

SOCIAL MEDIA DATA AND COMPUTER VISION IN SOCIAL IMPACT  
ASSESSMENT: UNDERSTANDING HUMAN DIMENSIONS AND CULTURAL  
ECOSYSTEM SERVICES IN HYDROELECTRIC LANDSCAPES

by

Yan Chen

Submitted in partial fulfilment of the requirements  
for the degree of Doctor of Philosophy

at

Dalhousie University

Halifax, Nova Scotia

August 2024

© Copyright by Yan Chen, 2024

## Table of Contents

<b>List of Tables</b> .....	v
<b>List of Figures</b> .....	vi
<b>Abstract</b> .....	vii
<b>List of Abbreviations and Symbols Used</b> .....	viii
<b>Acknowledgements</b> .....	ix
<b>Chapter 1 Introduction</b> .....	1
<b>1.1 Rationale</b> .....	1
<b>1.2 Research objectives</b> .....	3
<b>1.3 An interdisciplinary study</b> .....	3
<i>1.3.1 Energy Transitions and Energy Landscapes</i> .....	4
<i>1.3.2 Social Media Data</i> .....	4
<i>1.3.3 Artificial Intelligence and Computer Vision</i> .....	5
<b>1.4 Methods</b> .....	5
<i>1.4.1 Theory</i> .....	6
<i>1.4.2 Models</i> .....	6
<i>1.4.3 Hydroelectric dam cases</i> .....	7
<b>1.5 Organization of dissertation</b> .....	8
<b>Chapter 2 From theory to practice: Insights and hurdles in collecting social media data for social science research</b> .....	10
<b>2.1 Introduction</b> .....	12
<b>2.2 Methods</b> .....	14
<b>2.3 Results</b> .....	16
<i>2.3.1 Data philanthropy organizations</i> .....	16
<i>2.3.2 Data repositories</i> .....	16
<i>2.3.3 Data donation</i> .....	17
<i>2.3.4 Third-party data companies</i> .....	17
<i>2.3.5 Homegrown tools</i> .....	17
<i>2.3.6 Web scraping tools – commercial</i> .....	18
<i>2.3.7 Web scraping tools – non-commercial</i> .....	18
<i>2.3.8 Web Scraping scripts (single-purpose)</i> .....	18

<b>2.4 Discussion: A better solution?</b> .....	19
<b>Chapter 3 Image auto-coding tools for social impact assessment: Leveraging social media data to understand human dimensions of hydroelectricity landscape changes in Canada</b> .....	23
<b>3.1 Introduction</b> .....	25
<b>3.2 Methods</b> .....	27
3.2.1 <i>Study areas</i> .....	27
3.2.2 <i>Data Collection</i> .....	29
3.2.3 <i>Landscape image auto-coding</i> .....	31
3.2.4 <i>Topic Clustering</i> .....	32
<b>3.3 Results</b> .....	33
3.3.1 <i>Descriptive Statistics</i> .....	33
3.3.2 <i>LDA landscape clusters</i> .....	34
3.3.3 <i>Energy landscape</i> .....	35
<b>3.4 Discussion</b> .....	38
3.4.1 <i>Key landscape features and activities</i> .....	38
3.4.1.1 <i>Dams increase the prominence of water in the landscape</i> .....	38
3.4.1.2 <i>Losing landscape features and traditional practices to dams</i> .....	39
3.4.1.3 <i>Lifestyle can be reshaped by damming</i> .....	40
3.4.1.4 <i>Hydropower infrastructure has limited visual impact</i> .....	41
3.4.2 <i>Implications for study cases</i> .....	42
3.4.3 <i>Opportunities and limitations of AI tools to analyze social media data for SIA</i> .....	43
3.4.3.1 <i>Opportunities</i> .....	43
3.4.3.2 <i>Limitations</i> .....	43
<b>3.5 Conclusion</b> .....	45
<b>Chapter 4 Using computer vision to assess cultural ecosystem services relating to hydropower landscapes in Canada</b> .....	47
<b>4.1 Introduction</b> .....	49
<b>4.2 Methods</b> .....	52
4.2.1 <i>Study areas</i> .....	52
4.2.2 <i>Data collection</i> .....	52

4.2.3 CES coding themes and training dataset preparation .....	54
4.2.4 Model training and performance, and CES prediction.....	58
<b>4.3 Results .....</b>	<b>59</b>
4.3.1 Model training results .....	59
4.3.2 CES prediction results.....	59
4.3.2.1 Comparison among study areas .....	59
4.3.2.2 Results for Mactaquac.....	61
<b>4.4 Discussion.....</b>	<b>62</b>
4.4.1 CES provision .....	62
4.4.2 Implications for hydroelectric landscapes .....	65
4.4.3 CES assessment via social media images and custom-trained CV tools .....	66
<b>4.5 Conclusions.....</b>	<b>70</b>
<b>Chapter 5 Conclusion .....</b>	<b>71</b>
5.1 SIA methodological innovation .....	71
5.2 Hydroelectric social impacts .....	74
5.3 Summary.....	75
<b>References .....</b>	<b>76</b>
<b>Appendix A: Copyright Permission – Frontiers .....</b>	<b>97</b>
<b>Appendix B: Copyright Permission – Elsevier.....</b>	<b>100</b>
<b>Appendix C: Location name and ID for data collection (X: longitude, Y: latitude).....</b>	<b>101</b>
<b>Appendix D: Latent Dirichlet Allocation model for label analysis.....</b>	<b>102</b>
<b>Appendix E: Codebook for Coding Themes.....</b>	<b>105</b>
<b>Appendix F: Prediction results for each study areas by geo-tags.....</b>	<b>107</b>

## List of Tables

Table 2.1	Advantages and limitations of social media data collection approaches. ....	15
Table 3.1	Landscape topic clusters across cases. Asterisks denote the weight of the words in the topic: *** >0.1, **>0.05, *>0.025. ....	36
Table 4.1	Coding themes and cultural ecosystem services (adapted from MEA, 2005; more details can be found in Appendix E). ....	57
Table 4.2	Average percentage of coding themes in study areas (top 3 bolded for each case; highest site for each coding theme shaded). Note that the columns will not sum to 1 because images could be categorized in more than one theme (or none). ....	61

## List of Figures

Figure 3.1 Study areas (adapted from Natural Resources Canada, 2001). .....	27
Figure 3.2 Workflow of data collection (shaded background) and analysis (white background). .....	31
Figure 3.3 An example of Vision API results (the image was generated by ChatGPT, and the Vision API results were generated in February 2024). .....	32
Figure 3.4 Top 15 labels for each study case. ....	34
Figure 4.1 Workflow of data collection (shaded background) and analysis including model training and prediction (white background). .....	54

## **Abstract**

Social media data has proven to be a valuable resource for assessing social impacts, especially when combined with rapid advancements in artificial intelligence technologies. This combination enables comprehensive analyses of larger datasets than traditional methods allow. Additionally, visual content from images can uncover patterns of landscape changes that other data types may miss. However, the exploration of social impacts related to landscapes through social media images has not been sufficiently addressed in current literature or practices. This dissertation assessed the social impacts of hydroelectric dams and their reservoirs by employing two computer vision models to analyze large social media imagery datasets.

Chapter One is an introduction. Chapter Two examined the opportunities and challenges associated with collecting social media image data for research purposes, advocating for a stronger role for governments in access for public good research. Chapter Three utilized a ready-to-use pre-trained computer vision model to label landscape images sourced from Instagram and geo-tagged to study areas. A topic clustering model was applied to find patterns from the labels to interpret the landscape changes and perceptions in areas with pre-dam, 32-year, and 56-year dammed landscapes. The analysis identified common topics such as plant life, water, energy infrastructure, sunset/sunrise, winter scenes, pets and wildlife, people in nature, roads and vehicles, and recreational activities. Chapter Four explored the feasibility of training a computer vision model to classify images based on cultural ecosystem services (CES) coding themes. Two coders categorized a subset of Instagram landscape images into eleven themes, achieving an overall accuracy of 93.8% for model training. The model was subsequently used to predict CES provision for the entire valid dataset of the three study areas. Chapter Five is a conclusion.

The results showed that the expansion of water bodies in the reservoir-based landscape can reshape local lifestyles by increasing water accessibility and recreational opportunities, although it also results in the loss of agricultural lands and traditional cultures. The findings also revealed compelling patterns of aesthetic value, place identity, recreation, and cultural heritage. Damming a place does not necessarily lead to a reduction in values and CES for those who stay or visit. Some social impacts varied from case to case, such as impacts to agriculture. Therefore, it would be beneficial to explore enriching social impact assessment with social media analysis in planning and managing future projects. The findings also suggested that computer vision technologies can significantly enhance the social impact assessment toolkit, revealing meaningful patterns and implications for projects in practice.

## List of Abbreviations and Symbols Used

%	Percent
#	Number
AI	Artificial Intelligence
API	Application Programming Interface
BC	British Columbia (Canadian Province)
CES	Cultural Ecosystem Services
CNN	Convolutional Neural Networks
CV	Computer Vision
DSA	Digital Services Act (European Commission)
EIA	Environmental Impact Assessment
ES	Cultural Ecosystem Services
GIS	Geographic Information System
IP	Internet Protocol
LDA	Latent Dirichlet Allocation
NB	New Brunswick (Canadian Province)
NLP	Natural Language Processing
PPGIS	Public Participation Geographic Information System
SIA	Social Impact Assessment
Vertex AI	Google Cloud Vertex AI
Vision API	Google Cloud Vision API



## **Acknowledgements**

Completing this PhD thesis has been a challenging yet incredibly rewarding journey, and I owe thanks to a great many people throughout it. It would not have been possible without their support, guidance, and encouragement.

First and foremost, I would like to express my deepest gratitude to my supervisors, Dr. Kate Sherren and Dr. Mike Smit, for their invaluable guidance, patience, and unwavering support throughout my research. Their expertise and insightful feedback have been instrumental in shaping this thesis. Meanwhile, they also provided me with the freedom to explore new ideas and helped me to make these attempts become true. Dr. Kate Sherren, thank you for insightful conversations and patient guidance, especially in revising the manuscripts. Dr. Mike Smit, thank you for the immense support during my parental leave.

I would also like to extend my heartfelt thanks to the members of my dissertation committee, Dr. Kyung Young Lee and Dr. Lori McCay-Peet, for their constructive comments and suggestions, which have significantly improved the quality of this work. Additionally, I am grateful to Dr. Mona Holmlund, who was on the committee during the first two years and helped me learn many things.

Special thanks to my colleagues, Keahna Margeson, for helping with preparing the training dataset, and Shan Xue, for building the data collection tool to support my research.

I am grateful to the Interdisciplinary PhD Program at Dalhousie University for providing the facilities and resources necessary for my research. The administrative and technical staff, particularly the former Director Dr. Lynne Robinson and the former secretary Khoi Truong, as well as Dr. Peter Tyedmers and Emma Zang, deserve special mention for their assistance and support. Also, thank you to the School for Resource and Environmental Studies and the Department of Information Science.

