Regional-Scale Permafrost Mapping by Scalable Machine Learning

Ghislain de Laplante

Thesis submitted to the University of Ottawa
in partial fulfillment of the requirements for the
Master of Science in Earth Sciences

Department of Earth and Environmental Sciences
Faculty of Sciences
University of Ottawa

Thesis supervisors:
Dr. Clément Bataille (Department of Earth and Environmental Sciences, University of Ottawa, ON, Canada)
Dr. Elliott Skierszkan (Department of Earth Science, Carleton University, Ottawa, ON, Canada)

Preface

When I began my studies with Dr. Clément Bataille, I envisaged this effort taking the usual, semi-linear pattern typical of an M.Sc.: courses, field studies, publication, and subsequent graduation, with time for leisure here and there. However, as with most endeavors so far in my life, this path has been anything but linear. I started this degree with a very young child, soon returned to the Yukon mid-COVID pandemic, and was quickly offered work with the Yukon government in the field of aqueous data science. Thankful for this employment and aware of the short-term nature of the opportunity, I may have spent more time working than studying (oops!). This exposed an ongoing need for the position as well as my innate aptitude for the work which led to more work, and I soon found myself the successful applicant to a full-time, permanent position as Climate Change and Water Data Scientist. I was not expecting to find this quality of employment at least until after completion of my degree, but I certainly wasn't about to turn down such an opportunity! Amongst all of this, my wife continued a highly demanding position creating the Yukon's midwifery legislation and implementing its first program – and we had a second child. In retrospect, pushing forward with my studies was not the wisest path if I was concerned with maintaining my sanity, but alas – here I am!

Despite the difficulty of pushing through with studies, a young family, and full-time work, I can categorically state that the outcome was worth the effort. Through my studies, I have helped advance Drs. Bataille and Skierszkan's research into weathering processes and geochemistry, published meaningful contributions towards advances in permafrost mapping, and developed an open-source coding framework to facilitate spatial predictions using machine learning. I have gained invaluable skills in the process besides the data science turn of my degree, including experience in using clean laboratory techniques, understanding the chemistry behind ion chromatography and the niche analytical techniques used in isotopic research. Perhaps most importantly, I've leveraged this experience into my current and highly diversified work, wearing the many hats of data scientist, GIS specialist, aqueous geochemist, hydrologist, and database and application developer, while also representing my jurisdiction on the Canadian Council of Ministers of the Environment in the development of water quality guidelines. I owe these opportunities in no small part to this degree and to my supervisor's support while I juggled both student and government employee roles.

I'd also like to briefly mention that my studies began with the goal of answering questions regarding weathering processes using stable isotopes and evolved to have the very different focus presented in this thesis. At first glance, weathering processes and permafrost mapping may seem distantly related, yet they are inextricably linked in cold regions. Without accurate knowledge of permafrost presence or absence over our study area, it was impossible to know how the weathering signals we discerned in isotopic data were (or were not) influenced by perennially frozen ground. At Clément's behest, I therefore endeavoured to apply machine learning techniques to this question, thinking it would be a simple side question. I only understood later that this was no simple question to fully address. Most importantly, this work output was perhaps the most important contribution that I could make (in the scope of my studies) to Earth Sciences in my adopted home, the Yukon, and across the circumpolar world.

I plan to continue advancing the twin topics of my studies from my current professional position as scientist specializing in water chemistry and hydrology: the clear influence of weathering processes on aqueous chemistry as well as the rapid changes to hydrological processes brought about by thawing permafrost make this an imperative area in which to foster research. This thesis may mark the end of my studies with the University of Ottawa but begins new academic collaborations aimed at continuing progress on these ever-important topics.

Acknowledgements

First and foremost, I would like to thank Dr. Bataille for his infinite patience with my progress, always halting and sometimes absent, as I directed my attention to career or family. My co-supervisor, Dr. Elliott Skierszkan, has also been involved near constantly with my research and has provided invaluable access to knowledge, data, field planning expertise, and insightful reviews of my written work. I would not have achieved what is presented in this document without the patience and generosity of these two gentlemen.

I'd also like to thank Myunghak (Julian) Khang, my predecessor in Dr. Bataille's research group whose research was focussed on lithium isotope geochemistry. Julian was instrumental in teaching me how to work under ultra-clean laboratory conditions and generous in sharing the innovative methods he developed for lithium ion chromatography. The laboratory component of my studies was made entirely possible by his and his collaborator's preceding scientific contributions. Though I did not, in the end, work with isotope data, learning the ultra-clean laboratory techniques necessary for trace metal isotope analyses has proved useful in my professional life.

Of course, my scientific mind did not spring from nowhere. I owe a debt of gratitude to my parents for their patience in answering a never-ending barrage of 'why' questions, for fostering my young mind's inquisitiveness with countless projects, educational (and fun!) toys, and for exposing me to diverse experiences. Maman et Papa, je sais que je vous ai peut-être causé quelques mini attaques de cœur avec mes déplacements constants, ma tendance prononcée de vouloir habiter dans ma tente peu importe la saison, mon travail outre-mer, et mes longs voyages, mais regardez : ça en a valu la peine!

Lastly and most importantly, I am grateful for the enduring support of my dear wife. We first began dating as I was returning to school to finish my undergraduate degree, and for better or worse, I kept on going, even as life got increasingly busy. We've supported each other through big career changes and challenges, raising kids, and inter-provincial moves, but her support for my studies has no equivalent. Thank you, Elizabeth!

Table of Contents

List of Figures

List of Tables

Abstract

This thesis presents an integrated approach to mapping permafrost at high-resolution using scalable machine learning (ML) methods and is composed of two interrelated components. The first component is a research paper outlining the development and validation of a ML framework and model to predict permafrost presence and ice content across west-central Yukon, Canada (Chapter 2). The second component is a published, open-source R package developed in support of the aforementioned permafrost prediction model (Appendix), created with the goal of facilitating future applications of ML techniques to spatial predictions of permafrost or other spatial phenomena with built in support for model optimization.

In Chapter 2, permafrost distribution and relative ice content is modelled across west-central Yukon using as a training set a detailed field map of permafrost conditions across a proposed mine site (the Coffee Gold Project). Permafrost distribution was constructed using ML and a series of spatial predictors derived primarily from an open-access Canadian high-resolution (16 m) digital elevation model (e.g., elevation, slope, aspect, curvature, solar radiation). Using random forest (RF) classification, we predicted permafrost distribution across the training area with a balanced accuracy of approximately 0.86. We also validated the model's performance by accurately reproducing a permafrost distribution map produced at a second, 35 km distant mining project (Casino Mine Project) to a 73 % agreement. This high accuracy underscores the potential for ML to bridge the scale gap between detailed local field observations and broader regional permafrost mapping efforts and substantially improves upon the previously existing permafrost models over Yukon in terms of resolution and accuracy.

Complementing the research paper, the SAiVE R package (Appendix) offers a robust, reproducible, and user-friendly tool to streamline the creation of optimized ML models predicting spatially dependent variables. The package streamlines tasks ranging from predictor variable extraction and spatial sampling, model training, testing, and comparison and selection, and (if desired) the extrapolation of predictions over extensive geographic areas. By integrating functionalities for hyperparameter tuning and variable importance ranking, SAiVE not only facilitates advanced spatial modelling for permafrost studies but also provides a flexible framework adaptable to other environmental applications.

Together, these contributions offer a scalable solution for permafrost mapping that supports environmental management, infrastructure planning, and climate adaptation strategies in rapidly changing Arctic and sub-Arctic regions. I expect that the modelling approach presented here will lead to improved representations of permafrost in Yukon and beyond, with direct applications for research,

hydrology, mining exploration, and infrastructure projects looking to extrapolate local permafrost measurements to broader scales.

Chapter 1: Introduction

1.1 Thesis structure

I wrote this thesis in an article format, with an introduction and background written to support an independent research paper (Chapter 2) intended for publication to a scientific journal. In addition, I have provided documentation for a published R-language software package in the Appendix as this development formed a core component of my research. The remainder of this introduction aims to provide the reader with background information regarding my research topic, permafrost modelling, and my study area, defines my research objectives, provides a brief overview of recent advances in permafrost modelling, and explains the basics of machine learning (ML) and how I have used these tools to predict permafrost at the regional scale in Canada's Yukon territory. The conclusion (Chapter 3) discusses some advantages and limitations of this modelling approach, and suggests future improvements to refine the model's performance, allow the application of this approach to broader spatial scales, and incorporation of climate change predictions in model inputs.

1.2 Background

1.2.1 Permafrost in a changing climate

Permafrost, defined as ground that remains at or below 0°C for at least two consecutive years (ACGR, 1988), is a crucial component of the Earth's cryosphere. Accurate modelling of permafrost is essential for predicting the impacts of climate change, managing natural resources, and planning sustainable development in permafrost-affected areas. Its dynamics significantly influence global climate systems (Schaefer et al., 2020; Schuur et al., 2015), aquatic and terrestrial ecosystems (Schuur & Mack, 2018; Vonk et al., 2015), and the stability of infrastructure (Hjort et al., 2022; Schneider von Deimling et al., 2021) in cold regions. As global temperatures rise, thawing permafrost contributes to the release of stored carbon (Tarnocai et al. 2009; Hugelius et al. 2014; Conant 2019 and others), alters the chemistry (Colombo et al., 2018) and quantity (Walvoord & Kurylyk, 2016) of affected waters, and poses challenges for engineering and land use in affected areas. Mapping permafrost is therefore essential: it not only enhances our understanding of its spatial variability and dynamic behavior but also informs predictive models that can guide risk mitigation, adaptation strategies, and policy decisions.

Given the high importance of permafrost to physical and biological processes and infrastructure in affected regions, permafrost maps have been generated at many scales and resolutions. Local-scale (up to 10s of km^2) maps are typically informed by field surveys and constitute important tools for resource extraction and infrastructure projects (e.g. Tetra Tech EBA 2016; Western Copper and Gold 2024). These maps can be highly accurate and detailed, distinguishing between frozen and unfrozen land only meters apart, but are often limited in spatial coverage and require expensive fieldwork for data collection. Continental to hemisphere-scale maps (e.g. Gruber, 2012; Obu et al., 2019; O'Neill et al., 2020; Westermann et al., 2024), on the other hand, rely on sparse observations compared to their scales and rely on low-resolution (km scale) climate and land classification variables coupled with numerical models. While this approach can help predict present and future permafrost cover at large spatial scales, it is of limited use at small scales due to the low resolution of global climate datasets as well as by computational limitations which restrict the model's final resolution to 1 km^2 or more. This in turn limits the applicability of these maps to answer local scale questions, especially in topographically complex regions. Attempts to generate higher-resolution maps at intermediate regional scales (province/state/territory, ecoregions) have primarily relied on field observations, remote sensing datasets (Pastick et al., 2015), or empirical models (Bonnaventure et al., 2012) by building associations between permafrost cover and spatiotemporal predictor variables. However, these attempts still lack the resolution and/or precision needed for local infrastructure projects or small-scale scientific research.

Modelling approaches that can replicate the accuracy and granularity of local-scale maps at the regional or larger spatial scale hold promise for applications in land-use planning and development, scientific research, and public safety efforts such as flood forecasting or geohazard planning and mitigation. Developing such models would reduce the required extent and cost of field surveys for mapping permafrost, and open-access high-resolution maps of permafrost cover would allow resource-limited organizations such as local or indigenous governments, NGOs, and researchers to better assess the risks and potential benefits associated with proposed infrastructure or mining development. Similarly, these models would also support the assessment of permafrost-related risks and costs for industry and government projects without the need to develop costly permafrost maps, thereby facilitating sustainable development in circumpolar regions.

1.2.2 History and evolution of permafrost modelling

Over the past century, permafrost modelling has evolved from simple empirical methods to sophisticated physics-based simulations, and more recently, to data-driven approaches that incorporate ML. Early permafrost modelling efforts were predominantly empirical, relying on observed correlations between permafrost presence and environmental variables. One of the earliest approaches involved using mean annual air temperature (MAAT) to map permafrost distribution in Russia based on MAAT isotherms (Wild, 1882), providing a broad estimation of lateral permafrost extent. Degree-day models were also employed, calculating cumulative freezing and thawing indices to estimate ground temperature regimes and active layer thickness using methodology outlined in Nelson and Outcalt (1987). The Stefan equation (Stefan, 1891), originally conceived to predict the growth of ice over water, was later used to estimate the depth of freezing or thawing by relating it to thermal properties of the ground and cumulative temperature conditions. While useful, these empirical models often failed to capture some of the underlying physical characteristics governing permafrost dynamics, such as soil composition, moisture content, and the insulating effects of snow and vegetation.

The advent of numerical or physics-based models marked a significant advance in permafrost modelling. These models incorporate soil thermal properties and latent heat effects due to phase changes between ice and water, and more complex models simulate the movement of water and energy along a two- or three-dimensional representation of the land surface (see Flato 2011). Taking this approach further, earth system models can simulate the exchange of air, water, and energy both within and between the atmosphere and Earth's surface, and can even incorporate hydrological processes into permafrost simulations to form complex representations of earth systems (for example, Elshamy et al. 2024). Coupled with climatological models, these permafrost models allow for the assessment of future climate predictions on permafrost presence and properties such as active layer thickness or proportion of water ice (Matthes et al., 2025).

Physics-based models improved and continue to advance the understanding of permafrost processes by explicitly representing thermal and hydrological mechanisms, but require extensive and accurate input data on soil properties, vegetation, and meteorological conditions for accurate predictions. Uncertainties in the quality of these inputs or their unavailability limits their applicability, especially in data-sparse regions. These complex models are also computationally intensive, which either limits their resolution to cell sizes measured in square kilometers for broad scale modelling or limits their spatial extent for high resolution modelling. These constraints limit current applications of such models, as they may be inaccurate in regions with sub-kilometer variations in topography and geomorphology (e.g., mountainous regions), where topography and insolation play dominant roles on permafrost distribution (Gruber, 2012) or be severely limited in spatial extent.

To address the input data and scale (or conversely computing power) limitations of physics-based models, alternative methods were developed that balanced complexity and data requirements. For example, the Basal Temperature of Snow method first described by Wilfried Haeberli in 1973 uses measurements of temperature at the snow-ground interface in late winter as a proxy for underlying permafrost conditions. This method was notably applied in the Yukon Territory of Canada (Bonnaventure et al., 2012) to map permafrost distribution in rugged terrain with variable snow distribution conditions. While practical for regional-scale mapping and accounting for microclimatic variations due to snow cover, the basal temperature of snow method in its current implementations over Yukon does not adequately capture the effects of parameters such as solar radiation, vegetation cover, soil properties, or slope characteristics which govern permafrost presence (Etzelmüller & Frauenfelder, 2009; Luetschg et al., 2008) nor does it describe permafrost ice richness, active layer thickness, or other relevant characteristics.

1.2.3 Machine learning for spatial variables

In recent years, ML techniques have emerged as powerful approaches for spatial modelling. ML, a type of artificial intelligence or computer self-learning first proposed in the late 1950s (Samuel, 1959), enables computers to learn patterns and associations rather than relying solely on explicit logic (if-then statements). Once trained, ML models apply these learned associations to predict outcomes from new predictor variables. Approaches range from simple techniques, such as linear regression or nearest neighbor classification, to more complex and flexible approaches such as artificial neural networks (ANNs) and random forest (RF) models. These and other advanced algorithms iteratively adjust their internal parameters to minimize prediction error, and have been applied across diverse domains including speech recognition, automated text classification, image analysis, and environmental modelling.

The use of ML methods for spatial prediction dates at least to the 1990s (Haupt et al., 2022), with the earliest examples applying ANNs – computational models inspired by the brain's interconnected neurons that learn patterns from training data – to terrain and land cover classification problems (Benediktsson et al., 1990; Decatur, 1989). Later spatial applications include the use of ANNs and the introduction of decisions trees or random forests - which classify training data by sequentially splitting it using simple rules to build a 'tree' of decision points for later classification of new data - in image classification (Hansen et al., 1996; Paola & Schowengerdt, 1994), yielding superior performance compared to the rules-based classification techniques (if-then logic) in use at the time.

Around the same time, meteorologists and environmental scientists began to experiment with various ML techniques, achieving early success in streamflow modelling with ANNs (Crespo & Mora, 1993; Hsu et al., 1995; Karunanithi et al., 1994). Weather predictions using decision trees became important in the mid to late 2000s (Gagne et al., 2009; Williams et al., 2008), as ever larger sets of satellite and ground-based data, combined with improved computing performance, favored methods capable of extracting outcomes from large and continuous data streams.

The growth in use of ML for spatial modelling is driven by the increasing number of libraries available in R and Python such as the *caret* (Kuhn, 2008) and *scikit-learn* (Pedregosa et al., 2011) packages, the development of improved spatial manipulation libraries, the availability of large remote sensing datasets via open-access platforms (e.g., Google Earth Engine, ERA-5), and the growing body of literature streamlining the application of ML to predictions of all types (e.g. Du et al. 2020; Kopczewska 2022; Jemeļjanova et al. 2024). These techniques offer a potentially useful approach to mapping permafrost distribution because they can represent nonlinear relationships between permafrost occurrence and environmental variables without explicit assumptions about underlying processes. This can reduce or obviate the need for expensive and difficult-to-scale observations of ground temperatures, local climate, or other site-specific data except for permafrost presence/absence or qualities such as active layer thickness or water ice content.

However, despite the similarity of these aforementioned spatial predictions to permafrost prediction, there have been limited applications of ML to predict permafrost presence/absence or qualities such as active layer thickness or proportion of water ice, with limited research beginning in the mid 2010s (e.g. Pastick et al. 2015; Deluigi et al. 2017; Baral and Haq 2020; Thaler et al. 2023). These recent developments demonstrate that ML is a promising approach to permafrost modelling, with reported accuracies in the 70 to high 80% range using a variety of ML methods. The best performing methods differed between these approaches, with support vector machines, classification trees, and neural networks proving useful. This disparity suggests that various approaches may be suitable depending on the types of input data sets and local environmental conditions.

A concern in applying ML to permafrost mapping is the interpretability of the resulting statistical models. ML models can yield accurate predictions based on variables that lack clear physical relationships, and the process by which decision trees or mathematical formulas are developed during training further complicates the interpretation of these predictions (McGovern et al., 2019; Molnar, 2025). Building and interpreting ML-based permafrost predictions requires careful consideration of the selected predictors, building upon the literature of mechanistic models. Another concern when using ML-based approaches for mapping permafrost is their reliance on the quality and representativeness of the training data, as well as their inability to extrapolate predictions to regions with environmental conditions outside the training range (Beery et al., 2018; Gray et al., 2024). As such, ML-based models of permafrost distribution require spatially diverse training data to support extrapolation across varying landscapes.

1.2.4 Integration of permafrost prediction methods

The current state of permafrost modelling encompasses a spectrum of approaches. Field-based mapping continues to form the basic building blocks and validation for permafrost models and remains critical to the further development of numerical models. While often limited in extent and resolution and requiring accurate data on climactic variables (which can limit their use in many regions), physics-based models are essential for integrating process-level dynamics of permafrost distribution into land surface models for predicting the present and future distribution of permafrost. These models play a central role in informing predictions of vegetation changes or carbon biogeochemical cycling but do so at broad scales and low resolutions. ML-based modeling could offer a complementary and promising avenue to generate high-resolution permafrost distribution maps over large expanses to accommodate the range of applications which require meter-scale permafrost maps without needing to model permafrost evolution. These models also have the advantages of rapid computation (facilitating rapid iteration) and are not directly mechanistic, which could complement physics-based models by revealing environmental predictors which are important for permafrost distribution but have yet to be integrated into physics-based models.

1.2.5 Target machine learning methods

A wide range of ML algorithms can model permafrost distribution, but I chose to focus on two families – random forests (RF) and multivariate adaptive regression splines (MARS) – because they combine strong predictive power with interpretability.

Random forests, originally introduced by Breiman (2001) with several variations now in use, have already been used successfully to predict permafrost (e.g. Pastick et al. 2015; Baral and Haq 2020; Thaler et al. 2023) as well as other phenomena and are particularly well suited to this task. These methods are non-parametric, meaning that they make no assumptions about the relationship between predictor and outcome variables, and can learn complex interactions between variables which would be difficult, if not impossible, to discover otherwise. Predictions are made by 'growing' ensembles of decision trees trained on a bootstrap sample of the training data (a random sample drawn with replacement from the original data), a random subset of predictor variables, and the known outcomes associated with these predictors. For prediction, the forest (ensemble of decision trees) is pooled: each tree's outcome is a 'vote', and the majority class (for classification tasks) or the average value (regression tasks) becomes the final prediction for any given point. Though sometimes criticized as yielding ‘black box’, inscrutable models, modern diagnostic tools address this criticism: variable-importance scores rank predictors by their contribution to predictive accuracy, and partial-dependence plots reveal how the predicted outcome changes as one predictor varies while others are held constant, exposing key thresholds and interactions.

MARS functions (Friedman, 1991) are also potentially suitable for classification of permafrost presence/absence and are easily interpreted. MARS builds models from *piece-wise linear splines*: functions made up of straight-line segments joined smoothly at points called *knots* (break-points), with both the number and location of knots selected from the data. These splines allow the model to change slope at specific predictor values, capturing local thresholds and non-linearities. The resulting equation is explicit, so critical thresholds – such as a sharp change in permafrost probability when solar radiation exceeds a given value – appear directly as polynomial coefficients tied to the knot locations. Model construction uses an iterative, stepwise search that retains only splines that improve performance, providing automatic variable selection and reducing the risk of over-fitting. Because the final models typically contains only a few dozen mathematical terms, they are easy to interpret and communicate.

1.3 Study site

1.3.1 Regional setting

My research is focused on the Dawson Range region of south-eastern Yukon, Canada, a mountainous region near the Yukon-Alaska border situated within the discontinuous permafrost zone (Bond & Lipovsky, 2011; Obu et al., 2019) and without significant history of past glaciation (Bond, 2019; Duk-Rodkin, 1999). This region is widely known for its prolific placer gold deposits, themselves a consequence of a suitable geological setting and weathering over millennia without the presence of glaciers to grind and distribute the gold concentrated in valley bottoms (Lowey, 2006).

Owing to its history of gold production, the region (depicted in Figure 1.1) has been the subject of exploration interest by mining companies interested in both placer and hard-rock deposits (YESAB, 2024). Two of these prospects, the Casino and Coffee Gold projects (respectively Western Copper and Gold, Vancouver, Canada and Newmont Mining, Denver, Colorado) have advanced to the point of requesting or receiving regulatory approval for open-pit mining (Western Copper and Gold, 2024; YESAB, 2022). As part of the regulatory process, these companies have invested significant resources into climate and ecological monitoring and developed detailed permafrost maps (Tetra Tech EBA, 2016; Western Copper and Gold, 2024) across their areas of interest. Their willingness to share these data with researchers, coupled with the pressing need for improved permafrost mapping across the Dawson Range, was instrumental in selecting this region for the development of the permafrost prediction model described here.

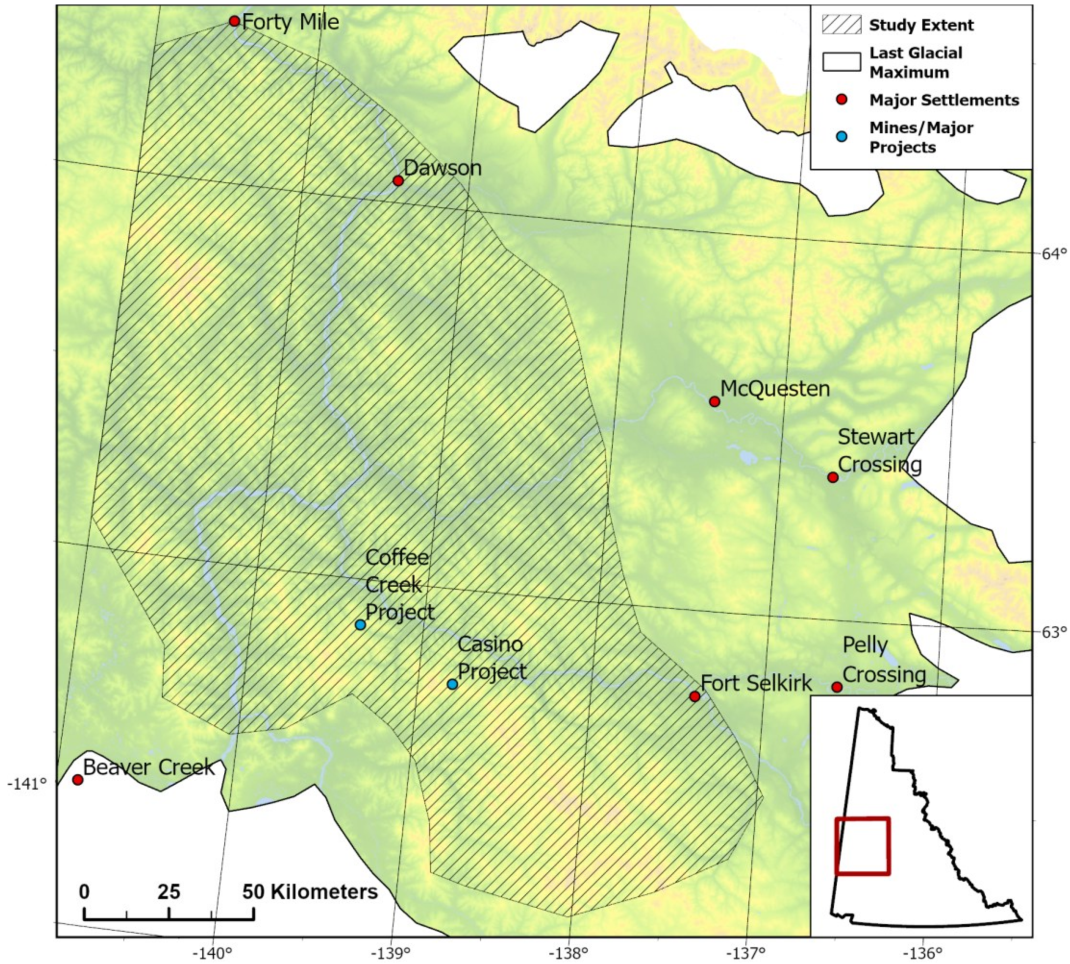

Figure 1.1: Study area, situated in south-eastern Yukon, Canada. The area surrounding and to the south of the city of Dawson has been the site of much placer mining activity in the past 130 years, and mineral exploration continues in the region for associated hard-rock deposits.

Despite interest in the area's mineral resources dating back to the late 1800s, extraction to date has been limited to placer deposits in river valleys. Consequently, little permanent infrastructure exists beyond the Klondike River valley except for roads west to Alaska and north to Inuvik, Northwest Territories, with only temporary gravel roads and camps built to access the gold bearing streams to the south. Interest from mining companies makes it likely that the region will see significant development pressure in the form of year-round access roads, mining infrastructure such as tailings and waste-rock storage facilities, wastewater treatment plants, and camps, and increased civil infrastructure demands stemming from population growth. Knowing where permafrost lies as well as its proportion of water ice and other properties is essential to the construction of infrastructure and to water-resource management and regulation (McKenzie et al., 2021; Skierszkan, Dockrey, et al., 2024). Predicting the fate of permafrost will empower governments to better manage their assets for resiliency and will facilitate research into pressing questions surrounding changing water quality in affected regions.

1.3.2 Ecological and climatic setting

The Dawson Range is within the Boreal Cordillera Ecozone (Geomatics Yukon, 2019) with typical vegetation varying widely between slope aspects and elevation (Reid et al., 2022). Dry, south aspects are dominated by aspen (*Populus tremuloides)* with understory species such as kinnikinnick *(Arctostaphylos uva-ursi)* and creeping juniper *(Juniperus horizontalis);* valleys with abundant sunshine and well-drained soils are dominated by white spruce *(Picea glauca)* with an understory of mosses and shrubs such as soapberry *(Shepherdia canadensis)* and red currant *(Ribes triste);* less sun exposed valley bottoms and north-facing aspects feature smaller white spruce (where active layer thickness is greater) to increasingly sparse and stunted black spruce *(Picea mariana)* as active layer thickness decreases and soil moisture increases. At greater elevations (~1000 m and above), trees give way to shrubs such as dwarf birch *(Betula nana)* and willow *(Salix spp.)* on non-north-facing aspects, and a mostly treeless, moss-dominated landscape on north aspects and mountaintops.

