'BoundingBox'
MergedRTS	String	Auto-Generated	UIDs of intersecting RTS that merged into one RTS
SplitRTS	String	Auto-Generated	UID of RTS that split into multiple RTS
NewRTS	String	Auto-Generated	UIDs of intersecting RTS that formed on top of a stabilized RTS scar
StabilizedRTS	String	Auto-Generated	UIDs of intersecting stabilized RTS scars
UnknownRelationship	String	Auto-Generated	UIDs of intersecting RTS with unknown relationship
ContributionDate	String	Auto-Generated	Date of contribution to the ARTS main file in YYYY-MM-DD
UID	36-character alphanumeric string	Auto-Generated	Unique identifier generated using uuid5 by concatenating all 'Required-True' fields as a single string
CustomColumn1 (e.g. Area)	Custom (e.g. Number)	Optional	e.g. Area of the RTS polygon in km ²
CustomColumn2 (e.g. State)	Custom (e.g. String)	Optional	e.g. Active state of the RTS polygon

Table 2. Metadata Format Summary.

Metadata Formatting. Merging the metadata of the raw data sets presented challenges including inconsistent units, repeated or redundant IDs, and inconsistent or ambiguous attribute names. Here we present a standard metadata system that regulates critical information for each RTS digitisation (Table 2). This will facilitate creating, merging, and utilising future data sets. The first nine columns (Required-True) are mandatory attributes used for UID generation that provide essential metadata with a fixed entry format and order. The required metadata will be used as inputs for the data compilation code to produce auto-generated metadata. Creators can customise extra columns to add task-specific information after the auto-generated columns, for example, whether field verification has been conducted. The default entry for not-applicable optional metadata is *null*.

In Table 2, the first column 'FieldName' lists the unified metadata names in the ARTS. The following columns 'Format', 'Required' and 'Description' are detailed standards and text descriptions that define the metadata. 'CentroidLat' and 'CentroidLon' are the latitude and longitude of the RTS centroid in decimal degrees, rounding off to 5 decimal places provides < 1m accuracy which is adequate for RTS features. This is useful for spatial distribution analysis or filtering based on coordinates. 'RegionName' is useful when filtering based on pre-defined regions or comparing regions. 'CreatorLab' records the data creator and affiliated organisation. This is useful when filtering based on creator or institution, and to credit the creator. 'BaseMapDate', 'BaseMapSource' and 'BaseMapResolution' record critical information about the base map used to create the polygons. This is useful for adapting polygons to another base map source. 'TrainClass' indicates if the data entry is a genuine RTS polygon or a non-RTS background tile, which is essential for providing both positive and negative examples to train deep learning models. 'LabelType' indicates the type of RTS digitisation, commonly provided as polygons (vectors) but could also be provided as bounding boxes or points. 'MergedRTS' and 'StabilisedRTS' are columns used to link the replacement relationship of new and old RTS.

Finally, all 20 data sets originally adopted sequential numbers to index the polygons. Ordinal number indices are ideal for small and simple data sets. However, for a large data set that expects frequent editing, updating and merging operations, numeric indices bear problems related to repetition, insertion and deletion at a massive scale. Therefore, a more systematic and robust indexing system is needed. This system should be 1) immune to adding, deleting, shuffling, or mixing, 2) reproducible, 3) unique to each RTS instance and consistent through time, and 4) easy to generate and maintain. We therefore utilised, and propose for future use, the Unique Identifier (UID) system which generates a unique 36-character alphanumeric ID string from a given input. The RTS UID is generated from a seeded random generator process, and the likelihood of repeated UID is extremely small and can be safely ignored. This guarantees unique IDs can be generated by different organisations without communication. To achieve reproducibility, all 'Required=True' metadata entries should be concatenated to a single string to be used as a generation seed. To ensure that the same RTS instance has the same UID across polygons that were delineated at different points in time, we require generating a UID for the earliest polygon and applying this UID to all subsequent polygons. Recognising the possibility that new contributions to the dataset could include polygons that correspond to features that already exist in the dataset, we will require that all new contributions be checked for intersections with the current version of the dataset, and the UID of any features which already exist in the dataset will be used.