I would also like to acknowledge the financial support from the Social Sciences and Humanities Research Council of Canada through Insight Grant 435-2018-1018 (2018-2022, MS as PI, KS CI) and the Nova Scotia Research and Innovation Graduate

Scholarship (2018-2024), which made this research possible. In addition, thanks to Insight Grant 435-2021-0221 (2021-2025, KS as PI) for supporting the publication of one manuscript.

My deepest appreciation goes to my family, especially my parents, my partner, and my daughter, for their unconditional love, encouragement, and sacrifices. Finally, to my friends and loved ones, thank you for your understanding, patience, and unwavering support during this long journey. Your faith in me has kept me going, even in the most challenging times.

This thesis is dedicated to all of you.

# Chapter 1 Introduction

## 1.1 Rationale

In Canada, hydroelectricity is a key source of energy, accounting for over 61% of total energy production and 55% of the total installed generation capacity in 2019 (International Hydropower Association, 2020). Although hydropower offers several environmental benefits, it faces criticism from social and cultural viewpoints. Hydropower stands out among renewable energy sources due to its maturity, large scale, long-term impacts, significant effects on landscapes, and the resulting human resettlements and lifestyle changes (Hough, 1990; Keilty, Beckley, & Sherren, 2016). According to a report by the World Commission on Dams (2001), dam constructions have diverted many rivers and displaced millions of people globally. However, conflicts at the end of a dam's lifespan can be equally controversial. Jørgensen and Renöfält (2012) examined cases in Sweden, revealing that supporters and opponents often frame the impacts of dam removal differently: proponents emphasize ecosystem services and river fishing, while opponents highlight cultural services related to recreation, aesthetics, and heritage. Therefore, understanding the social impact, particularly the long-term and cumulative effects, is essential in the decision-making process for hydroelectric projects (Arnold et al., 2022).

Social Impact Assessment (SIA) has become an effective tool to evaluate such impacts related to projects (Imperiale & Vanclay, 2023; Vanclay & Esteves, 2011). Since its emergence alongside Environmental Impact Assessment (EIA) in the 1970s (Gramling & Freudenburg, 1992), SIA has evolved from a regulatory tool into a research field and a management instrument used by project proponents (Vanclay, 2020). While traditional EIA remains a regulatory requirement in many regions, social impacts have increasingly been integrated into the assessment framework, highlighting a broader societal trend toward greater respect for human rights, as endorsed by the United Nations's (2011) Guiding Principles on Business and Human Rights. Unlike biophysical impacts, social impacts can be complex and prolonged, often beginning even before the development of a mature project proposal. Mere rumors about a project can cause anxiety and fears (Vanclay, 2020). Additionally, the long-term impacts of large public projects, such as

hydroelectric dams, become more pronounced than short-term impacts due to local environmental changes and residential relocations (Gramling & Freudenburg, 1992). Even accepted changes can cause emotional impacts like nostalgia for former landscapes and solastalgia or grief around losses (Galway et al., 2019). In both public and industrial projects, social costs and risks are real and must be effectively managed (Grieco, 2018). Despite the long history of SIA, gaps still exist, as highlighted by previous research. Vanclay and Esteves (2011) identified limitations in traditional SIA methods, including inadequate stakeholder engagement and failures to predict residual impacts and consequent harm to communities. Vanclay (2020) further updated the gaps, noting a limited understanding of the complexities involved in restoring pre-existing livelihoods or implementing alternative ones, the non-market costs of these impacts (such as psychological, social, or cultural costs), and the inadequacy of one-size-fits-all solutions for valuing SIA and compensating for losses. Grieco (2018) pointed out the difficulty of fully grasping the complexities of how a project can socially and culturally affect local communities. These observations underscore the need for improvements in SIA practices. Many practitioners feel they lack the resources to conduct SIA effectively due to barriers to accessing available tools (Grieco, 2018). Additionally, Pimental da Silva et al. (2021) found a lack of innovation in methods and a gap between scholarly literature and SIA practice after reviewing 37 hydroelectric SIA reports.

In the past decade, social media data has proven valuable in various social science research fields, indicating understanding social phenomena and assessing social impacts (Chen et al., 2023). For example, Sottini et al. (2019) used a filtered dataset of 9,304 Flickr photos from 2005 to 2017 to map rural landscapes in Italy, illustrating how changes in infrastructure, crops, and environmental factors affect visitors' use of agricultural land. Likewise, innovations in SIA methods have emerged, leveraging alternative data sources from social media (Sherren et al., 2023). For instance, Chen et al. (2018) utilized Instagram images to study hydroelectric landscapes as experienced by young people. Aaen et al. (2018) highlighted the significant potential of social media to engage citizens in assessing social impacts, though they noted that more efforts are needed to effectively integrate social media into the SIA process.

However, processing and analyzing the vast amount of social media data and their metadata often requires significant human labor. For instance, Cortese et al. (2018) needed six well-trained coders to categorize 5,721 profile images over three months, even though the categories involved simple binary judgments. To address this challenge, artificial intelligence-based image mining and automatic analysis technologies are increasingly essential. The concept of machine learning was introduced in the 1940s, but the challenge of detecting visual data was not overcome until Fukushima's idea of convolutional neural networks (CNNs) in 1980, further improved by LeCun et al. in 1998 (Heaton, 2015). Models based on CNNs, commonly referred to as computer vision, are now increasingly used to classify and code images in social science studies (Karpathy & Li, 2017; Vinyals et al., 2016). Many commercial companies, such as Google and Microsoft, offer pre-trained and customizable models that social scientists can use. However, neither social media data nor computer vision technology has been fully explored for SIA purposes.

## **1.2 Research objectives**

This dissertation has two main objectives: firstly, to explore methodological innovation for SIA by retrieving alternative data sources and applying computer vision-based analysis models; and secondly, to understand social impacts caused by large hydroelectric dams and their reservoirs, providing insights for decision-making processes in similar projects.

## **1.3 An interdisciplinary study**

This study spans multiple disciplines, including SIA, energy landscapes, data mining and analysis, and computer vision (a subfield of artificial intelligence). With the widespread use of social media platforms and recent advancements in artificial intelligence, it has become possible to innovate new research methods by using alternative data sources and automatic analysis tools for large datasets. These new methods can offer fresh perspectives on landscape, renewable energy, and SIA research due to the unique characteristics of social media content, such as its younger demographic, online behaviors, and user-generated posts. Therefore, an interdisciplinary approach was

necessary to explore how technological advancements can drive innovation in traditional SIA.

### *1.3.1 Energy Transitions and Energy Landscapes*

The world is facing a rapid increase in energy demand, exacerbated by recent supply disruptions due to the Russian invasion of Ukraine and the lingering effects of the pandemic (International Energy Agency, 2022). Conventional energy sources like coal and fossil fuels are plagued by depletion and significant environmental impacts, including pollution and climate change. Renewable energy sources, such as hydroelectricity, wind, and solar power, are seen as the future of energy. However, while these sources may cause less harm to the biophysical environment, if compared to fossil fuels, they can also cause social disturbances and impacts such as visual disruption, displacement, changes in landscape use, the loss of traditional lifestyles and practices, and social detachment (Kirchherr & Charles, 2016). Therefore, understanding these social impacts is crucial for informing the decision-making process.

### *1.3.2 Social Media Data*

Despite being less than 20 years old, social media has rapidly developed into a valuable and rich data source for research, particularly in social sciences such as sociology, politics, public health, and environmental studies (Chen et al., 2023). The value of social media data lies in three key aspects: first, it represents a younger cohort that is often difficult to access through conventional research methods but who will be impacted longest by energy decisions (Chen et al., 2018); second, the data is not created for research purposes and is thus less influenced by researcher bias (unlike interviews and surveys where participants can be guided by preset questions and answers) (Li et al., 2019); third, it offers a larger dataset at a lower cost compared to traditional approaches and could offset collapses in response rates for those traditional approaches (Azevedo et al., 2022; Stedman, 2019). Exploring social media data, therefore, holds significant potential for research innovation. However, after the Cambridge Analytica Scandal in 2018 (Confessore, 2018), many social media platforms have restricted their data access (Freelon, 2018). Although some platforms (e.g., Meta, X, TikTok) have opened access for research purposes more recently, there are still limitations regarding regions, study topics,

and data amount and types (Chen et al., 2024). Using social media for social impact assessment will require a reliable supply for public good inquiry.

### *1.3.3 Artificial Intelligence and Computer Vision*

Recent advancements in artificial intelligence have created numerous opportunities in data analysis, applicable across various research and practical areas (Johnson et al., 2021; Manley et al., 2022). While computer vision has seen significant improvements, it still faces challenges that require further experimentation to fully explore its capabilities (Zou & Schiebinger, 2018). Unlike language processing models, where texts are more organized and information-dense, visual materials like images often contain noise, such as irrelevant objects in the background. This noise raises concerns about the accuracy of computer vision models, particularly for more subjective concepts (Vigl et al., 2021). Testing different computer vision models to understand lifestyles and social impacts in hydroelectric landscapes contributes to methodological innovation, such as the expansion of the social impact assessment toolkit.

## **1.4 Methods**

The methods in this dissertation include data access, data filtering, and data analysis applied in three hydroelectricity case studies across Canada. Data access was subject to challenges in retrieving social media data due to API closures after 2018, toward which we tested eight different data collection approaches, offering experiences and insights valuable to other scholars. Our final approach involved two-step custom designed web scraping scripts to extract Instagram data for case study areas. Instagram is a platform where people share their day-to-day activities, by contrast with tourist photography sites like Flickr or opinion-driven sites like X. Extensive data filtering is required in social media studies given the noisy dataset, and in our case included manual filtering of landscape-based Instagram posts. It was conducted because the accuracy of computer vision was insufficient to achieve this task due to the complexity of the data and the limitations of machine power in 2021.

Data analysis includes two computer vision models, the pre-trained and the custom-trained models, to assess social impacts related to hydroelectric dams focusing on human

dimensions of the landscape and the provision of CES within the landscape. Mixed methods are employed, including both qualitative and quantitative approaches. Computer vision and natural language processing are quantitative, relying on statistical analysis, while image content coding and the interpretation of clustered topics are qualitative.

#### *1.4.1 Theory*

Landscape perception theory and cultural ecosystem services (CES) are two frameworks utilized to understand social impacts in this dissertation.

Landscape perception theory discusses how landscapes are used and valued by people based on physical features (e.g., trees, waterbodies), activities (e.g., boating, dog walking), and values (e.g., aesthetics, recreation) (Chen et al., 2018; Chen et al., 2019). This was originally inspired by the work of Taylor, Zube, and Sell (1987), who summarized previous studies from the 1960s to 1970s, laying the groundwork for later landscape studies on topics such as landscape preference, aesthetics, symbolism, and sense of place (Jacobsen, 2007). This theory was used in my previous research to understand the social and cultural values of the hydroelectric landscape (Chen et al., 2018; Chen et al., 2019); and, it is now used in this dissertation to interpret the human dimensions labeled by the pre-trained computer vision model, connecting the landscape physical features to the values.