The climate is continental, with the nearest long-term climate monitoring station situated in Dawson City, Yukon, ~130 km to the north. Recent statistics calculated for the 1991-2020 period show a mean annual air temperature (MAAT) of -3.8 ºC (Environment and Climate Change Canada, 2024). Monthly mean air temperatures range from -25.7 ºC in January to 15.9 ºC in July, though extremes range from -52 ºC to 34.5 ºC. Precipitation averages 319 mm/year, with ~3/5 as rain and the rest as snow. As with other northerly locations, temperatures are warming, both on an annual average basis and with fewer extreme cold events (AMAP 2021; IPCC 2023). These changes in MAAT (Figure 1.2) are in keeping with observations of an average annual air temperature increase of three times the global annual rate since the late 1970s (AMAP 2021; Li et al. 2023)

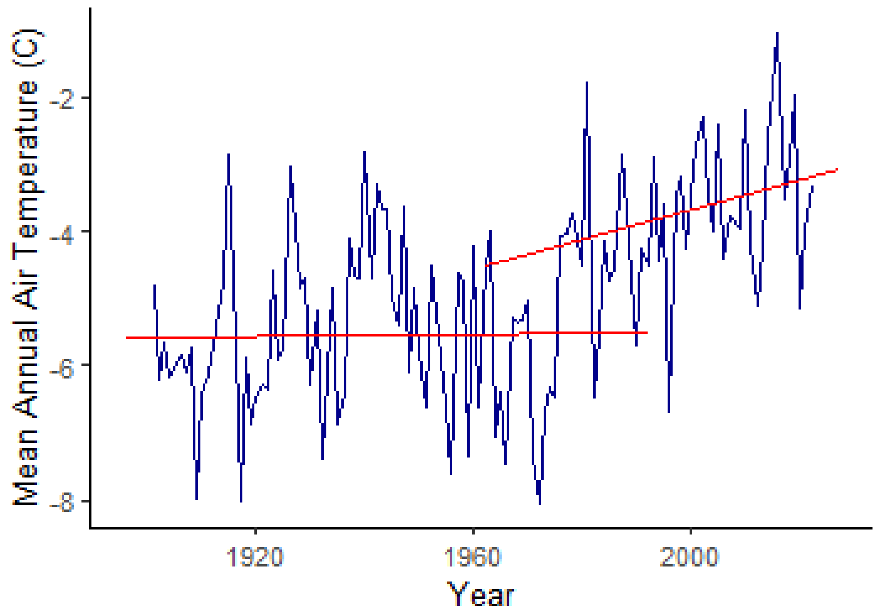

Figure 1.2: Change in MAAT at Dawson City, Yukon from 1898 to 2024. A clear change in trend is evident in the late 70's, consistent in timing with global observations but greater in magnitude (IPCC 2023). This represents a compound climate record from three locations within the Klondike River Valley, with historical records adjusted using periods of overlap to align with the most recent monitoring location.

Underscoring this change, the most recent 10 summers in Dawson City (2013-2022) with complete records had much greater accumulated degree days of thaw than the historical mean (all exceeding or at the 75th percentile), while all but one of the past 10 winters have been warmer than the historical (Figure 1.3).

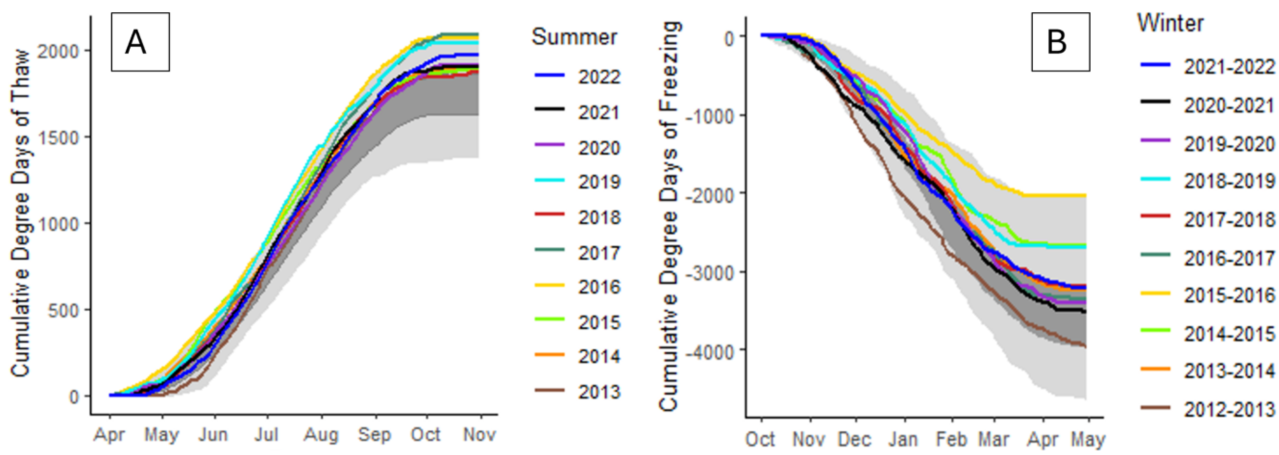

Figure 1.3: Cumulative degree days of thaw (A) and freezing (B) at Dawson City. The light grey ribbon represents the minimum and maximum bounds of cumulative degree days dating back to 1901, while the dark grey band represents the 25th and 75th percentiles.

These warming temperatures have important implications for the state of permafrost in the region, especially given its already warm status: near the Coffee Gold mining project, permafrost temperatures range from -2 to 0 ºC (Tetra Tech EBA, 2016), while borehole temperature records from the Klondike River valley (near Dawson) indicate similar temperatures (Yukon Geological Survey, Energy, Mines and Resources, Government of Yukon, 2025). This warm, discontinuous permafrost is at high risk of degradation and of eventual disappearance.

1.3.3 Existing permafrost maps

The Dawson Range region is largely classified as being within the discontinuous permafrost zone (Brown et al., 1997; Obu et al., 2019), which is reflected in observations that permafrost presence/absence is tightly linked to slope aspect and incoming solar radiation resulting in its presence on shaded slopes and absence elsewhere (Bond & Lipovsky, 2011; Tetra Tech EBA, 2016; Western Copper and Gold, 2024). Global permafrost models at 1 km or greater resolution (e.g. Gruber, 2012; Obu et al., 2019; Westermann et al., 2024), and a higher resolution (30m) model (Bonnaventure et al., 2012) exist for the region, yet both hold limitations. The global models are of limited use at small scales as their resolution prevents the differentiation of permafrost status on adjacent slopes with different local controls, such as incoming solar radiation, while the higher resolution model fails to capture these controlling factors despite its finer scale. These models apply a classification of discontinuous to extensive discontinuous permafrost to both south and north facing slopes, while it is well established that that permafrost cover is regionally negligible on south-facing slopes and extensive on north-facing slopes (Bond & Lipovsky, 2011). For example, near the Coffee Gold mining project we can observe clear distinctions between permafrost-free south-facing slopes and permafrost affected north facing slopes (Tetra Tech EBA, 2016; Western Copper and Gold, 2024, and shown in Figure 2.1), while previous models show these areas as extensively discontinuous. These discrepancies between field observations and existing models restrict their usefulness for site-specific infrastructure and environmental planning and limit our understanding of regional baseline hydrological and hydrogeochemical conditions in the Dawson Range.

In nearby Alaska, ML approaches have been used with some success for permafrost mapping (Pastick et al., 2015; Thaler et al., 2023). However, the approaches used in these two studies also have limitations for these purposes: Pastick et al. made use of several variables which are significantly downscaled or only cover the state of Alaska (wetland maps) and used a potential incident solar radiation layer which does not factor the surrounding topography, limiting the applicability of their approach in mountainous regions. On the other hand, Thaler et al. made use of high-resolution predictors but trained their models on permafrost observations derived from geophysical surveys which are very costly to obtain at large regional scales. These two studies suggest that mapping permafrost using ML approaches in the Dawson Range could be feasible, but this would require developing a more generalizable framework applicable to regions with fewer and lower resolution geophysical and remote sensing data set.

1.4 Research objectives and hypotheses

Given the combination of warming permafrost and melting ground ice and ongoing mineral exploration in the Klondike and Dawson Range region, it is imperative to build models which accurately depict permafrost in this region at present and into the future. The primary objective of my research is to develop and validate a ML framework for high-resolution mapping of permafrost distribution and proportion of water ice. This work is focussed on the Dawson Range region of south-western Yukon, Canada, but can be easily adapted to different spatial resolutions, geographic extents, predictor variables, and outcome variables such as active presence/absence, relative ice content, or active layer thickness. Current permafrost maps in the region either overlook the fine-scale variability present at the sub-catchment level (Bonnaventure et al., 2012) or lack the resolution required for effective local applications (Gruber, 2012; Obu et al., 2019). By leveraging detailed field-survey data obtained from proposed mine sites and high-resolution DEM-derived predictors, my research aims to bridge the scale gap between localized field observations and regional permafrost mapping efforts.

With the applicability of ML methods demonstrated for similar spatial prediction problems such as mapping flood extents (Bhuiyan et al., 2022; Wang et al., 2015), vegetation classes (Furuya et al., 2020), geomorphic disturbances (Perry & Dickson, 2018), and permafrost in other settings (e.g. Pastick et al. 2015; Thaler et al. 2023), I hypothesize that freely available high-resolution environmental and topographic geospatial variables derived from open-access sources can be (1) Combined with a field-based permafrost map to accurately train a ML classification to predict permafrost across the training and (2) that this ML classification map can be accurately extended beyond the training region as far as predictor variables remain within the training range.

As predictions of permafrost in my region of interest are not materially different in process than predictions elsewhere, and considering that most of the time involved in training ML models lies in the pre-processing stage, I also acknowledge that a flexible modelling framework which can work at different scales, with different predictors, and with different outcome variables is desirable. My research focusses on two broad machine learning method families, RFs and MARS due to their interpretability and demonstrated suitability for spatial predictions, but many other methods exist which may be suitable or even out-perform my tested methods. To achieve my research goals in a reproducible manner (thus facilitating future testing of new methods or parametrization), I have aimed to develop a modelling framework by which ML models can be trained on outcome variables in formats common for permafrost (points and polygons), with any number of predictor variables, and which can iterate on several ML methods and parametrizations to find the best performing combination.

Chapter 2: Regional-Scale Permafrost Mapping by Scalable Machine Learning

Regional-Scale Permafrost Mapping by Scalable Machine Learning

Ghislain T. de Laplante[1,3], Elliott K. Skierszkan[2], Clément P. Bataille[1,4]
[1]Department of Earth and Environmental Sciences – University of Ottawa, Ottawa, ON, Canada
[2]Department of Earth Science – Carleton University, Ottawa, ON, Canada
[3]Department of Environment – Yukon Government, Whitehorse, YT, Canada
[4]Department of Forestry and Natural Resources – Purdue University, West Lafayette, IN, USA

2.1 Abstract

Mean annual air temperatures at northern latitudes are warming at two to four times the global average and are driving rapid recent permafrost degradation. These changes have significant implications for northern ecosystems, water resources, and infrastructure. Effective management of this growing environmental challenge requires improved predictions of permafrost distribution. Current approaches face limitations: low-resolution (kilometer-scale or coarser) probabilistic permafrost maps derived from remote sensing data lack the precision necessary for local to regional scale scientific research, infrastructure development, and regulatory oversight, while high-resolution (10-meter scale or finer) maps generated from intensive field surveys are cost-prohibitive and thus restricted in geographic scope. To address this gap between resolution and scalability, we present a novel machine learning framework in R software and distributed as a CRAN-validated package. We tested this framework in the Dawson Range, a discontinuous permafrost zone of west-central Yukon, Canada, where a high-resolution 87 km² permafrost map derived from extensive field investigations in support of a proposed mineral development (Coffee Gold Project) was available for model training. Predictor variables were derived from globally available high-resolution elevation and vegetation cover maps. Our model's parametrization, including algorithm selection and training dataset size, was optimized to achieve a balanced cross-validated accuracy of 0.857 (95% CI: 0.855–0.859) and validated by accurately reproducing a 500 km² permafrost map produced from field observations at a second proposed mine site located approximately 35 km from the training area. We used the optimized model to predict permafrost distribution at a 16-meter resolution across a 30,000 km² area of the Dawson Range with similar geomorphology and climate to the training region. Our new permafrost map substantially improves the resolution and accuracy of existing regional-scale permafrost maps. Beyond the Dawson Range, the framework presented here offers a cost-effective, scalable, and adaptable solution to map permafrost distribution at high-resolution by integrating

existing field surveys with remote sensing datasets. The resultant high-resolution permafrost predictions and coding framework will support the sustainable development and adaptability of northern communities and infrastructure and improve our ability to monitor and protect terrestrial and aquatic ecosystems impacted by permafrost thaw.

Keywords: Permafrost, machine learning, classification, random forest, high-resolution, Yukon, sustainability, Arctic, climate change.

2.2 Introduction

Permafrost, defined as frozen ground persisting for two or more consecutive years (ACGR, 1988), shapes circumpolar and alpine landscapes, communities, and ecosystems. It is warming and degrading faster now than at any time in the Holocene Epoch (Biskaborn et al., 2019; Camill, 2005; Payette et al., 2004; Smith et al., 2022): as much as 90% of the world's permafrost is projected to disappear by the year 2100 (Lawrence & Slater, 2005). Permafrost provides a structural substrate for infrastructure such as roads, buildings, and mining facilities (Grosse et al., 2011; Hugelius et al., 2014; Intergovernmental Panel On Climate Change, 2023a; Turetsky et al., 2020) and acts as a vast reservoir of carbon (Tarnocai et al., 2009), nutrients (Keuper et al., 2012) and metal(loid)s (O'Donnell et al., 2024; Schaefer et al., 2020; Skierszkan et al., 2021; Skierszkan, Schoepfer, et al., 2024; Tarbier et al., 2021). It also regulates hydrological and weathering processes in northern hydrosystems, influencing surface and subsurface flow paths as well as solute fluxes (Brabets & Walvoord, 2009; Hinzman et al., 2005; Jin et al., 2022; Walvoord & Kurylyk, 2016; Walvoord & Striegl, 2007). Changes in hydrology and water quality can in turn disrupt aquatic ecosystems (Frey & McClelland, 2009; Walvoord & Kurylyk, 2016) and terrestrial ecosystems (Mack et al., 2011). Rapid permafrost degradation under a warming climate has widespread implications for infrastructure stability, water quality, hydrology and biodiversity.

Climatic conditions, such as mean annual air temperature, directly control the thermal regime of the ground, determining in large part whether permafrost can persist (Gruber, 2012; T. Zhang et al., 2008). Elevation further influences permafrost through temperature gradients and precipitation patterns, in turn affecting vegetation cover and the availability of moisture. Topographic factors such as slope and aspect control incoming solar radiation and play a particularly important role in mountainous regions (Etzelmüller & Frauenfelder, 2009; Gruber, 2012). Topography also influences water retention, the distribution of soil types, as well as the deposition of moisture and wind patterns, which can vary across the landscape (Johnson et al., 2013). Curvature, both tangential and profile, influence the relative ice content of permafrost by altering how water interacts with the surface (Giuseppe et al., 2016; Pachepsky et al., 2001; Pishvaei et al., 2020). Vegetation has direct effects through its insulating effect on the ground surface and as well as indirect effects by affecting soil moisture and temperature dynamics (Heijmans et al., 2022; Pastick et al., 2014; Stuenzi et al., 2021). Direct and indirect solar radiation modulates surface temperatures and energy transfer, further influencing permafrost dynamics (Julián & Chueca, 2007; Stuenzi et al., 2021). Geology, including glaciation history, contributes by influencing the physical properties of the substrate such as grain size and the water retention capacity, affecting thermal conductivity and moisture availability (Johnson et al., 2013). Summer and winter precipitation also impact permafrost: snow depth and density provide seasonal insulation, with greater accumulation reducing ground heat loss in winter (O'Neill & Burn, 2017) while summer precipitation enhances energy transfer to the subsurface, accelerating thawing in some regions (Douglas et al., 2020).

Together, these factors combine to create intricate variations in permafrost distribution and proportion of water ice. In a context of climate change, predictions of permafrost cover and evolution are essential to anticipate and mitigate the local impacts of permafrost thaw on northern communities and infrastructures (Gruber et al., 2023; Schneider von Deimling et al., 2021) as well as its impacts on hydrology, climate, biogeochemical cycles, and ecological processes (Colombo et al., 2018; Walvoord & Kurylyk, 2016). However, mapping permafrost often requires a compromise between scale and resolution. Decameter scale maps generated using field surveys, ground thermistor measurements, and geophysical surveys such as ground penetrating radar are accurate but resource intensive, and usually target small geographical extents of economic or natural resource interest. Conversely, continental-scale permafrost maps are essential to integrate permafrost in large-scale studies on climate, biogeochemical cycling, and ecological processes in northern regions (Grosse et al., 2011; Schaefer et al., 2014; Schuur et al., 2015; Skierszkan et al., 2020; Skierszkan, Schoepfer, et al., 2024; Turetsky et al., 2020). These can be produced at low cost using remote sensing and modelling approaches but often lack the resolution needed to support regional-scale studies.

Mechanistic or physics-based models of permafrost distribution are rapidly advancing with promising potential to predict permafrost distribution in the future (i.e. Elshamy et al. 2024). While the development of these models is key to developing integrated land-surface models capable of incorporating feedback mechanisms and predicting coupled evolution of earth systems, this approach is currently limited by the availability of high-resolution, high-quality data (e.g., ground temperature measurements) and computational limitations require the use of large model cell sizes resulting in severe limitations in mountainous regions (Gruber, 2012). Consequently, there is currently limited availability of high-resolution permafrost maps at suitably broad spatial scales to support the sustainable management of natural resources in permafrost regions.

Past work has demonstrated that machine learning (ML) classification can be a rapid, practical, and cost-effective approach to extrapolate field data to regional scales for a range of environmental science applications. Recent developments have applied these methods to spatial prediction problems such as integrating and expanding flood maps (Bhuiyan et al., 2022; Wang et al., 2015), mapping vegetation (Furuya et al., 2020), and predicting geomorphic disturbances (Perry & Dickson, 2018) as well as permafrost probability (Pastick et al., 2015; Thaler et al., 2023). ML thus provides a cost-effective means of producing high resolution maps of physical conditions and could be used in concert with these mechanistic methods to expand their geographic scope or resolution. However, even though ML may be useful to predict permafrost distribution and fill this gap, its implementation has remained constrained by inflexible methodologies and poor suitability to sparse point observations.

In the present study, we aimed to address the compromise between permafrost mapping resolution and scale through development and application of a ML methodology that pairs high density and fidelity field observations rendered as an interpreted map with high-resolution remote sensing variables to predict permafrost cover. We implemented our machine learning methodology in an open-access R package (de Laplante, 2024) to streamline its application to diverse regions and types of predictions with the goal of facilitating sustainable resource management and decision making across permafrost affected landscapes.

2.3 Background

2.3.1 Study area

This study focuses on the Dawson Range and surrounding regions of west-central Yukon, Canada, a mountainous region contained within the discontinuous permafrost zone (Obu et al., 2019). Most of this region has seen relatively limited direct human disturbance other than localized placer mining, yet hosts several economic mineral deposits in early to late stages of development (MacWilliam, 2018; Skierszkan et al., 2020; YESAB, 2024). These projects include the Coffee Gold project and the Casino Mining Project, both of which have produced extensive site-specific permafrost maps as part of their baseline environmental assessments (Tetra Tech EBA, 2016; Western Copper and Gold, 2024). The limited human disturbance and availability of detailed baseline permafrost maps make the Dawson Range an ideal study area for permafrost modelling.

Local topography ranges from ~400 to ~1,400 m asl and is shaped by weathering and erosion and which escaped major Pleistocene glaciations (Duk-Rodkin 1999; Bond 2019; see Figure 2.1). In the absence of recent glaciation, V-shaped valleys dominate the landscape with dendritic drainage patterns. Vegetation differences are pronounced between slope aspects and are strongly correlated to permafrost presence/absence: for example, north-facing permafrost-rich slopes support strands of black spruce (*Picea mariana)* while trembling aspen (*Populus tremuloides*) and white spruce *(Picea glauca)* are more common on south-facing, permafrost-poor slopes (Reid et al., 2022). Local geology features metamorphic and igneous rocks of the Yukon-Tanana terrane (MacWilliam, 2018; Yukon Geological Survey, 2020).

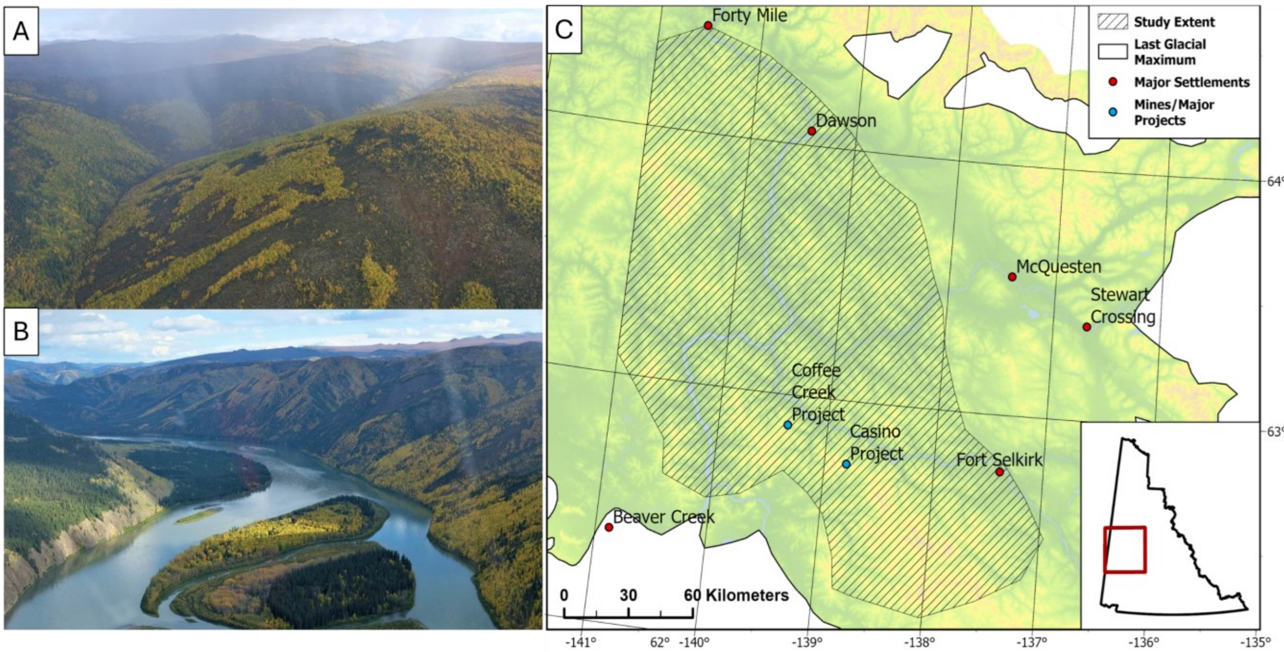

Figure 2.1: Representative field photos of Dawson Range topography and vegetation coverage at high elevation (A) and low elevation (B) near the Newmont Coffee Gold Project. (C) shows our training location (the Coffee Gold Project) within the Western Yukon territory, Canada, along with past glaciation extents. The nearby Casino Project is where a similar permafrost distribution map was used to further validate our model. The hatched area represents the area where we extrapolated our model to produce a regional permafrost map. Photo credit Elliott Skierszkan.

Climate is modulated by topography, with mean annual temperature at ridge tops circa –2.6°C (–19°C in December and +13°C in July) while winter inversions produce up to 10°C higher temperatures in ridgetops relative to valley bottoms (Lorax Environmental 2017). Climatological monitoring around the Coffee Gold Project (see Figure 2.1 for location) shows orographic effects with altitudinal gradient of + 4 % per 100 m for rainfall and + 9 % per 100 m for snowfall. The mean annual precipitation is 485 mm at ridge top, of which 65 % falls as rain and the balance as snow (Lorax Environmental 2017). The nearest-available multidecadal air temperature records are for Dawson City, at the northern margin of our study area, and shows accelerated warming since the 1960s with ~3 °C in warming since the start of the 20th Century (see Supplementary Information (SI) Figure 1). This large magnitude of change is triple the global average over the same time period (IPCC 2023b).

Permafrost cover in the Dawson Range is controlled by the interplay of these varied topographic, geomorphological, and climatic conditions, with continuous permafrost dominating high-elevations and north-facing slopes, discontinuous permafrost in valley bottoms, and little to no permafrost on south-facing slopes (Bond & Lipovsky, 2011; Tetra Tech EBA, 2016). Permafrost is locally warm, ranging from -2 to 0°C (Tetra Tech EBA, 2016), further underscoring the need for accurate permafrost mapping in this region highly vulnerable to permafrost degradation.

2.4 Methodology

Briefly summarized, our methodology involved:

1. Gathering and processing a series of high-resolution remote-sensing datasets for use as predictor variables in modeling permafrost distribution and relative ice-content (Table 2.1).
2. Creating a set of spatial points overlain on a high resolution, field survey-based map of permafrost at the Coffee Gold Project.
3. Extracting the corresponding permafrost type and the value of predictor variables at each point generated across the Coffee Gold region training map to create a tabular data set.
4. Partitioning the data set into training and testing sub-sets.
5. Training and testing ML classification methods and selecting the best-performing method.
6. Optimizing the size of the training data set and the classification method parameterization to obtain the best performing results for permafrost distribution.
7. Testing model performance through cross-validation across the Coffee Gold Project map.
8. Using the developed model to map permafrost regionally and cross-validate these predictions against an independent map from the nearby Casino mineral exploration project.
9. Production of a 30,000 km^2 permafrost cover and relative proportion of water ice map spanning the Canadian part of the Dawson Range without recent glaciation history.

2.4.1 Training and testing data

Model training relied on an 87 km^2 permafrost map prepared for the Newmont Coffee Gold Project (Tetra Tech EBA, 2016). This map classifies permafrost into four classes (unfrozen, ice-poor [no visible ice], ice-rich [some visible ice], very ice-rich [visible ice crystals, massive ice, ice lenses]) with data acquired between 2011 and 2016. The map authors leveraged Newmont's extensive mineral exploration data set, including records from hundreds of boreholes, thousands of soil samples, dozens of trenches, thermistor strings, and supplemented these data with additional test pits and hand probing observations. These combined datasets were used to create a 3-dimensional model of terrain overlaying field point observations, geophysical observations, and 1:20 000 scale aerial orthophotos using the PurVIEW ArcGIS extension (I.S.M. International Systemap Corp, Vancouver, Canada), enabling accurate stereoscopic image interpretation and digitizing of overlaying polygons. The resulting synthesis produced a highly accurate and high-resolution permafrost that is well suited to ML approaches.

We opted to use this interpreted product instead of the point observations because its comprehensive coverage of all terrain possibilities within our training region is much more suitable for training data-intensive ML models than point observations alone. ML models do not perform well on data values outside of their training set (Beery et al., 2018; Gray et al., 2024) and require thousands of data points for optimal performance and model stability (A. Ramezan et al., 2019; Figueroa et al., 2012). Using the thermistor observations, test pits, and trenches (the only completely reliable observations) would preclude model training on the full range of predictor variable values and result in a poorly performing, high-uncertainty, and poorly reproducible model.

2.4.2 Predictor variables

Predictor variables used in our ML classification approach consisted of geomorphological and environmental variables potentially influencing permafrost distribution and ground ice content directly or indirectly and for which we could acquire or derive high-resolution rasters: slope, aspect, elevation, tangential and profile curvature, vegetation/land cover, and direct + indirect solar radiation (see Table 2.1).

Table 2.1: Selected predictor variables for model training.