RTS features exhibit complex spatial and temporal relationships, as new RTS can develop within stabilised scars and adjacent RTS may merge over time^{41–43}. These dynamics necessitate a robust system for tracking RTS relationships in the geospatial database⁴⁴. The UIDs for intersected RTS instances will be categorised by their relation. For example, in cases where multiple RTS features merge into one, the merged feature should receive a

FieldName	Value
CentroidLat	−77.508333
CentroidLon	164.754167
RegionName	'Yamal-Gydan, NW Russia'
CreatorLab	'JDean, Woodwell'
BaseMapDate	'2020-12-31,2022-11-29'
BaseMapSource	'WorldView-2'
BaseMapResolution	4
TrainClass	'Positive'
LabelType	'Polygon'
MergedRTS	'a6cf10e7-6515-5ab7-b7bb-66606b16ef10', 'dd137c86-6db6-57da-bb46-961c89845a97'
StabilizedRTS	null
ContributionDate	'2023-12-31'
UID	'acde070d-8c4c-4f0d-9d8a-162843c10333'
Area	10
State	'Active'

Table 3. Example of valid FieldName and Values for mandatory and optional columns.

new UID, and the UIDs of the component features should be included in the 'MergedRTS' column. Additionally, a newly initiated RTS feature on an old, stabilised scar should receive a new UID, and the UID of the old RTS should be included in the 'StabilisedRTS' column. To make this process easy for contributors, we have created publicly available automation scripts in Python and R to ensure that the UIDs meet all requirements and that the metadata formatting is correct. In total, there are five different categories of RTS relations in the ARTS, see Table 2. Using RTS IDs and required metadata columns, it will be possible to easily differentiate 1) the digitisation of the same RTS by multiple groups using different base maps, 2) the digitisation of the same RTS instance at different times, 3) the digitisation of a new RTS which initiated on an older, stabilised RTS, and 4) the digitisation of a coalesced RTS which formed when two adjacent RTS merged.

Intersecting digitisations and UID generation rule. Intersecting digitisations are due to mainly three reasons: 1) multiple groups digitising the same RTS with different definitions or standards, 2) a series of digitisations of RTS development through time, i.e. a time-series, and 3) digitisation on images acquired on different dates, mosaics or different sensors.

All new entries need to be checked for repeat delineations of the same RTS feature. For this purpose, we have published scripts on GitHub (<https://github.com/whrc/ARTS>) in Python and R that check all new RTS polygons for intersections with polygons in the published RTS data set. In brief, the scripts allow contributors to easily create a GeoJSON that enumerates all overlapping polygons between the new and published data sets. This file can then be visualised over imagery so that the contributor can determine whether polygons with intersections in the published data set are new RTS features or repeat delineations of previously observed RTS features. Using this information, UIDs can then be assigned to new RTS features, while repeat RTS features inherit the UID of the previously delineated polygon. In cases where a new RTS feature has initiated on top of an old, stabilised RTS scar or multiple RTS features have merged, the polygon is considered a new RTS feature for UID generation purposes.

Naming Custom Columns. The full names of custom columns should be named using Pascalcase: complete words delimited by capital letters. This is called the FieldName, which lists the column names presented in the data set (e.g. Table 3). Before adding a new custom column, contributors need to check with the late.

Units. In the Metadata Format Summary (Table 2) Description column, we specified the type of default International System of Units (SI units) that should be used for scalar metadata, therefore unit should not be repeatedly specified in the metadata entries. For a new custom column without an existing description in the Metadata Format Summary, a suitable SI unit should be specified and consistently used for future entries. For instance, CustomColumn for RTS areas is defined in Table 2 row 14 and specified its unit in km², therefore the unit is omitted in the actual metadata entry (Table 3 row 14).

RTS-present and RTS-absent data. The ARTS data set collected 23,529 manually digitised RTS-present polygons (Fig. 1), as well as 20,434 RTS-absent digitisations (Fig. 2), sampled across the whole Arctic region. RTS-present (positive) and RTS-absent (negative) data are verified and indexed with UIDs. Positive data should ideally be added as polygons, but other digitisation types such as bounding boxes or points are also accepted. Verified negative data should be added as bounding boxes or polygons excluding any genuine RTS.