CES is a more recently developed concept used to understand the intangible benefits that an ecosystem provides to people. The CES categories investigated in this study include aesthetic values, sense of place and place identity, recreation, social relations, and cultural heritage value, based on the Millennium Ecosystem Assessment report (2005) and precedent research in similar research areas (e.g., Cardoso et al., 2022; Mouttaki et al., 2022; Richards & Friess, 2015). These CES categories were adapted in this study to identify relevant physical features in photo content and classify images into different CES-based theme coding categories.

#### *1.4.2 Models*



Two different computer vision models, a pre-trained model and a custom-trained model, were used to assess perceived landscapes via landscape perception and CES-related themes, respectively.

A pre-trained model can identify subjective features from landscape images based on millions of preset categories, including biophysical features (e.g., trees, water bodies, mountains) and human appearances and activities (e.g., recreation). In this study, we used the Google Cloud Vision API, a tool known for its ease of use and cost efficiency. However, it has limitations as its labels cannot be customized, and important features may be overlooked if they appear small. This model is useful for identifying a large number of physical feature labels, and the landscape perception theory provides a valuable framework for interpreting these features based on their values. Additionally, a natural language processing model, Latent Dirichlet Allocation, was used to cluster the large set of labels. This revealed prominent and co-occurring features, helping to understand how people perceive, use, and value the landscape.

The custom-trained model required users to provide a training dataset, allowing for customized classification categories based on specific research aims, datasets, and case studies. While this approach demands more initial labor and some knowledge of artificial intelligence, it offers the potential to capture more complex concepts such as CES. In this study, the custom-trained model was developed using Google Cloud Vertex AI. Coders manually assigned images to CES-based coding themes, such as natural landscapes, humans in nature, recreation, social relationships, and historical features. The trained model was then used to predict the valid dataset to understand CES provision in the three study areas.

The rationale for choosing these models was to explore the advantages and limitations of the two different types of computer vision models in practice and to understand how they can help indicate social impacts in different ways.

#### *1.4.3 Hydroelectric dam cases*

This dissertation investigates three study areas across Canada with (or facing) large hydroelectric dams, including the in-progress Site C Dam (Site C) in British Columbia

(BC), the Oldman River Dam (Oldman) built in Alberta in 1991, and the Mactaquac Dam (Mactaquac) built in New Brunswick (NB) in 1968 (more details can be found in chapter 3 and 4). Innovating methods for social impact assessment and leveraging social media data is one of the key objectives of this research, but the earliest social media data available dates to July 18, 2011, making it impossible to use it to study a long-dammed landscape like the 56-year-old Mactaquac. While combining other data types (e.g., archival data, news data) could provide a longitudinal perspective, it would detract from the primary focus on social media data and computer vision technologies. The research approach of comparing three dams at different stages of their lifespan, rather than examining longitudinal data from a single dam, is an alternative to long-term studies. Such space-for-time substitution is described by Pickett (1989) to allow “the extrapolation of a temporal trend from a series of different-aged samples (p.111)”. Thus, comparing three hydroelectric dams at different stages (pre-dam, 32 years old, and 56 years old) is a better approach for understanding potential changes and social impacts over a dam’s lifespan while focusing on online data sources and analysis tools.

## **1.5 Organization of dissertation**

This dissertation is organized into five chapters, with Chapters 2, 3, and 4 presented in their respective journal publication formats. The committee members of this dissertation and other contributors are listed as co-authors for Chapter 2, 3, and 4 as journal submissions. When writing in the first person, therefore, the term “we” is used rather than “I”. I have taken the lead role in research design, data collection, data analysis, results interpretation, and draft writing for all three manuscripts.

Chapter 1 is the introductory chapter, explaining the overarching theme of social impact assessment, the rationale of conducting this research, outlining the research objectives, describing the choice of study areas, and highlighting the interdisciplinary nature of this study.

Chapter 2 examines various approaches to collecting social media data, particularly in the post-API era when platforms have become stricter in their data policies. This chapter discusses the challenges and experiences associated with these approaches, along with their practical pros and cons. It also proposes the creation of a dedicated API for research

purposes to address the issue of data scarcity among scholars. This has been published in *Frontiers in Big Data*.

Chapter 3 employs a pre-trained computer vision model (Google Cloud Vision API) to identify labels in landscape images retrieved from Instagram. These labels are then clustered using a natural language processing model (Latent Dirichlet Allocation). The clustered topics reveal the prominence and co-occurrence of landscape features and human activities in the study areas, showing patterns of change in dammed landscapes and the resulting social impacts. This is in review with *Landscape and Urban Planning*.

Chapter 4 explores the feasibility of using a custom-trained computer vision model to assess cultural ecosystem services. A classification model was trained using Google Cloud Vertex AI by learning from a training dataset of images randomly selected from the valid dataset, assigning them to 11 CES-related coding themes. The model achieved an average precision of 93.8%, demonstrating high accuracy and feasibility. The trained model was then deployed to predict the entire dataset, with coding results identifying CES provision in the study areas and allowing interpretation of the social impacts of damming a landscape. This chapter also discusses the integration of social media data and the custom-trained computer vision model into SIA. This is in preparation for *Ecosystem Services*.

Chapter 5 concludes the dissertation with a summary of the main findings and implications, the contributions of this study, and directions for future research.

## **Chapter 2 From theory to practice: Insights and hurdles in collecting social media data for social science research**

A version of this chapter has been published by Frontiers in the journal *Frontiers in Big Data*. Refer to Appendix A for the copyright agreement to reproduce this material.

Chen, Y., Sherren, K., Lee, K, Y., McCay-Peet, L., & Xue, S., Smit, M. (2024). From theory to practice: Insights and hurdles in collecting social media data for social science research. *Frontiers in Big Data*, 7, 1379921. <https://doi.org/10.3389/fdata.2024.1379921>

Statement of Authors' Contributions: YC is responsible for conducting the experiments of all different data collection approaches, the literature review, and writing all sections. KS and MS supervised the writing development including providing feedback and editorial revisions. KL and LMP provided feedback and editorial revisions on a final draft. SX built the code for Instagram data collection, and YC amended the code.

## **Abstract**

Social media has profoundly changed our modes of self-expression, communication, and participation in public discourse, generating volumes of conversations and content that cover every aspect of our social lives. Social media platforms have thus become increasingly important as data sources to identify social trends and phenomena. In recent years, academics have steadily lost ground on access to social media data as technology companies have set more restrictions on Application Programming Interfaces (APIs) or entirely closed public APIs. This circumstance halts the work of many social scientists who have used such data to study issues of public good. We considered the viability of eight approaches for image-based social media data collection: data philanthropy organizations, data repositories, data donation, third-party data companies, homegrown tools, and various web scraping tools and scripts. This paper discusses the advantages and challenges of these approaches from literature and from the authors' experience. We conclude the paper by discussing mechanisms for improving social media data collection that will enable this future frontier of social science research.

## 2.1 Introduction

Social media has profoundly changed our modes of self-expression, communication, receipt and dissemination of information, construction of social bonds, and participation in public discourse and events (Lazer et al., 2009; Acquisti, Brandimarte, & Loewenstein, 2015). In the first decade of the flourishing of social media, the potential value of social media data also caught the attention of researchers. Over the subsequent years, it has been consistently demonstrated that this data assists in our understanding of society and human behavior (Chen et al., 2023; Sherren et al., 2023). Early studies on the use of social media in politics confirmed the meaningfulness and power of the data, followed by research in various areas like business, communication, health, environment, and sociology (Chen et al., 2018; Edwards et al., 2013; Procter, Vis, & Voss, 2013; Savage & Burrows, 2007), leveraging platforms such as Twitter, Flickr, Weibo, Panoramio, YouTube, Facebook, and Instagram (Chen et al., 2023; Ghermandi & Sinclair, 2019; Gone et al., 2023).

The Cambridge Analytica Scandal in 2018 – triggered when The New York Times reported the data of millions of Facebook users were fraudulently accessed by a consulting company (Confessore, 2018) – was a major turning point that led to the current “post-API” (application programming interface) age, with social media platforms restricting or paywalling access to public search APIs, fine-grained location data, and more (Freelon, 2018). The majority of the most widely used platforms have transitioned towards a stricter and more commercialized policy, closing access to such data for public good research. Instagram closed their less-restricted access; instead, they issued two types of API for business app use only (Meta for Developers, 2023). This has an impact on research areas that place greater value on image data, such as landscape studies. Meta recently launched a new Content Library API in November 2023, an access-controlled space to work on Facebook and Instagram data rather than downloading complete copies, yet it has not been widely used (Meta, 2023). X, previously known as Twitter, closed their academic research API in 2023 after Elon Musk’s acquisition and created three new versions of paid-API access (X Developer Platform, 2023). By contrast, the most popular short video platform, TikTok, has launched an application-based research API. Currently,

the application is only open to US- and Europe-based researchers, but it may become available to all researchers in the future (TikTok for developers, 2023).

Researchers are in an increasingly weak position with respect to social media data access (Zuckerman, 2023). John and Nissenbaum (2019) wrote that “researchers are ultimately dependent on tech companies for data and have to find a way to collaborate while serving the public interest and avoiding bias” (p.3). The increased restrictions in APIs leave social media researchers grappling with non-public, legally ambiguous, and ethically grey approaches to collecting data, or push them toward impermanent types of data that hamper the detection of trends (Kinder-Kurlanda & Weller, 2020; Weller & Kinder-Kurlanda, 2015). Business-oriented users of data continue unencumbered, while access to data for public good is curtailed (Acker & Kreisberg, 2020; Bruns, 2019; John & Nissenbaum, 2019). Poletti and Gray (2019) mentioned that “academic research is now competing with market research, and it is no longer the dominant party when it comes to providing interpretations of society” (p.265).

Social media imagery data, along with its geo-tags, is recognized for its value across diverse social science fields (e.g., environment, sociology, politics, health, etc.), though its collection can be complex due to the necessity of retrieving additional image files (Chen et al., 2023). In this paper, we considered eight approaches to image-based social media data collection: data philanthropy organizations, data repositories, data donation, third-party data companies, homegrown tools, and various web scraping tools and scripts. To manage the scope, we considered the viability of each approach for an energy landscape study in rural Canada. The case leverages our engagement in longitudinal research which helped us to understand the challenges after 2018. While the case study might not interest a broad research community, the insights garnered from the data collection process are universal because: 1) the framework of the eight approaches is consistent for all social media data, and 2) various tools examined in this paper can extract diverse data types, such as texts and videos, from different social media platforms. From the analysis of the approaches, this paper offers three main contributions. First, it tests these approaches for their feasibility in gathering Instagram data by geographic locations, providing insights for social scientists who are interested in leveraging social

media data for place-specific questions. Second, it details advantages and challenges from the literature and from the authors' experience. Third, it raises the idea of a forward-looking solution for a research API, building on nascent efforts undertaken by social media companies and regulatory frameworks.

## 2.2 Methods

The energy landscape project we used to assess the viability of these data collection approaches necessitated gathering Instagram posts depicting images of outdoor landscape use around hydroelectric dams and reservoirs in rural Canada. We identified eight social media data collection approaches from literature and practice to assess for their effectiveness, benefits, and drawbacks for the project. These approaches include the following:

***Data philanthropy organizations:*** Entities that distribute data without charge. Many of them are based on industry-academic partnerships.