Variable	Source/derived from	Derivation/acquisition method
Elevation	Canadian Digital Elevation Model (Natural Resources Canada, 2013)	Orthophotos
Slope	Digital elevation model	*terra::slope* function (Hijmans, 2023)
Aspect	Digital elevation model	*terra::terrain* function (Hijmans, 2023)
Tangential curvature	Digital elevation model	*spatialEco::curvature* function (Evans et al., 2024)
Profile curvature	Digital elevation model	*spatialEco::curvature* function (Evans et al., 2024)
Direct solar radiation	Digital elevation model	Area Solar Radiation tool (ESRI, 2023)
Indirect solar radiation	Digital elevation model	Area Solar Radiation tool (ESRI, 2023)
Vegetation/land cover	Land Cover map of Canada (Natural Resources Canada, 2022)	Spectral imagery, up sampled from 30 m resolution to match other variables (~16 m)

The Canadian Digital Elevation Model (CDEM), used for the derivation of most other variables, was created from aerial orthophotography and features a horizontal resolution of ~16 m and a published vertical accuracy of 15-30 meters across much of the area we modelled. Stitching artefacts are present at orthophoto stitching boundaries and are expected to result in erroneous modelling of permafrost along these boundaries when extrapolating the predictions beyond our training area. Importantly, no stitching boundaries are present over our model training region, and the relation from one elevation pixel to another *within* each orthophoto is highly representative of the lay of the land, limiting major concerns about vertical accuracy to stitching boundary artefacts.

2.4.3 Training and optimization of the ML framework across the Coffee Gold Project region

Machine learning training and optimization was performed using the function *spatPredict* from the R package *SAiVE* developed in part for this study (de Laplante, 2024), with a graphical representation of the function's workflow in Figure 2.11 of the SI. First, we generated a training dataset of point values from observations provided as mapped polygons within the Coffee Gold Project permafrost map, using a stratified random sampling approach to generate points with an even representation across permafrost classes to improve classification performance on classes with low spatial representation (A. Ramezan et al., 2019; Shetty et al., 2021). We then extracted the intersecting value of rasterized predictor variables with our sample points to create a tabular representation of outcome and predictor variables.

To balance model training time with iteration across multiple methods and parameters while ensuring accurate, cross-comparable results across methods, we generated 10,000 sample points per permafrost class during our method selection step based on best practices from previous research comparing ML algorithms(Perlich et al., 2003; Ramezan et al., 2021). We then partitioned these points into spatially distinct training and testing sub-sets. Given the spatial autocorrelation of the data and the need for independent training and testing data sets (Bahn & McGill, 2013; Wenger & Olden, 2012), we used spatial blocking to split the training region into rectangular polygons and assigned 70% of polygons at random to the 'training' group and the remainder to the 'testing' group in a method similar to that used in the *blockCV* R package (Valavi et al., 2019). Sample points were assigned to either group based on their spatial intersection with either polygon class.

To identify the best performing machine learning method for classification of permafrost cover, the *spatPredict* function tested which combination of methods and parameterization performed best on the training data. Knowing that random forest (RF) and similar classification methods were likely to perform well on spatial classification tasks (as demonstrated by Furuya et al. 2020; Furuya et al. 2020;

Bhuiyan et al. 2022; Thaler et al. 2023 among others), we identified several RF and multivariate adaptive regression splines (MARS) algorithms suitable for our classification (see SI section 2.9.4). We selected 10 of these identified methods for testing to limit model training and optimization times, noting that this may not capture the best possible algorithm. These 10 methods were passed to the *spatPredict* function along with the training sub-set of randomly generated points and generic training control parameters (see SI section 2.9.5). Automatic optimization of each model's hyperparameters (parameters which control the learning process of the model) within *spatPredict* is performed internally on a random training/testing split of the training data. Finally, the retained, spatially distinct testing sub-set was used to determine the best performing algorithm using balanced accuracy. This metric, the average of the model's accuracy [sensitivity (true positive rate) + specificity (true negative rate) /2] across classes, is the most unbiased metric for assessing model performance when class imbalances are high (Brodersen et al., 2010; Carrillo et al., 2014) and provides a realistic accuracy metric where simple accuracy (lumped class performance) would be overly optimistic or yield biased results.

Using the best performing model obtained from this initial set of 40,000 total observations we optimized the size of the training set to maximize balanced accuracy, determining the ideal size of the training set by fitting a curve to the balanced accuracy of each iteration to identify a peak (as demonstrated by Perry and Dickson (2018)). We used the best performing model (i.e., top algorithm and optimized training set size) to quantify the accuracy of the method within the Coffee Gold Project region before applying the approach across the broader Dawson Range.

2.4.4 Regional validation and application to the broader Dawson Range

We quantified the model's performance at the regional scale by testing the model predictions against an independent permafrost created for the Casino Mining Project (Western Copper and Gold, 2024), located ~35 km south-east of the training region (see Figure 2.1Figure 2.5). The methodology used in creating this map differs from that of the Tetra Tech map as it classifies permafrost as present/absent vs the four classes of ice richness used in the Coffee Creek region training map Most importantly, it labels regions of discontinuous permafrost as entirely frozen; the map we use to derive training data breaks these regions into distinct frozen/unfrozen polygons while our model represents these regions as a mix of frozen/unfrozen pixels. Despite these uncertainties and differences, the Casino map is the only other independent permafrost distribution testing set available in the region. To compare our predictions to this data set, we reclassified our model's four classes into presence/absence classes by assigning ice-poor, ice-rich, and very ice rich classes to the 'frozen' class and keeping the original 'unfrozen' class. We intersected these two classes with the Casino map polygons and report the % accuracy from this comparison under these caveats.

Knowing that the entire unglaciated, south-western portion of Yukon has similar geomorphology, climate, vegetation, and permafrost controls, we applied our best classification model to a ~30,000 km^2 area roughly bound by the Yukon-Alaska border to the east, the Ruby Range mountains to the south, the Tintina Trench and Ogilvie Mountains to the north, and the Klondike Highway to the east. The selected region has consistent temperature, precipitation, and snowpack conditions as reported by Environment and Climate Change Canada stations at the north, south-east, and south-west margins of our study area (Dawson, Carmacks, and Beaver Creek) – an essential distinction as our model is not trained on these parameters and thus cannot account for their variation. These bounds also delimit an area without significant history of recent glaciation and with similar geomorphology as the CCR, including regions at the margins of old (ca. 3 Ma) glaciation extents with no obvious distinctive features such as wetlands or lakes. To confirm the suitability of this extrapolation region for prediction using a model trained at the CCR, we calculated the multivariate environmental similarity of the two regions using the *dismo* R package (Hijmans et al., 2010) This allowed us to identify pixels with predictor values outside of the range of the training data, showing that dissimilar pixels were concentrated in the broad river valleys as well as in higher elevation regions to the south of our extrapolation region (Figure 2.2), as expected since these regions have elevations outside of the training set.

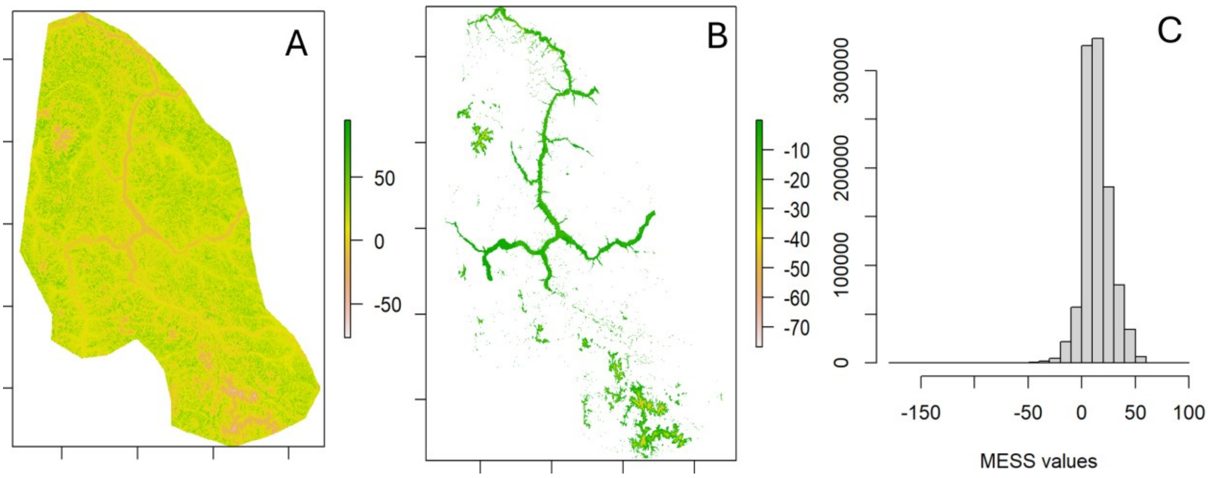

Figure 2.2: Multivariate environmental similarity values comparing the similarity of the Coffee Creek region with the broader extent of predictions. Figure A shows all calculated similarity values across the prediction region while Figure B shows only the negative (dissimilar) values. Figure C plots the distribution of these values.

2.5 Results

2.5.1 Method selection and parameterization

The best performing method was the *ranger* RF implementation (Wright & Ziegler, 2017). Comparison to other methods is provided in Section 2.9.4 of the SI. To identify the optimal training set size, we iteratively ran the *ranger* RF implementation at increasing training set sizes while using *spatPredict*'s default testing methodology to assess model performance. The balanced accuracy metric shows increasing performance as the training set size increases before plateauing (Figure 2.3). An optimal training set size was reached at approximately 80 000 random points per class, or a total of 320 000 points in total. This optimal training set size was used for final model training, performance evaluation, and regional-scale extrapolation to the Dawson Range.

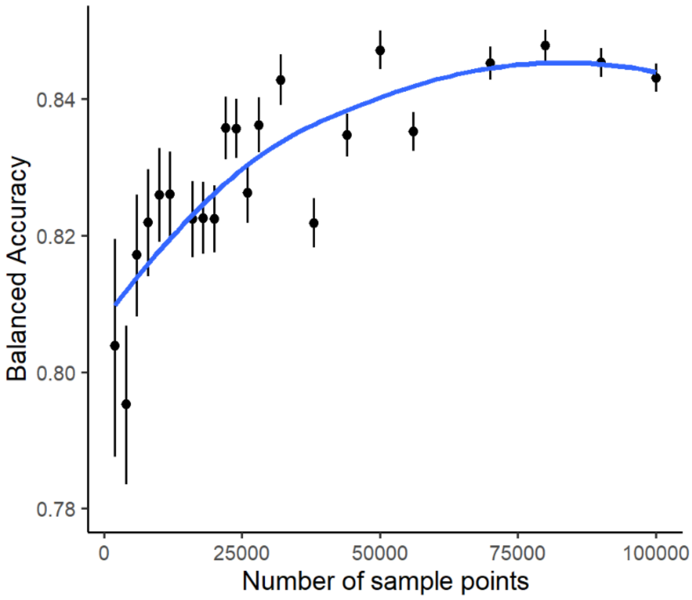

Figure 2.3: Ranger model balanced accuracy vs number of sample points per class. Bars represent the 95% CI of the balanced accuracy as output by function spatPredict.

2.5.2 Final model performance

Our best performing ML model successfully predicted permafrost distribution and ice content with a calculated overall (across all classes) balanced accuracy of 0.857 when tested against spatially partitioned data using the *spatPredict* function. Evaluation metrics, including sensitivity, specificity, positive predictive value, negative predictive value, and balanced accuracy (Table 2.2) indicate that evaluation metrics are highest in the very ice-rich and unfrozen permafrost classes, and slightly lower in the ice-poor and ice-rich classes (Table 2.3). This trend is reflected in classification test outcomes viewed as a confusion matrix, where very ice-rich ground was correctly classified at a rate of 0.98, decreasing to a rate of 0.78 for ice-rich ground.

Table 2.2: Model test statistics by class. Sensitivity measures the proportion of positives correctly identified by the model (true positive rate). Specificity measures the proportion of negatives correctly identified (true negative rate). Positive Predictive Value measures the proportion of positive results that are true positives, while Negative Predictive Value measures the proportion of results that are true negatives. Prevalence is the proportion of each class in the dataset. Detection Rate is the proportion of true positives in the entire dataset, and Detection Prevalence is the proportion of predicted positives. Balanced Accuracy is the average of sensitivity and specificity.

	Ice-poor	Ice-rich	Very Ice-rich	Unfrozen
Balanced accuracy	0.866	0.879	0.948	0.925
Sensitivity	0.780	0.840	0.902	0.896
Specificity	0.943	0.917	0.994	0.954
Positive Pred Value	0.819	0.778	0.976	0.876
Negative Pred Value	0.932	0.943	0.971	0.962
Prevalence	0.247	0.257	0.231	0.265
Detection Rate	0.195	0.216	0.208	0.238
Detection Prevalence	0.238	0.278	0.213	0.271

Table 2.3: Confusion matrix showing the tested performance of the final ranger *RF model trained on 80 000 sample points per class. Each cell in the matrix shows the proportion of sample points classified into each class by the model. The rows represent the actual (observed) classes, while the columns represent the predicted classes. For example, ice-poor ground was correctly identified at a rate of 0.82 and mis-classified as ice-rich at a rate of 0.12. A confusion matrix with the number of observations rather than percentages is provided in Figure 2.12 of the SI.*

	Ice-poor	Ice-rich	Very Ice-rich	Unfrozen
Ice-poor	0.82	0.12	0.002	0.06
Ice-rich	0.12	0.78	0.06	0.04
Very Ice-rich	0.001	0.009	0.98	0.01
Unfrozen	0.06	0.04	0.03	0.88

2.5.3 Regional model performance

Our classification model was additionally validated through reproduction of mapped permafrost extents (lateral spatial coverage) near the Casino Project (Figure 2.4), yielding a 73% agreement. Importantly, most disagreements between our model output and the Casino map were in regions where we predict patchy or discontinuous permafrost. The Casino permafrost map labels regions of discontinuous permafrost as contiguous 'frozen' zones, resulting in overrepresentation of permafrost coverage relative to our model (see Figure 2.4). This lowers the calculated accuracy of the comparison and based on the Casino labelling methodology and our field observations we anticipate our representation of discontinuous permafrost distribution to be more accurate than the Casino Project's map.

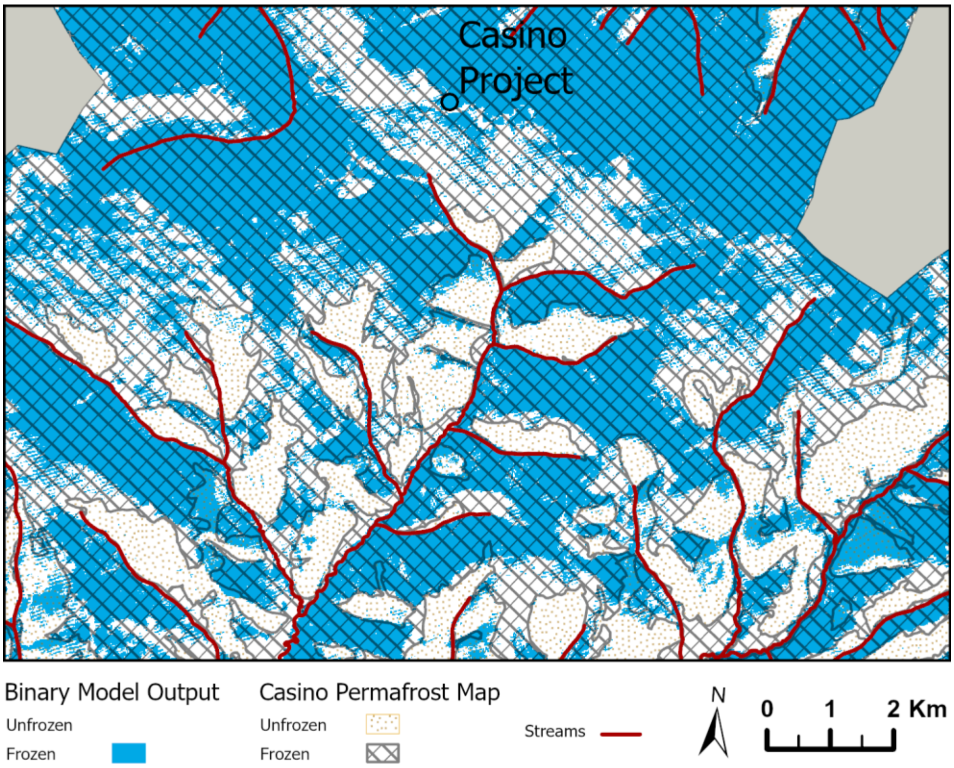

Figure 2.4: Subset of the Casino Project permafrost map overlain on our model output. The grey regions are not defined in the Casino map. The Casino map classifies discontinuous permafrost regions as 'frozen' while this simplified representation of our model labels each pixel as frozen/unfrozen. Blue/white mixed regions of our model output, indicating discontinuous permafrost, should therefore often agree with the Casino map's classification of these regions as 'frozen'.

We then applied the model across a broader area encompassing ~30,000 km^2 of the Dawson Range and surrounding regions (Figure 2.5), delineating our prediction extent based on assumptions discussed in our methodology.

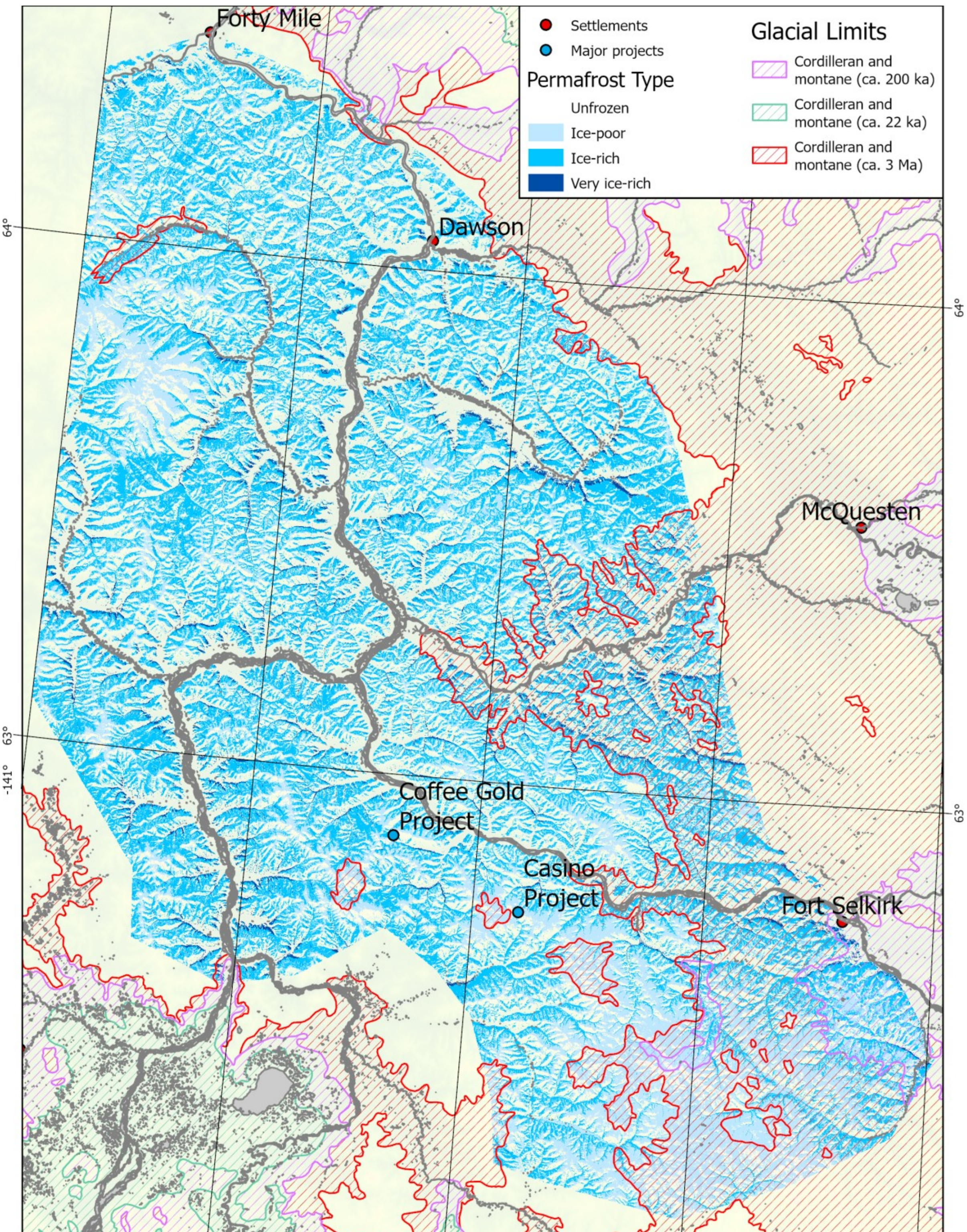

Figure 2.5: Extrapolation of permafrost model to the full extent of predictor variables and similar geomorphology and climate across the Dawson Range. Areas with an abundance of lakes and wetlands were excluded as these regions differ significantly in geomorphology. Glaciation margins to the east and south-east were included where geomorphology was subjectively similar to the training area, likely owing to the age (< 3Ma in most regions) and transience of last glaciation. Similar and otherwise suitable terrain in Alaska (>141◦ W) is excluded because the elevation model we used does not include it.

2.6 Discussion

2.6.1 Model performance

Although our model performance was validated by adequate reproduction of mapped permafrost cover outside of the training region (i.e., at the Casino mine project), quantification of model accuracy remains limited by the lower level of detail at the Casino site, and the absence of other high-resolution permafrost maps elsewhere in the Dawson Range. At the same time, the deterministic nature of our model makes it challenging to quantitatively compare against other, probabilistic models, including the Bonnaventure et al. (2012) Yukon permafrost model with a similar resolution. We also note that the best classification model we tested performed well at classifying permafrost by ice-richness class, but that this underestimates the ability of the model to make binary, presence/absence of permafrost predictions which may be more relevant to some end uses.

Despite these limitations to our model's performance validation, our testing data broken down by ice richness class provides useful insights into the limitations of our training dataset and points to additional predictors which could improve model accuracy. Every tested model exhibited high accuracy in distinguishing between the very ice rich and unfrozen ground classes, but performance decreased when classifying the intermediate ice rich and ice poor classes. Some of this inaccuracy is likely a reflection of the inaccuracy of the training data itself, with these two intermediate classes being more challenging to classify systematically using ground observations and orthophotos, but some of the remaining inaccuracy likely reflects the lack of predictors accounting for geology, soil particle size distribution or clay content, all of which are known to influence permafrost cover and ice content (e.g., Kanevskiy et al. (2013); Johnson et al. (2013); Fan et al. (2023)). While unfrozen and very ice-rich ground are well explained by solar radiation and topographic (for very ice-rich ground) parameters such as curvature and elevation, other intermediate classes of permafrost cover may additionally depend on soil characteristics such as particle size, which influences ice content (Fan et al., 2023; Lacelle et al., 2022). Improved geospatial mapping of soil would be needed to integrate those characteristics into our framework and improve classification of intermediate ice-richness classes.

2.6.2 Variable Importance

Partial dependence and variable importance plots can be used to explain associations between permafrost class and predictor variables by extracting and synthesizing information about our machine learning model's decision-making logic. Figure 2.6 shows the relative importance of each predictor variable for each permafrost class, with large differences in variable importance between ice richness classes. Solar radiation was most important to classifying ground as unfrozen or ice-poor with secondary importance to elevation, but little to no dependence on aspect or vegetation; in contrast, the ice-rich and very ice-rich categories showed more complex dependencies on elevation, curvature, and diffuse solar radiation.

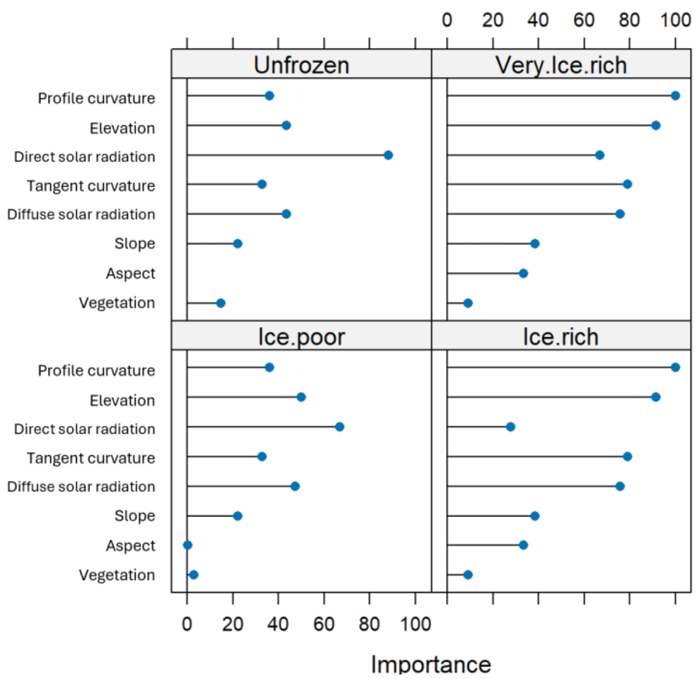

Figure 2.6: Relative importance of each variable for each permafrost category. Direct solar radiation is most important in predicting unfrozen and ice-poor ground, while profile curvature followed by elevation are most important in predicting very ice-rich and ice-rich ground. The 0-100 relative importance scale reflects each variable's relative importance to the global maximum, showing that profile curvature and elevation are the most important metrics to the model's performance overall.

Figure 2.7 further explains the effect of each variable. For example, greater direct solar radiation increases the likelihood of unfrozen ground whereas lower direct solar radiation increases the likelihood of ice-rich ground (Figure 2.6 A). Ice-poor ground, however, shows an association to high as well as low direct solar radiation. This appears counter-intuitive but can be explained by other variables: ice-poor ground predominates at high elevations (Figure 2.6 B) where slope is high (Figure 2.6 C) and where profile curvature is positive (Figure 2.6 G) despite relatively high solar radiation – in other words, the top of hills.