Data Storage Format. Most of the original data sets were provided as Esri shapefiles - a vector data storage format that stores the location, shape and custom attributes of geographic features, only one data set is in the format of Geopackage and two in .tab files. Although shapefile has been the most widely used format for geographic vector data, it suffers from several limitations, such as being slow, restricted lengths for column names, and being

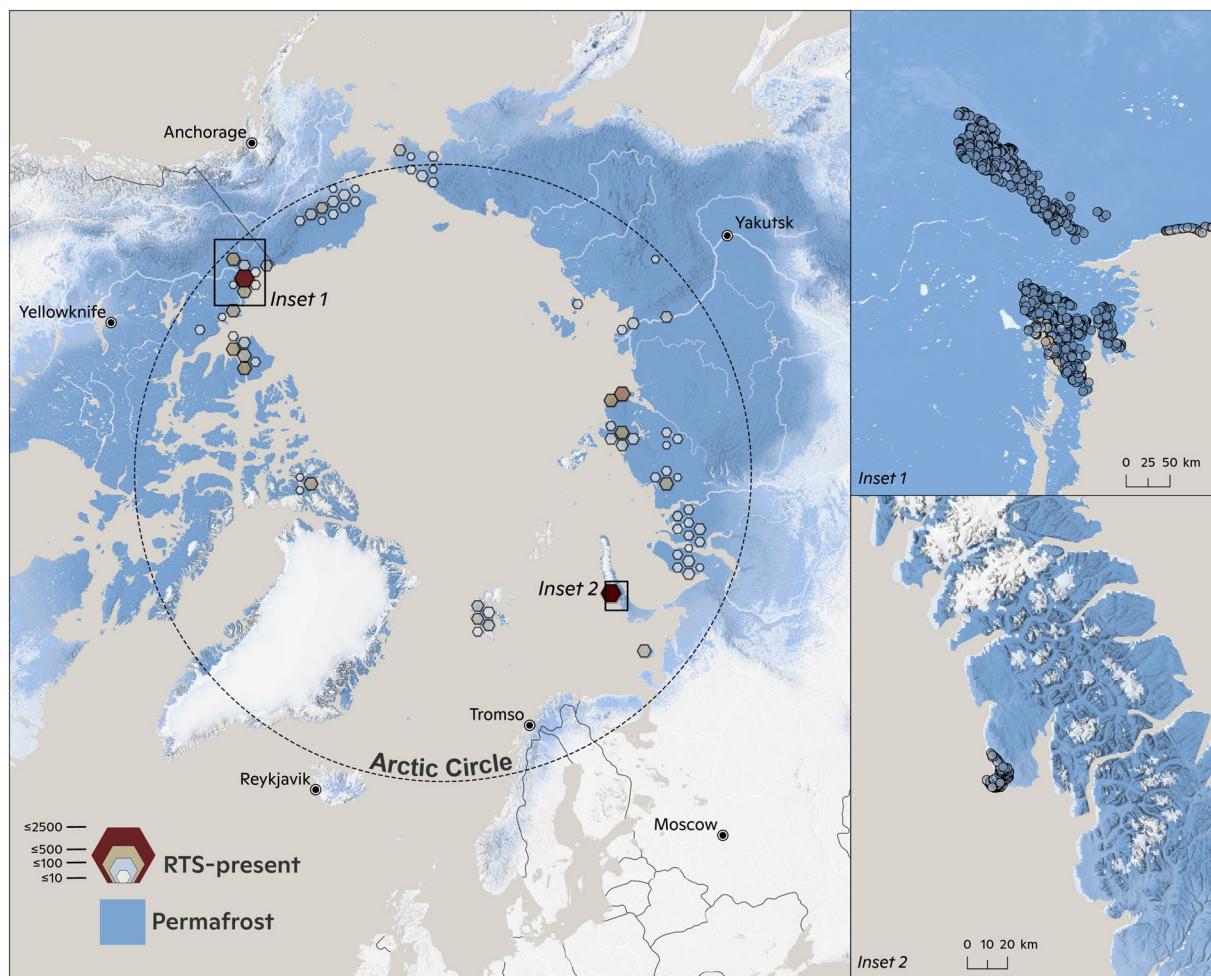


Fig. 1 Visualisation of the distribution of RTS-present digitisations in the ARTS data set. Inset 1,2 shows the highly heterogeneous RTS density distribution where RTS occurrence tend to be highly clustered in hot-spot regions.

composed of multiple files, which can result in version control and file sharing problems. We chose GeoJSON as the data storage format because it is an open standard vector geospatial data format that is widely used and easily integrated with many desktop geospatial packages and web platforms such as QGIS, ArcGIS, and Google Earth Engine. Additionally, raw GeoJSON is directly readable and editable by humans, making it easy to inspect and maintain. Our team also considered using the GeoPackage format but ultimately rejected it mainly because of its relatively lower adoption rate and higher complexity.