***Data repositories:*** Entities that host data contributed by scholars for broader re-use.

***Data donation:*** A relatively new practice in which individual users can request their personal social media archives and donate them to data repositories or research projects.

***Third-party data companies:*** Commercial tools developed by third-party companies to monitor, retrieve, and analyze social media data.

***Homegrown tools:*** Software tools for extracting social media data that are rooted in academic soil (i.e., made for researchers by researchers) and tend to charge more affordable rates and provide data in researcher-friendly formats.

***Web scraping tools – commercial:*** Commercial web scraping tools provide the service of automatically extracting content from social media posts for profit.

***Web scraping tools – non-commercial:*** Non-commercial web-scraping tools are collaboratively built and shared as open-source software online.

***Web Scraping scripts (single-purpose):*** Software scripts developed by researchers for a specific research project, perhaps using a library or template.



Table 2.1 Advantages and limitations of social media data collection approaches.

Approach	Examples	Advantages	Limitations	Hurdles from authors' experiences
Data philanthropy organization	Facebook Ad Library Social Science One	<ul style="list-style-type: none"> <li>• Full access to a more complete dataset</li> <li>• Low or no legal risk</li> <li>• No monetary cost to the researcher</li> </ul>	<ul style="list-style-type: none"> <li>• Limited social media platforms</li> <li>• Delivery delays</li> <li>• Veto right reserved by social media companies</li> <li>• Requires an application</li> <li>• Limited research topics</li> <li>• Deadline restriction</li> </ul>	<ul style="list-style-type: none"> <li>• Instagram does not have such organizations to provide data</li> <li>• Landscape research is not a prioritized topic</li> </ul>
Data repositories	Inter-university Consortium for Political and Social Research	<ul style="list-style-type: none"> <li>• No monetary cost to the researcher</li> </ul>	<ul style="list-style-type: none"> <li>• Legal and ethical concerns of sharing data</li> <li>• No existing data available</li> <li>• Data disconnected from original context</li> <li>• Requires sustainable funding for the repository</li> </ul>	<ul style="list-style-type: none"> <li>• No existing data for our study cases</li> </ul>
Data donation	Breuer et al., 2020	<ul style="list-style-type: none"> <li>• Data include a wide range of user activities</li> <li>• No or low legal and ethical risk</li> </ul>	<ul style="list-style-type: none"> <li>• Requires recruiting participants (time-consuming and ethical review)</li> <li>• Complicated process</li> <li>• Limited size of data and response bias</li> </ul>	<ul style="list-style-type: none"> <li>• Limited time and budget to collect large-sized data</li> </ul>
Third-party data companies	HootSuite Sprout Social	<ul style="list-style-type: none"> <li>• User-friendly</li> <li>• Low legal risk</li> </ul>	<ul style="list-style-type: none"> <li>• High cost</li> <li>• Ill-suited data formats for research purposes</li> </ul>	<ul style="list-style-type: none"> <li>• Difficult to find a provider (they are business-oriented)</li> <li>• Limited budget</li> </ul>
Homegrown tools	Netlytic communalytic	<ul style="list-style-type: none"> <li>• Well-suited data formats for research purposes</li> <li>• User-friendly</li> <li>• Affordable price</li> </ul>	<ul style="list-style-type: none"> <li>• Heavily depend on platforms' APIs</li> <li>• Not always well-maintained or self-sustaining</li> </ul>	<ul style="list-style-type: none"> <li>• Unable to collect data by geographic index due to API limitations</li> </ul>
Web Scraping tools (commercial)	ScrapeStorm Apify	<ul style="list-style-type: none"> <li>• Affordable price</li> <li>• User-friendly</li> </ul>	<ul style="list-style-type: none"> <li>• Incomplete dataset (export limit)</li> <li>• Ill-suited data formats for research purposes</li> <li>• Ethical and legal risk</li> </ul>	<ul style="list-style-type: none"> <li>• The tool we attempted had a maximum data export limit</li> </ul>
Web Scraping tools (non-commercial)	Instagram Scraper	<ul style="list-style-type: none"> <li>• No monetary cost</li> </ul>	<ul style="list-style-type: none"> <li>• Not user-friendly and requiring programming skills</li> <li>• Inflexible and unstable (can stop working when the social media interface changes)</li> <li>• Ethical and legal risk</li> <li>• Time consuming</li> </ul>	<ul style="list-style-type: none"> <li>• The tool we chose became non-functional halfway through</li> </ul>
Web Scraping scripts (single-purpose)	github.com/Titration /Ins-Scraping	<ul style="list-style-type: none"> <li>• Well-suited data formats for research purposes</li> <li>• Flexible</li> <li>• Low up-front cost</li> </ul>	<ul style="list-style-type: none"> <li>• Ethical and legal risk</li> <li>• Time consuming</li> <li>• Requires programming skills</li> <li>• Must be customized for each research project and platform.</li> <li>• Unstable</li> </ul>	<ul style="list-style-type: none"> <li>• It took us 5 months to collect around 80,000 posts</li> </ul>

## 2.3 Results

Many approaches proved unsuitable for collecting Instagram data in our landscape research for diverse reasons (Table 2.1). This section outlines each approach's strengths and weaknesses, as per literature and our experiments, and clarifies why most failed.

### *2.3.1 Data philanthropy organizations*

The advantages include that researchers can have full access to a more complete dataset than would be available through other means without any legal risk, because the organization helps to build industry-academic partnerships through which social media researchers can obtain data directly from the company (Social Science One, 2022). To qualify, the study topics must be narrowly related to specified areas, such as the effect of social media on democracy for Social Science One, which eliminates most environmental and landscape research like ours. The scrutiny on applications is strict, and the veto right reserved by social media companies casts a long shadow over research independence (Bruns, 2019). Such organizations often only receive application submissions by deadlines, also leaving the data collection work less flexible and incompatible with fast-changing environmental and social issues.

### *2.3.2 Data repositories*

Data repository is an alternative that can reduce the influence of social media companies (Acker & Kreisberg, 2020; Borgman, 2019). However, largely due to legal and ethical concerns about sharing social media data, many researchers are cautious. Also, the specifications required for data collection for one study may make the data useless to others. Few scholars use the entirety of social media data during a period; most are looking for subsets in specific locations or referencing specific keywords or hashtags. Our research project is an example: there were no previous studies we could find that collected and shared Instagram posts from our target case areas. Keeping and delivering huge social media data can be financially, ethically and technologically difficult (Borgman, 2019), especially when most social science users will use a small (but unpredictable) fragment of that data (Chen et al., 2023). In repositories, data can become disconnected from their context and dealing with data duplicates can be troublesome

(Weller & Kinder-Kurlanda, 2015). As such, data repositories may not yet be a solution for most research projects.

### *2.3.3 Data donation*

Data donation can provide a wide range of user activities including private messages and ephemeral content (van Driel et al., 2022). However, since the donation system has not been fully developed, it typically requires researchers to find, contact and potentially compensate people first and ask them to follow the donation steps (Breuer et al., 2020). This may not benefit all kinds of research, especially those requiring large-sized data or not focusing on specific actors in a system. In addition, for research using social media data as a substitute for conventional approaches like survey and interview, data donation follows a more complicated but less mature process. It also introduces response bias on top of biases inherent to social media data, and there is not a clear research ethics regime in place for encouraging donations to a repository.

### *2.3.4 Third-party data companies*

Although free of legal concerns, purchasing data from these third-party companies can be expensive and yet not provide data well-suited for research purposes. We inquired with two Instagram partner companies to collaborate on data collection for our research project on the topic of hydropower landscape. Neither responded to our emails or web submission forms. It is easy to understand why: first, both companies are large and likely prefer large customers who can bring sizable revenue; second, their business is focused on marketing and advertising analytics and their tools are less applicable for research purposes.

### *2.3.5 Homegrown tools*

Homegrown tools for data collection can provide data in researcher-friendly formats at a reasonable cost. For instance, Netlytic is free for small datasets and has been used by many social scientists to extract social media data, especially pre-2018. A related product communalytic is affordably priced and includes access to historical Reddit data and has limited abilities to import other social media data (e.g. comments on a specific YouTube video). However, any such tools are heavily dependent on social media APIs which grant

them no higher levels of access than the public. We used Netlytic in our original cross-sectional landscape studies to collect Instagram posts by geographic coordinates. When Instagram stopped supporting the geo-location index in their API, Netlytic terminated the service for Instagram data.

### *2.3.6 Web scraping tools – commercial*

Commercial scraping tools are often more affordable than third-party companies but can still be a big investment if a large-sized dataset is required. In our review of widely used scraping tools for Instagram data, most provided services by subscription at prices ranging from \$5 to hundreds of US dollars per month, and there were limitations in terms of exported data size and formats (e.g., ScrapeStorm and Apify). Downloading images, which are increasingly critical to social media research (Chen et al., 2023) and to landscape studies, would result in additional charges. There is another concern on the completeness of datasets because most of these tools have a cap on data export amounts.

### *2.3.7 Web scraping tools – non-commercial*

Another type of scraping tool is the non-commercial ones, such as Instagram Scraper (GitHub, 2022). This tool is not as user-friendly as the commercial tools that operate on a graphical interface. Instead, it is code-based which requires users to have basic knowledge of and experience with Python to operate it. Instagram Scraper operated properly when we started to collect data in September 2020; however, it became quite unstable from February 2021 and there was no update of the tool until one year later. Open-source software is community-supported, which means a developer needs to be willing to contribute their time to ensure the software stays up-to-date with rapidly changing social media platforms.

### *2.3.8 Web Scraping scripts (single-purpose)*

For researchers with (or with access to) sufficient technical expertise, self-developed scraping scripts can provide more flexibility and collect data that is well-suited to the scholarly research analysis they have planned. In our case, a local software developer agreed to help develop custom scraping scripts based on two Python packages –Selenium and Instascape (our Instagram scraping program code is posted on GitHub at

<https://github.com/Titration/Ins-Scraping>). However, data collection with the scripts was time-consuming for several reasons. First, the scripts needed to be maintained and updated frequently to cope with the platform's changes in terms of APIs or the anti-auto-data-scraping strategies. We retrieved around 80,000 Instagram posts for our study, but it took five months. While we have released our scripts as open-source, and it worked at the time of release, it too will soon require further development effort to match changes to the social media platform.

Additionally, an extensive list of available Instagram accounts and IP addresses (Internet Protocol address, provided by Virtual Private Networks which can establish a digital connection between a computer and a remote server) are also necessary to respond to blocks by the platform. Once any suspicious actions (e.g., excessive visits) are detected and identified as an auto-scraping bot, the IP address and account can be banned, temporarily or permanently, from making further requests. According to our experience, on average, we changed IP addresses three times per day and switched accounts two times per day to download 1,000 posts. Although IP addresses and social media accounts can be changed or replaced, the data delay and gaps caused by successive blocks can impact research results to different degrees (Freelon, 2018). It is also alarming for academics, particularly those early in their careers, to worry about being perceived as operating outside of a platform's legal terms and conditions.