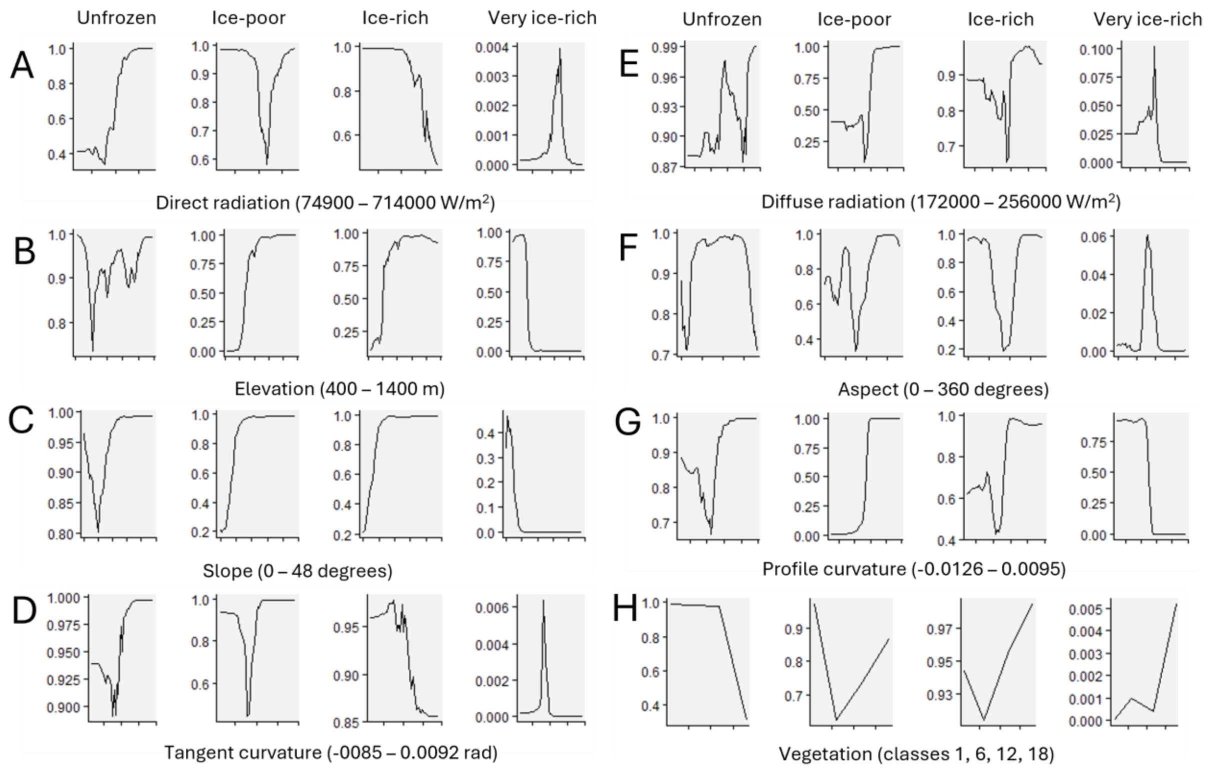

Figure 2.7: Partial dependence plots for the 8 predictor variables. Each plot represents the association between a predictive feature in isolation and a permafrost class. For example, the four top-left plots (A) show the association with direct solar radiation. As energy input to the ground increases (while other variables are kept constant), the probability of unfrozen ground increases; conversely for ice-rich ground. Note that y-axis values do not *all range from 0 to 1. Vegetation classes 1, 6, 12, 18 (H) correspond to "temperate or sub-polar needleleaf forest", "mixed forest", "sub-polar or polar grassland-lichen-moss", and "barren land", respectively.*

Profile curvature proved to be the most important variable for ice-rich and very ice-rich regions. Ice-rich regions are more likely in areas with positive (convex) curvature, while very ice-rich ground was only present on negative (concave) curvature regions (Figure 2.6 G). Regional groundwater flow dynamics (Tóth, 1963) help explain this relationship: very ice-rich regions are only present at the lowest elevations where steep slopes give way to flat valley bottoms and where groundwater elevations approach ground surface. The slope's convexity and flattening results in concentration of overland flows along with an abundance of sub-surface moisture, resulting in suitable conditions for the development of very ice-rich permafrost in cold, shaded valley bottoms. In contrast, ice-rich regions were more common at higher elevations because they overlie thicker unsaturated zones and are well-drained, leading to lower ground ice content. To our knowledge, curvature measures have not been used as predictor variables in other permafrost modelling efforts, but our findings suggest that its addition could be valuable.

Aspect was a relatively unimportant predictor for all classes relative to radiation and topographic variables likely because its effects are largely accounted for by direct solar radiation calculations (with the latter additionally incorporating terrain masking). However, when training a model without aspect, we noticed a significant decrease in accuracy suggesting that aspect plays a role beyond simply controlling solar energy input. Aspect can affect precipitation, wind loading of snow-covered slopes, and vegetation type and growth habits through the effect of prevailing winds and precipitation. We therefore posit that the additional importance of aspect lies in its control of such parameters not otherwise accounted for in our model.

Although vegetation and geobotanical indicators are commonly used to map permafrost distribution (Tetra Tech EBA 2016; Grünberg et al. 2020; Heijmans et al. 2022; Yang et al. 2024 among others), it was the least important predictor variable in our model (Figure 2.6) and did not show strong associations with any permafrost class (Figure 2.7). Two factors explain this observation: first, vegetation distribution in mountainous regions is highly influenced by the terrain characteristics already accounted for in our model (e.g., solar radiation, elevation, slope curvature). Second, the vegetation layer we used (Natural Resources Canada, 2022) is derived from multi-spectral satellite imagery and does not differentiate between the spectrally similar white spruce and black spruce (Richardson et al., 2003). These species are widespread in the region but occupy different niches, with black spruce dominant on frozen ground and white spruce dominant in well-drained, unfrozen soils, decreasing the power of vegetation as a permafrost indicator. For example, satellite imagery shows dark green vegetation dominated by black spruce on frozen ground in the lower portions of Figure 2.8A, while similar dark green regions in the middle portion of Figure 2.8B are dominated by white spruce on unfrozen ground yet both of these regions are uniformly labelled as 'needleleaf forest'. The presence of deciduous trees in other classes differentiates these regions spectrally from the spruce-dominated stands, with the predominant deciduous tree (poplar) reliably absent from permafrost. The small performance gains observed by using the vegetation layer (+~2 % balanced accuracy) therefore appear to be the result of its ability to make broad differentiations between needleleaf stands and deciduous/mixed stands underlain by unfrozen ground. Further permafrost mapping improvements would be possible with improved spectral techniques to distinguish morphologically similar species such as white and black spruce.

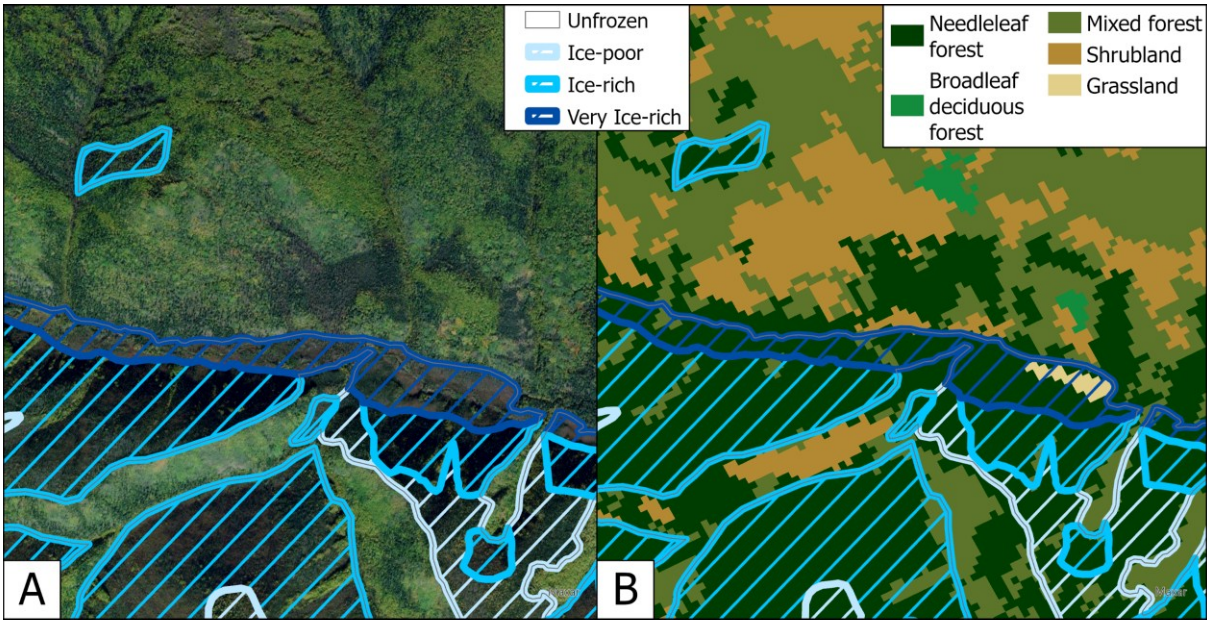

Figure 2.8: Satellite imagery of a hillslope on Coffee Creek (A) and classified vegetation output according to Natural Resources Canada (2022) with simplified class labels (B) on identical map extents. The Tetra Tech mapped permafrost classes overlay both figures for comparison. Minor vegetation classes (shrubland, grassland) not employed by the model are not represented in the partial dependence plots (Figure 2.7); conversely, 'barren land' is not present in this image but is shown in Figure 2.7.

2.6.3 Comparison to existing models

Our new regional permafrost model of the Dawson Range presents substantial improvements in detail and accuracy relative to previously available larger-scale maps. The 30-m resolution Bonnaventure et al. map classifies most of the Coffee Creek region as extensively discontinuous permafrost with some continuous permafrost on northern aspects, while our model identifies unfrozen ground along southern aspects consistent with field mapping and observations (Figure 2.9). The recent 1-km resolution global-scale permafrost zonation map of Obu et al (2019) similarly classifies the Dawson Range as containing mostly discontinuous permafrost but provides even less granularity due to the coarser resolution. In contrast, our map now provides much improved distribution of permafrost status and ground ice conditions at a 16 m scale, providing more detail, resolution, and accuracy relative to preexisting regional permafrost maps.

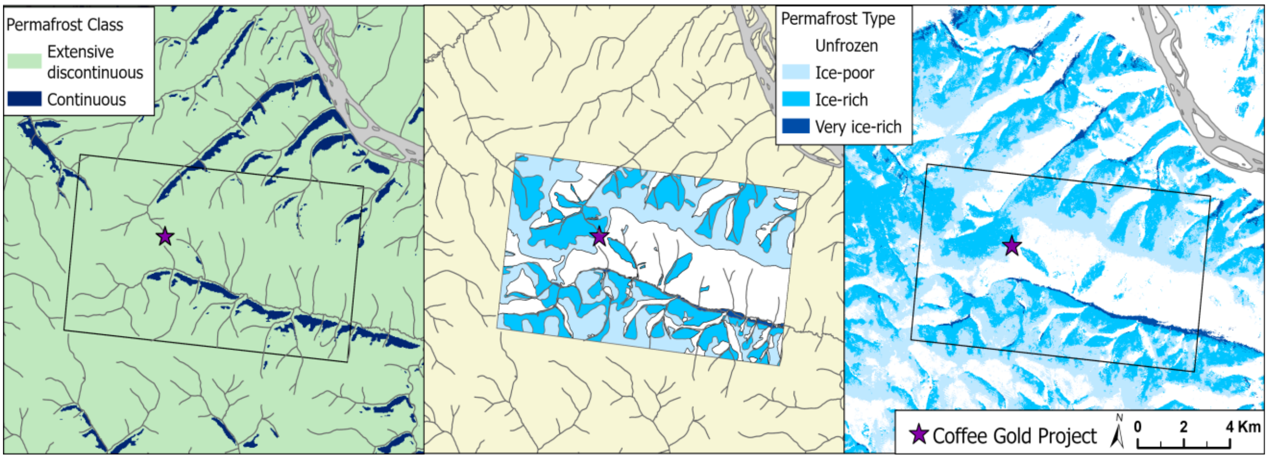

Figure 2.9: Comparison of the existing best permafrost map of Yukon (Bonnaventure et al. (2012), left), the Coffee Creek region mapped extent (Tetra Tech (2016), center), and our model (right). Note that the center and right maps share a legend while the left map has its own legend.

Our framework also provides a less-resource intensive and more flexible approach to permafrost mapping relative to a similar ML permafrost model of western Alaska (Thaler et al., 2023). In their approach, Thaler et al. predicted permafrost distribution at 3-meter resolution, testing three machine learning methods (extremely randomized trees, support vector machines, and ANNs) on LiDAR-derived elevation data, geophysical surveys supplemented by field observations, and high-resolution spectral imagery. Their approach marked significant advancements in spatial estimation, with classification accuracy similar to our classification (70%–90% vs. ~86). However, our approaches differ in several important ways. We opted to formalize our coding framework into an open-source package, designing the code to be adaptable to any number of predictor variables; future refinements could rapidly incorporate additional or more specialized predictors as well as geographically distant training data. We also limited our predictor variables to freely available or derivable variables, removing the requirement for field work beyond the creation of training data, and incorporated calculations of solar radiation which proved to be the key control of permafrost presence/absence in our region. We anticipate that the SAiVE coding framework will facilitate high-resolution permafrost modeling across permafrost affected regions by leveraging existing field data, and permit iteration as increasingly precise remote sensing products become available.

2.6.4 Model limitations

Our framework simplifies the extrapolation of the model to the regional scale with satisfactory results, but we caution that our model's performance is not tested beyond the Coffee Creek and Casino mining projects. We expect its accuracy to decrease with distance from the training area not only because climatic conditions may differ but also because our model's most important variable – solar radiation calculations – are centered on the Coffee Creek region's latitude due to a limitation of the

Area Solar Radiation tool we used (ESRI, 2023). Within our extrapolated area this results in an overestimation of solar energy input at northern latitudes and an underestimation to the south. A comparison of our model with satellite imagery and personal knowledge of permafrost distribution around Dawson (130 km north of our study area) indicates that permafrost is under-represented in the northernmost mapped extent. Specifically, some north facing slopes in the Klondike valley known to be extensively frozen are labelled with only sparse frozen pixels, and most of Dawson City's buildings are built on permafrost without corresponding representation on our map.

We also notably train our model on an interpreted continuous map rather than direct point observations. This approach is practical and beneficial for training a ML model because all predictor variable conditions across our training region can be used for training data, and we can generate a large dataset of random points to optimize model accuracy and stability. However, this approach inherently means that our model's accuracy is dependent upon the accuracy of the Tetra Tech map used for training. While this permafrost map was created using a comprehensive set of field surveys, remote sensing inputs, and extensive local knowledge, inaccuracies will inevitably be present and these inaccuracies will be propagated to our predictions.

While our model is extrapolated across the Dawson Range within an area of high similarity (Figure 2.5) we caution that our model's performance is not tested beyond the Coffee Creek and Casino mining projects. We expect its accuracy to decrease with distance from the training area not only because climatic conditions may differ but also because our model's most important variable – solar radiation calculations – are centered on the Coffee Creek region's latitude due to a limitation of the Area Solar Radiation tool we used (ESRI, 2023). Within our extrapolated area this results in an overestimation of solar energy input at northern latitudes and an underestimation to the south. A comparison of our model with satellite imagery and personal knowledge of permafrost distribution around Dawson (130 km north of our study area) indicates that permafrost is under-represented in the northernmost mapped extent. Specifically, some north facing slopes in the Klondike valley known to be extensively frozen are labelled with only sparse frozen pixels, and most of Dawson City's buildings are built on permafrost without corresponding representation on our map.

Additionally, we did not incorporate climatic and geological parameters into our model despite their importance to permafrost mapping, which limits the extrapolation accuracy in regions beyond the climatic conditions of the training area. Climatic parameters (snow depth and density, summer precipitation, temperature, and environmental lapse rate) were excluded for three reasons: 1) few ground observations exist in Yukon to create rasters of these parameters or to validate reanalysis products (e.g. ERA5, Muñoz-Sabater et al. 2021); 2) reanalysis product cell sizes are at best roughly equivalent to our training data set's spatial extent, limiting training to a single value per parameter; 3) our training data is limited to a small region, and as such does not permit training on a broad range of climatic parameters.

Geological parameters (bedrock geology and surficial geology) are thoroughly mapped over the extent of our training data set by Newmont prior property owners, but the surrounding area is mapped at much coarser and variable scales (1:15 000 to 1:250 000) or not at all (Figure 2.10), over many decades, and with different methodologies and classifications. A model trained on Newmont's local, high-resolution map units would fail to generalize to regional map units; also, a model trained on either maps would 'see' only a few geological units and fail to generalize to new units beyond the training region. One way to integrate geology and improve the model's performance would be to focus only on a narrow set of surficial geology parameters such as texture and clay content, re-classifying the existing surficial geology maps to a standard classification schema. As many areas do not have the required data for this classification, machine learning method selection would be limited to approaches which can work with missing data.

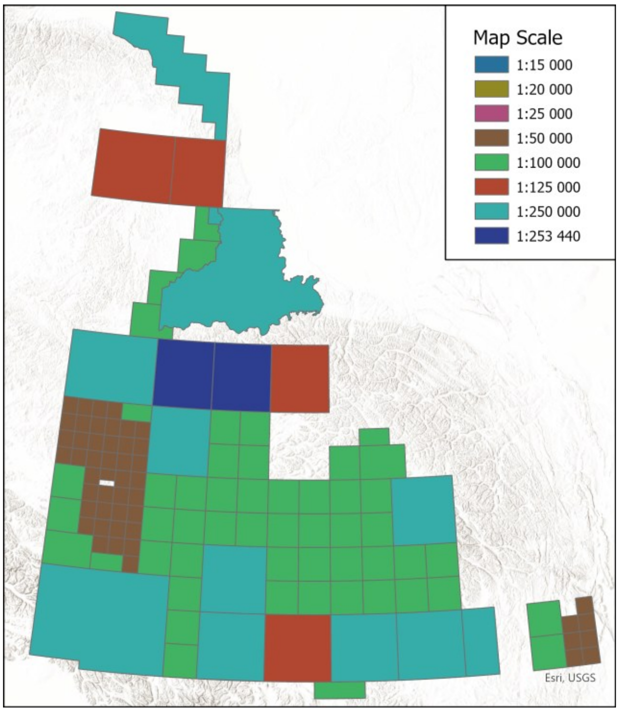

Figure 2.10: Mapped surficial geology extents in Yukon at various scales.

Improvement or expansion of our mapping effort will be possible with the addition of spatially distant permafrost observations for model training along with predictor variables of temperature, snow cover, precipitation, and surficial geology. Additionally, solar radiation would need to be calculated in multiple latitudinal tranches or using a different approach than ours, as the ESRI tool we used does not factor varying latitude in its calculations.

Integration of transient changes in permafrost coverage into our model, currently trained on permafrost status in 2016, will be necessary to ensure its applicability in the face of continuing permafrost degradation. If permafrost responded evenly to changes in thermal regimes, this could be addressed by incorporating modelled climate variables into a model's prediction step. However, permafrost does not degrade at even rates even when face with identical thermal condition, as soil's thermal inertia is heavily influenced by water ice content and advective or convective heat transfer via fluids or the varying insulating/conducting effects of vegetation can differ widely. Numerical (physics-based) models can model such changes explicitly, provided that evolving conditions are mapped or modelled at sufficient resolution and key ground properties (e.g., active layer thickness, ice content, permafrost thickness) are known or accurately modelled. By contrast, ML needs alternative strategies to handle the differential responses of soils to changing conditions. Two possible options could be envisioned to maintain the accuracy of high-resolution maps: (1) post-processing of ML outputs using physics-informed adjustments (e.g., conditioning on permafrost temperature, ice content, and advective transfer of energy) to alter model outputs using physical constraints; and (2) hybrid ML-numerical modeling, wherein ML provides high-resolution spatial predictions, while lower-resolution numerical simulations inform and constrain those predictions so that projected changes remain physically plausible under evolving climate and ground conditions.

2.7 Conclusion

Our study demonstrates the feasibility and utility of machine learning classification for mapping of permafrost distribution and characteristics (such as ice content) in permafrost affected regions. By combining field observations and remote-sensing data with an easy to use, scalable, and reproducible code framework, our approach could be used to advance the understanding of permafrost dynamics and their implications for ecosystems and communities affected by permafrost change, for example by closely linking stream geochemistry to permafrost or ecological niches at all trophic levels. Notably, our approach performed well despite not using variables typically thought of as important to permafrost distribution, such as snow cover, precipitation, and temperature. We expect that the addition of reliable and high-resolution datasets of these climatic variables could further improve the mapping accuracy, especially if this approach is applied to broad scales (10,000 km^2 or greater) or incorporates field observations from multiple regions.

We demonstrate that it is possible to quickly, inexpensively, and reproducibly map permafrost at large scales *and* at high resolution, with map accuracy limited only by suitable predictor variables and reliable training data. Our map has immediate applications for land use planning, ecological research, mining infrastructure development, maintenance, and decommissioning, and will be used to develop improved hydrological models for permafrost-affected regions in planned research by the Yukon Department of Environment and affiliated researchers. Non-governmental organizations such as conservation societies and First Nation governments will also benefit from these maps when developing land use agreements (a process which is currently underway in Yukon) and will be empowered to provide informed input into resource development applications.

By bridging the gap between local surveys and global models, our methodology empowers researchers and stakeholders to make informed decisions regarding infrastructure development, environmental management, and community resilience in Arctic and sub-Arctic regions. Moving forward, continued research and innovation in permafrost mapping are essential to address the challenges posed by climate change and ensure the sustainability of northern ecosystems.

2.8 Acknowledgments

This work was supported by the National Sciences and Engineering Research Council (NSERC) Discovery Grant and Northern Research Supplement (RGPIN-2019-05709) awarded to Clément Bataille. We also wish to thank Newmont Gold and Western Copper and Gold for their generosity as hosts during field work and willingness to share information and staff resources with us.

We acknowledge that this work pertains to the traditional territories of the White River, Selkirk, Na-Cho Nyak Dun, and Tr'ondëk Hwëch'in First Nations.

2.9 Supplementary Information

2.9.1 Software used and versioning

R software

We used the open-source R programming language versions 4.3.1 to 4.4.1 to pre-process variables, interface with the *spatPredict* function, and to analyze results using tabular and graphical means such as confusion matrices, variable importance plots, and partial dependence plots. All packages used and their dependencies were download from CRAN, the central repository of R packages, except for the SAiVE package which was developed and published to CRAN for this work.

ESRI software

We used ArcGIS Pro version 3.1 (ESRI, 2023) with a Spatial Analyst licence to calculate and pre-process rasters of aspect, slope, tangential and profile curvature from the orthophoto-derived Canadian Digital Elevation Model (Natural Resources Canada, 2013) to the best resolution available of ~16 m. The vegetation layer used was acquired as a 30 m raster layer, which we resampled using *R* to match the resolution and cell alignment of the other raster layers.

Inputs of solar radiation (diffuse, direct) were calculated for a whole year using the ESRI proprietary tool *Area Solar Radiation*, based on the work of Fu and Rich (1999 and 2002). This tool uses terrain inputs to calculate terrain masking and atmospheric diffusion at multiple times throughout each day for the period specified; solar radiation totals in Wh/m^2 are reported in our case for an entire non-leap calendar year.

2.9.2 Machine learning modelling workflow

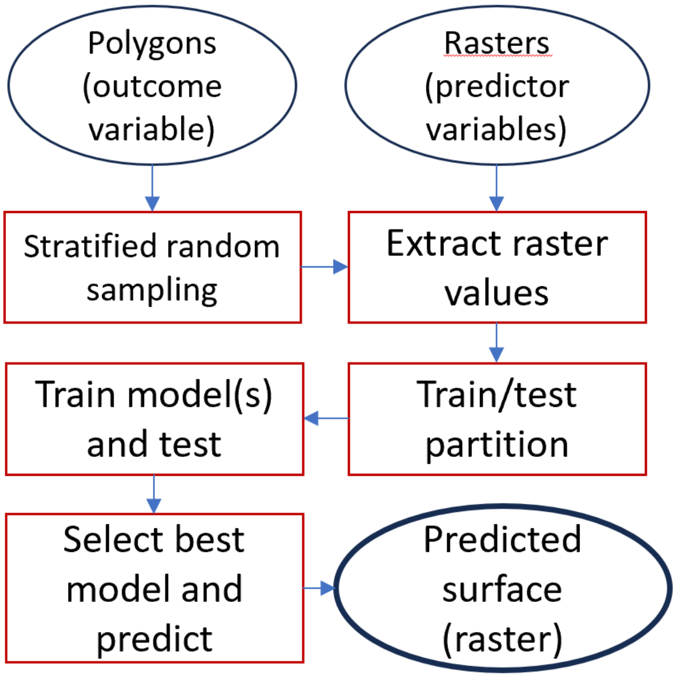

Figure 2.11: Simplified workflow of spatPredict function. The input "Polygons (outcome variables)" is the permafrost classification map used for model training, while the input "Rasters (predictor variables)" are the rasters used to predict permafrost class.

2.9.3 Model performance statistics

Calculation of statistics

The *spatPredict* function outputs several model performance statistics (sensitivity, specificity, balanced accuracy, and overall balanced accuracy) derived from the retained testing dataset which are presented in the paper's main text. The derivation of these statistics is performed as such:

$$sensitivity_i = \frac{true\ positives\ for\ class\ i}{total\ actual\ positives\ for\ class\ i}$$

$$specificity_i = \frac{true\ negatives\ for\ class\ i}{total\ actual\ negatives\ for\ class\ i}$$

N.B.: for multi-class problems, true negatives refer to instances correctly identified as not belonging to class i; total negatives are all instances that do not belong to class i.

$$balanced\ accuracy_i = \frac{sensitivity_i + specificity_i}{2}$$

$$overall\ balanced\ accuracy = \sum_{i=1}^{N} \frac{balanced\ accuracy_i}{N}$$

Confusion matrix

Table 2.4: Confusion matrix showing the tested performance of the final ranger *RF model trained on 80 000 sample points per class; testing was performed on ~30% of this data set, i.e. ~24 000 points per class. Each cell in the matrix shows the number of sample points classified into each class by the model. The rows represent the actual (observed) classes, while the columns represent the predicted classes.*

	Ice-poor	**Ice-rich**	**Very Ice-rich**	**Unfrozen**
Ice-poor	18266	2719	43	1263
Ice-rich	3255	20214	1465	1054
Very Ice-rich	29	181	19456	260
Unfrozen	1579	946	614	22245

2.9.4 Model/algorithm selection and performance

The 10 machine learning methods used for comparison testing are listed below in Table 2.5, and the resultant comparison metrics are shown in Figure 2.12. We were not able to run the two bagged multivariate adaptive regression splines methods (*bagged MARS* and *bagged MARS using gCV pruning)* to completion due to software compatibility issues between the R version we used (4.4.1) and the *earth* package.

Table 2.5: Methods selected for performance comparison.

Method	**Interfaced via R function/package**
***Ranger* RF implementation**	Function *ranger* from package *ranger* (Wright & Ziegler, 2017)
***Rborist* RF implementation**	Function *rf* from package *Rborist* (Seligman, 2024)
Conditional Inference Random Forests	Function *cforest* from package *party* (Hothorn et al., 2023)
Random Forest by Randomization	Function *extraTrees* from package *extraTrees* (Simm & de Abril, 2014)
Parallel Random Forests	Function *parRF* from package *BAGofT* (J. Zhang et al., 2022)
***Bagged MARS* (multivariate adaptive regression splines)**	Function *bagEarth* from package *earth* (Milborrow, 2024)
Bagged MARS using gCV Pruning	Function *bagEarthGCV* from package *earth* (Milborrow, 2024)
Random Forest Rule-Based Model	Function *rfRules* from package *randomForest* (Liaw & Wiener, 2022)
Regularized Random Forest	Function *RRF* from package *RRF* (Deng et al., 2022)

Global Regularized Random Forest	Function *RRFglobal* from package *RRF* (Deng et al., 2022)

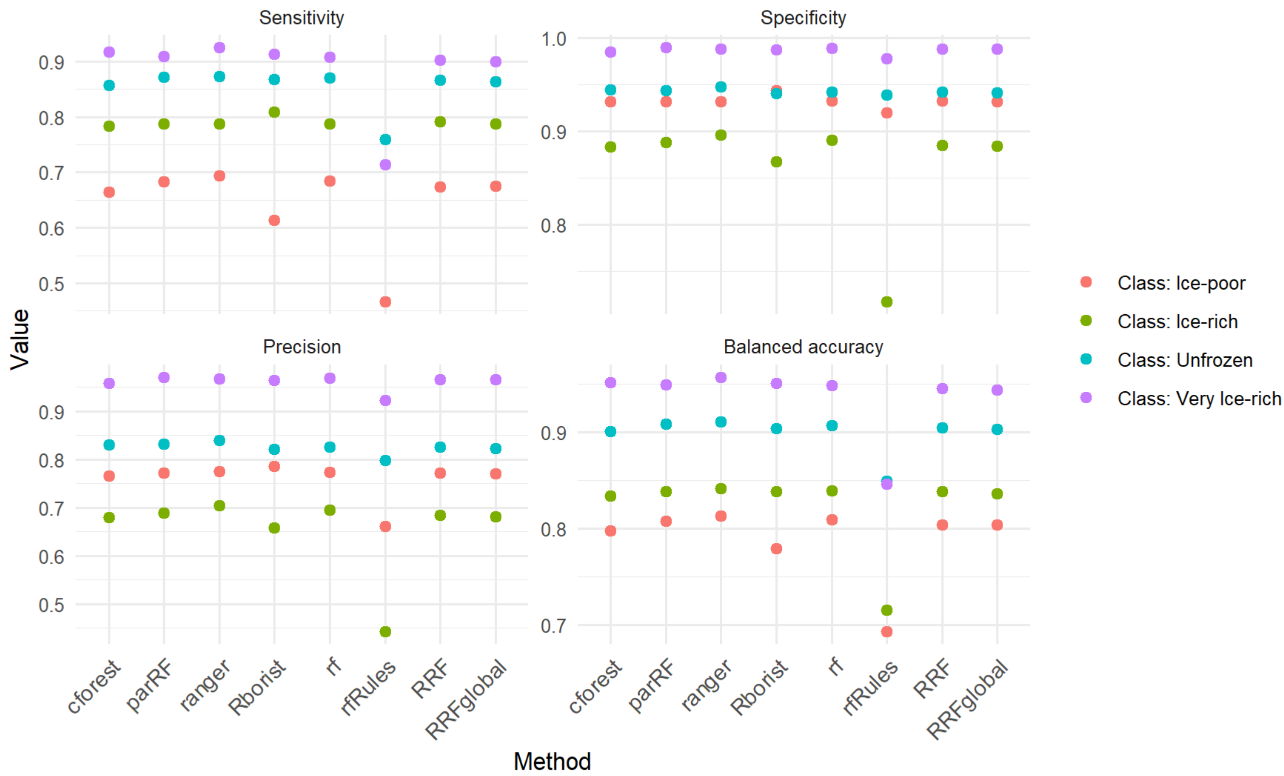

Figure 2.12: Performance metrics of machine learning methods compared in this study. The ranger method outperformed other methods on balanced accuracy (arithmetic mean of sensitivity and specificity).

