Tracking Rapid Permafrost thaw Through Time: Exploring the Potential of Convolutional Neural Network based Models

TRACKING RAPID PERMAFROST THAW THROUGH TIME: EXPLORING THE POTENTIAL OF CONVOLUTIONAL NEURAL NETWORK BASED MODELS

Felix Rustemeyer[1,3], Julia Barrott[1], Matthew Fielding[1], Adam Wickenden, Gustaf Hugelius[2], Alexia Briassouli[3]

Stockholm Environment Institute[1], Stockholm University[2], Maastricht University[3]

1. ABSTRACT

This paper presents the novel use of convolutional neural network (CNN)-based machine learning models for remotely detecting and monitoring retrogressive thaw slumps (RTS) in high latitude northern permafrost using open-source Sentinel-2 satellite data. RTS are indicative of rapid permafrost thaw (RPT), the accelerated release of greenhouse gases (GHG) and potentially runaway changes in the cryosphere. Attempts to quantify GHG emissions from RTS are inhibited by a lack of information on RTS incidence and area affected. We show that site-specific CNN models can be used to produce time series data on rapid RTS development that allow for the approximation of associated GHG emissions. For the sites assessed we achieve good model precision, recall and F1 values of $>$ 0.8. The short time series studied so far do not reveal clear trends in RTS development. These limitations arise from the low resolution of Sentinel-2 data (10 m) and limited availability and diversity of validated training data. The capability shown here is the first step towards achieving automated monitoring of rapid environmental change in permafrost using satellite data. This work highlights the need for ready access to open-source high resolution satellite data and permafrost field data if the potential of such approaches is to be fully realized.

2. INTRODUCTION

Soils and deep sediments in northern high latitude permafrost, an area of 13.9 million km^2, hold approximately 1460 – 1,600 billion tonnes of carbon – roughly twice as much carbon as in the atmosphere and approximately 42 per cent of all soil carbon globally [1]. Climate change is driving rapid warming in these regions, which in turn is driving permafrost thaw and loss [2]. As permafrost thaws, microbial decomposition of stored organic carbon increases, resulting in the release of influential and long-lived greenhouse gases (GHG) (Strauss et al., 2021). This is a positive climate feedback loop: as temperatures rise and more permafrost thaws, more GHG are released, driving further climate change (IPCC, 2021). Knowing how much and when these GHGs are likely to be released is critical to determining the extent of climate change mitigation required under the UNFCCC.

Estimates and projections of GHG contributions from permafrost thaw are not well constrained (IPCC, 2021). This is especially the case for RPT, which could release 60–100 billion tonnes of carbon by 2030, adding to the 200 billion tonnes expected to be released via more gradual thaw [3]. Currently, GHG emission and climate models used for climate policy and planning only account for gradual permafrost thaw over decades and starting from the surface downwards (Turetsky et al., 2020). To better understand RPT at scale and incorporate associated GHG emissions into these models, we need to be able to track and quantify RPT through time [4].

RPT results in visibly distinct landforms that can be identified using machine learning [4, 5]. These include RTS, which develop in areas of loosely consolidated ice-rich sediments such as glacial moraines and have been increasing in frequency over recent decades in response to environmental changes [6]. Understanding the incidence and evolution of RTS is critical as they shift significant volumes of sediments over short time frames, exposing substantial quantities of previously buried carbon and disturbing the surrounding permafrost [4].

This project uses computer vision to remotely identify and monitor RTS to understand the drivers behind their development and quantify their GHG emissions. We combine convolutional neural network (CNN)-based models with open-source Sentinel-2 satellite data to detect and characterize RTS development in six sites over the past six years. This paper focuses on the results from two of these sites. We show that there is significant potential from this approach for building time series data (including RTS incidence and area) that can be used to quantify GHG emissions over time. However, we find that accurate year-to-year monitoring of RTS development requires a high level of performance of machine learning models that cannot be achieved with the comparably low resolution of open-source imagery and readily available observed data on RTS. This work underscores the potential benefits of centralized curation of available permafrost field data and increasing the accessibility of high resolution satellite data.