Data Records

The ARTS data set v.1.0.0 is available at the Arctic Data Center⁴⁵ (arcticdata.io/catalog/view/doi:10.18739/A2PK0738B).

Technical Validation

All raw data sets are associated with at least one peer-reviewed publication except one in preparation (Elia *et al.*, in prep.). Beyond the data quality assurance by publications, we also actively verified the data quality by partially inspecting each raw data set. Randomly sampled RTS polygons in the data sets were manually examined to our best ability to verify the genuineness based on a 4 m Maxar base map and a 0.5 m Esri base map. Repeat polygons drawn on the same RTS instance were preserved for use in time-series analyses. We are aware that due to the fast-developing nature of RTS and the different image acquisition times of base maps one can access, there are unavoidable discrepancies between polygons and base maps when used in different studies (Fig. 3). We therefore did not aim for perfect physical alignment between polygons and the base map, but rather focused on verifying that the RTS feature did exist at some point in time. As a result, we note that the data set is designed to require users' pre-processing through careful filtering and adaption to ensure the best possible alignment between the data set and specific imagery sources.

Data bias. We acknowledge that the ARTS database incorporates data from various research groups, inevitably introducing differences in visual interpretations⁴⁶. This diversity is crucial for supporting future automated mapping procedures aimed at consistent and multi-temporal RTS mapping. However, users should be aware that



Fig. 2 Visualisation of the distribution of RTS-absent digitisations in the ARTS data set. The RTS absent data is relatively homogeneously sampled across the Arctic permafrost regions compared with the RTS-present data.

potential biases from individual inventories could influence automated mapping models. For example, if a particular group mapped RTSs at specific spatial scales, a complex deep learning algorithm might become sensitive to this preference, potentially affecting results. Similar considerations apply to RTS shape preferences and other morphometric characteristics. We believe the breadth of our global database should help mitigate individual biases with minimal bias-removal techniques required. Nevertheless, we encourage future users to consider bias removal strategies in their models or at least assess their necessity when utilising the ARTS database.

Data quality check. RTS polygons in the data sets were examined to the best ability to verify the genuineness based on a 4 m Maxar base map and a 0.5 m Esri base map. Repeat polygons drawn on the same RTS instance were preserved for use in time-series analyses. There are some repetitions of RTS delineations due to different groups working in the same area, e.g. Banks Island, in addition to repetitions due to RTS development through time and repeat delineations on images acquired on different dates. We are aware that due to the fast-developing nature of RTS and the different image acquisition times of base maps one can access, there are unavoidable discrepancies between polygons and base maps when used in different studies (Fig. 3). We therefore did not aim for perfect physical alignment between polygons and the base map, but rather focused on verifying that the RTS feature did exist at some point in time. As a result, we note that the data set is designed to require users' pre-processing through careful filtering and adaption to ensure the best possible alignment between the data set and specific imagery sources.

Due to the cooperative nature of the ARTS data set, the consistency of the RTS digitisation criteria across (and even within) contributors is likely heterogeneous. Therefore, for studies that are highly sensitive to digitisation quality, we recommend a thorough data quality check, filtering, or adaptation before data processing. Second, due to the fast-developing nature of RTS, we recommend a walkthrough of RTS digitisations in use with a base map underlay to check for recent RTS developments that affect the accuracy of the digitisation. Likely, negative data may also be affected by future RTS development and therefore requires verification before use with more recent imagery.

Generating training-validation-testing patches for DL models. We observed from the data set that the occurrence of RTS features can be discrete, but more frequently they tend to be clustered or adjacent.



Fig. 3 Selected examples of digitised RTS features from the source data sets. The underlying base map is Maxar 0.5 m high-resolution satellite imagery. This shows a range of RTS types, appearances and development stages from various locations across the Arctic.