2.9.5 R scripts

This section contains R scripts useful for re-creating our final model output, which also includes testing statistics as part of the final returned object. Another script is also included demonstrating our methodology for comparison of our model output to the published Casino Mining Project map.

Model parametrization and sample size selection

```
library(SAiVE)

# Find the 10 models most similar to ranger (11 because one of these is not able to run)
models <- modelMatch("ranger", similarity = 0.7)
# Order the result, select top 10
models <- models[order(models$`Similarity to Random Forest (ranger)`, decreasing = TRUE),
][1:11,]

# Resample vegetation raster to the same resolution and extent as other layers
vegetation <- terra::rast("path/CAN_NALCMS_landcover_2020_30m.tif")
DEM <- terra::rast(“path/DEM.tif”)
```

```
terra::crs(vegetation) <- DEM #Assign the same crs as other layers

vegetation <- terra::resample(vegetation, Elevation, method = "near") # resample to the same
resolution and extent as other layers
# terra::writeRast(vegetation, "path/vegetation_resampled.tif", overwrite = TRUE)
# vegetation <- terra::rast("path/vegetation_resampled.tif")

# Load the predictor and outcome variables, make trainControl
features <- c(aspect = terra::rast("path/Aspect.tif"),
        DEM = DEM,
        difrad = terra::rast("path/DiffuseRad.tif")
        dirrad = terra::rast("path/DirectRad.tif"),
        profcurv = terra::rast("path/ProfileCurv.tif"),
        tancurv = terra::rast("path/TangentCurv.tif.tif"),
        slope = terra::rast("path/SlopeDeg.tif"),
        veg = vegetation)
outcome <- terra::vect("path//Tetra_pmfst.shp")[,2]
outcome$Type <- as.factor(outcome$Type)

trainControl <- caret::trainControl(
  method = "repeatedcv",
  number = 10, # 10-fold Cross-validation
  repeats = 10, # repeated ten times
  verboseIter = FALSE ,
  returnResamp = "final",
  savePredictions = "all",
  allowParallel = TRUE)

# Run spatPredict using the 10 models to find the best one
res <- spatPredict(features,
          outcome,
          10000, # 10 000 sample points just to find the best performing method
          trainControl = trainControl, # Same trainControl for all models (otherwise matched
one to one)
          methods = models$`Model Abbreviation`,
          predict = FALSE, # Don't bother predicting across the whole raster for now
          fastCompare = FALSE, # We don't want to reduce the training set size further
          thinFeatures = FALSE # Don't attempt to trim features with VSURF (already trimmed,
and just tries to leave everything in anyways)
          )
# Accuracy vs sample size
test <- list()
samp <- c(thousand = 1000,
     twothousand = 2000,
     fourthousand = 4000,
    sixthousand = 6000,
    eightthousand = 8000,
    ... and so on)
for (i in 1:length(samp)) {
  test[[names(samp[i])]] <- spatPredict(features,
                   outcome,
```

```
                    samp[i],
                    trainControl = trainControl,
                    methods = 'ranger', # Only use best performing model from previous step
                    thinFeatures = FALSE)
}

# Save the output in case anything goes wrong... the previous step is very long
save(test, file = "path.Rdata")

# Reload
# load("path.Rdata")

df <- data.frame(points = samp,
               accuracy = c(test$thousand$selected_model_performance$overall[1],
               test$twothousand$selected_model_performance$overall[1],
               test$fourthousand$selected_model_performance$overall[1],
               test$sixthousand$selected_model_performance$overall[1],
               test$eightthousand$selected_model_performance$overall[1],
               ... and so on
                )

# Plot the result
ggplot2::ggplot(df, ggplot2::aes(x = points, y = accuracy)) +
 ggplot2::geom_point() +
 ggplot2::labs() +
 ggplot2::geom_smooth(method = "loess")
```

Comparison of model output to Casino Mining Project map

```
# This script was used to compare the model output to the Casino Mining Project's map. To
replicate, replace the paths in the script with the appropriate paths on your system.

casino <- terra::vect("path_to_casino_map.shp")

# Turn the Permafrost column into a binary column
# column pmfts_int gets 1 if column Permafrost == 'yes' else 0
casino$pmfst_int <- ifelse(casino$Permafrost == 'yes', 1, 0)

# Write the vector to disk for later use
# terra::writeVector(casino, "path_to_write_reclassed_vector", overwrite = TRUE)

# Load the predicted surface
model <- terra::rast("path_to_predicted_surface")

# Make the casino vector file into a raster of same resolution as 'model'
casino <- terra::project(casino, model)
casino_rast <- terra::rasterize(casino, model, field = "pmfst_int")
# terra::writeRaster(casino_rast, "path_to_write_raster", overwrite = TRUE)

tbl <- data.frame(id = c(1:4), c(1, 1, 0, 1))
```

```
model_bin <- terra::classify(model, tbl)
# terra::writeRaster(model_bin, "path_to_write_reclassed_model", overwrite = TRUE)


# Determine the agreement between the model and the casino_rast data
agreement <- model_bin == casino_rast
# Find the percentage of cells that agree
agreement_perc <- sum(values(agreement), na.rm = TRUE) / ncell(agreement) * 100

# Print the agreement percentage
agreement_perc
```

Calculation of environmental similarity

Chapter 3: Discussion and conclusion

3.1 Summary

A random-forest-based classification of permafrost distribution was successfully developed and applied to an area roughly 350 times greater than the training region, with measured accuracy ranging from 0.86 (cross-validation within the training region) to 0.73 (on an independent validation region outside of the training region). The resulting map provides predictions of permafrost presence/absence as well as relative ice richness to a resolution of 16 meters. This performance was achieved by training a model on an existing detailed field-survey map and by leveraging primarily high-resolution DEM-derived predictors which are broadly available across the world. Notably, predictions were not dependent on climatic variables and no physical process calculations were required. The high accuracy of the resulting model demonstrates that ML-approaches can be successfully applied to map permafrost distribution at high-resolution across large areas as far as some local survey data are available. The use of open-access remote sensing and DEM-derived geospatial dataset also makes this framework transferable to other regions with the possibility to develop regional scale maps of permafrost distribution within topographically and climatically similar areas. Expanding permafrost predictions at larger scale would require inclusion of other high-resolution climate predictor variables and the incorporation of training data from a range of climatic and environmental conditions.

The coding framework developed through this effort as an R-language software package is designed to facilitate future model improvements (via additional predictor variables or permafrost observations) and permit the rapid creation of predictive models from a given training region to as far as suitable predictor variables and sufficient field-based permafrost observations are available. Expansion of the presented model domain, while more challenging, is also supported by this framework, requiring the integration of suitable climatic variables and field observations at new locations to capture the regional effect of climate in addition to geomorphological controls. The approach’s flexibility and adaptability facilitate leveraging various algorithms and parametrization methods to identify the best performing method, parametrized to obtain the most robust model performance in diverse environmental settings.

3.2 Potential applications

The regional permafrost map developed in this study is immediately applicable to research questions in hydrology or the development of infrastructure in the Yukon's Dawson Range region. Mineral exploration companies, local First Nation governments, NGOs and regulators will benefit from having access to a more accurate and higher resolution permafrost prediction map than the previously available products to have more informed discussions about the feasibility, environmental impact, and cost of projects. Similarly, researchers focusing on the Dawson Range region could leverage this map to advance our understanding of the hydrology, contaminant fluxes, fire behavior, or ecosystem impacts of permafrost cover.

Our modelling framework also has broader but less immediate applicability. With suitable training data from field observations and the availability of open-access geospatial and remote sensing variables, it is possible to use this framework and the co-developed R package to train and test permafrost models in *any* region and using any available predictor variables. However, caution should be exercised when extrapolating these ML-based models to regions with different geomorphology or where climatic variables differ from the training region, as ML algorithms are notoriously poor at extrapolating beyond the predictor range used for model training (see Ben-David et al., 2010 or Sugiyama et al., 2007 for theoretical reasons, or Duan et al., 2024, Elith et al., 2010 Feng et al., 2019 for practical examples). Additionally, as ice formation or degradation is a slow process due to the large enthalpy of changes in state, using old training data may not capture the differential rate of decay (or growth) in permafrost with varying ice content into the future and would require the use of mechanistic models.

While ML-based permafrost classifications are useful to provide cost-effective and practical predictions of permafrost extent over large scales, they certainly do not replace mechanistic modeling but could complement the advance of process-based approaches. As seen in this study, ML-based approaches provide a degree of interpretability of the predictors controlling permafrost extent through importance and partial dependence plots. As such, ML-based models could serve to identify, test, and parameterize poorly represented processes in physics-based models.

3.2.1 Map and model framework availability

The modelling framework is published as an R package to the CRAN repository (de Laplante, 2024), and code development is viewable on GitHub (https://github.com/UO-SAiVE/SAiVE) with contributions encouraged. Map files can be requested from the author, and will ultimately be published to the Yukon Geological Survey's permafrost database website (https://service.yukon.ca/permafrost/index.html).

3.3 Model limitations and future directions

Besides the known extrapolation issues of ML-based modeling that limit the accuracy of predictions beyond the range present in training data, there are several other mechanisms not incorporated into the modeling that limit applicability to broader region. In contrast to other permafrost modelling efforts, this map did not include any climate variables as predictors. Climate, especially mean annual air temperature and snow cover, are established drivers of permafrost distribution but these variables were not included into the modeling framework because of the lack of available high-resolution data. The highest resolution climate reanalysis product, ERA-5 Land (Muñoz-Sabater et al., 2021) has a cell size of 81 km^2 which is the same order of magnitude as the entire training area 87 km^2. As a result, there would not be sufficient variability in climatic conditions over the training area to properly incorporate climate as a predictor. Satellite data from Sentinel and Landsat missions (e.g., thermal infrared imager and operational land imager) exist at high spatial (~30 m) and temporal resolution (~days) and could be leveraged along with topography to develop Earth-Observation driven permafrost time-series incorporating temperature, vegetation and land cover predictors. This idea is currently being pursued by the European Space Agency Permafrost Climate Change Initiative which has generated novel 1 km resolution monthly time series of permafrost cover across the Arctic during the timeframe of our study (Westermann, Barboux, Bartsch, Delaloye, Grosse, et al., 2024) and is planning a release of permafrost cover at 10 m resolution (Westermann, Barboux, Bartsch, Delaloye, Hugelius, et al., 2024)

Another potential issue that limits the applicability of my work to broader regions is the use of solar radiation products which are latitude dependent. The tool used to calculate solar radiation (Area Solar Radiation, ESRI 2023) uses a single latitudinal input to calculate the sun's path. In practice, this means that only locations along the tool's latitude input have precise solar radiation calculation. While the error is small for short distances (i.e. within a few tenths of a latitude degree), it becomes more problematic when using these calculated surfaces for predictions at the regional scale with error growing when moving away from the training area used to calculate solar radiation inputs, with regions to the north receiving an overestimation of solar energy input and vice-versa. To solve this issue, it would be important to incorporate solar radiation variables which are accurate across the entire extrapolation area.

Despite the lack of climate variables and the limitations of the solar radiation tool used, the model's performance was good when tested on an independent map (the Casino Project map). However, further away from the training area, the model was not tested and appears to under-predict permafrost coverage at its northernmost extent when compared to surficial geology maps factoring permafrost cover near Dawson (McKenna & Lipovsky, 2014). This poorer performance is not unexpected and could be explained by either of the limitations discussed above. First, MAAT is likely lower at Dawson than at the Coffee Gold project: the nearest station to the south (Carmacks) has a reported MAAT of -2.1 ^{0}C vs Dawson's -3.8 ^{0}C (ECCC 2024). Second, solar radiation was calculated with a fixed latitudinal input at the Coffee Gold project away from Dawson, which resulted in an overestimation of solar energy received in Dawson. The combination of these two factors should result in the model's underestimation of permafrost cover at its northern extent and an overestimation at its southern extent. As a result, we caution users that the permafrost distribution map generated in this work for the broader Dawson Range (Figure 2.5) should be used with caution in areas distant from the Coffee Creek training area.

3.4 Potential improvements

Improvements to the current methodology fall under two broad categories: those aimed at improving the model's performance locally and those aimed at broadening the model's geographic scope. Improving accuracy would require working with more accurate training data to avoid biases or inaccuracies related to field classification methodology, the incorporation of additional, more refined, and/or higher-resolution predictors to resolve remaining classification inaccuracies, and improvements of the ML algorithms and parametrization used to train the model. Broadening the geographical scope of our model (and coincidentally improving performance at the margins of our mapped extent) will require the incorporation of training data from a broader spatial area along with the incorporation of large-scale climate variables such as mean annual temperature (at minimum), snow characteristics, precipitation, and geology that are known to influence permafrost cover at the continental scale.

3.4.1 Refining geospatial predictors

The limitations of the Area Solar Radiation tool, used to calculate solar radiation across the modelled portion of the Dawson Range could be resolved by the recent release of a new tool, Raster Solar Radiation, in ArcGIS Pro 3.4 (ESRI, 2024). This tool reduces the distortion in calculated solar flux at Earth's surface facilitating its use at broad spatial scale. As the DEM-derived surfaces are central to the accuracy of the classification of permafrost distribution, it is critical to ensure that those DEM-derived predictors have the maximum precision and accuracy. The Canadian Elevation Model (Natural Resources Canada 2013) has a vertical accuracy of 15-30 meters or worse within the study area (Figure 3.1) and relies on an aging dataset (Figure 3.2), which may become problematic in topographically-dynamic river/stream valleys. In addition, the CDEM features numerous stitching artefacts at orthophoto boundaries, resulting in erroneous calculations of derived parameters such as slope, aspect, or solar radiation.

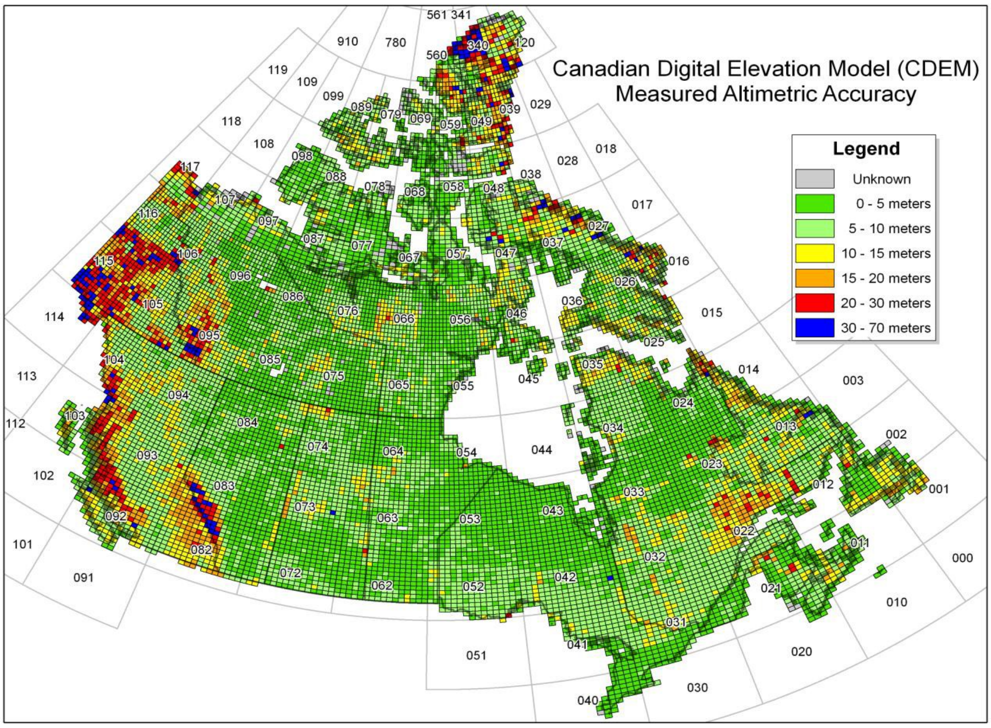

Figure 3.1: Canadian Digital Elevation Model measured altimetric accuracy, reproduced from the CDEM's product specifications (Natural Resources Canada, 2013). Most of the permafrost model's domain shows an accuracy of 15-30 meters or worse.

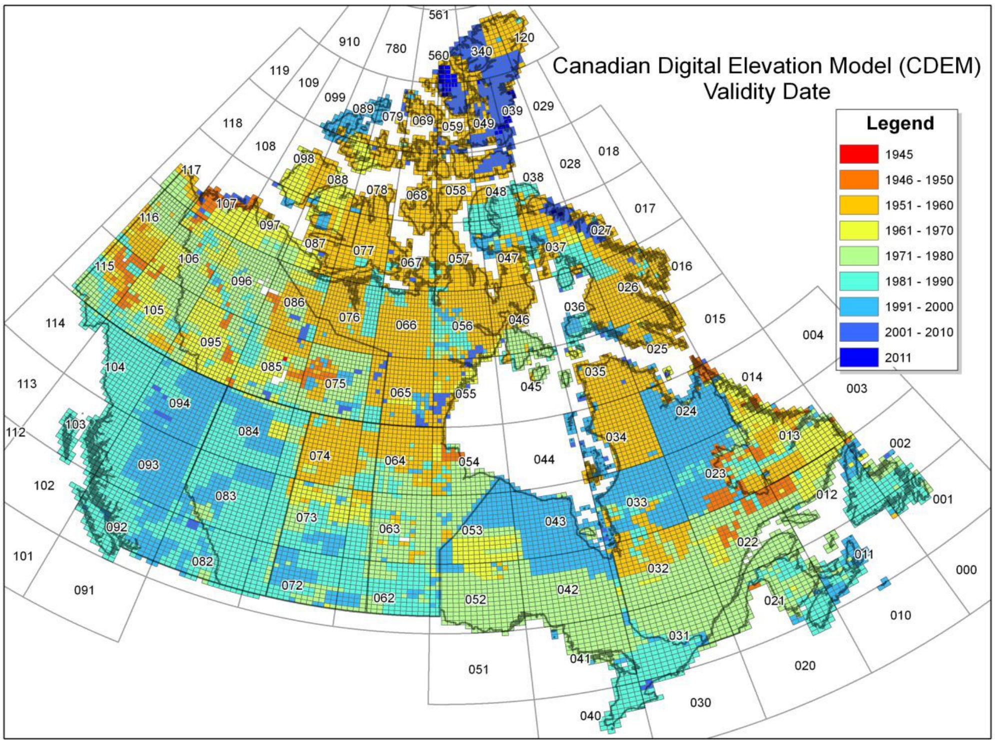

Figure 3.2: Canadian Digital Elevation Model data validity date, reproduced from the CDEM's product specifications (Natural Resources Canada, 2013). The data used in this model's training and extrapolation largely dates to the 1946-1960 range.

The ArcticDEM (Porter et al., 2023) is a newer and higher resolution product (2 meters vs 16 meters) being developed for the Arctic and sub-Arctic, promising a vertical accuracy of 4 meters. Unfortunately, this DEM is still incomplete with multiple sections of missing data throughout our model's domain and cannot be used directly. However, as those data gaps are progressively filled, the ArcticDEM should become the basis for modeling permafrost distribution across the Arctic and sub-Arctic.

Improving the performance of the model might also require use of an improved vegetation/land cover layer. In this study, we leveraged the Land Cover Map of Canada, (NRCAN 2022) which classifies landcover in broad classes. Instead of using this landcover product which incorporate its own classification inaccuracies, it may be more appropriate to use the continuous remote sensing satellite data used to generate this landcover product such as LANDSAT imagery (U.S. Geological Survey) or the Normalized Difference Vegetation Index derived from these products (as used by Pastick et al. 2015 in permafrost modelling in Alaska). As discussed in Chapter 2, the Land Cover Map of Canada does not distinguish between white spruce and black spruce, which are very similar in appearance. As black spruce is more likely to be found on frozen soil than white spruce, this limits the usefulness of this layer to the permafrost model. However, the spectral signatures of both spruce species do appear to be clearly differentiable in laboratory conditions (Richardson et al., 2003), which may allow for differentiation in raw (unclassified) satellite spectral imagery. At the same time, direct use of spectral imagery may help in differentiating between treeless frozen, ice-rich regions covered by moss and accumulated organic matter and more ice-poor regions covered by lichens and rock/soil.

3.4.2 Additional predictor variables

Another approach to improve the permafrost classification accuracy and precision would require the addition of new predictors which are not accounted for in the current model. Additional predictor variables can be seamlessly incorporated into the modelling framework presented in Chapter 2 if they are a gridded product with a resolution matching the existing predictors. Possible new variables include remote sensing spectral products, surficial geology, and various climate variables acquired from reanalysis (e.g. ERA-5 Land) or forecast products. Many remote sensing products are available at cell resolutions approaching that used in this present permafrost model, but climate variables have much larger cell sizes than the DEM-derived variables, limiting their usefulness unless they can be properly resampled; in addition, temperature predictions from reanalysis products tend to be overly warm (Batrak & Müller, 2019) and may require adjustment using ground observations.

A complicating factor to using these reanalysis products is that climate models in Canada's north incorporate sparse ground and atmospheric observations and are consequently prone to inaccuracies in modelled temperature and precipitation (Loeb et al., 2022; Pernov et al., 2024; Tao et al., 2023). The performance of reanalysis product should be verified against field observations and, if possible, adjusted/calibrated to local conditions/observations. For example, models of snow water equivalent, depth, or density such as CaLDAS (Environment and Climate Change Canada, n.d.) or ERA-5 Land (Muñoz-Sabater et al., 2021) could be compared against a network of snow survey locations in Canada and Alaska (58 in Yukon, 63 in the Northwest Territories, and 100 + in Alaska) which could be used to validate and calibrate snow-related variables. However, this task is time consuming as many local observational data sets (especially precipitation and temperature observations) are not widely available or may be difficult to access from government or industry.

3.4.3 Improving model selection and parametrization

In addition to these structural improvements to the model it bears mention that some easy but time-consuming improvements are possible to the methodology by which I selected the best performing machine learning method. Most importantly, choosing from the 10 methods most similar to the *ranger* RF implementation, which limits the tested methods to RF and MARS variations, may exclude better algorithms. More extensive testing of machine learning methods is warranted, including convolutional neural networks, support vector machines, or gradient boosted trees. I also selected the best performing method with 40 000 sample points for each method, but machine learning methods do not respond identically to changes in sample size. With more computing resources, it may be beneficial to run the model selection step with a larger selection of methods and using several different sample sizes.

3.4.4 Expanding the extent of permafrost predictions

Expanding the model's domain beyond the Dawson Range would significantly advance knowledge of permafrost conditions in Yukon and beyond, and would represent a useful product to governments, industry, and non-profit organizations involved in conservation or interventions in development project consultations. As presented here, model performance and confidence in its classifications would decrease as climatic conditions become increasingly different from the training region since these are not incorporated as predictors. It is therefore necessary to incorporate, at minimum, a temperature raster, but this in turn requires the expansion of the model's training data set to capture sufficient variability in temperature conditions.

Expanding the training data set could be done by intensive field studies at multiple locations, but some datasets already exist which could be used. Major infrastructure projects in the Yukon as well as in nearby Alaska and Northwest Territories often map permafrost lateral extent (and often active layer and permafrost thickness) where they propose impacts, whether for their own planning purposes or to satisfy regulatory requirements, and these data sets are often representative of undisturbed regions. To minimize their costs and improve planning processes, these project proponents would benefit from sharing their permafrost data with researchers aiming to generate permafrost maps at the regional scale. Local researchers are also potential sources of valuable information, often in undisturbed regions. Finally, territorial/state governments hold data sets of permafrost data in the form of borehole logs and thermistor string timeseries, often along highway corridors and in populated areas (for example, the Yukon Permafrost Database (Yukon Geological Survey 2025)). These data are often representative of disturbed areas but could be used with other data if given careful consideration.

The modelling framework presented in Chapter 2 (and supported by the R packaged presented in the Appendix) is adaptable to new data, both in terms of additional or different predictor variables and in terms of training data (outcome variable). Some pre-processing of training data would be necessary to homogenize data provided as separate files or as different spatial data types (i.e. point, polygon, or raster outcome data, polygon and raster predictor variables), and some refinements may be necessary to the spatial blocking technique I used to separate training from testing data sets. Nonetheless, expanding the present model's spatial coverage is an achievable goal with sufficient time and effort dedicated to acquiring and homogenising permafrost observations at a range of sites and incorporating climatic predictor variables.

3.5 Incorporation of climate change scenarios

The incorporation of climate change scenarios into this type of ML-based permafrost modeling is more challenging. It first requires that the trained ML-based permafrost model incorporates climate variables (temperature, precipitation, snow cover, etc.) and solar insolation variables susceptible to change with climate change (e.g., cloud cover). Once these variables are properly integrated into a ML-based permafrost modeling using present-day climate conditions, the effect of climate change on permafrost distribution could be incorporated using future climate predictions in the modeling framework. Provided that the present-day training set of the model featured data with a range of values covering most of the range of future-state prediction data, predictions should reflect the future state of permafrost, albeit without considering the thermal inertia of permafrost via the enthalpy of ice fusion. Caution should be exercised, however, as long term simulations (CMIP 5 and 6, Copernicus Climate Change Service 2021) have tended to under-predict historical temperature rises while over-predicting precipitation (Cai et al., 2024; Schetselaar et al., 2023)

Another approach to making predictions of permafrost distribution into the future would be to couple the output of ML-based permafrost classification trained on present-day data with a physics-based model. For example, an extensive ML-based permafrost map covering much of the Northern hemisphere's permafrost-affected regions could be used to calibrate an earth system model's run from past to present times, allowing modellers to tune model parameters and initial state conditions to match a much more detailed and accurate map than presently available. This could result in more appropriate model parametrization, which would improve the accuracy of predictions as a model is set to run into the future. Earth system models are typically run over long simulation periods to build steady numerical representations of soil and hydrological layers (see Elshamy et al. 2024), but incorporating accurate permafrost representations over the entire domain of such a model could be used to reduce required simulation periods by setting more accurate initial state conditions. At present, the resolution of Earth system models (1 km^2 or more) remains too low to leverage the high-resolution layers generated via machine learning approaches. This limits the benefit of using high resolution validation or initial condition layers, but this type of high-resolution validation layers will be critical to enhance the resolution and accuracy of Earth system models in the future.

3.6 Perspectives

Permafrost research, as with research on many issues of relevance to northern and less inhabited regions, is a developing field with potential for further discoveries and improved methodologies. Though most of the physical processes which govern permafrost growth or decay are now well understood, large challenges remain in predicting where permafrost is and where it will be in the future, in assessing how these changes will affect the landscape and communities which depend on it, and in assessing how changes in permafrost will affect global biogeochemical cycles. A persisting limitation in improving permafrost prediction is the lack of high-resolution climate data and the scarcity of field data of permafrost conditions. Solving this issue will require a better integration of stakeholders in affected regions. Future researchers will benefit from partnerships with local governments and industry, who often hold valuable proprietary climate or permafrost data sets outside the public domain. This thesis has demonstrated how such partnership could greatly enhance our collective knowledge of permafrost distribution in remote regions with benefit for all stakeholders.