#### **2.4 Discussion: A better solution?**

The available solutions for academics to access social media data under current restrictions are making the research field highly uneven and heterogenous (Acker & Kreisberg, 2020; Bruns, 2019). The amount of social media research is increasing, and it is easy to have an illusion that we are getting more data than ever before. But while the potential supply is growing—users are creating more data every day—we have access to a smaller proportion of the corpora or must rely on data collected pre-API closure, concerning researchers around issues like data representativeness, currency and generalization of results (King & Persily, 2020). Researchers using social media appear unwilling to articulate the details of their data collection process (Poletti & Gray, 2019; Weller & Kinder-Kurlanda, 2015), either because of cumbersomeness (given the

patchwork of tools available) or legal and ethical concerns. The lack of detail in method discussion is increasingly pervasive due to some new considerations: (i) published methods may not be useful for long given platform policy changes; (ii) the details may be too technical for social science audiences to follow; and (iii) there is a motivation to protect data and skills to maintain researcher competitiveness (Kinder-Kurlanda & Weller, 2014; Weller & Kinder-Kurlanda, 2015).

While the practices of data philanthropy organizations and repositories are still developing and the cost of third-party data companies are high enough to scare many researchers away (even if they can be enticed into collaboration), scraping tools and scripts may be the most feasible option but remain an imperfect one – the legal status of scraping is inherently problematic in addition to the privacy concerns (Bruns, 2019). Freelon (2018) noted that “researchers should bear in mind the potential (if unlikely) consequences of even small-scale terms of service violations (p.667).” Scraping also introduces more general challenges (Weller & Kinder-Kurlanda, 2015). First, data quality is problematic in most cases where data is collected with tools without sufficient documentation, leading to opaque processes and thus weak replicability. Second, platforms may have limits on the type or the amount of data the public can access during a given period, which may result in sample biases. Third, ephemerality is a perennial challenge of social media research: platform policies can be updated at any minute, and the data can be altered or deleted (Walker, 2017). Fourth, platforms with highly restrictive APIs (e.g., Instagram and Facebook) might be avoided and more permissive APIs (e.g., Flickr and Reddit) might be preferred by researchers, causing either under- or over-representation of certain social media platforms over others, and thus certain user demographics (Barnhart, 2023).

The profound impact of social media on society suggests we should not leave addressing this problem solely to the creativity and innovation of researchers. We advocate that a better solution is a separate public research API that is not based on social media companies imposing an application-approval process. In the long run, we do not believe that data philanthropy organizations will be effective because they are often simply a distribution center and cannot guarantee the integrity and delivery of the data. Social

media companies could change their one-size-fits-all approach to APIs on social media platforms to multiple ones that better serve data users with different motivations (Acker & Kreisberg, 2020; Shtern et al., 2013). However, a research API where researchers apply to the social media company for access, such as currently offered by TikTok is not favored, either. This grants the company complete authority to decide who can access how much data and when they can receive it. Another option is a research API gatekept by a third organization, like the Inter-university Consortium for Political and Social Research (ICPSR) at the University of Michigan Institute for Social Research (ISR) for Meta Content Library API (Facebook and Instagram). It is unclear how different this is than the practice of data philanthropy organization, Social Science One. Decisions about access to such a portal should not be based on an application and review process which can favor certain research fields, regions, and researchers. However, a simple process to verify researcher status will be necessary. The platform should ideally include fact-based verification, such as verifying profile pages or email addresses of academic and research institutions, without making the topic vulnerable to rejection if it is clearly public good rather than those favored by the platforms such as commercial benefits and online democracy. As highlighted by Rieder and Hofmann (2020), it is necessary to broaden the analytical scope: data for public good should serve broader societal interests like cultural production, beyond just critical algorithm studies.

Social media companies clearly have little incentive to facilitate such a public research API (Steen-Johnsen & Enjolras, 2015). It might fulfill the company's social responsibility expectations in the social and government sphere but will not benefit (and might potentially weaken) its revenue in terms of data selling. Thus, a new governance model is required to enforce the public good values. Government and regulators need to intervene with laws or policies and, ideally, processes that support data sharing from social media companies to verified researchers (Vogus, 2022). An example is the Digital Services Act that was approved by the European Parliament in 2022, which provides rules to establish a mechanism for researchers to gain data access to large social media platforms and search engines (Joint Research Centre, 2023). At this stage, the effectiveness of this Act is not completely known, though X has taken early actions to allow EU researchers to access licensed data for DSA-related research purposes by the end of 2023 (X Developer

Platform, 2024). However, the European Commission has opened formal proceedings to assess whether X may have breached the DSA, including concerns of suspected shortcomings in giving researchers access to X's publicly accessible data (European Commission, 2023). Nevertheless, an open gateway in Europe could facilitate transnational partnerships, allowing non-EU researchers to access data via EU collaborators.

In general, there must be clear guidelines for researchers, including how to use, store, protect, and (possibly) share data, along with the corresponding consequences for violations. Currently, such datasets sit outside of the purview of most human research ethics boards since the data is notionally available publicly. A new form of research ethics review should be developed, including setting the boundary of public data, defining fair and public-good use of social media data, and estimating the effectiveness of anonymity strategies (Chen et al., 2023; Taylor & Pagliari, 2018). A risk analysis review may also be necessary to estimate and monitor the potential harm to individual users of using social media data in specific ways and for specific purposes.

Until such a public research API can be achieved, researchers have a long and potentially dark journey ahead. There are many data collection methods at our disposal, but none of them are reliable and all come with risks such as personal legal and research quality. Researchers should speak frankly about the data collection process and challenges they experience, such as adding a supplementary document to disclose their detailed steps of data collection and any developed code, if applicable. Collaborative data repositories could become a feasible solution only if researchers are willing and able to share social media data with other researchers, and critically if the legal and ethical grounds can be safely and legally addressed. The precedent *Sandvig v. Barr* (2020) may provide an example: a district court in Columbia in the US granted researchers freedom to use data from employment websites to conduct their study. It is in the public interest to give public-good researchers legal access to high-quality social media data that is at least comparable to what commercial users have; we believe most of those contributing content to social media platforms would agree. The next thing we should do is ask them.



### **Chapter 3 Image auto-coding tools for social impact assessment: Leveraging social media data to understand human dimensions of hydroelectricity landscape changes in Canada**

A version of this chapter has been submitted to the journal *Landscape and Urban Planning* (by Elsevier). Refer to Appendix B for the copyright agreement to reproduce this material.

Chen, Y., Smit, M., Lee, K, Y., McCay-Peet, L., & Sherren, K. (In review). Image auto-coding tools for social impact assessment: Leveraging social media data to understand human dimensions of hydroelectricity landscape changes in Canada. *Landscape and Urban Planning*.

Statement of Authors' Contributions: YC is responsible for conducting the literature review, data collection and analysis, and writing all sections. KS and MS supervised the writing development including providing feedback and editorial revisions. KL and LMP provided feedback and editorial revisions on a final draft.

## **Abstract**

Social media data has been shown to be a valuable data source for assessing social impacts, particularly when paired with the swift progress in artificial intelligence technologies, allowing comprehensive analyses of larger datasets than is possible using conventional approaches. We sought to understand the social impacts of hydropower-related landscape changes based on a quasi-chronosequence of three study cases in Canada, using social media images in conjunction with machine learning to conduct image and textual analysis. We employed the Google Cloud Vision API, a pre-trained deep learning model, to detect labels from over 19,000 landscape images of the relevant regions sourced from Instagram. This yielded a comprehensive set of over 188,000 labels. We used a generative probabilistic model (Latent Dirichlet Allocation, an unsupervised machine learning algorithm) to create clusters based on the labels. These clusters revealed prevalent landscape features, human activities, and animate and inanimate objects—as well as which frequently co-occurred—allowing us to understand and predict some of the social impacts of the landscape changes caused (or that might be caused) by hydroelectric dams and reservoirs. This provides a novel example of integrating large-sized social media data and automated analysis tools powered by machine/deep learning into social impact assessment. Notably, such pre-trained models and “off-the-shelf” unsupervised algorithms require minimum programming skills, benefiting scholars and practitioners who are less versed in technical domains. The insights gained from hydroelectricity case studies can also inform decisions about energy transition.

### 3.1 Introduction

The development of hydroelectric dams since the mid-20th century has stirred controversy, attracting attention not only at the local level but also often provincial and nationwide scrutiny (McElroy, 2016). Hydroelectric projects, such as the Oldman River Dam in Alberta, have emerged as milestones in Canadian history, sparking public awareness of environmental protection and driving legislative initiatives for environmental impact assessment (Muldoon et al., 2020; Shpyth, 1991), and such trends were also observed in other regions like Tasmania and New Zealand (Rainbow, 1992). There is a growing recognition, however, of the social impacts associated with the transformation of landscapes and lifestyles due to dams and their reservoirs (Chen et al., 2018; Chen et al., 2019). This encompasses various aspects, including displacement, social networks and status, happiness, social stress and safety concerns, social costs of uncertainty, recreation, cultural and historical heritage, and landscape loss (Pimental da Silva et al., 2021; Kirchherr & Charles, 2016).

The limitations of traditional research methods, such as surveys, interviews, and focus groups, have become increasingly evident as we continue to live more of our lives online, a phenomenon accelerated by the global pandemic (Kulanthaivel et al., 2017; Sherren et al., 2023). Although traditional methods are well-established both theoretically and empirically, their advantages are now often counteracted by drawbacks like prohibitive cost, low response rate, and systematic bias in engagement that may exclude specific social groups such as young people (Chen et al., 2019; Stedman et al., 2019). Over the past two decades, researchers have leveraged a valuable new data source – social media data – to fill the gaps left by conventional research approaches (Azevedo et al., 2022), including to understand energy development impacts. For example, Chen et al. (2018) utilized Instagram images and captions to understand young people’s perceptions of hydropower landscapes, and Mohammadi et al. (2023) assessed the visual impacts of wind turbines and solar panels in amenity vineyard landscapes.

While social media data can serve as a reliable proxy for social phenomena, its sheer volume often poses challenges and dilemmas for researchers. In a review conducted by Chen et al. (2023), 78% of studies using social media image data in social science

research from 2015 to 2019 primarily relied on manual analysis approaches, and only 18% analyzed more than 10,000 posts – a small fraction of the data available. The full potential of social media data has remained largely untapped in prior research, in part because of the challenge of manual coding of such amounts of data, particularly for images. It wasn't until recently that advancements in artificial intelligence (AI) technologies ushered in a new era and opportunities such as pre-trained models for the non-programming research community (Manley et al., 2022). Leveraging AI tools such as machine learning and deep learning models in Big Data analysis has become one of the focal points of social science research using social media data (Dangi et al., 2022; Milusheva et al., 2021). Relatively few researchers, however, have conducted research using larger scale social media image data leveraging AI. For instance, Vigl et al. (2021) employed the image auto-annotation engine Clarifai to elicit visual-sensory landscape values from 100,000 Flickr images. Mouttaki et al. (2022) developed and trained a model based on convolutional neural networks to classify 29,000 Flickr photos by cultural ecosystem service categories.