3. METHODOLOGY

In this section, we present the entire workflow for our approach, from data collection to model selection and train-

Site	Coordinates	# Images
Herschel Island	139.00°W , 69.60°N	4
Horton 01	126.75°W , 69.75°N	3
Horton 02	126.60°W , 69.64°N	4
Kolguev 01	48.35°E , 69.22°N	5
Kolguev 02	48.51°E , 69.35°N	7
Lena	124.40°E , 69.12°N	18

Table 1. Sites with centre point coordinates and number of multi-dimensional images used in this study.

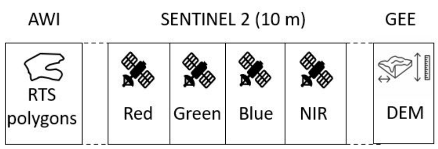

Fig. 1. The data sources of the multi-dimensional images.

ing. Experiments to test the proposed approaches have been conducted in a Python environment in Microsoft's Planetary Computer, consisting of a CPU with 4 cores and 32 GB of memory. Planetary Computer is a platform providing tools for research in environmental sustainability that is currently in preview stage and free use for interested parties. The data and repository of this study are stored in Microsoft Azure, using Blob storage and DevOps. The neural networks are implemented and trained using Keras.

3.1. Data collection

The raw ground truth data, provided by Ingmar Nitze (Alfred Wegener Institute, Bremerhaven, Germany) and used in Nitze et al. (2021)[4], consists of 1203 polygons representing RTS across six arctic sites (see Table 1), each with a size of 100 km^2. These polygons are grouped together based on site and collection date, and are all gathered in 2018 or 2019. Thereafter, the polygon-groups are processed into multi-dimensional images using a bespoke data collection pipeline that obtains spectral bands and elevation data for the study sites. Planetary Computers integrated data access API is employed for the gathering of four spectral bands (blue, red, green and near infrared) with a resolution of 10 m. For every polygon group, we select a Sentinel-2 acquisition with a date as close as possible to the collection date of the polygons and cloud cover below 10%. A slope layer is added to the multi-dimensional images by applying a boxcar filter on the Digital Elevation Model (DEM) obtained through Google Earth Engine (GEE). The final layer of the images is the binary ground truth mask of the site indicating the locations of the positives, the RTS pixels. The result of the data collection pipeline is a collection of 41 six-dimensional images containing spectral, slope and ground truth information from the six sites across different dates.

3.2. Data pre-processing

A series of pre-processing steps have taken place, in order to obtain objective and representative results. Before use, the images were appraised and those with missing spectral bands and erroneous data structures removed. Thereafter, the images were split into 64×64-sized training arrays, and then split into arrays containing only negatives, and arrays containing at least one positive pixel. Only 7.9% of the arrays contained positive pixels. As the aim of the model is to segment positive pixels, training the model using this imbalanced dataset resulted in poor performance. This data imbalance was resolved by undersampling to obtain 1416 arrays with a 1:1 ratio of arrays containing only negatives and arrays containing at least one positive pixel. To increase the number of training instances, and consequently the model's performance and generalizability, data augmentation was applied to the training data. After experiments with several data augmentation techniques — including rotation, Gaussian blur and flipping — it was found that Gaussian blur led to the highest performance. Finally, the arrays are normalized by scaling all the values between zero and one.

3.3. Network architecture

This study uses a U-Net, as it is the State-of-the-Art in biomedical semantic segmentation, and can also be employed for other segmentation tasks, like predicting of RTS [4]. The U-Net consists of a contracting path and an expanding path, each containing a five times repetition of two convolutions followed by a Scaled Exponential Linear Unit (SELU). After this convolution block, the contracting path applies downsampling using 2×2 max pooling, and the expansive path employs 2×2 upsampling and an additional 2×2 up-convolution layer. Unlike the classic U-Net, proposed by Ronneberger et al. [7], the repeating two convolutional layers have an alternating kernel size. For this study, increasing kernel sizes in the contracting path and decreasing kernel sizes in the expansive path, ranging between three and six, led to the best performance. To improve stability, generalizability and training time of the model, dropout and batch normalization layers were later added after multiple experiments.

3.4. Model training

Hyperparameter tuning has been conducted to find an optimal configuration of hyperparameters. The binary cross entropy dice loss, a sum of binary cross entropy and dice loss, has been selected as loss function. The Adam algorithm with a learning rate of 0.0001 is the best performing optimizer. A sigmoid layer was used as the output activation layer, and the model was trained for 100 epochs with a batch size of 32.