I would like to conclude by highlighting some of the current challenges of not having better permafrost maps that I experience as a northern data scientist focused on hydrology and water quality. The hydrological models we use in my office are based on the MESH hydrological model developed for the Yukon River basin (Elshamy et al., 2020). Our original hydrological models have shown increasingly worsening fidelity to observed conditions in the northern regions of the Yukon (less 'flashy' hydrograph responses compared to modelled responses) while hewing closer to true conditions in more southern catchments. We suspect that this discrepancy is linked to changing permafrost distribution and active layer thickness: without much permafrost in the southern catchments, these changes have less impact on their hydrology. These suspected changes are reflected in observations of greater increases in baseflow as well as dissolved loads at northern vs southern latitudes (Figure 3.3 and Figure 3.4), which in concert suggest deepening flow paths and increasing aquifer recharge to the detriment of supra-permafrost flow.

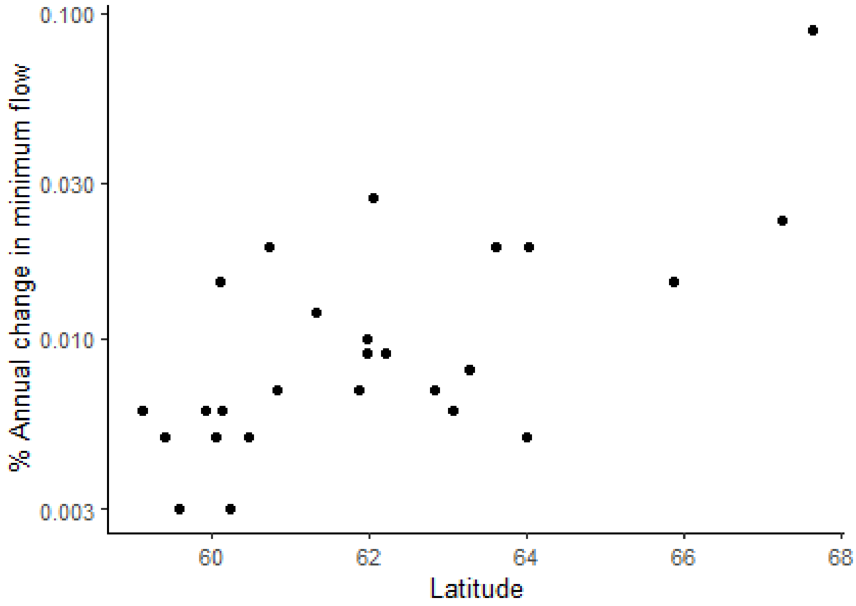

Figure 3.3: Percent annual change in minimum (base) flow at locations within Yukon monitored by the Water Survey of Canada and the Yukon Department of Environment. Locations influenced by hydroelectric power generation, affected by glacier melt redirection, and with fewer than 40 years of records were excluded. From that subset, only those where the p-value of the trend is < 0.05 were plotted.

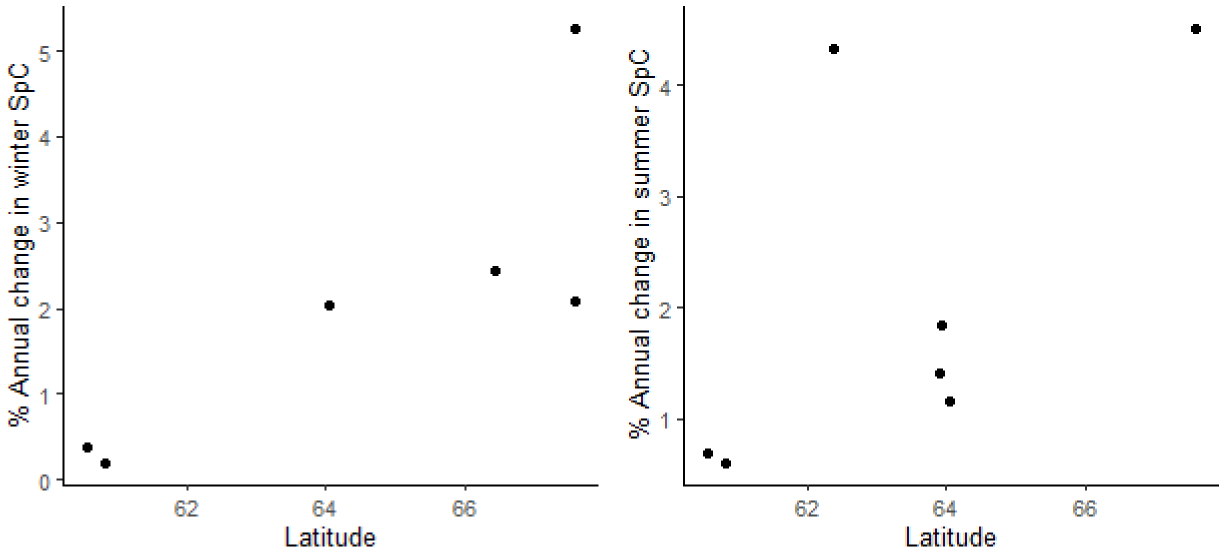

Figure 3.4: Percent annual change in maximum winter (left) and summer (right) specific conductance from ECCC long-term monitoring sites in Yukon. Due to data scarcity, only points with a p-value of < 0.2 and > 10 years of records are shown.

Another example is rapid changes in water chemistry that are difficult to predict including decreases in pH (see Figure 3.5), possibly caused by degrading permafrost leading to exposure and oxidation of sulphides (Van Stempvoort et al., 2023). Likewise, the increases in specific conductance (a rough proxy for dissolved load) as shown in Figure 3.4 likely result from increased surface-water groundwater connectivity as demonstrated in the Yukon River basin and elsewhere (Toohey et al., n.d.; Walvoord et al., 2012). Understanding these changes is critical from a scientific perspective because they can have broad consequences on local human and ecological communities creating potential hazards. For example, increasing winter flows (possibly resulting from increased ground water recharge through degrading permafrost) can mean thinner ice, dryer ground which is more prone to fires, and unstable slopes prone to landslides.

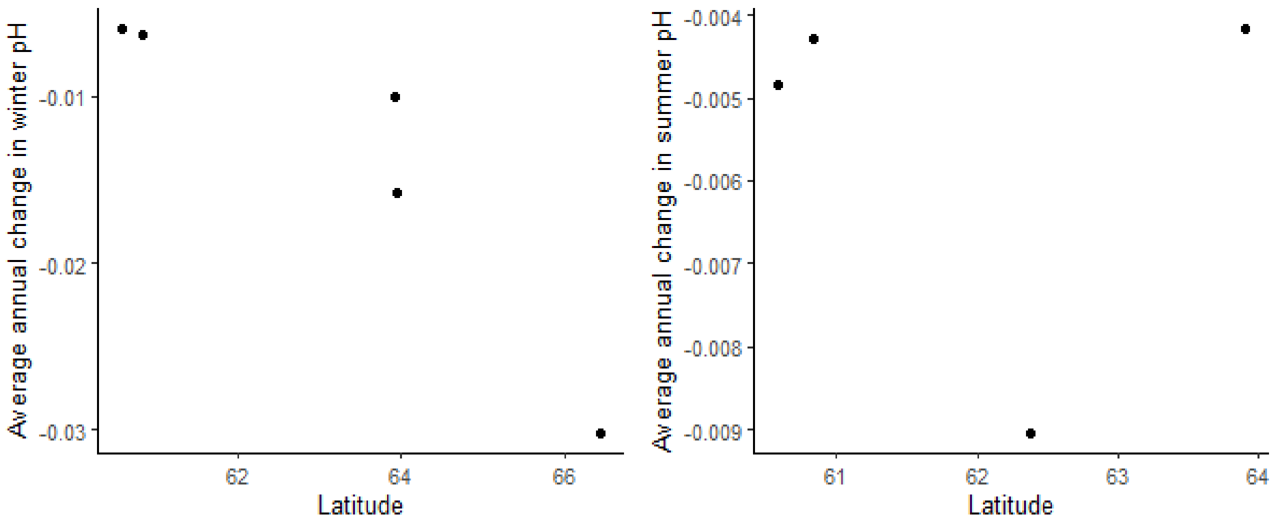

Figure 3.5: Annual change in average winter (left) and summer (right) pH from ECCC long-term monitoring sites in Yukon. Due to data scarcity, points with a p-value of < 0.2 and > 10 years of records are shown.

It is my hope that the present work advances our knowledge of and ability to know more about permafrost, through its contribution of a high-resolution permafrost map as well as via the published modelling framework. In a Yukon context, I have had interest from local environmental consultants seeking to use this approach to expand their own permafrost maps for mining project proponents and have had discussions about generating maps in the South Klondike River region in support of hydrological model development. The map I have produced along with applications and ideas like these will form the basis for improved economic productivity, decrease human suffering and property loss, and support better management of natural resources – all the while empowering local First Nations, NGOs, and other stakeholders with improved knowledge about their surroundings.

Chapter 4: Appendix

4.1 SAiVE R Package

An integral part of the development of the permafrost map presented in Chapter 2 is the published R-language code developed for this purpose. This is presented in the form of a software package called *SAiVE* and published on the Comprehensive R Archive Network (CRAN), the official repository of the R Project for Statistical Computing packages and source code (The R Foundation, n.d.). Publication of software to CRAN requires a peer-review process and a series of automated tests on the developed software; in addition, each function must be complemented by worked examples and unit tests. Examples facilitate the use of the software by potential users, while unit tests ensure that the software works as expected regardless of changes to its own code, to that of other dependencies, or even structural changes to the programming language itself.

Documentation for R packages is built into the coding framework, and a user manual is created for publication and updated at every version change by CRAN. The published manual is therefore the authoritative source of information for the *SAiVE* R package, and I advise readers of this manuscript to consider the manual published on CRAN as superseding the one reproduced below.

As publication of this software package formed an integral part of my studies, I have included the full manual in the next section as published. The manual's original formatting has been preserved as best as possible, but some minor differences will be evident if comparing the present version to the CRAN-published version owing to conversion of the pdf manual. The manual's page numbers, table of contents, and index have been kept as published.

Package 'SAiVE'

October 23, 2024

Type Package

Title Functions Used for SAiVE Group Research, Collaborations, and Publications

Version 1.0.7

Date/Publication 2024

Description Holds functions developed by the University of Ottawa's SAiVE (Spatio-temporal Analysis of isotope Variations in the Environment) research group with the intention of facilitating the re-use of code, foster good code writing practices, and to allow others to benefit from the work done by the SAiVE group. Contributions are welcome via the 'GitHub' repository <https://github.com/UO-SAiVE/SAiVE> by group members as well as non-members.

License MIT + file LICENSE

URL https://github.com/UO-SAiVE/SAiVE

BugReports https://github.com/UO-SAiVE/SAiVE/issues

Depends R (>= 4.2)

Imports caret,
crayon,
doParallel
parallel,
proxy,
rlang,
stats,
terra,
utils,
VSURF

Suggests ranger,
testthat (>=
3.0.0),
vdiffr,
whitebox

Config/testthat/edition 3

Encoding UTF-8

LazyData true

Roxygen list(markdown = TRUE)

RoxygenNote 7.3.2

R topics documented:

---

aspect *Raster of aspect*

Description

Raster of aspect

Usage

```
aspect
```

Format

aspect:

A tif file loaded as a terra spatRaster

Source

Derived from the Canadian Digital Elevation Model DEM

basin_dem	*Raster of elevation for basin delineation testing*

Description

Raster of elevation for basin delineation testing

Usage

basin_dem

Format

basin_dem**:**
A tif file loaded as a terra spatRaster

Source

Small subset of the Canadian Digital Elevation Model DEM

basin_pts	*Points for basin delineation*

Description

Points for basin delineation

Usage

basin_pts

Format

basin_pts**:**
A geopackage file loaded as a terra spatVector

Source

Created by package developer in ArcGIS Pro createStreams *Create stream network from DEM*

Description

[Stable]

Creates a stream network from a provided DEM. In most cases it is advisable to first hydro-process the DEM (see hydroProcess()) to remove depressions which preclude continuous flow from one DEM cell to the next.

Usage

```
createStreams(
  DEM,
  threshold,
  vector =
NULL,
  save_path =
NULL,
  n.cores =
NULL,
  force_update
_wbt = FALSE,
  silent_wbt =
TRUE
)
```

Arguments

DEM	The path to a digital elevation model file with .tif extension, or a terra spatRaster object. It is usually advisable to have already hydro-processed the DEM to remove artificial depressions. See hydroProcess().
threshold	The accumulation threshold in DEM cells necessary to start defining a stream.
vector	Output file specifications. NULL for no vector file at all; "env" for only a variable returned to the R environment; or, to return to environment and save to disk, "gpkg" for a geopackage file, "shp" for a shapefile.
save_path	An optional path in which to save the newly created stream network. If left NULL will save it in the same directory as the provided DEM or, if the DEM is a terra object, return only terra objects.
n.cores	The maximum number of cores to use. Leave NULL to use all cores minus 1.
force_update_wbt	Whitebox Tools is by default only downloaded if it cannot be found on the computer, and no check are performed to ensure the local version is current. Set to TRUE if you know that there is a new version and you would like to use it.
silent_wbt	Should Whitebox tools messages be suppressed? This function prints messages to the console already but these messages can be useful if you need to do some debugging.

Details

This function is essentially a convenient wrapper around three WhiteboxTools geospatial tools: whitebox::wbt_d8_flow_accumulation(), whitebox::wbt_d8_pointer(), and whitebox::wbt_extract_streams Terra functions are used to manipulate raster and vector files before and after WhiteboxTools functions are called.

Value

A raster representation of streams and, if requested, a vector representation of streams. Returned as terra objects and saved to disk if save_path is not null.

Author(s)

Ghislain de Laplante

Examples

```
hydroDEM <- hydroProcess(elev, 200, streams, n.cores = 2)
res <- createStreams(hydroDEM, 50, n.cores = 2)

terra::plot(res$streams_derived)
```

---

drainageBasins *Watershed/basin delineation*

Description

[Stable]

Hydro-processes a DEM, creating flow accumulation, direction, and streams rasters, and (option-ally) delineates watersheds above one or more points using Whitebox Tools. To facilitate this task in areas with poor quality/low resolution DEMs, can "burn-in" a stream network to the DEM to en-sure proper stream placement (see details). Many time-consuming raster operations are performed, so the function will attempt to use existing rasters if they are present in the same path as the base DEM and named according to the function's naming conventions. In practice, this means that only the first run of the function needs to be very time consuming. See details for more information.

NOTE 1: This tool can be slow to execute and will use a lot of memory. Be patient, it might take several hours with a large DEM.

NOTE 2: ESRI shapefiles, on which the Whitebox Tools functions depend, truncate column names to 10 characters. You may want to save and re-assign column names to the output terra object after this function has run.

NOTE 3: If you are have already run this tool and are using a DEM in the same directory as last time, you only need to specify the DEM and the points (and, optionally, a projection for the points output). Operations using the optional streams shapefile and generating flow accumulation direction, and the artificial streams raster do not need to be repeated unless you want to use a different DEM or streams shapefile.

NOTE 4: This function is very memory (RAM) intensive. You'll want at least 16GB of RAM, and to ensure that most of it is free. If you get an error such as 'cannot allocate xxxxx bytes', you probably don't have the resources to run the tool. All rasters are un-compressed and converted to 64-bit float type before starting work, and there needs to be room to store more than twice that uncompressed raster size in memory.

Usage

```
drainageBasins(
  DEM,
  streams = NULL,
  breach_dist = 10000,
  threshold = 500,
  overwrite = FALSE,
  projection = NULL,
  points = NULL,
  points_name_col = NULL,
  save_path = NULL,
  snap = "nearest",
  snap_dist = 200,
  burn_dist = 10,
  n.cores = NULL,
  force_update_wbt = FALSE,
  silent_wbt = TRUE
)
```

Arguments

DEM	The path to a DEM including extension from which to delineate watersheds/ catchments. Must be in .tif format. Derived layers such as flow accumulation, flow direction, and streams will inherit the DEM coordinate reference system.
streams	Optionally, the path to the polylines shapefile/geopackage containing lines, which can be used to improve accuracy when using poor quality DEMs. If this shapefile is the only input parameter being modified from previous runs (i.e. you've found a new/better streams shapefile but the DEM is unchanged) then specify a shapefile or geopackage lines file here and overwrite = TRUE.
breach_dist	The max radius (in raster cells) for which to search for a path to breach depressions, passed to whitebox::wbt_breach_depressions_least_cost(). This value should be high to ensure all depressions are breached. Note that the DEM is *not* breached in order of lowest elevation to greatest, nor is it breached sequentially (order is unknown, but the raster is presumably searched in some grid pattern for depressions). This means that flow paths may need to cross multiple depressions, especially in low relief areas.
threshold	The accumulation threshold in DEM cells necessary to start defining a stream. This streams raster is necessary to snap pout points to, so make sure not to make this number too great!
overwrite	If applicable, should rasters present in the same directory as the DEM be over-written? This will also force the recalculation of derived layers.
projection	Optionally, a projection string in the form "epsg:3579" (find them here). The derived watersheds and point output layers will use this projection. If NULL the projection of the points will be used.
points	The path to the points shapefile (extension .shp) containing the points from which to build watersheds. The attribute of each point will be attached to the newly created drainage polygons. Leave NULL (along with related parameters) to only process the DEM without defining watersheds.
points_name_col	The name of the column in the points shapefile containing names to assign to the

	watersheds. Duplicates *are* allowed and are labelled with the suffix _duplicate and a number for duplicates 2 +.
save_path	The directory where you want the output shapefiles saved.
snap	Snap to the "nearest" derived (calculated) stream, or to the "greatest" flow accumulation cell within the snap distance? Beware that "greatest" will move the point downstream by up to the 'snap_dist' specified, while nearest might snap to the wrong stream.
snap_dist	The search radius within which to snap points to streams. Snapping method depends on 'snap' parameter. Note that distance units will match the projection, so probably best to work on a meter grid.
burn_dist	If specifying a streams layer polyline layer, the number of DEM units to depress the DEM along the stream trace.
n.cores	The maximum number of cores to use. Leave NULL to use all cores minus 1.
force_update_wbt	Whitebox Tools is by default only downloaded if it cannot be found on the computer, and no checks are performed to ensure the local version is current. Set to TRUE if you know that there is a new version and you would like to use it.
silent_wbt	Should Whitebox tools messages be suppressed? This function prints messages to the console already but these messages can be useful if you need to do some debugging.

Details

This function uses software from the Whitebox geospatial analysis package, built by Prof. John Lindsay. Refer to this link for more information.

Creating derived raster layers without defining watersheds:
This function can be run without having any specific point above which to define a watershed. This can come in handy if you need to know where the synthetic streams raster will end up to ensure that your defined watershed pour points do not end up on the wrong stream branch, or if you simply want to front-load work while you work on defining the watershed pour points. To do this, leave the parameter points and associated parameters as NULL.

Explanation of process:
Starting from a supplied DEM, the function will fill single-cell pits, burn-in a stream network depression if requested (ensuring that flow accumulations happen in the correct location), breach depressions in the digital elevation model using a least-cost algorithm (i.e. using the pathway resulting in minimal changes to the DEM considering distance and elevation) then calculate flow accumulation and direction rasters. Then, a raster of streams is created where flow accumulation is greatest. The points provided by the user are then snapped to the derived streams raster and watersheds are computed using the flow direction rasters. Finally, the watershed/drainage basin polygons are saved to the specified save path along with the provided points and the snapped pour points.

Using a streams shapefile to burn-in depressions to the DEM:
Be aware that this part of the function should ideally be used with a “simplified” stream shapefile. In particular, avoid or pre-process stream shapefiles that represent side-channels, as these will burn-in several parallel tracks to the DEM. ESRI has a tool called "simplify hydrology lines" which is great if you can ever get it to work, and WhiteboxTools has functions whitebox::wbt_remove_short_stream to trim the streams raster, and whitebox::wbt_repair_stream_vector_topology() to help in converting a corrected streams vector to raster in the first place.

Value

A list of terra objects. If points are specified: delineated drainages, pour points as provided, snapped pour points, and the derived streams network. If no points: flow accumulation and direction rasters, and the derived streams network. If points specified, also saved to disk: an ESRI shapefile for each drainage basin, plus the associated snapped pour point and the point as provided and a shapefiles for all basins/points together. In all cases the created or discovered rasters will be in the same folder as the DEM.

Author(s)

Ghislain de Laplante

Examples

```
# Must be run with file paths as well as a save_path

# Interim raster are created in the same path as the DEM

file.copy(system.file("extdata/basin_rast.tif", package = "SAiVE"),
  paste0(tempdir(), "/basin_rast.tif"))

basins <- drainageBasins(save_path = tempdir(),
  DEM = paste0(tempdir(), "/basin_rast.tif"),
  streams = system.file("extdata/streams.gpkg", package = "SAiVE"), points =
  system.file("extdata/basin_pts.gpkg", package = "SAiVE"), points_name_col =
  "ID",
  breach_dist = 500,
  n.cores = 2)

terra::plot(basins$delineated_basins)
```

---

elev *Raster of elevation*

Description

Raster of elevation

Usage

elev

Format

elev:
A tif file loaded as a terra spatRaster

Source

Small subset of the Canadian Digital Elevation Model DEM

hydroProcess	*Hydro-process a DEM*

Description

[Stable]

Takes a digital elevation model and prepares it for hydrological analyses, such as basin delineation. Modifies the input DEM by breaching single cell pits/depressions and then breaching remaining depressions using a least cost algorithm (where cost is a function of distance plus elevation change to the DEM).

If a streams layer is specified, a depression will be "burned-in" to the DEM along the stream path (after converting the vector file to a raster). This is very useful when trying to delineate basins with a poor resolution DEM. You can control the depth of this depression with parameter 'burn_dist'.

Usage

```
hydroProcess(
  DEM,
  breach_dist,
  streams = NULL,
  burn_dist = 10,
  save_path = NULL,
  n.cores = NULL,
  force_update_wbt = FALSE,
  silent_wbt = TRUE
)
```

Arguments

DEM	The path to a digital elevation raster file with .tif extension, or a terra spatRaster object.
breach_dist	The max radius (in raster cells) in which to search for a path to breach depressions, passed to whitebox::wbt_breach_depressions_least_cost(). This value should be high to ensure all depressions are breached, keeping in mind that greater distance = greater computing time. Note that the DEM is *not* breached in order of lowest elevation to greatest, nor is it breached sequentially (order is unknown, but the raster is presumably searched in some grid pattern for de-pressions). This means that flow paths may need to cross multiple depressions, especially in low relief areas.
streams	Optionally, the path to the polylines shapefile or geopackage file containing streams, which can be used to improve hydrological accuracy when using poor quality DEMs but decent accuracy stream networks.
burn_dist	The number of units (in DEM units) to use for burning-in the stream network.
save_path	An optional path in which to save the processed DEM. If left NULL will save it in the same directory as the provided DEM or, if the DEM is a terra object, return only terra objects.
n.cores	The maximum number of cores to use. Leave NULL to use all cores minus 1.

force_update_w

Whitebox Tools is by default only downloaded if it cannot be found on the computer, and no check are performed to ensure the local version is current. Set to TRUE if you know that there is a new version and you would like to use it.

silent_wbt

Should Whitebox tools messages be suppressed? This function prints messages to the console already but these messages can be useful if you need to do some debugging.

Details

Relies on two WhiteboxTools functions: whitebox::wbt_fill_single_cell_pits() and whitebox::wbt_breach_depressions_least_cost(). If the parameter streams is specified, a depression is burned into the DEM after running fill_single_cell_pits and before breaching depressions.

Value

A hydro-processed DEM returned as a terra object and saved to disk if save_path is not null.

Author(s)

Ghislain de Laplante

Examples

```
# Running with terra objects:
res <- hydroProcess(DEM = elev,
   breach_dist = 500,
   streams = streams,
   n.cores = 2)

terra::plot(res)

# Running with file paths:
res <- hydroProcess(DEM = system.file("extdata/dem.tif", package = "SAiVE"),
   breach_dist = 500,
   streams = system.file("extdata/streams.gpkg", package = "SAiVE"),
   n.cores = 2)

terra::plot(res)
```

---

modelMatch *Find machine learning models for use in caret*

Description

[Experimental]

As of 2023-06-15, there are 238 different machine learning models which can be used with the CARET package. As evaluating model performance is time consuming, selecting a subset of models to test prior to deciding on which model to use is essential. This function aims to facilitate this process by matching models according to their Jaccard similarity, in a process inspired by this section in the CARET e-book. Model data is fetched from here. The result of this function can then be passed to spatPredict() to further refine model selection.

Usage

```
modelMatch(model, type = "match", similarity = 0.7)
```

Arguments

model	The abbreviation or short name of the model you'd like to match, taken from here.
type	The type of model. You can match the input model type with "match", or select from dual-purpose models ("dual"), regression models only ("regression"), or classification models only ("classification").
similarity	The similarity threshold to use as a numeric value from 0 to 1. Models with a similarity score greater than this will be returned.

Details

This function requires internet access to get an up-to-date list of models.

Value

A data.frame of models meeting the requested similarity threshold along with the model abbrevia-tions that can be passed to caret::train() or to function spatPredict().

Author(s)

Ghislain de Laplante

Examples

```
# Find models similar to 'ranger'
modelMatch("ranger")

# Find only models with a similarity > 0.8 to 'ranger'
modelMatch("ranger", similarity = 0.8)
```

---

permafrost *Permafrost data*

Description

A small data set of permafrost type with correlated terrain attributes

Usage

```
permafrost
```

Format

permafrost:

A data frame with 400 rows and 8 columns

Source

Permafrost type classification from central Yukon, with corresponding values derived from the Canadian Digital Elevation Model.

---

permafrost_polygons *Polygons of permafrost occurrence and type*

Description

Polygons of permafrost occurrence and

type

Usage

permafrost_polygons

Format

permafrost_polygons:

A geopackage file loaded as a terra spatVector.

Source

Permafrost classification polygons created from ground observations and geophysics in central Yukon.

---

slope *Raster of slope angle*

Description

Raster of slope angle

Usage

slope

Format

slope:

A tif file loaded as a terra spatRaster

Source

Derived from the Canadian Digital Elevation Model DEM solrad *Raster of solar radiation*

Description

Raster of solar radiation

Usage

solrad

Format

solrad**:**
A tif file loaded as a terra spatRaster

Source

Derived from the Canadian Digital Elevation Model DEM using the ArcGIS Area Solar Radiation tool

spatPredict *Predict spatial variables using machine learning*

Description

[Stable]

Function to facilitate the prediction of spatial variables using machine learning, including the selec-tion of a particular model and/or model parameters from several user-defined options. Both classi-fication and regression is supported, though please ensure that the models passed to the parameter methods are suitable.

Note that you may need to acquiesce to installing supplementary packages, depending on the model types chosen and whether or not these have been run before; this function may not be 'set and forget'.

It is possible to specify multiple machine learning methods (the methods parameter) as well as method-specific parameters (the trainControl parameter) if you wish to test multiple options and select the best one. To facilitate method selection, refer to function modelMatch(). If you are unsure of the best model to use, you can use the fastCompare parameter to quickly compare models and select the best one based on accuracy. If you wish to use a single model and/or trainControl object, you can pass a single string to methods and a single trainControl object to trainControl.

Warning options are changed for this function only to show all warnings as they occur and reset back to their original state upon function completion (a test is done first to ensure it can be reset). This is to ensure that any warnings when running models are shown in sequence with the messages indicating the progress of the function, especially when running multiple models and/or trainControl options.