In this paper, we use two AI models, the image object detector Google Cloud Vision API and the generative probabilistic model Latent Dirichlet Allocation (LDA) for topic clustering, to analyze Instagram images collected by geo-location tags in research areas associated with three hydroelectric projects in Canada. We will probe four research questions: (1) how is the current landscape perceived and used by people in terms of landscape features and human activities; (2) what do the resulting patterns say about how landscape changes related to hydroelectric projects impact places and people in terms of social and cultural values; (3) what are the implications for the study cases and generalized insights; and (4) what are the opportunities and limitations of using the AI models for working with social media data to explore hydroelectric social impacts? In the following sections, the Methods section will encompass study cases, data collection, and details about the two AI models; the Results section will show the landscape clusters that consist of landscape features and human activities; and the Discussion and Conclusion sections will provide case-based insights and implications for decision-makers and identify opportunities and limitations of using social media data and AI tools for future research.

## 3.2 Methods

### 3.2.1 Study areas

This research has three study cases across Canada, including the Site C Dam (Site C) in British Columbia (BC), the Oldman River Dam (Oldman) in Alberta, and the Mactaquac Dam (Mactaquac) in New Brunswick (NB) (see Figure 3.1). The three dams are at various stages of their lifespan – Site C is under construction, Oldman was built in the early 1990s, and Mactaquac dates to the late 1960s and was approved for a refurbishment plan in 2016. The comparison of various cases will offer insights into the social impacts of hydroelectric dam projects during their lifespans.

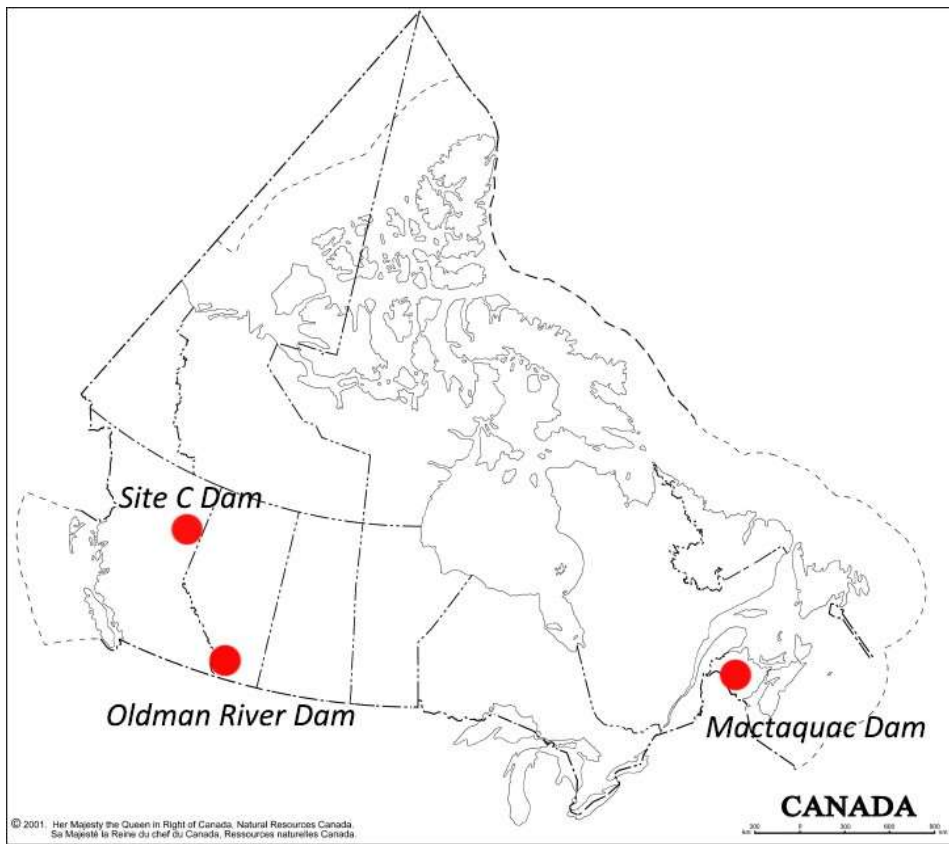


Figure 3.1 Study areas (adapted from Natural Resources Canada, 2001)

Site C will be the third mega-hydroelectric infrastructure on the Peace River in north BC (Clarke, 2014). The river has watered the Peace River Valley for over ten-thousand years, supporting human habitation from prehistoric times to modern farming and ranching practices (Imrie, 1991). Site C is number 3 (hence “C”) of a string of five dams initially

proposed by the former premier William Andrew Cecil (W.A.C.) Bennett in the 1950s (Cox, 2018). It was designed to produce up to 1,100 megawatts of power for over a century, forming a reservoir 83 km long and up to 55 metres deep (Canadian Environmental Assessment Agency, 2014), and inundating a total of 5,550 hectares of riparian lands (BC Hydro, 2018). The project has continued to be controversial in public discourse and the courts since its approval, including concerns of irreversible negative impacts on the biophysical environment, social disturbances, and devastating harms on Indigenous communities (Cox, 2018; Muir, 2018). According to proponents, Site C is on schedule to start reservoir filling in fall 2024 (BC Hydro, 2023).

Oldman was built at Three Rivers in southern Alberta where the Oldman River is joined by its main tributaries, the Castle and the Crowsnest Rivers (Glenn, 1999). The original rationale to dam the river was agricultural irrigation and water management in the province due to drought caused by seasonal variations in flows (Rojas et al., 2009). The construction started in 1986 and the required access and use of the river and riparian lands for dam construction and operation triggered frequent fights and negotiations between the government and the closest Indigenous group, the Piikani Nation (Fabris, 2023; Glenn, 1999). Despite these disputes and without a legal or social license, Oldman was finished in late 1991 and started to form its reservoir in the next year (Daschuk & Marchildon, 1993). A 32-megawatt Hydroelectricity plant was added in 2003, in collaboration with the Piikani Nation, which owns a quarter of the project (Clarion Energy Content Directors, 2003).

Mactaquac was constructed on the St. John River in the late 1960s, in New Brunswick, as the largest hydroelectric dam in Atlantic Canada (Bourgoin, 2013). The dam has an installed capacity of 670 megawatts, and the reservoir inundated around 5000 hectares of riparian land. Its construction bolstered employment and in-migration in the first couple of years followed by development based on the reservoir as a recreational site, but the biophysical environment was considerably altered and many people suffered from displacement, disconnections to social relationships and communities, destruction of historical records and cultural heritage, and losing traditional ways of life such as farming and grazing (Chen et al., 2018; Lawson et al., 1985; Reilly & Adamowski, 2017). The

dam faced a premature end of service life by 2030 due to an alkali-aggregate reaction that weakened the concrete, so in 2015, the province invited public input about the dam's future around three feasible options: refurbishment, retaining the reservoir without power generation, and river restoration (Stantec, 2016). Despite the initial trauma experienced by the pre-dam generation, most local people expressed their preference to keep the dam and reservoir intact (Sherren et al., 2016), especially among the cohorts who grew up with the reservoir landscape or moved into it as adults (Keilty et al., 2016). The operator decided to extend the dam's current lifespan.

### *3.2.2 Data Collection*

We collected data from Instagram, one of the most popular visual content-sharing platforms (Instagram, n.d.). It was launched in 2010 and had over 2 billion users as of January 2023 (Dixon, 2023a). Slightly over 60% of Instagram users fall into the age group of 18 to 34 (Dixon, 2023b). To retrieve Instagram images, we developed a scraping tool which contains two main Python packages: Selenium and Instascape (Chen et al., 2024). Selenium was employed to login to Instagram, search Instagram posts by location IDs, and save post links into a log file (Titration, 2022a). The Instascape package provides a flexible API for scraping Instagram data including images (or videos) and metadata based on the post links in the log file (Titration, 2022b).

Place names of cities, towns, parks, and points of interest near the three hydropower sites and along their existing or planned reservoirs were manually identified through Google Maps and then searched on Instagram to see: 1) whether there was a related geo-location tag that had been created; and 2) whether there were posts geo-tagged to this geo-location. If both criteria were satisfied, location IDs were collected from Instagram and geographic coordinates from Google Maps. We identified 5 in Site C area, 4 valid locations in the Oldman case, and 25 in Mactaquac (Appendix C). The uneven distribution of locations in the three cases stems from Mactaquac being in a more densely populated area with a reservoir flanked by numerous towns.

The location IDs were used to search geo-tagged posts and store post links in log files from May 30 to June 2, 2021. The earliest post retrieved was published on July 18, 2011. We identified 49,351 posts in the Site C case, 8,727 in Oldman, and 28,393 in Mactaquac

(Figure 3.2). Later, from June 7 to July 25, 2021, images/videos and metadata were retrieved using the links. Two percent of the posts could not be properly scraped due to errors (e.g., posts were deleted, or accounts were changed to private). We downloaded 48,301 images/videos with metadata for Site C, 8,555 for Oldman, and 27,891 for Mactaquac. Video data were removed from the datasets, leaving 44,875 images in the Site C case, 7,980 in Oldman, and 26,178 in Mactaquac. Among the raw datasets, many of the images had little value to our research purpose of understanding landscape use and value. Thus, we manually filtered out images in which the portion of landscape features were less than 60% of the entire content, such as indoor photos and selfies without clearly visible landscape. We attempted to train an AI model to automatically filter the raw datasets; however, the accuracy was insufficient due to the complexity of the data and the limitations of machine power in 2021. As a result, the Site C case had 6,817 images left as valid data, 3,060 in Oldman, and 9,292 in Mactaquac.