To address the circa 30:1 data imbalance between negative and positive pixels, the loss function takes class weights into account. This is implemented by adding sample weighting to

Precision	Recall	F1
0.86	0.80	0.83

Table 2. Performance of the primary model on the test set.

the ImageGenerator class in Keras. For every batch, the ratio of positive and negative pixels is calculated and considered in the calculation of the loss. This places more focus on the underrepresented positive pixels.

Seven different models were trained, each with a different training, validation and test set. We first trained a model, subsequently called the 'primary model', on images from every site and evaluated it using one selected image per site. Thereafter, we conducted regional cross validation by training six models on all images except the images from a specific site. The images of this site were excluded from the training phase and only used for evaluating the model. Regional cross validation enables the testing of the generalizability of the model and the performance on unseen regions.

3.5. Application

Using the data collection pipeline and one of the trained models, thaw slumps can be predicted from 2016 (the launch of Sentinel-2) onwards, for any chosen area in the arctic without needing data source licenses. Here, we demonstrate the potential and functionality of the developed tool by applying it to characterise RTS development in the six sites, over the last six years. We find that the inconsistent presence of snow, ice and clouds in images heavily influences model performance. We therefore use summertime images to minimize the amount of snow, ice and cloud cover, so that year-to-year images are similar. Using these images, the primary model is used to predict the amount, extent and changes of RTS in these sites.

4. RESULTS

4.1. Evaluation

The significantly high precision, recall and F1 scores of the primary model, displayed in Table 2, indicate the good performance of the model in predicting RTS in these sites. This is supported when comparing the RTS ground truth and predictions, as shown for Horton 01 in Figure 2. For all the six sites, the model is capable of identifying the largest RTS clusters with a high level of accuracy. However, some predicted RTS clusters contain gaps of false negatives, contributing to a decreased recall value. Additionally, the model's precision is decreased by small false positive RTS clusters.

The regional cross validation models, evaluated solely on images of a site not seen during training, have significantly lower performance, as shown in Table 3, suggesting that the trained models are not able of accurately detecting RTS when being employed on region not seen during training. This is not unexpected, as a relatively low amount of training data is available, limiting the generalisability of our approach.

(a) Ground truth (b) Predictions

Fig. 2. Horton 01 satellite images with RTS ground truths and predictions.

Herschel Island	Horton 01	Horton 02	Kolguev 01	Kolguev 02	Lena
0.10	0.02	0.07	0.03	0.01	0.04

Table 3. F1 values of the regional cross validation models.

4.2. Detecting RTS over time

Figure 3 shows the total area and amount of RTS predicted by the model for the sites Horton 01 and Kolguev 02 for the last six years. The model is able to identify the main RTS cluster of the site for every year, and tracks small distinctions in area and amount of RTS between years. Because of the imperfect accuracy of the model, differences in image quality and only small distinctions in detected RTS, we cannot draw a general conclusion about the change of RTS in these areas. Using data from Turetksy et al 2020 [3], and assuming that pixels not classified as RTS are undisturbed tundra we calculate that, as a result of RTS development, Horton 01 and Kolguev 02 are seeing net CO_2 losses to the atmosphere of 1810 and 850 tons C/yr respectively. Following Turetsky et al. 2020 [3], these fluxes include CO_2 release from net ecosystem exchange as well as particulate and dissolved organic carbon losses, were two thirds of losses are assumed to be mineralized to CO_2. Hillslope thermokarst has negligable effects on CH_4 fluxes so only CO_2 is included.

5. DISCUSSION AND CONCLUSIONS

This work presents a novel application of CNNs for identifying and characterizing RTS in permafrost regions, adding to a growing literature demonstrating the power of CNNs for computer vision applications. Tackling data imbalance and enlarging the training set using data augmentation play a key role in maximizing model performance. In particular, the use of sample weighting — to lay more focus on the underrepresented pixels — has improved performance remarkably.

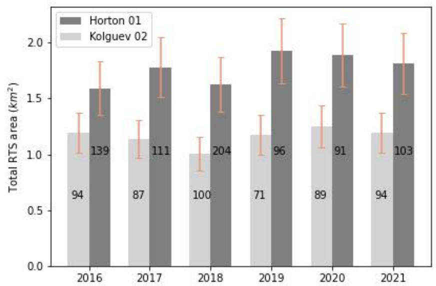

Fig. 3. Predicted total area of RTS from 2016 - 2021. 15% error bars are included to represent uncertainties in the model output based on performance factors listed in Table 2. The number of detected RTS is displayed in the bars.