This nature of RTS leads to a caveat that often more than one RTS can appear in a single training image patch. Therefore, when using the data set in DL model training, a random training-validation-testing split is often problematic. Two or more adjacent RTS can be randomly assigned to the training and testing set while their distance is so close that the testing RTS has already appeared in the training RTS patch, causing data leakage and overestimated model accuracy. Some possible solutions to this issue could be 1) manually selecting the validation and testing set and making sure no training RTS appears in the same scene; 2) using algorithms to ensure that RTS within a certain distance of other RTS are included in the same training/testing set; or 3) using region-based cross-validation. We have developed an automatic data set splitting algorithm which will guarantee completely

separated training-validation-testing subsets generated for a given ratio of split. The algorithm is implemented in our ARTS GitHub repository with a tutorial and use instructions. We recommend using this tool to do data-splitting for DL model training.

Contribution Guideline. The database is open to high-quality contributions regardless of size. There are four or five steps to contribution depending on whether or not new metadata columns are being added:

1. Generate UIDs for all RTS digitisation entries (Automated, see section Metadata Formatting)
2. Check polygon intersection (Automated)
3. Update UIDs using the intersection information (Automated)
4. Request data contribution via GitHub (recommended) or share directly with the correspondence author.

For a data set that has new columns that do not exist in the latest version of the Metadata Format Summary (<https://github.com/whrc/ARTS/blob/main/Metadata%20Format%20Summary.csv>):

5. Add a new metadata entry type to the Metadata Format Summary, including FieldName, Format and Description.

Data Versioning. For a cooperative and evolving data set, it is important to track changes over time and avoid overwriting accidents. Therefore we defined a data versioning scheme to help users understand the data set evolution and allow rolling back to previous versions. We adopted a three-part semantic version number convention consisting of three numbers connected by dots (e.g. 3.0.2). Where the first number indicates the incorporation of a new set of RTS entries from a new data source. The second number indicates batch changes or additions to the metadata (new rows or new columns) without introducing a new data source. The last number indicates minor changes or fixes to the existing data or metadata, such as editing existing metadata or adjusting polygons' vertices to the existing data. Each version update will require a concise update note describing the changes to the data set.

Code availability

Code described in this manuscript can be accessed at the GitHub repository⁴⁷ (<https://github.com/whrc/ARTS>).

Received: 18 July 2024; Accepted: 1 January 2025;

Published online: 06 January 2025

References

1. Hugelius, G., Kuhry, P. & Tarnocai, C. Ideas and perspectives: Holocene thermokarst sediments of the yedoma permafrost region do not increase the northern peatland carbon pool. *Biogeosciences* **13**, 2003–2010 (2016).
2. Schuur, E. A. *et al.* Permafrost and climate change: Carbon cycle feedbacks from the warming arctic. *Annual Review of Environment and Resources* **47**, 343–371 (2022).
3. Jorgenson, M. T. *et al.* Drivers of historical and projected changes in diverse boreal ecosystems: fires, thermokarst, riverine dynamics, and humans. *Environmental Research Letters* **17**, 045016 (2022).
4. Li, G. *et al.* Changes in permafrost extent and active layer thickness in the northern hemisphere from 1969 to 2018. *Science of The Total Environment* **804**, 150182 (2022).
5. Chadburn, S. *et al.* An observation-based constraint on permafrost loss as a function of global warming. *Nature climate change* **7**, 340–344 (2017).
6. Turetsky, M. R. *et al.* Carbon release through abrupt permafrost thaw. *Nature Geoscience* **13**, 138–143 (2020).
7. Treharne, R. *et al.* Comprehensive assessment of permafrost carbon emissions indicates need for urgent action to keep paris agreement temperature goals within reach. *Nature Climate Change* (in review).
8. Bernhard, P., Zwieback, S., Leinss, S. & Hajnsek, I. Mapping retrogressive thaw slumps using single-pass tandem-x observations. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **13**, 3263–3280 (2020).
9. Wang, B., Paudel, B. & Li, H. Behaviour of retrogressive thaw slumps in northern canada—three-year monitoring results from 18 sites. *Landslides* **13**, 1–8 (2016).
10. Nicu, I. C., Lombardo, L. & Rubensdotter, L. Preliminary assessment of thaw slump hazard to arctic cultural heritage in nordenskiöld land, svalbard. *Landslides* **18**, 2935–2947 (2021).
11. Lacelle, D., Bjornson, J. & Lauriol, B. Climatic and geomorphic factors affecting contemporary (1950–2004) activity of retrogressive thaw slumps on the aklavik plateau, richardson mountains, nwt, canada. *Permafrost and Periglacial Processes* **21**, 1–15 (2010).
12. Van der Sluijs, J., Kokelj, S. V., Fraser, R. H., Tunnicliffe, J. & Lacelle, D. Permafrost terrain dynamics and infrastructure impacts revealed by uav photogrammetry and thermal imaging. *Remote Sensing* **10**, 1734 (2018).
13. Hjort, J. *et al.* Impacts of permafrost degradation on infrastructure. *Nature Reviews Earth and Environment* **3**, 24–38 (2022).
14. Langer, M. *et al.* Thawing permafrost poses environmental threat to thousands of sites with legacy industrial contamination. *Nature Communications* **14**, 1721 (2023).
15. Levenstein, B., Culp, J. M. & Lento, J. Sediment inputs from retrogressive thaw slumps drive algal biomass accumulation but not decomposition in arctic streams, nwt. *Freshwater Biology* **63**, 1300–1315 (2018).
16. Levenstein, B., Lento, J. & Culp, J. Effects of prolonged sedimentation from permafrost degradation on macroinvertebrate drift in arctic streams. *Limnology and Oceanography* **66**, S157–S168 (2021).
17. Abbott, B. W. & Jones, J. B. Permafrost collapse alters soil carbon stocks, respiration, ch 4, and n2o in upland tundra. *Global Change Biology* **21**, 4570–4587 (2015).
18. Kokelj, S. *et al.* Thermokarst transformation of permafrost preserved glaciated landscapes. In *AGU Fall Meeting Abstracts*, vol. 2017, C23F–01 (2017).
19. Littlefair, C. A., Tank, S. E. & Kokelj, S. V. Retrogressive thaw slumps temper dissolved organic carbon delivery to streams of the peel plateau, nwt, canada. *Biogeosciences* **14**, 5487–5505 (2017).
20. Cassidy, A. E., Christen, A. & Henry, G. H. Impacts of active retrogressive thaw slumps on vegetation, soil, and net ecosystem exchange of carbon dioxide in the canadian high arctic. *Arctic Science* **3**, 179–202 (2017).