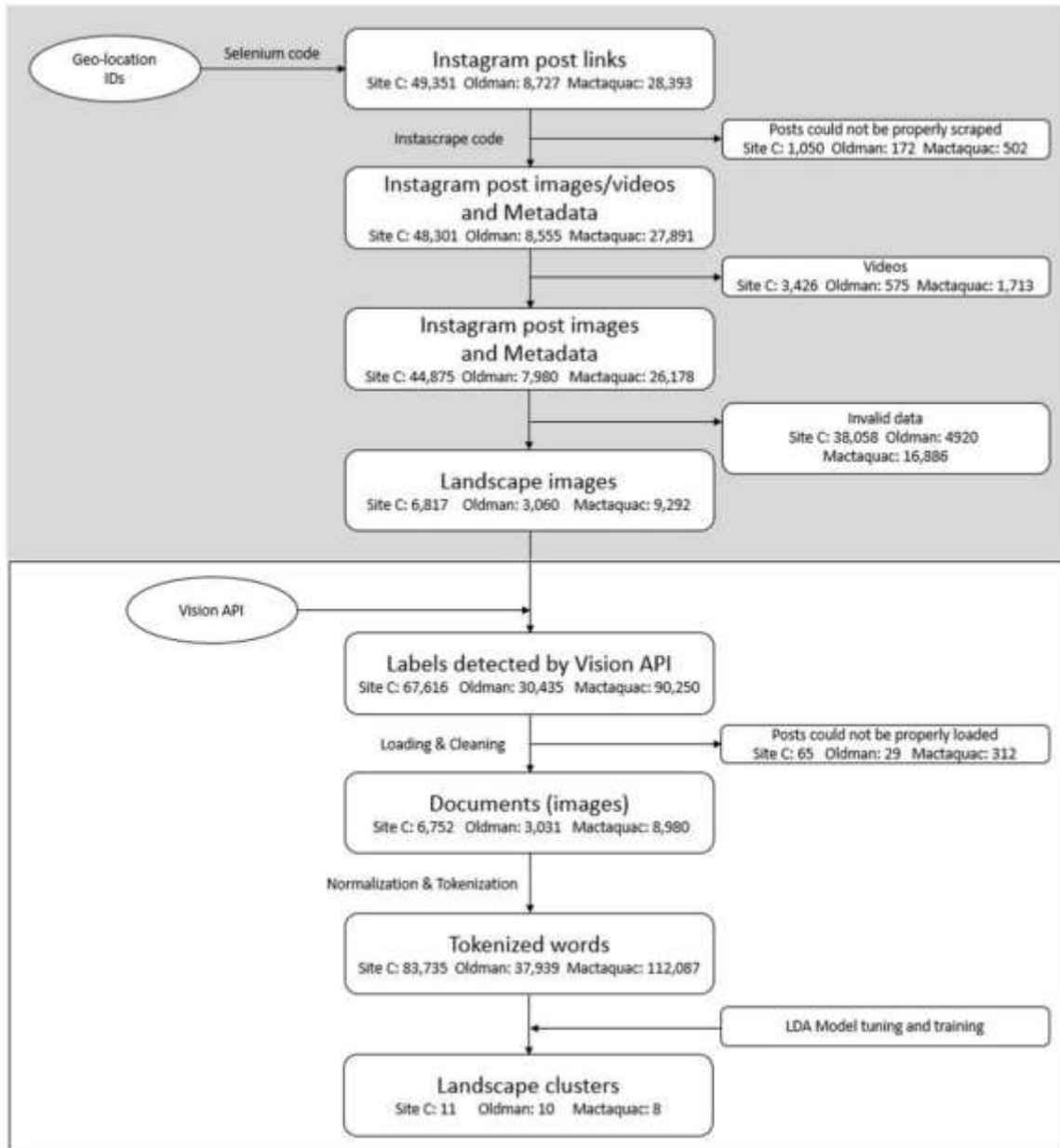


Figure 3.2 Workflow of data collection (shaded background) and analysis (white background).

### 3.2.3 Landscape image auto-coding

We used Google Cloud Vision API (Vision API) to label the images, detecting landscape features and human activities (see Figure 3.3 as an example). Vision API is a pre-trained machine learning model that can assign labels to images using millions of predefined categories (Google Cloud, 2023a). The fundamental algorithm of image detection models is based on the convolutional neural network (CNN) (Fukushima, 1980; LeCun et al.,

1998). The process of utilizing the Vision API (or comparable pre-trained models) is straightforward, uploading and receiving predicted labels. However, it still requires basic programming skills for batching prediction tasks for multiple images at a time (Google Cloud, 2023b). For each image, the model returns a maximum of 10 labels with the highest confidence scores. Owing to this limitation, energy facilities like dams, wind turbines, and solar panels, if they appeared small in the background, were not accurately detected, leading us to manually code valid images for these energy infrastructures. During the process, image contents were viewed, and our observations will assist the discussion.

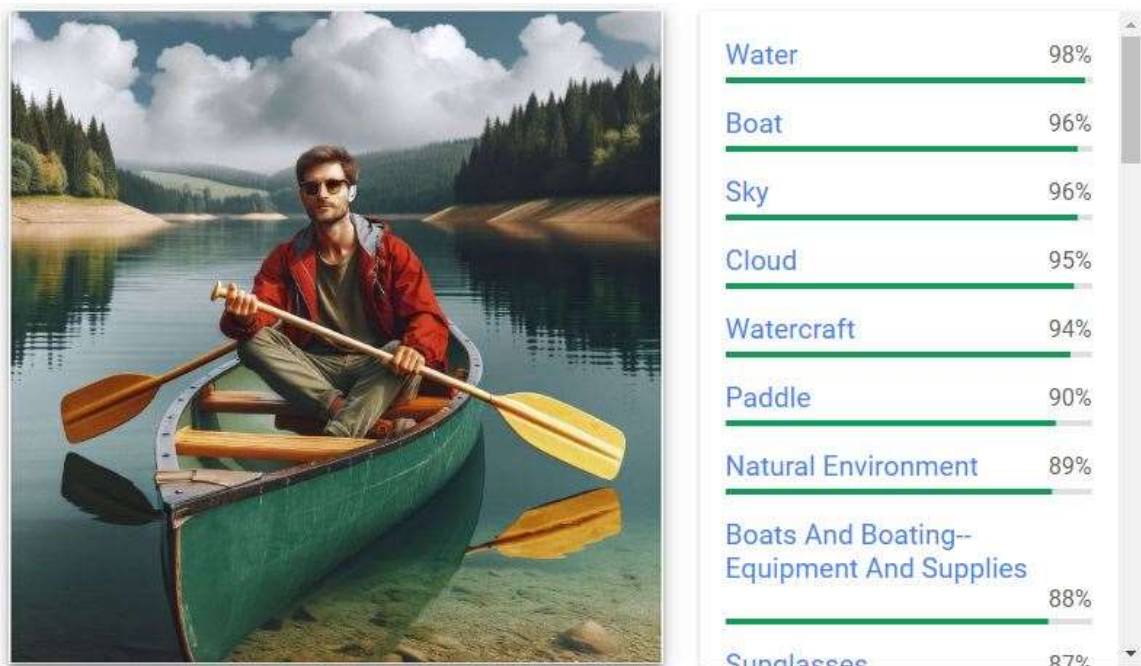


Figure 3.3 An example of Vision API results (the image was generated by ChatGPT, and the Vision API results were generated in February 2024)

### 3.2.4 Topic Clustering

Latent Dirichlet Allocation (LDA) is a generative probabilistic Bayesian model that automatically identifies topics based on the co-presence of certain words in textual corpora (Blei et al., 2003). We used the Gensim library, a free open source set of topic models. Topic models are unsupervised models that automatically discover statistical co-occurrence patterns within a corpus (Gensim, 2022), giving the ability to find patterns in a dataset. The analysis was conducted in Jupyter Notebook using Python. The script was

developed based on Kapadia's (2019) work and adapted to meet the needs of this study. More details can be found in Appendix D.

We used LDA for topic clustering, to identify patterns within the 188,301 labels generated by Vision API. There were four main phases, including label data loading and cleaning, data normalization and tokenization, model tuning, and model training (Figure 3.2). First, labels from the same image were treated as a single document in subsequent LDA model training. There were 6752 documents in the Site C case, 3031 in Oldman, and 8980 in Mactaquac (2% of images were excluded because they were not appropriately loaded). Second, data normalization (removing stop words and lower casing the text) and tokenizing (identifying individual words in a document) was completed. Third, the LDA model was tuned, the process of identifying the parameters that govern its learning process to optimize performance. Finally, the model was trained to identify the topic clusters. We received results of clustered labels as topics which were then interpreted (Table 3.1)– giving a descriptive title for each topic – by authors through examining the labels and related image contents.

### **3.3 Results**

#### *3.3.1 Descriptive Statistics*

Using Vision API, there were 67,616 total labels returned for the Site C case, 30,435 for Oldman, and 90,250 for Mactaquac, adding to a total of 188,301 labels and 1,704 unique ones. Figure 3.4 shows the top 15 labels in each study case. Eleven labels were identified from all cases including sky, cloud, natural landscape, plant, tree, grass, landscape, people in nature, atmosphere, water, and snow. Site C and Mactaquac shared the label of wood, while Oldman and Site C shared ecoregion. Unique labels were happy and vehicle for Site C, mountain, highland, and grassland for Oldman, and lake, branch, and water resources for Mactaquac.

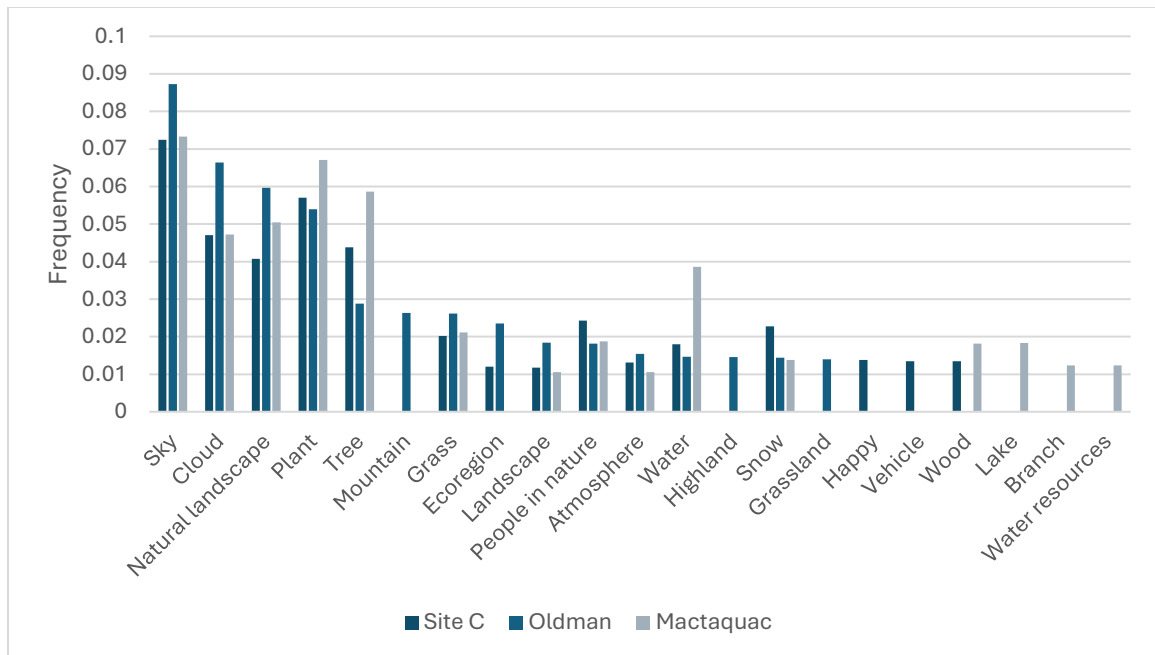


Figure 3.4 Top 15 labels for each study case

### 3.3.2 LDA landscape clusters

Table 3.1 shows the clustered topics according to the LDA model, and the topics were interpreted by authors according to key landscape features. The most prominent features in the study areas were plants and water. Mountains appeared more frequently with plants (including grass and grassland) in Oldman, and with water in Site C, but not at all in Mactaquac. Diurnal or seasonal views were clustered such as sunset, sunrise, and winter, in all three study cases. However, water only served as a highly weighted element in the sunset/sunrise landscape in Mactaquac. In terms of winter landscapes, Site C images included more recreation activities (top words include recreation and outdoor); Oldman mostly included pure landscape views with key elements like snow, ice, ice cap, and freezing; and in Mactaquac buildings frequently occurred with winter elements like snow.