Overall, this work demonstrates that machine learning models using imbalanced, medium-resolution, open-source data can be used to identify and monitor the incidence and area of RTS in permafrost regions. The insights gained using this approach — including the total area of RTS – can be used to produce more accurate estimates of GHG emissions arising from RTS development in these sites over time. Our calculations show even the small areas of RTS seen now are enough to turn the study areas from sinks to sources of CO_2. This approach has significant potential applications for monitoring GHG emissions from these natural processes, which are accelerating under climate change, and for better understanding the drivers behind their development.

Our evaluation factors greater than 0.8 for detecting RTS in sites seen during training, are comparable to those achieved by Nitze et al. (2021) and Huang et al. (2020) [4, 5]. Although this indicates the model is performing well, we show that the uncertainty in the model outputs have a big impact on results and trends (Figure 3). Higher model performance is required for models to be sensitive enough to accurately track changes in RTS year to year for the sites studied. The performance of our model may be limited by the resolution of the open source Sentinel-2 spectral data. RTS in Canada extend at a rate of between 7.2 and 26.7 m per year [8]. It is likely that these small changes in RTS are difficult to detect because Sentinel-2 has a 10 m resolution.

Our regional cross validation results demonstrate the general difficulties of predicting RTS in unseen sites across the arctic and replicates issues encountered by Nitze et al. (2021) [4]. These challenges likely arise from the heterogeneity of landscape and varying site properties, such as ground ice content and sediment composition. To address these challenges, more and more diverse RTS ground truth data is needed to train the models. Models developed by Nitze et al. (2021) [4] using PlanetScope data and a similar deep learning approach perform better for unknown sites; this may be due to much higher resolution data (3 m) compared to this study (10 m).

Retrogressive thaw slumps are a canary in the coalmine, indicating areas of rapid thaw and high GHG emissions. This work sets the foundation for an immediate use of computer vision for tracking RPT at scale, using free data and software. Improved access to high resolution satellite data and more extensive and diverse RTS ground truth data are needed to realize full the potential of this and similar approaches.

6. ACKNOWLEDGEMENTS

We are grateful to Ingmar Nitze and Guido Grosse from the Alfred Wegener Institut for their valuable input. We worked in partnership with Accenture UKI Applied Intelligence who provided a team of Cloud Architects, Engineers and Data Scientists pro bono to build the initial CNN model and cloud based solution to identify permafrost thaw. We are especially thankful to Leonor Furtado, Jeffrey M. Nogosek, David E. Boston, and Mark Smith from Accenture. The access granted to Microsoft's Planetary Computer made this work possible.

7. REFERENCES

[1] Jens Strauss et al., "Permafrost," *Recarbonizing global soils–A technical manual of recommended management practices: Volume 2–Hot spots and bright spots of soil organic carbon*, p. 130, 2021.

[2] V Masson-Delmotte et al., "Climate change 2021: The physical science basis. contribution of working group I to the sixth assessment report of the intergovernmental panel on climate change," *IPCC*, 2021.

[3] Merritt R Turetsky et al., "Carbon release through abrupt permafrost thaw," *Nature Geoscience*, vol. 13, no. 2, pp. 138–143, 2020.

[4] Ingmar Nitze et al., "Evaluating a deep-learning approach for mapping retrogressive thaw slumps across the arctic," American Geophysical Union, 2021.

[5] Lingcao Huang et al., "Using deep learning to map retrogressive thaw slumps in the beiluhe region (tibetan plateau) from cubesat images," *Remote Sensing of Environment*, vol. 237, pp. 111534, 2020.

[6] Jing Luo et al., "Recent acceleration of thaw slumping in permafrost terrain of qinghai-tibet plateau: An example from the beiluhe region," *Geomorphology*, vol. 341, pp. 79–85, 2019.

[7] Ronneberger Olaf et al., "U-net: Convolutional networks for biomedical image segmentation," 2015.

[8] Denis Lacelle et al., "Distribution and growth of thaw slumps in the Richardson Mountains–Peel Plateau region, northwestern Canada," *Geomorphology*, vol. 235, pp. 40–51, 2015.