21. St. Pierre, K. A. *et al.* Unprecedented increases in total and methyl mercury concentrations downstream of retrogressive thaw slumps in the western canadian arctic. *Environmental Science & Technology* **52**, 14099–14109 (2018).
22. Sisneros-Kidd, A. M., Monz, C., Hausner, V., Schmidt, J. & Clark, D. Nature-based tourism, resource dependence, and resilience of arctic communities: Framing complex issues in a changing environment. *Journal of Sustainable Tourism* **27**, 1259–1276 (2019).
23. Hollesen, J. *et al.* Climate change and the deteriorating archaeological and environmental archives of the arctic. *Antiquity* **92**, 573–586 (2018).
24. Nicu, I. C. & Fatorić, S. Climate change impacts on immovable cultural heritage in polar regions: A systematic bibliometric review. *Wiley Interdisciplinary Reviews: Climate Change* e822 (2023).
25. Lique, C., Holland, M. M., Dibike, Y. B., Lawrence, D. M. & Screen, J. A. Modeling the arctic freshwater system and its integration in the global system: Lessons learned and future challenges. *Journal of Geophysical Research: Biogeosciences* **121**, 540–566 (2016).
26. Rantanen, M. *et al.* The arctic has warmed nearly four times faster than the globe since 1979. *Communications Earth & Environment* **3**, 168 (2022).
27. Liu, Z. *et al.* Widespread deepening of the active layer in northern permafrost regions from 2003 to 2020. *Environmental Research Letters* **19**, 014020 (2023).
28. Kokelj, S. *et al.* Increased precipitation drives mega slump development and destabilization of ice-rich permafrost terrain, northwestern canada. *Global and Planetary Change* **129**, 56–68 (2015).
29. Nicu, I. C., Elia, L., Rubensdotter, L., Tanyaş, H. & Lombardo, L. Multi-hazard susceptibility mapping of cryospheric hazards in a high-Arctic environment: Svalbard Archipelago. *Earth System Science Data* **15**, 447–464 (2023).
30. Lewkowicz, A. G. & Way, R. G. Extremes of summer climate trigger thousands of thermokarst landslides in a high arctic environment. *Nature communications* **10**, 1329 (2019).
31. Nitze, I., Heidler, K., Barth, S. & Grosse, G. Developing and testing a deep learning approach for mapping retrogressive thaw slumps. *Remote Sensing* **13**, 4294 (2021).
32. Huang, L. *et al.* Accuracy, efficiency, and transferability of a deep learning model for mapping retrogressive thaw slumps across the canadian arctic. *Remote Sensing* **14**, 2747 (2022).
33. Witharana, C. *et al.* Automated detection of retrogressive thaw slumps in the high arctic using high-resolution satellite imagery. *Remote Sensing* **14**, 4132 (2022).
34. Yang, Y. *et al.* Mapping retrogressive thaw slumps using deep neural networks. *Remote Sensing of Environment* **288**, 113495 (2023).
35. Huang, L., Luo, J., Lin, Z., Niu, F. & Liu, L. Using deep learning to map retrogressive thaw slumps in the beiluhe region (tibetan plateau) from cubesat images. *Remote Sensing of Environment* **237**, 111534 (2020).
36. Brooker, A., Fraser, R. H., Olthof, I., Kokelj, S. V. & Lacelle, D. Mapping the activity and evolution of retrogressive thaw slumps by tasselled cap trend analysis of a landsat satellite image stack. *Permafrost and Periglacial Processes* **25**, 243–256 (2014).
37. Cray, H. A. & Pollard, W. H. Vegetation recovery patterns following permafrost disturbance in a low arctic setting: case study of herschel island, yukon, canada. *Arctic, antarctic, and alpine research* **47**, 99–113 (2015).
38. Xia, X., Lu, Q. & Gu, X. Exploring an easy way for imbalanced data sets in semantic image segmentation. In *Journal of Physics: Conference Series*, 022003 (IOP Publishing, 2019).