Usage

```
spatPredict(
  features,
  outcome,
  poly_sample = 1000,
  trainControl,
  methods,
  fastCompare = TRUE,
  fastFraction = NULL,
  thinFeatures = TRUE,
  predict = FALSE,
  n.cores = NULL,
  save_path = NULL
)
```

Arguments

features	Independent variables. Must be either a NAMED list of terra spatRasters or a multi-layer (stacked) spatRaster (c(rast1, rast2)). All layers must all have the same cell size, alignment, extent, and crs. These rasters should include the training extent (that covered by the spatVector in outcome) as well as the desired extrapolation extent.
outcome	Dependent variable, as a terra spatVector of points or polygons with a single attribute table column (of class integer, numeric or factor). The class of this column dictates whether the problem is approached as a classification or regression problem; see details. If specifying polygons, stratified random sampling will be done with poly_sample number of points per unique polygon value.
poly_sample	If passing a polygon SpatVector to outcome, the number of points to generate from the polygons for each unique polygon value.
trainControl	Parameters used to control training of the machine learning model, created with caret::trainControl(). Passed to the trControl parameter of caret::train(). If specifying multiple methods in methods you can use a single trainControl which will apply to all methods, or pass multiple variations to this argument as a list with names matching the names of methods (one element for each model specified in methods).
methods	A string specifying one or more classification/regression methods(s) to use. Passed to the method parameter of caret::train(). If specifying more than one method they will all be passed to caret::resamples() to compare method performance. Then, if predict = TRUE, the method with the highest overall accuracy will be selected to predict the raster surface across the exent of features. A different trainControl parameter can be used for each method in methods.
fastCompare	If specifying multiple methods in methods or one method with multiple trainControl objects, should the points in outcome be sub-sampled for the comparison step? The selected method will be trained on the full outcome data set after selection. This only applies if methods is length > 3, with behavior further modified by fastFraction.
fastFraction	The fraction of points to use for the method comparison step (final training and testing is always done on the full data set) if fastCompare is TRUE and multiple methods. Default NULL ranges from 1 for 5000 or fewer points to 0.1 for 50 000 or more points. You can also set this to any value between 0 and 1 to override this behavior.

thinFeatures	Should random forest selection using VSURF::VSURF() be used in an attempt to remove irrelevant variables?
predict	TRUE will apply the trained model to the full extent of features and return a raster saved to save_path.
n.cores	The maximum number of cores to use. Leave NULL to use all cores minus 1.
save_path	The path (folder) to which you wish to save the predicted raster. Not used unless predict = TRUE.

Details

This function partly operates as a convenient means of passing various parameters to the caret::train() function, enabling the user to rapidly trial different model types and parameter sets. In addition, pre-processing of data can optionally be done using VSURF::VSURF() (parameter thinFeatures) which can decrease the time to run models by removing superfluous parameters.

Model testing, comparison, and reported metrics:

After extracting raster values at *n* points from the features rasters the point values are split spatially into training and testing sets along a 70/30 split. This is accomplished by creating a grid (200*200) of polygons over the extent of the points and randomly assigning polygons to training or testing sets. Points within these polygons are then assigned to the corresponding set, ensuring that the training and testing sets are spatially independent.

Method for selecting the best model:

When specifying multiple model types in methods, each model type and trainControl pair (if trainControl is a list of length equal to methods) is run using caret::train(). To speed things up you can use fastCompare = TRUE. Models are then compared on their 'accuracy' metric as output by caret::resamples() when run on the testing partition, and the highest-performing model is selected. If fastCompare is TRUE, this model is then run on the complete data set provided in outcome. Model statistics are returned upon function completion, which allows the user to select their own 'best performing' model based on other criteriaif desired.

Balancing classes in outcome (dependent) variable:

Models can be biased if they are given significantly more points in one outcome class vs others, and best practice is to even out the number of points in each class. If extracting point values from a vector or raster object and passing a points vector object to this function, a simple way to do that is by using the "strata" parameter if using terra::spatSample(). If working directly from points, caret::downSample() and caret::upSample() can be used. See this link for more information. Note that if passing a polygons object to this function stratified random sampling will automatically be performed.

Classification or regression:

Whether this function treats your inputs as a classification or regression problem depends on the class attached to the outcome variable. A class factor will be treated as a classification problem while all other classes will be treated as regression problems.

Value

If passing only one method to the method argument: the outcome of the VSURF variable selection process (if thinFeatures is TRUE), the training and testing data.frames, the fitted model, model performance statistics, and the final predicted raster (if predict = TRUE).

If passing multiple methods to the method argument: the outcome of the VSURF variable selection process (if thinFeatures is TRUE), the training and testing data.frames, character vectors for failed methods, methods which generated a warning, and what those errors and warnings were, model performance comparison (if methods includes more than one method), the selected method, the trained model performance statistics, and the final predicted raster (if predict = TRUE).

In either case, the predicted raster is written to disk if save_path is specified.

Author(s)

Ghislain de Laplante

Examples

```
# These examples can take a while to run!

# Install packages underpinning examples
rlang::check_installed("ranger", reason = "required to run example.")
rlang::check_installed("Rborist", reason = "required to run example.")

# Single model, single trainControl

trainControl <- caret::trainControl(
                        method = "repeatedcv",
                        number = 2, # 2-fold Cross-validation
                        repeats = 2, # repeated 2 times
                        verboseIter = FALSE,
                        returnResamp = "final",
                        savePredictions = "all",
                        allowParallel = TRUE)

  outcome <- permafrost_polygons
  outcome$Type <- as.factor(outcome$Type)

result <- spatPredict(features = c(aspect, solrad, slope), outcome =
   outcome,
   poly_sample = 100,
   trainControl = trainControl,
   methods = "ranger",
   n.cores = 2,
   predict = TRUE)

terra::plot(result$prediction)


# Multiple models, multiple trainControl

trainControl <- list("ranger" = caret::trainControl(
                                    method = "repeatedcv",
                                    number = 2,
                                    repeats = 2,
                                    verboseIter = FALSE,
                                    returnResamp = "final",
                                    savePredictions = "all",
                                    allowParallel = TRUE),
                        "Rborist" = caret::trainControl(
                                    method = "boot",
                                    number = 2,
                                    repeats = 2,
```

```
                         verboseIter = FALSE,
                         returnResamp = "final",
                         savePredictions = "all",
                         allowParallel = TRUE)
                         )

result <- spatPredict(features = c(aspect, solrad, slope), outcome =
  outcome,
  poly_sample = 100,
  trainControl = trainControl,
  methods = c("ranger", "Rborist"),
  n.cores = 2,
  predict = TRUE)

terra::plot(result$prediction)
```

---

streams *Lines representing streams*

Description

Lines representing streams

Usage

```
streams
```

Format

streams:
A geopackage file loaded as a terra
spatVector

Source

A subset of the Yukon CANVEC water flow lines found here

---

thinFeatures *Remove irrelevant predictor variables*

Description

[Stable]

Uses VSURF::VSURF() to build random forests and remove irrelevant predictor variables from a data.frame containing an outcome variable and 2 or more predictor variables.

Usage

```
thinFeatures(data, outcome_col, n.cores = NULL)
```

Arguments

data	A data.frame containing a column for the outcome variable and *n* columns for predictor variables.
outcome_col	The name of the outcome variable column.
n.cores	The maximum number of cores to use. Leave NULL to use all cores minus 1.

Value

A list of two data.frames: the outcome of the VSURF algorithm and the data after applying the VSURF results (rows removed if applicable)

Author(s)

Ghislain de Laplante

Examples

```
# thinFeatures on 'permafrost' data set

data(permafrost)
res <- thinFeatures(permafrost, "Type", n.cores = 2)

# Results will vary due to inherent randomness of random forests!
```

---

veg *Raster of vegetation types*

Description

Raster of vegetation types

Usage

veg

Format

veg:
A tif file loaded as a terra spatRaster

Source

A small subset of the North American Land Cover dataset produced by the Commission for Environmental Cooperation, resampled to match cell size of other rasters in this package.

wbtCheck	*Check WhiteboxTools binaries installation*

Description

[Stable]

Checks for the existence of WhiteboxTools in its default directory and installs it if necessary or if force = TRUE.

Usage

```
wbtCheck(force = FALSE, silent)
```

Arguments

force	Set TRUE to force update of WhiteboxTools binaries.
silent	TRUE to suppress console messages other than those explaining that Whitebox Tools binaries are being installed.

Value

Returns the version number of the installed binaries and (if necessary) installs WhiteboxTools in its default location.

Author(s)

Ghislain de Laplante

Examples

```
#Check if WhiteboxTools binaries are installed. If not, install latest version.
wbtCheck()

# Update WhiteboxTools binaries if they are already installed.
wbtCheck(force = TRUE)
```

Index

References

A. Ramezan, C., A. Warner, T., & E. Maxwell, A. (2019). Evaluation of Sampling and Cross-Validation Tuning Strategies for Regional-Scale Machine Learning Classification. *Remote Sensing*, *11*(2), 185. https://doi.org/10.3390/rs11020185

AMAP. (2021). *Arctic climate change update 2021: Key trends and impacts. Summary for policy-makers.* Arctic Monitoring and Assessment Programme (AMAP). https://www.amap.no/documents/doc/arctic-climate-change-update-2021-key-trends-and-impacts.-summary-for-policy-makers/3508

Associate Committee on Geotechnical Research, National Research Council of Canada, Permafrost Subcommittee (Ed.). (1988). *Glossary of permafrost and related ground-ice terms*. National Research Council of Canada.

Bahn, V., & McGill, B. J. (2013). Testing the predictive performance of distribution models. *Oikos*, *122*(3), 321–331. https://doi.org/10.1111/j.1600-0706.2012.00299.x

Baral, P., & Haq, M. A. (2020). Spatial prediction of permafrost occurrence in Sikkim Himalayas using logistic regression, random forests, support vector machines and neural networks. *Geomorphology*, *371*, 107331. https://doi.org/10.1016/j.geomorph.2020.107331

Batrak, Y., & Müller, M. (2019). On the warm bias in atmospheric reanalyses induced by the missing snow over Arctic sea-ice. *Nature Communications*, *10*(1), 4170. https://doi.org/10.1038/s41467-019-11975-3

Beery, S., Van Horn, G., & Perona, P. (2018). *Recognition in terra incognita*. 1–18. https://openaccess.thecvf.com/content_ECCV_2018/papers/Beery_Recognition_in_Terra_ECCV_2018_paper.pdf

Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., & Vaughan, J. W. (2010). A theory of learning from different domains. *Machine Learning*, *79*(1–2), 151–175. https://doi.org/10.1007/s10994-009-5152-4

Benediktsson, J. A., Swain, P. H., & Ersoy, O. K. (1990). Neural network approaches versus statistical methods in classification of multisource remote sensing data. *IEEE Transactions on Geoscience and Remote Sensing*, *28*(4), 540–552. https://doi.org/10.1109/TGRS.1990.572944

Bhuiyan, S. A., Bataille, C. P., & McGrath, H. (2022). Harmonizing and Extending Fragmented 100 Year Flood Hazard Maps in Canada's Capital Region Using Random Forest Classification. *Water*, *14*(23), Article 23. https://doi.org/10.3390/w14233801

Biskaborn, B. K., Smith, S. L., Noetzli, J., Matthes, H., Vieira, G., Streletskiy, D. A., Schoeneich, P., Romanovsky, V. E., Lewkowicz, A. G., Abramov, A., Allard, M., Boike, J., Cable, W. L., Christiansen, H. H., Delaloye, R., Diekmann, B., Drozdov, D., Etzelmüller, B., Grosse, G., … Lantuit, H. (2019). Permafrost is warming at a global scale. *Nature Communications*, *10*(1), 264. https://doi.org/10.1038/s41467-018-08240-4

Bond, J. D. (2019). *Paleodrainage map of Beringia* [Map]. Yukon Geological Survey, Energy, Mines and Resources, Government of Yukon. https://data.geology.gov.yk.ca/Reference/81642#InfoTab

Bond, J. D., & Lipovsky, P. S. (2011). *Surficial geology, soils and permafrost of the northern Dawson Range* (Yukon Exploration and Geology, pp. 19–32). Yukon Geological Survey, Energy, Mines and Resources, Government of Yukon. https://data.geology.gov.yk.ca/Reference/43189#InfoTab

Bonnaventure, P. P., Lewkowicz, A. G., Kremer, M., & Sawada, M. C. (2012). A Permafrost Probability Model for the Southern Yukon and Northern British Columbia, Canada. *Permafrost and Periglacial Processes*, *23*(1), 52–68. https://doi.org/10.1002/ppp.1733

Brabets, T. P., & Walvoord, M. A. (2009). Trends in streamflow in the Yukon River Basin from 1944 to 2005 and the influence of the Pacific Decadal Oscillation. *Journal of Hydrology*, *371*(1–4), 108–119. https://doi.org/10.1016/j.jhydrol.2009.03.018

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Brodersen, K. H., Ong, C. S., Stephan, K. E., & Buhmann, J. M. (2010). The Balanced Accuracy and Its Posterior Distribution. *2010 20th International Conference on Pattern Recognition*, 3121–3124. https://doi.org/10.1109/ICPR.2010.764

Brown, J., Ferrians Jr., O. J., Heginbottom, J. A., & Melnikov, E. S. (1997). *Circum-Arctic map of permafrost and ground-ice conditions* [Map]. United States Geological Survey. https://doi.org/10.3133/cp45

Cai, Z., You, Q., Chen, H. W., Zhang, R., Zuo, Z., Chen, D., Cohen, J., & Screen, J. A. (2024). Assessing Arctic wetting: Performances of CMIP6 models and projections of precipitation changes. *Atmospheric Research*, *297*, 107124. https://doi.org/10.1016/j.atmosres.2023.107124

Camill, P. (2005). Permafrost Thaw Accelerates in Boreal Peatlands During Late-20th Century Climate Warming. *Climatic Change*, *68*(1), 135–152. https://doi.org/10.1007/s10584-005-4785-y

Carrillo, H., Brodersen, K. H., & Castellanos, J. A. (2014). Probabilistic Performance Evaluation for Multiclass Classification Using the Posterior Balanced Accuracy. In M. A. Armada, A. Sanfeliu, & M. Ferre (Eds.), *ROBOT2013: First Iberian Robotics Conference* (Vol. 252, pp. 347–361). Springer International Publishing. https://doi.org/10.1007/978-3-319-03413-3_25

Colombo, N., Salerno, F., Gruber, S., Freppaz, M., Williams, M., Fratianni, S., & Giardino, M. (2018). Review: Impacts of permafrost degradation on inorganic chemistry of surface fresh water. *Global and Planetary Change*, *162*, 69–83. https://doi.org/10.1016/j.gloplacha.2017.11.017

Conant, R. (2019). Old Arctic carbon stability. *Nature Climate Change*, *9*(6), 437–438. https://doi.org/10.1038/s41558-019-0483-z

Copernicus Climate Change Service. (2021). *CMIP6 predictions underpinning the C3S decadal prediction prototypes* [Dataset]. ECMWF. https://doi.org/10.24381/CDS.C866074C

Crespo, J. L., & Mora, E. (1993). Drought estimation with neural networks. *Advances in Engineering Software*, *18*(3), 167–170. https://doi.org/10.1016/0965-9978(93)90064-Z

de Laplante, G. (2024). *SAiVE: Functions Used for SAiVE Group Research, Collaborations, and Publications* (Version 1.0.6) [Computer software]. https://github.com/UO-SAiVE/SAiVE

Decatur. (1989). Application of neural networks to terrain classification. *International Joint Conference on Neural Networks*, 283–288 vol.1. https://doi.org/10.1109/IJCNN.1989.118592

Deluigi, N., Lambiel, C., & Kanevski, M. (2017). Data-driven mapping of the potential mountain permafrost distribution. *Science of The Total Environment*, *590–591*, 370–380. https://doi.org/10.1016/j.scitotenv.2017.02.041

Deng, H., Guan, X., Liaw, A., Breiman, L., & Cutler, A. (2022). *RRF: Regularized Random Forests* (Version 1.9.4) [Computer software]. https://cran.r-project.org/package=RRF

Douglas, T. A., Turetsky, M. R., & Koven, C. D. (2020). Increased rainfall stimulates permafrost thaw across a variety of Interior Alaskan boreal ecosystems. *Npj Climate and Atmospheric Science*, *3*(1), 28. https://doi.org/10.1038/s41612-020-0130-4

Du, P., Bai, X., Tan, K., Xue, Z., Samat, A., Xia, J., Li, E., Su, H., & Liu, W. (2020). Advances of Four Machine Learning Methods for Spatial Data Handling: A Review. *Journal of Geovisualization and Spatial Analysis*, *4*(1), 1–25. https://doi.org/10.1007/s41651-020-00048-5

Duan, S., Ullrich, P., Risser, M., & Rhoades, A. (2024). Using Temporal Deep Learning Models to Estimate Daily Snow Water Equivalent Over the Rocky Mountains. *Water Resources Research*, *60*(4), e2023WR035009. https://doi.org/10.1029/2023WR035009

Duk-Rodkin, A. (1999). *Glacial Limits Map of Yukon* [Map]. Indian & Northern Affairs Canada, Department of Indian & Northern Development. https://data.geology.gov.yk.ca/Reference/42050#InfoTab

Elith, J., Kearney, M., & Phillips, S. (2010). The art of modelling range-shifting species. *Methods in Ecology and Evolution*, *1*(4), 330–342. https://doi.org/10.1111/j.2041-210X.2010.00036.x

Elshamy, M., Loukili, Y., Princz, D., Richard, D., & Tesemma, Z. (2020). *Yukon River Basin Streamflow Forecasting System.* (No. Report No. 16). Centre for Hydrology, University of Saskatchewan.

Elshamy, M., Pomeroy, J., Pietroniro, A., Wheater, H., Abdelhamed, M., & Davison, B. (2024). *Land Surface Hydrological Modelling of the Mackenzie River Basin: Parametrization to Simulate Streamflow and Permafrost Dynamics*. https://doi.org/10.2139/ssrn.4791947

Environment and Climate Change Canada. (n.d.). *Canadian Land Data Assimilation System in the National Surface and River Prediction System (CaLDAS-NSRPS)* [Model]. https://open.canada.ca/data/en/dataset/3959c86b-b555-4ad8-9fcc-8fecfb79918c

Environment and Climate Change Canada. (2024). *Canadian Climate Normals 1991-2020*. https://climate.weather.gc.ca/climate_normals/index_e.html

ESRI. (2023). *ArcGIS Pro 3.1* (Version 3.1.0) [Computer software]. Environmental Systems Research Institute (ESRI). https://www.esri.com/en-us/arcgis/products/arcgis-pro/overview

ESRI. (2024). *ArcGIS Pro 3.4* (Version 3.4.0) [Computer software]. Environmental Systems Research Institute (ESRI). https://www.esri.com/en-us/arcgis/products/arcgis-pro/overview

Etzelmüller, B., & Frauenfelder, R. (2009). Factors Controlling the Distribution of Mountain Permafrost in the Northern Hemisphere and their Influence on Sediment Transfer. *Arctic, Antarctic, and Alpine Research*, *41*(1), 48–58. https://doi.org/10.1657/1523-0430-41.1.48

Evans, J. S., Murphy, M. A., & Ram, K. (2024). *spatialEco: Spatial Analysis and Modelling Utilities* (Version 2.0-2) [Computer software]. https://cran.r-project.org/package=spatialEco

Fan, X., Wang, Y., Niu, F., Li, W., Wu, X., Ding, Z., Pang, W., & Lin, Z. (2023). Environmental Characteristics of High Ice-Content Permafrost on the Qinghai–Tibetan Plateau. *Remote Sensing*, *15*(18), 4496. https://doi.org/10.3390/rs15184496

Feng, X., Park, D. S., Liang, Y., Pandey, R., & Papeş, M. (2019). Collinearity in ecological niche modeling: Confusions and challenges. *Ecology and Evolution*, *9*(18), 10365–10376. https://doi.org/10.1002/ece3.5555

Figueroa, R. L., Zeng-Treitler, Q., Kandula, S., & Ngo, L. H. (2012). Predicting sample size required for classification performance. *BMC Medical Informatics and Decision Making*, *12*(1), 8. https://doi.org/10.1186/1472-6947-12-8

Flato, G. M. (2011). Earth system models: An overview. *WIREs Climate Change*, *2*(6), 783–800. https://doi.org/10.1002/wcc.148

Frey, K. E., & McClelland, J. W. (2009). Impacts of permafrost degradation on arctic river biogeochemistry. *Hydrological Processes*, *23*(1), 169–182. https://doi.org/10.1002/hyp.7196

Friedman, J. H. (1991). Multivariate adaptive regression splines. *The Annals of Statistics*, *19*(1), 1–67. https://doi.org/10.1214/aos/1176347963

Fu, P., & Rich, P. (1999). Design and implementation of the Solar Analyst: An ArcView extension for modeling solar radiation at landscape scales. *Proceedings of the Nineteenth Annual ESRI User Conference*.

Fu, P., & Rich, P. M. (2002). A geometric solar radiation model and its applications in agriculture and forestry. *Proceedings of the Second International Conference on Geospatial Information in Agriculture and Forestry*, 357–364.

Furuya, D. E. G., Aguiar, J. A. F., Estrabis, N. V., Pinheiro, M. M. F., Furuya, M. T. G., Pereira, D. R., Gonçalves, W. N., Liesenberg, V., Li, J., Marcato Junior, J., Prado Osco, L., & Ramos, A. P. M. (2020). A Machine Learning Approach for Mapping Forest Vegetation in Riparian Zones in an Atlantic Biome Environment Using Sentinel-2 Imagery. *Remote Sensing*, *12*(24), 4086. https://doi.org/10.3390/rs12244086

Gagne, D. J., McGovern, A., & Brotzge, J. (2009). Classification of Convective Areas Using Decision Trees. *Journal of Atmospheric and Oceanic Technology*, *26*(7), 1341–1353. https://doi.org/10.1175/2008JTECHA1205.1

Geomatics Yukon. (2019). *Ecoregions 2014—1M* [Map]. https://yukon.maps.arcgis.com/home/item.html?id=5e576674c89e4abbb8a7cb5dfde0c03e

Giuseppe, F., Simoni, S., Godt, J. W., Lu, N., & Rigon, R. (2016). Geomorphological control on variably saturated hillslope hydrology and slope instability. *Water Resources Research*, *52*(6), 4590–4607. https://doi.org/10.1002/2015WR017626

Gray, P. C., Boss, E., Prochaska, J. X., Kerner, H., & Demeaux, C. B. (2024). The Promise and Pitfalls of Machine Learning in Ocean Remote Sensing. *Oceanography*, *37*(3), 52–63. https://doi.org/10.5670/oceanog.2024.511

Grosse, G., Harden, J., Turetsky, M., McGuire, A. D., Camill, P., Tarnocai, C., Frolking, S., Schuur, E. A. G., Jorgenson, T., Marchenko, S., Romanovsky, V., Wickland, K. P., French, N., Waldrop, M., Bourgeau-Chavez, L., & Striegl, R. G. (2011). Vulnerability of high-latitude soil organic carbon in North America to disturbance. *Journal of Geophysical Research*, *116*, G00K06. https://doi.org/10.1029/2010JG001507

Gruber, S. (2012). Derivation and analysis of a high-resolution estimate of global permafrost zonation. *The Cryosphere*, *6*(1), 221–233. https://doi.org/10.5194/tc-6-221-2012

Gruber, S., Hayley, J., Karunaratne, K., King, J., MacLean, T., Marshall, S., & Moore, D. (2023). Considerations toward a vision and strategy for permafrost knowledge in Canada. *Arctic Science*, *9*(4), iv–viii. https://doi.org/10.1139/as-2023-0016

Grünberg, I., Wilcox, E. J., Zwieback, S., Marsh, P., & Boike, J. (2020). Linking tundra vegetation, snow, soil temperature, and permafrost. *Biogeosciences*, *17*(16), 4261–4279. https://doi.org/10.5194/bg-17-4261-2020

Haeberli, W. (1973). Die Basis-Temperatur der winterlichen Schneedecke als moglicher indikator fur die Verbreitung von permafrost in den Alpen. *Zeitschrift Für Gletscherkunde Und Glazialgeologie*, *9*, 221–227.