39. Makopoulou, E. *et al.* Retrogressive thaw slump susceptibility in the northern hemisphere permafrost region. *Earth Surface Processes and Landforms* (2024).
40. Luo, J. *et al.* Machine learning-based predictions of current and future susceptibility to retrogressive thaw slumps across the northern hemisphere. *Advances in Climate Change Research* **15**, 253–264 (2024).
41. van der Sluijs, J., Kokelj, S. V. & Tunnicliffe, J. F. Allometric scaling of retrogressive thaw slumps. *The Cryosphere* **17**, 4511–4533 (2023).
42. Lantuit, H. & Pollard, W. Fifty years of coastal erosion and retrogressive thaw slump activity on herschel island, southern beaufort sea, yukon territory, canada. *Geomorphology* **95**, 84–102 (2008).
43. Leibman, M., Nesterova, N. & Altukhov, M. Distribution and morphometry of thermocirques in the north of west siberia, russia. *Geosciences* **13**, 167 (2023).
44. Nesterova, N. *et al.* Retrogressive thaw slump characteristics and terminology. *The Cryosphere* **18**, 4787–4810 (2024).
45. Yang, Y., Rodenhizer, H. & Dean, J. Arctic retrogressive thaw slumps (ARTS): digitisations of pan-arctic retrogressive thaw slumps, 1985–2021, <https://doi.org/10.18739/A2PK0738B> (2024).
46. Nitze, I. *et al.* A labeling intercomparison of retrogressive thaw slumps by a diverse group of domain experts. *Permafrost and Periglacial Processes* (2024).
47. Yang, Y. *et al.* Arts: a scalable data set for arctic retrogressive thaw slumps. Tech. Rep., Copernicus Meetings (2024).
48. Segal, R., Lantz, T. & Kokelj, S. Inventory of active retrogressive thaw slumps on eastern banks island, northwest territories (2015).
49. Huang, L. *et al.* Identifying active retrogressive thaw slumps from arcticdem. *ISPRS Journal of Photogrammetry and Remote Sensing* **205**, 301–316 (2023).
50. Bernhard, P., Zwieback, S., Bergner, N. & Hajnsek, I. Assessing volumetric change distributions and scaling relations of retrogressive thaw slumps across the arctic. *The Cryosphere* **16**, 1–15 (2022).
51. Bernhard, P., Zwieback, S. & Hajnsek, I. Accelerated mobilization of organic carbon from retrogressive thaw slumps on the northern taymyr peninsula. *The Cryosphere* **16**, 2819–2835 (2022).
52. Lin, Y. & Knudby, A. J. A transfer learning approach for automatic mapping of retrogressive thaw slumps (rtss) in the western canadian arctic. *International Journal of Remote Sensing* **44**, 2039–2063 (2023).
53. Ramage, J. L. *et al.* Terrain controls on the occurrence of coastal retrogressive thaw slumps along the yukon coast, canada. *Journal of Geophysical Research: Earth Surface* **122**, 1619–1634 (2017).
54. Barth, S., Nitze, I., Juhls, B., Runge, A. & Grosse, G. Vector dataset of manually mapped retrogressive thaw slumps from very high-resolution multispectral imagery in the Russian High Arctic between 2011 and 2020 (2023).
55. Runge, A., Nitze, I. & Grosse, G. Remote sensing annual dynamics of rapid permafrost thaw disturbances with landtrendr. *Remote Sensing of Environment* **268**, 112752 (2022).
56. Zwieback, S. *et al.* Sub-seasonal thaw slump mass wasting is not consistently energy limited at the landscape scale. *The Cryosphere* **12**, 549–564 (2018).
57. Swanson, D. K. Permafrost thaw-related slope failures in alaska's arctic national parks, c. 1980–2019. *Permafrost and periglacial processes* **32**, 392–406 (2021).
58. Noerling, C. *et al.* Permafrost disturbance in Central Yamal along the Bovanenkovo railway line and thermokarst lakes, link to files in different formats (2017). Supplement to: Noerling, Caroline (2017): Short-term changes of permafrost degradation triggered by anthropogenic impacts and climatic events in Western Siberia 2010–2013. Master Thesis, University of Potsdam, 74 pp, hdl:10013/epic.51479.d001.

Acknowledgements

We acknowledge members of the RTSIntrain working group for early-stage discussions for this work, in particular Chunli Dai and Anna Liljedahl. This work was supported by the Heising-Simons Foundation (grant #2021-3040) and funding catalysed by the Audacious Project (Permafrost Pathways) for Woodwell-affiliated coauthors. Ionut Cristi Nicu and Lena Rubensdotter were partly supported by the Fram Centre project PermaRICH (Advanced

Mapping and Monitoring for Assessing Permafrost Thawing Risks for Modern Infrastructure and Cultural Heritage in Svalbard). Ionut Cristi Nicu, Lena Rubensdotter, and Luigi Lombardo were also supported by the 'SIOS-Planet cooperation project proposal call to demonstrate the usability of high-resolution Planet data in the Arctic' project MACROS (Mapping cryospheric hazards towards a spatio-temporal multi-hazard susceptibility modelling). Guido Grosse, Ingmar Nitze, and Sophia Barth were supported by the HGF AI-CORE, BMWK ML4EARTH, and NSF Permafrost Discovery Gateway (awards #1927872, #2052107). Alexandra Runge was supported by the ESA CCI Postdoctoral Fellowships contract No. 4000134121/21/I-NB. Nina Nesterova was supported by a DAAD scholarship (#57588368). Eirini Makopoulou was supported by the KVANTUM Institute within the program Changing Climate and the Northern Environment 'HYPERISK, Hybrid Modelling for Improved Permafrost Risk Assessments (2021-2024)'.

Author contributions

Y.Y.: Conceptualisation, Data curation, Methodology, Investigation, Project administration, Software, Supervision, Writing-original draft. H.R.: Conceptualisation, Data curation, Methodology, Software, Writing-review & editing. B.M.R.: Conceptualisation, Project administration, Funding acquisition, Supervision, Writing-review & editing. S.P.: Conceptualisation, Writing-review & editing. J.D.: Data curation. J.W.: Conceptualisation. S.Z.: Writing-review & editing. R.S.: Data curation. G.F.: Conceptualisation, Funding acquisition, Visualisation. A.P.: Data curation. T.W.: Data Curation. L.H.: Resources, Writing-review & editing. C.W.: Resources. I.N.: Resources, Writing-review & editing. N.N.: Resources. S.B.: Resources. G.G.: Resources. T.L.: Resources. A.R.: Resources. L.L.: Resources, Writing-review & editing. I.C.N.: Resources, Writing-review & editing. L.R.: Resources. E.M.: Resources. S.N.: Conceptualization, Supervision, Funding acquisition.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Y.Y., H.R. or B.M.R.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2025