Hansen, M., Dubayah, R., & Defries, R. (1996). Classification trees: An alternative to traditional land cover classifiers. *International Journal of Remote Sensing*, *17*(5), 1075–1081. https://doi.org/10.1080/01431169608949069

Haupt, S. E., Gagne, D. J., Hsieh, W. W., Krasnopolsky, V., McGovern, A., Marzban, C., Moninger, W., Lakshmanan, V., Tissot, P., & Williams, J. K. (2022). The History and Practice of AI in the Environmental Sciences. *Bulletin of the American Meteorological Society*, *103*(5), E1351–E1370. https://doi.org/10.1175/BAMS-D-20-0234.1

Heijmans, M. M. P. D., Magnússon, R. Í., Lara, M. J., Frost, G. V., Myers-Smith, I. H., Van Huissteden, J., Jorgenson, M. T., Fedorov, A. N., Epstein, H. E., Lawrence, D. M., & Limpens, J. (2022). Tundra vegetation change and impacts on permafrost. *Nature Reviews Earth & Environment*, *3*(1), 68–84. https://doi.org/10.1038/s43017-021-00233-0

Hijmans, R. J. (2023). *terra: Spatial Data Analysis* (Version 1.7-39) [Computer software]. https://CRAN.R-project.org/package=terra

Hijmans, R. J., Phillips, S., Leathwick, J., & Elith, J. (2010). *dismo: Species Distribution Modeling* (p. 1.3-16) [Dataset]. https://doi.org/10.32614/CRAN.package.dismo

Hinzman, L. D., Yoshikawa, K., & Kane, D. L. (2005). Permafrost Controls on Hydrological Processes in a Changing Climate. *Impacts of Global Climate Change*, 1–2. https://doi.org/10.1061/40792(173)255

Hjort, J., Streletskiy, D., Doré, G., Wu, Q., Bjella, K., & Luoto, M. (2022). Impacts of permafrost degradation on infrastructure. *Nature Reviews Earth & Environment*, *3*(1), 24–38. https://doi.org/10.1038/s43017-021-00247-8

Hothorn, T., Hornik, K., Strobl, C., & Zeileis, A. (2023). *party: A Laboratory for Recursive Partytioning* (Version 1.3-14) [Computer software]. https://CRAN.R-project.org/package=party

Hsu, K., Gupta, H. V., & Sorooshian, S. (1995). Artificial Neural Network Modeling of the Rainfall-Runoff Process. *Water Resources Research*, *31*(10), 2517–2530. https://doi.org/10.1029/95WR01955

Hugelius, G., Strauss, J., Zubrzycki, S., Harden, J. W., Schuur, E. A. G., Ping, C.-L., Schirrmeister, L., Grosse, G., Michaelson, G. J., Koven, C. D., O'Donnell, J. A., Elberling, B., Mishra, U., Camill, P., Yu, Z., Palmtag, J., & Kuhry, P. (2014). Estimated stocks of circumpolar permafrost carbon with quantified uncertainty ranges and identified data gaps. *Biogeosciences*, *11*(23), 6573–6593. https://doi.org/10.5194/bg-11-6573-2014

Intergovernmental Panel On Climate Change. (2023a). *Climate Change 2021 – The Physical Science Basis: Working Group I Contribution to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change* (1st ed.). Cambridge University Press. https://doi.org/10.1017/9781009157896

Intergovernmental Panel On Climate Change. (2023b). *Climate Change 2022 – Impacts, Adaptation and Vulnerability: Working Group II Contribution to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change* (1st ed.). Cambridge University Press. https://doi.org/10.1017/9781009325844

Jemeļjanova, M., Kmoch, A., & Uuemaa, E. (2024). Adapting machine learning for environmental spatial data—A review. *Ecological Informatics*, *81*, 102634. https://doi.org/10.1016/j.ecoinf.2024.102634

Jin, H., Huang, Y., Bense, V. F., Ma, Q., Marchenko, S. S., Shepelev, V. V., Hu, Y., Liang, S., Spektor, V. V., Jin, X., Li, X., & Li, X. (2022). Permafrost Degradation and Its Hydrogeological Impacts. *Water*, *14*(3), 372. https://doi.org/10.3390/w14030372

Johnson, K. D., Harden, J. W., David McGuire, A., Clark, M., Yuan, F., & Finley, A. O. (2013). Permafrost and organic layer interactions over a climate gradient in a discontinuous permafrost zone. *Environmental Research Letters*, *8*(3), 035028. https://doi.org/10.1088/1748-9326/8/3/035028

Julián, A., & Chueca, J. (2007). Permafrost distribution from BTS measurements (Sierra de Telera, Central Pyrenees, Spain): Assessing the importance of solar radiation in a mid-elevation shaded mountainous area. *Permafrost and Periglacial Processes*, *18*(2), 137–149. https://doi.org/10.1002/ppp.576

Kanevskiy, M., Shur, Y., Krzewinski, T., & Dillon, M. (2013). Structure and properties of ice-rich permafrost near Anchorage, Alaska. *Cold Regions Science and Technology*, *93*, 1–11. https://doi.org/10.1016/j.coldregions.2013.05.001

Karunanithi, N., Grenney, W. J., Whitley, D., & Bovee, K. (1994). Neural Networks for River Flow Prediction. *Journal of Computing in Civil Engineering*, *8*(2), 201–220. https://doi.org/10.1061/(ASCE)0887-3801(1994)8:2(201)

Keuper, F., Van Bodegom, P. M., Dorrepaal, E., Weedon, J. T., Van Hal, J., Van Logtestijn, R. S. P., & Aerts, R. (2012). A frozen feast: Thawing permafrost increases plant-available nitrogen in subarctic peatlands. *Global Change Biology*, *18*(6), 1998–2007. https://doi.org/10.1111/j.1365-2486.2012.02663.x

Kopczewska, K. (2022). Spatial machine learning: New opportunities for regional science. *The Annals of Regional Science*, *68*(3), 713–755. https://doi.org/10.1007/s00168-021-01101-x

Kuhn, M. (2008). Building Predictive Models in *R* Using the **caret** Package. *Journal of Statistical Software*, *28*(5), 1–26. https://doi.org/10.18637/jss.v028.i05

Lacelle, D., Fisher, D. A., Verret, M., & Pollard, W. (2022). Improved prediction of the vertical distribution of ground ice in Arctic-Antarctic permafrost sediments. *Communications Earth & Environment*, *3*(1), 31. https://doi.org/10.1038/s43247-022-00367-z

Lawrence, D. M., & Slater, A. G. (2005). A projection of severe near-surface permafrost degradation during the 21st century. *Geophysical Research Letters*, *32*(24). https://doi.org/10.1029/2005GL025080

Li, Z.-C., Sun, W.-B., Liang, C.-X., Xing, X.-H., & Li, Q.-X. (2023). Arctic warming trends and their uncertainties based on surface temperature reconstruction under different sea ice extent scenarios. *Advances in Climate Change Research*, *14*(3), 335–346. https://doi.org/10.1016/j.accre.2023.06.003

Liaw, A., & Wiener, M. (2022). *randomForest: Breiman and Cutler's Random Forests for Classification and Regression* (Version 4.7-1.1) [Computer software]. https://cran.r-project.org/package=randomForest

Loeb, N. A., Crawford, A., Stroeve, J. C., & Hanesiak, J. (2022). Extreme Precipitation in the Eastern Canadian Arctic and Greenland: An Evaluation of Atmospheric Reanalyses. *Frontiers in Environmental Science*, *10*. https://doi.org/10.3389/fenvs.2022.866929

Lorax Environmental Services Ltd. (2017, March 3). *Lorax Environmental Services Ltd, 2017. Coffee gold project – hydro-meteorology baseline report. Project No. A405–3*. https://yesabregistry.ca/projects/00036ef5-c28e-46be-bfc3-4c48808d9c3c/documents

Lowey, G. W. (2006). The origin and evolution of the Klondike goldfields, Yukon, Canada. *Ore Geology Reviews*, *28*(4), 431–450. https://doi.org/10.1016/j.oregeorev.2005.03.007

Luetschg, M., Lehning, M., & Haeberli, W. (2008). A sensitivity study of factors influencing warm/thin permafrost in the Swiss Alps. *Journal of Glaciology*, *54*(187), 696–704. https://doi.org/10.3189/002214308786570881

MacWilliam, K. R. G. (2018). *The geology and genesis of the Coffee gold deposit in west-central Yukon, Canada: Implications for the structural, magmatic, and metallogenic evolution of the Dawson Rage, and gold deposit models*. https://doi.org/10.14288/1.0371172

Matthes, H., Damseaux, A., Westermann, S., Beer, C., Boone, A., Burke, E., Decharme, B., Genet, H., Jafarov, E., Langer, M., Parmentier, F.-J., Porada, P., Gagne-Landmann, A., Huntzinger, D., Rogers, B. M., Schädel, C., Stacke, T., Wells, J., & Wieder, W. R. (2025). Advances in Permafrost Representation: Biophysical Processes in Earth System Models and the Role of Offline Models. *Permafrost and Periglacial Processes*, *36*(2), 302–318. https://doi.org/10.1002/ppp.2269

McGovern, A., Lagerquist, R., Gagne, D. J., Jergensen, G. E., Elmore, K. L., Homeyer, C. R., & Smith, T. (2019). Making the Black Box More Transparent: Understanding the Physical Implications of Machine Learning. *Bulletin of the American Meteorological Society*, *100*(11), 2175–2199. https://doi.org/10.1175/BAMS-D-18-0195.1

McKenna, K. M., & Lipovsky, P. S. (2014). *Surficial geology, Dawson Region, Yukon, parts of NTS 115O/14 & 15 and 116B/1, 2, 3, & 4* [Map]. Yukon Geological Survey, Energy, Mines and Resources, Government of Yukon. https://data.geology.gov.yk.ca/Reference/68408#CompilationsTab

McKenzie, J. M., Kurylyk, B. L., Walvoord, M. A., Bense, V. F., Fortier, D., Spence, C., & Grenier, C. (2021). Invited perspective: What lies beneath a changing Arctic? *The Cryosphere*, *15*(1), 479–484. https://doi.org/10.5194/tc-15-479-2021

Milborrow, S. (2024). *earth: Multivariate Adaptive Regression Splines* (Version 5.3.3) [Computer software]. https://cran.r-project.org/package=earth

Molnar, C. (2025). *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable.* (3rd ed.). [Online book]. https://christophm.github.io/interpretable-ml-book/

Muñoz-Sabater, J., Dutra, E., Agustí-Panareda, A., Albergel, C., Arduini, G., Balsamo, G., Boussetta, S., Choulga, M., Harrigan, S., Hersbach, H., Martens, B., Miralles, D. G., Piles, M., Rodríguez-Fernández, N. J., Zsoter, E., Buontempo, C., & Thépaut, J.-N. (2021). ERA5-Land: A state-of-the-art global reanalysis dataset for land applications. *Earth System Science Data*, *13*(9), 4349–4383. https://doi.org/10.5194/essd-13-4349-2021

Natural Resources Canada. (2013). *Canadian Digital Elevation Model* (1.1) [Map]. Government of Canada. https://open.canada.ca/data/en/dataset/7f245e4d-76c2-4caa-951a-45d1d2051333

Natural Resources Canada. (2022). *2020 Land Cover of Canada* [Map]. Government of Canada. https://open.canada.ca/data/en/dataset/ee1580ab-a23d-4f86-a09b-79763677eb47

Nelson, F. E., & Outcalt, S. I. (1987). A Computational Method for Prediction and Regionalization of Permafrost. *Arctic and Alpine Research*, *19*(3), 279–288. https://doi.org/10.1080/00040851.1987.12002602

Obu, J., Westermann, S., Bartsch, A., Berdnikov, N., Christiansen, H. H., Dashtseren, A., Delaloye, R., Elberling, B., Etzelmüller, B., Kholodov, A., Khomutov, A., Kääb, A., Leibman, M. O., Lewkowicz, A. G., Panda, S. K., Romanovsky, V., Way, R. G., Westergaard-Nielsen, A., Wu, T., … Zou, D. (2019). Northern Hemisphere permafrost map based on TTOP modelling for 2000–2016 at 1 km2 scale. *Earth-Science Reviews*, *193*, 299–316. https://doi.org/10.1016/j.earscirev.2019.04.023

O'Donnell, J. A., Carey, M. P., Koch, J. C., Baughman, C., Hill, K., Zimmerman, C. E., Sullivan, P. F., Dial, R., Lyons, T., Cooper, D. J., & Poulin, B. A. (2024). Metal mobilization from thawing permafrost to aquatic ecosystems is driving rusting of Arctic streams. *Communications Earth & Environment*, *5*(1), 268. https://doi.org/10.1038/s43247-024-01446-z

O'Neill, H. B., & Burn, C. R. (2017). Impacts of variations in snow cover on permafrost stability, including simulated snow management, Dempster Highway, Peel Plateau, Northwest Territories. *Arctic Science*, *3*(2), 150–178. https://doi.org/10.1139/as-2016-0036

O'Neill, H. B., Wolfe, S. A., & Duchesne, C. (2020). *Ground ice map of Canada* (No. 8713; version 1, p. 8713). https://doi.org/10.4095/326885

Pachepsky, Ya. A., Timlin, D. J., & Rawls, W. J. (2001). Soil Water Retention as Related to Topographic Variables. *Soil Science Society of America Journal*, *65*(6), 1787–1795. https://doi.org/10.2136/sssaj2001.1787

Paola, J. D., & Schowengerdt, R. A. (1994). Comparisons of neural networks to standard techniques for image classification and correlation. *Proceedings of IGARSS '94 - 1994 IEEE International Geoscience and Remote Sensing Symposium*, *3*, 1404–1406. https://doi.org/10.1109/IGARSS.1994.399452

Pastick, N. J., Jorgenson, M. T., Wylie, B. K., Nield, S. J., Johnson, K. D., & Finley, A. O. (2015). Distribution of near-surface permafrost in Alaska: Estimates of present and future conditions. *Remote Sensing of Environment*, *168*, 301–315. https://doi.org/10.1016/j.rse.2015.07.019

Pastick, N. J., Jorgenson, M. T., Wylie, B. K., Rose, J. R., Rigge, M., & Walvoord, M. A. (2014). Spatial variability and landscape controls of near-surface permafrost within the Alaskan

Yukon River Basin. *Journal of Geophysical Research: Biogeosciences*, *119*(6), 1244–1265. https://doi.org/10.1002/2013JG002594

Payette, S., Delwaide, A., Caccianiga, M., & Beauchemin, M. (2004). Accelerated thawing of subarctic peatland permafrost over the last 50 years. *Geophysical Research Letters*, *31*(18). https://doi.org/10.1029/2004GL020358

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.*, *12*(null), 2825–2830.

Perlich, C., Provost, F., & Simonoff, J. S. (2003). Tree Induction vs. Logistic Regression: A Learning-Curve Analysis. *Journal of Machine Learning Research*, *4*, 211–255.

Pernov, J. B., Gros-Daillon, J., & Schmale, J. (2024). Comparison of selected surface level ERA5 variables against in-situ observations in the continental Arctic. *Quarterly Journal of the Royal Meteorological Society*, *150*(761), 2123–2146. https://doi.org/10.1002/qj.4700

Perry, G. L. W., & Dickson, M. E. (2018). Using Machine Learning to Predict Geomorphic Disturbance: The Effects of Sample Size, Sample Prevalence, and Sampling Strategy. *Journal of Geophysical Research: Earth Surface*, *123*(11), 2954–2970. https://doi.org/10.1029/2018JF004640

Pishvaei, M. H., Sabzevari, T., Noroozpour, S., & Mohammadpour, R. (2020). Effects of hillslope geometry on spatial infiltration using the TOPMODEL and SCS-CN models. *Hydrological Sciences Journal*, *65*(2), 212–226. https://doi.org/10.1080/02626667.2019.1686636

Porter, C., Howat, I., Noh, M.-J., Husby, E., Khuvis, S., Danish, E., Tomko, K., Gardiner, J., Negrete, A., Yadav, B., Klassen, J., Kelleher, C., Cloutier, M., Bakker, J., Enos, J., Arnold, G., Bauer, G., & Morin, P. (2023). *ArcticDEM - Mosaics, Version 4.1* [Dataset]. Harvard Dataverse. https://doi.org/10.7910/DVN/3VDC4W

Ramezan, C. A., Warner, T. A., Maxwell, A. E., & Price, B. S. (2021). Effects of Training Set Size on Supervised Machine-Learning Land-Cover Classification of Large-Area High-Resolution Remotely Sensed Data. *Remote Sensing*, *13*(3), 368. https://doi.org/10.3390/rs13030368

Reid, K. A., Reid, D. G., & Brown, C. D. (2022). Patterns of vegetation change in Yukon: Recent findings and future research in dynamic subarctic ecosystems. *Environmental Reviews*, *30*(3), 380–401. https://doi.org/10.1139/er-2021-0110

Richardson, A. D., Berlyn, G. P., & Duigan, S. P. (2003). Reflectance of Alaskan black spruce and white spruce foliage in relation to elevation and latitude. *Tree Physiology*, *23*(8), 537–544. https://doi.org/10.1093/treephys/23.8.537

Samuel, A. L. (1959). Some Studies in Machine Learning Using the Game of Checkers. *IBM Journal of Research and Development*, *3*(3), 210–229. https://doi.org/10.1147/rd.33.0210

Schaefer, K., Elshorbany, Y., Jafarov, E., Schuster, P. F., Striegl, R. G., Wickland, K. P., & Sunderland, E. M. (2020). Potential impacts of mercury released from thawing permafrost. *Nature Communications*, *11*(1), 4650. https://doi.org/10.1038/s41467-020-18398-5

Schaefer, K., Lantuit, H., Romanovsky, V. E., Schuur, E. A. G., & Witt, R. (2014). The impact of the permafrost carbon feedback on global climate. *Environmental Research Letters*, *9*(8), 085003. https://doi.org/10.1088/1748-9326/9/8/085003

Schetselaar, A., Andersen, T., & Burn, C. R. (2023). Performance of Climate Projections for Yukon and Adjacent Northwest Territories, 1991 – 2020. *ARCTIC*, *76*(3), Article 3. https://doi.org/10.14430/arctic77263

Schneider von Deimling, T., Lee, H., Ingeman-Nielsen, T., Westermann, S., Romanovsky, V., Lamoureux, S., Walker, D. A., Chadburn, S., Trochim, E., Cai, L., Nitzbon, J., Jacobi, S., & Langer, M. (2021). Consequences of permafrost degradation for Arctic infrastructure – bridging the model gap between regional and engineering scales. *The Cryosphere*, *15*(5), 2451–2471. https://doi.org/10.5194/tc-15-2451-2021

Schuur, E. A. G., & Mack, M. C. (2018). Ecological Response to Permafrost Thaw and Consequences for Local and Global Ecosystem Services. *Annual Review of Ecology, Evolution, and Systematics*, *49*(Volume 49, 2018), 279–301. https://doi.org/10.1146/annurev-ecolsys-121415-032349

Schuur, E. A. G., McGuire, A. D., Schädel, C., Grosse, G., Harden, J. W., Hayes, D. J., Hugelius, G., Koven, C. D., Kuhry, P., Lawrence, D. M., Natali, S. M., Olefeldt, D., Romanovsky, V. E., Schaefer, K., Turetsky, M. R., Treat, C. C., & Vonk, J. E. (2015). Climate change and the permafrost carbon feedback. *Nature*, *520*(7546), 171–179. https://doi.org/10.1038/nature14338

Seligman, M. (2024). *Rborist: Extensible, Parallelizable Implementation of the Random Forest Algorithm* (Version 0.3-7) [Computer software]. https://cran.r-project.org/package=Rborist

Shetty, S., Gupta, P. K., Belgiu, M., & Srivastav, S. K. (2021). Assessing the Effect of Training Sampling Design on the Performance of Machine Learning Classifiers for Land Cover Mapping Using Multi-Temporal Remote Sensing Data and Google Earth Engine. *Remote Sensing*, *13*(8), 1433. https://doi.org/10.3390/rs13081433

Simm, J., & de Abril, I. M. (2014). *extraTrees: Extremely Randomized Trees (ExtraTrees) Method for Classification and Regression* (Version 1.0.5) [Computer software]. https://cran.r-project.org/package=extraTrees

Skierszkan, E. K., Dockrey, J. W., Helsen, J., Findlater, L.-L., Bataille, C. P., De Laplante, G., McBeth, J. M., Mayer, K. U., & Beckie, R. D. (2021). Persistence of Uranium in Old and Cold Subpermafrost Groundwater Indicated by Linking ^{234}U- ^{235}U- ^{238}U, Groundwater Ages, and Hydrogeochemistry. *ACS Earth and Space Chemistry*, *5*(12), 3474–3487. https://doi.org/10.1021/acsearthspacechem.1c00307

Skierszkan, E. K., Dockrey, J. W., & Lindsay, M. B. J. (2024). Metal Mobilization from Thawing Permafrost Is an Emergent Risk to Water Resources. *ACS ES&T Water*, *5*(1), 20–32. https://doi.org/10.1021/acsestwater.4c00789

Skierszkan, E. K., Dockrey, J. W., Mayer, K. U., & Beckie, R. D. (2020). Release of geogenic uranium and arsenic results in water-quality impacts in a subarctic permafrost region of granitic and metamorphic geology. *Journal of Geochemical Exploration*, *217*, 106607. https://doi.org/10.1016/j.gexplo.2020.106607

Skierszkan, E. K., Schoepfer, V. A., Fellwock, M. D., Dockrey, J. W., Hayatifar, A., Bondici, V. F., McBeth, J. M., & Lindsay, M. B. J. (2024). Arsenic Mobilization from Thawing Permafrost. *ACS Earth and Space Chemistry*, *8*(4), 745–759. https://doi.org/10.1021/acsearthspacechem.3c00355

Smith, S. L., O'Neill, H. B., Isaksen, K., Noetzli, J., & Romanovsky, V. E. (2022). The changing thermal state of permafrost. *Nature Reviews Earth & Environment*, *3*(1), 10–23. https://doi.org/10.1038/s43017-021-00240-1

Stefan, J. (1891). Ueber die Theorie der Eisbildung, insbesondere über die Eisbildung im Polarmeere. *Annalen Der Physik*, *278*(2), 269–286. https://doi.org/10.1002/andp.18912780206

Stuenzi, S. M., Boike, J., Cable, W., Herzschuh, U., Kruse, S., Pestryakova, L. A., Schneider Von Deimling, T., Westermann, S., Zakharov, E. S., & Langer, M. (2021). Variability of the surface energy balance in permafrost-underlain boreal forest. *Biogeosciences*, *18*(2), 343–365. https://doi.org/10.5194/bg-18-343-2021

Sugiyama, M., Krauledat, M., & Müller, K.-R. (2007). Covariate Shift Adaptation by Importance Weighted Cross Validation. *Journal of Machine Learning Research*, *8*, 985–1005.

Tao, J., Cheng, X., Zheng, L., Xiao, X.-X., Zhong, X.-Y., Liang, Q., Zhang, Z.-Q., & Lin, H. (2023). Performance of climate reanalyses in the determination of pan-Arctic terrestrial rain-on-snow events. *Advances in Climate Change Research*, *14*(4), 522–536. https://doi.org/10.1016/j.accre.2023.08.002

Tarbier, B., Hugelius, G., Kristina Sannel, A. B., Baptista-Salazar, C., & Jonsson, S. (2021). Permafrost Thaw Increases Methylmercury Formation in Subarctic Fennoscandia. *Environmental Science & Technology*, *55*(10), 6710–6717. https://doi.org/10.1021/acs.est.0c04108

Tarnocai, C., Canadell, J. G., Schuur, E. a. G., Kuhry, P., Mazhitova, G., & Zimov, S. (2009). Soil organic carbon pools in the northern circumpolar permafrost region. *Global Biogeochemical Cycles*, *23*(2). https://doi.org/10.1029/2008GB003327

Tetra Tech EBA. (2016, May 2). *Coffee gold project environmental baseline report, mine site study area: Surficial geology, permafrost, and terrain stability.* https://emrlibrary.gov.yk.ca/minerals/MajorMines/coffee/project-proposal/appendices/11-A_Coffee_20170331_Baseline_Surficial_Geol_Perm_Terrain.pdf

Thaler, E. A., Uhleman, S., Rowland, J. C., Schwenk, J., Wang, C., Dafflon, B., & Bennett, K. E. (2023). High-Resolution Maps of Near-Surface Permafrost for Three Watersheds on the Seward Peninsula, Alaska Derived From Machine Learning. *Earth and Space Science*, *10*(12), e2023EA003015. https://doi.org/10.1029/2023EA003015

The R Foundation. (n.d.). *The R Project for Statistical Computing* [Computer software]. Retrieved November 5, 2024, from https://www.r-project.org/

Toohey, R. C., Herman-Mercer, N. M., Schuster, P. F., Mutter, E. A., & Koch, J. C. (n.d.). *Multidecadal increases in the Yukon River Basin of chemical fluxes as indicators of changing flowpaths, groundwater, and permafrost*. Retrieved April 6, 2025, from https://repository.library.noaa.gov/view/noaa/62390

Tóth, J. (1963). A theoretical analysis of groundwater flow in small drainage basins. *Journal of Geophysical Research (1896-1977)*, *68*(16), 4795–4812. https://doi.org/10.1029/JZ068i016p04795

Turetsky, M. R., Abbott, B. W., Jones, M. C., Anthony, K. W., Olefeldt, D., Schuur, E. A. G., Grosse, G., Kuhry, P., Hugelius, G., Koven, C., Lawrence, D. M., Gibson, C., Sannel, A. B. K., & McGuire, A. D. (2020). Carbon release through abrupt permafrost thaw. *Nature Geoscience*, *13*(2), 138–143. https://doi.org/10.1038/s41561-019-0526-0

Valavi, R., Elith, J., Lahoz-Monfort, J. J., & Guillera-Arroita, G. (2019). blockCV: An R package for generating spatially or environmentally separated folds for k-fold cross-validation of species distribution models. *Methods in Ecology and Evolution*, *10*(2), 225–232. https://doi.org/10.1111/2041-210X.13107

Van Stempvoort, D. R., Spoelstra, J., Bickerton, G., Koehler, G., Mayer, B., Nightingale, M., & Miller, J. (2023). Sulfate in streams and groundwater in a cold region (Yukon Territory, Canada): Evidence of weathering processes in a changing climate. *Chemical Geology*, *631*, 121510. https://doi.org/10.1016/j.chemgeo.2023.121510

Vonk, J. E., Tank, S. E., Bowden, W. B., Laurion, I., Vincent, W. F., Alekseychik, P., Amyot, M., Billet, M. F., Canário, J., Cory, R. M., Deshpande, B. N., Helbig, M., Jammet, M., Karlsson, J., Larouche, J., MacMillan, G., Rautio, M., Walter Anthony, K. M., & Wickland, K. P. (2015). Reviews and syntheses: Effects of permafrost thaw on Arctic aquatic ecosystems. *Biogeosciences*, *12*(23), 7129–7167. https://doi.org/10.5194/bg-12-7129-2015

Walvoord, M. A., & Kurylyk, B. L. (2016). Hydrologic Impacts of Thawing Permafrost—A Review. *Vadose Zone Journal*, *15*(6). https://doi.org/10.2136/vzj2016.01.0010

Walvoord, M. A., & Striegl, R. G. (2007). Increased groundwater to stream discharge from permafrost thawing in the Yukon River basin: Potential impacts on lateral export of carbon and nitrogen. *Geophysical Research Letters*, *34*(12), L12402. https://doi.org/10.1029/2007GL030216

Walvoord, M. A., Voss, C. I., & Wellman, T. P. (2012). Influence of permafrost distribution on groundwater flow in the context of climate-driven permafrost thaw: Example from Yukon Flats Basin, Alaska, United States. *Water Resources Research*, *48*(7), 2011WR011595. https://doi.org/10.1029/2011WR011595

Wang, Z., Lai, C., Chen, X., Yang, B., Zhao, S., & Bai, X. (2015). Flood hazard risk assessment model based on random forest. *Journal of Hydrology*, *527*, 1130–1141. https://doi.org/10.1016/j.jhydrol.2015.06.008

Wenger, S. J., & Olden, J. D. (2012). Assessing transferability of ecological models: An underappreciated aspect of statistical validation. *Methods in Ecology and Evolution*, *3*(2), 260–267. https://doi.org/10.1111/j.2041-210X.2011.00170.x

Westermann, S., Barboux, C., Bartsch, A., Delaloye, R., Grosse, G., Heim, B., Hugelius, G., Irrgang, A., Kääb, A. M., Matthes, H., Nitze, I., Pellet, C., Seifert, F. M., Strozzi, T., Wegmüller, U., Wieczorek, M., & Wiesmann, A. (2024). *ESA Permafrost Climate Change Initiative (Permafrost_cci): Permafrost extent for the Northern Hemisphere, v4.0* (p. 25 Files, 320327171 B) [Application/xml]. NERC EDS Centre for Environmental Data Analysis. https://doi.org/10.5285/93444BC1C4364A59869E004BF9BFD94A

Westermann, S., Barboux, C., Bartsch, A., Delaloye, R., Hugelius, G., Heim, B., Irrgang, A., Kääb, A. M., Matthes, H., Seifert, F. M., Wieczorek, M., Wiesmann, A., Wegmüller, U., Strozzi, T., Pellet, C., Nitze, I., & Grosse, G. (2024). *Dataset Collection Record: ESA Permafrost Climate Change Initiative (Permafrost_cci): Permafrost version 4 data products* [Dataset]. NERC EDS Centre for Environmental Data Analysis. https://catalogue.ceda.ac.uk/uuid/7479606004d9465bad949671501e5f21/

Western Copper and Gold. (2024, January). *Casino Project | Proposal for Executive Committee Review*. https://casinomining.com/_resources/YESAA_Project_Proposal/Volume3/7C_2012_Hydrology_Baseline_Report.pdf

Wild, G. O. (1882). *Air Temperature in the Russian Empire*. Izdatel'stvo Russkogo Geograficheskogo Obshestva.

Williams, J. K., Ahijevych, D., Dettling, S., & Steiner, M. (2008). *Combining observations and model data for short-term storm forecasting* (W. F. Feltz & J. J. Murray, Eds.; p. 708805). https://doi.org/10.1117/12.795737

Wright, M. N., & Ziegler, A. (2017). ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *Journal of Statistical Software*, *77*(1). https://doi.org/10.18637/jss.v077.i01

Yang, Y., Wang, X., & Wang, T. (2024). Permafrost Degradation Induces the Abrupt Changes of Vegetation NDVI in the Northern Hemisphere. *Earth's Future*, *12*(10), e2023EF004309. https://doi.org/10.1029/2023EF004309

YESAB. (2022, March 4). *Yukon Environmental & Socio-economic Assessment Act Decision Document – Executive Committee Screening*. Yukon Environmental and Socio-economic Assessment Board. https://yesabregistry.ca/projects/00036ef5-c28e-46be-bfc3-4c48808d9c3c/documents

YESAB. (2024). *Yukon Environmental and Socio-economic Assessment Board Registry*. YESAB Registry. https://yesabregistry.ca/

Yukon Geological Survey. (2020). *A digital atlas of terranes for the northern Cordillera* [Map]. Yukon Geological Survey, Energy, Mines and Resources, Government of Yukon. https://data.geology.gov.yk.ca/Compilation/2

Yukon Geological Survey, Energy, Mines and Resources, Government of Yukon. (2025). *Yukon Permafrost Database*. Yukon Permafrost Database. https://service.yukon.ca/permafrost

Zhang, J., Ding, J., & Yang, Y. (2022). *BAGofT: A Binary Regression Adaptive Goodness-of-Fit Test (BAGofT)* (Version 1.0.0) [Computer software]. https://cran.r-project.org/package=BAGofT

Zhang, T., Barry, R. G., Knowles, K., Heginbottom, J. A., & Brown, J. (2008). Statistics and characteristics of permafrost and ground-ice distribution in the Northern Hemisphere. *Polar Geography*, *31*(1–2), 47–68. https://doi.org/10.1080/10889370802175895