There were landscape topic clusters involving human appearance and activities, such as people in nature, pets and animals, trip-related activities (road and vehicle), and recreation. They were common in all study cases but had some differences if comparing details of the top words in each set. Oldman had a unique cluster interpreted as people in nature, in which no top word indicated any activities. This might illustrate typical images showing that people were in the landscape largely to enjoy how it looks and take a photo.

Dog walking and playing were popular in all study areas, and it was more common in winter landscapes in the Mactaquac case. Birds (bird and beak) and fawn (young deer) were identified as prominent words in Oldman and Site C; while Mactaquac includes a more domesticated picture of animals (pets or wild animals) walking on asphalt roads. Road trips and vehicles were prominent key elements clustered by the model, which indicates that people often took pictures while driving or of their vehicles using the landscape as the background. In Site C only, electricity transmission or distribution lines were captured quite frequently within the road and vehicle landscape.

Recreational activities were different from case to case. The artistic photography cluster in Oldman was comparable to that of people in nature: people were taking pictures with the landscape in the background. In the Oldman and Site C case, artistic photography was clustered with smile and happy, indicating people's appreciation and enjoyment of the landscape. Besides photography and the winter recreation mentioned above, Site C had another cluster related to horseback riding and boating (based on observing photo contents, mostly kayaking and canoeing). In the Mactaquac case, smile and happy were also clustered with photography, while it had other recreational activities identified like biking and leisure activities in the water (e.g., swimming, beach visiting, yachting). Compared with Oldman and Site C, water was a more prominent landscape feature for recreational purposes in Mactaquac, mostly supported by the Mactaquac reservoir.

### *3.3.3 Energy landscape*

The LDA clusters only captured the landscape with wind turbines in the Oldman case and transmission line infrastructure in the Site C 'road and vehicle' topic, but no other energy-related topics for Mactaquac. The visibility of energy facilities, such as hydro dams, wind turbines, and solar panels, was also manually coded due to the inability of the pre-trained model to accurately recognize these objects when they appeared small in the background. The results indicate that the hydro dam itself was most frequently captured in the Mactaquac area (0.91%, compared with 0.16% for Oldman and 0.09% for Site C, the W.A.C. Bennett Dam), while wind turbines were prominent in the Oldman area (18%) but not much photographed elsewhere (0.03% in Site C and 0.01% in Mactaquac). Solar panels were rarely presented in the datasets: 0.04% in Site C and 0.07% in Oldman.

Table 3.1 Landscape topic clusters across cases. Asterisks denote the weight of the words in the topic: \*\*\* >0.1, \*\*>0.05, \*>0.025.

Topic interpreted based on key landscape features	Site C Dam, BC	Oldman River Dam, AB	Mactaquac Dam, NB
<b>Plant</b>	natural***, landscape***, plant***, sky**, tree**, cloud*, grass*, ecoregion*, wood*, environment*  plant***, twig**, tree*, flower*, wood*, trunk*, terrestrial*, grass*, branch*, shades	landscape***, natural**, sky**, cloud**, plant**, grass**, grassland*, mountain*, tree*, lot*	plant***, grass**, terrestrial*, flower*, tree*, leaf*, botany*, community, groundcover, flowering
<b>Water</b>	water***, sky**, cloud**, landscape*, lake*, landforms*, resources*, mountain*, natural*, body*	water***, landscape*, natural*, sky*, landforms*, resources*, lake*, fluvial*, streams*, plant	water***, natural**, landscape**, landforms**, resources*, plant*, watercourse*, tree*, streams*, fluvial*
<b>Energy</b>		wind***, natural**, windmill**, sky**, farm**, turbine**, landscape*, cloud*, ecoregion*, electricity*	
<b>Sunset/sunrise</b>	sky***, cloud**, atmosphere**, sunlight**, afterglow**, dusk**, landscape*, natural*, sunset, morning	sky**, landscape**, natural**, cloud**, atmosphere**, afterglow*, sunlight*, dusk*, ecoregion*, atmospheric	sky***, natural***, cloud***, landscape**, water**, atmosphere**, sunlight*, afterglow*, ecoregion*, dusk*
<b>Winter</b>	snow***, freezing**, slope**, tree**, sky**, recreation*, cap*, mountain, outdoor, ice	snow**, slope**, freezing**, sky**, mountain**, cloud*, cap*, tree*, landscape, ice	landscape***, tree**, natural**, plant**, wood**, branch*, sky*, snow*, building*, twig*
<b>Pet and wildlife</b>	dog***, carnivore**, animal**, breed**, fawn**, working*, mammal*, companion*, tree*, vertebrate*	dog**, carnivore*, animal*, bird*, sky*, breed, plant, fawn, terrestrial, beak	tree**, dog**, road**, sky**, surface*, asphalt*, plant*, animal*, carnivore*, snow*
<b>People in nature</b>		plant***, nature**, people**, natural**, sky*, landscape*, grass*, tree*, cloud*, flower*	
<b>Road and Vehicle</b>	road **, asphalt**, surface**, sky**, cloud*, tree*, light, electricity, line, building	road**, surface **, asphalt**, sky*, cloud*, landscape*, natural*, plant, ecoregion, line  vehicle***, tire**, automotive**, wheel**, sky**, car*, cloud*, motor*, lighting, hood	vehicle***, tire***, automotive***, sky**, wheel**, light*, cloud*, motor*, car*, lighting*

	vehicle***, tire***, automotive***, wheel**, sky**, car*, motor*, cloud*, lighting*, land		
<b>Recreation</b>	horse*, supplies*, equipment*, bird*, sky, beak, boating, boats, feather, watercraft  smile**, sleeve*, leg*, plant*, sky, hair, flash, photography, clothing, sunglasses  nature ***, people***, happy**, tree**, plant*, photography*, flash*, gesture*, sky*, smile*	sky*, nature*, smile*, people*, plant, happy, dress, photography, flash, cloud	nature***, people***, happy*, smile*, sky*, photography*, flash*, tree*, gesture*, bicycle*  sky***, water***, lake**, cloud**, tree*, equipment*, recreation*, blue*, leisure*, bridge*

### 3.4 Discussion

To answer the four research questions, we will discuss the results in terms of the key landscape features and activities identified (answering research question 1 and 2), their implications for the study cases (question 3), and the opportunities and limitations of using AI tools to analyze social media data for social impact assessment (question 4). Except for the results listed above, observations from viewing photo contents will also assist discussions.

#### 3.4.1 Key landscape features and activities

##### 3.4.1.1 Dams increase the prominence of water in the landscape

Our first insight is that a mega hydroelectric dam and its reservoir seems to increase the prominence of water in the dammed landscape, impacting uses and values. This is consistent with Zhang et al.'s (2009) observation in the Tree Gorges Reservoir area of Yangtze River from 1977 to 2005 when the waterbodies and built-up lands had a significant increase after damming the river by analyzing remote sensing images. In our case, among the landscape topic clusters, water was a key feature in Oldman and Mactaquac where the dams and reservoirs have existed for decades: they had various keywords such as water, watercourse, lake, fluvial, and streams. The increased visual impact is no surprise, as demonstrated in previous research that filling a reservoir raises the original river's water level and expands the width of a waterbody upstream by a significant amount (Liu et al, 2016), meaning the reservoir is perceived by people more like a lake than a river (Ioannidis & Koutsoyiannis, 2020). In Mactaquac, the reservoir provided a serene and lake-like view and motivated users to post water-related landscape photos, indicating their appreciation of aesthetics (also observed in Chen et al., 2018). This aligns with the findings in Sargentis et al. (2005) that water level changes in a reservoir did not negatively impact the aesthetic value of the landscape. However, the visual impacts can vary in different topographies. For instance, dam infrastructure was only captured in 0.16% of the posts in Oldman but 0.91% in Mactaquac, and others have shown that the visibility of dam infrastructure and reservoir can be confined in valley terrain (Dehkordi & Nakagoshi, 2004; Ioannidis & Koutsoyiannis, 2020): the Oldman area is more mountainous (e.g., mountain was identified as a keyword in the topic



clusters) than Mactaquac. In addition, transportation development (e.g., roads and bridges) can increase visibility by improving accessibility to the dam wall, reservoir, and related recreational facilities (Zhao et al., 2021). In Mactaquac, where transportation infrastructure is close to the reservoir and even crosses the dam (Lawson et al., 1985), water- and recreation-related keywords were clustered, consistent with previous studies in the same area (Chen et al., 2018; Chen et al., 2019).

#### 3.4.1.2 Losing landscape features and traditional practices to dams

Benefits brought by the development of hydropower and reservoir-based lifestyles often come at the cost of sacrificing important landscape features and undermining other social and cultural values. Dam-related changes in riparian lands can impact the traditional way of life by pushing out people or changing beyond recognition the places that enable it. Agricultural land losses have been a marked issue in previous studies (Swe et al., 2023; Tilt et al., 2009). This was also observed in the Mactaquac case with an aging dammed landscape, where more Instagram photos depicted farming scenes simulated at the tourism site built to house historic buildings that were relocated from the reservoir area, Kings Landing Historical Settlement (Lawson et al., 1985; Sherren et al., 2016). The Oldman area, though it has had a dam since the 1990s, remains less urbanized, continuing the traditional rural practices such as farming, grazing, hunting, and fishing identified in the Instagram photo contents, as described by Byrne et al.'s (2006) study of the Oldman River Basin. In Site C, where the landscape has not yet been completely altered by the dam, working animals and horseback riding were clustered in the topics. Farming will change significantly after impoundment in Site C, in ways it may not have in Oldman, in part because the good agricultural land along the Peace River is more constrained to low-lying riparian land in that high-topography area (Cox, 2018).

Another loss from damming rivers is that Indigenous people can be impacted significantly by giving up their ceremonial sites, seasonal camps, medicine collection areas, locations associated with oral histories, traplines, and fishing camps, as observed in Schapper and Urban's (2021) study in Canada, Aiken and Leigh's (2015) in Malaysia, and Jaichand and Sampaio's (2013) in Brazil. However, such losses may not be directly identified by the clustered landscape topics, in large part because of the biases inherent in

our data source and analysis method. In Site C, place names such as Attachie (one of the data collection geo-tags) have an important role in Indigenous oral history in the region and that continuity may be negatively impacted by inundation (Cox, 2018).

#### 3.4.1.3 Lifestyle can be reshaped by damming

Another insight relates to how lifestyles can be reshaped by damming in terms of leisure and rural-to-urban shifts in livelihood. Existing recreational activities in a pre-dam landscape can be impacted or limited when losing natural ecosystems or certain types of landscape features (Sæþorsdóttir & Ólafsson, 2010). Reservoir creation leading to flooding can often eliminate diverse landscape types and sites along the river such as gorges, beaches, and islands, reducing opportunities for some outdoor activities (Rodrigues & Silva, 2012). In Mactaquac, the once-famous beauty spot Pokiok Falls was submerged as the reservoir filled in the 1960s (Lawson et al., 1985), and a similar loss could occur in Site C following the dam construction and subsequent reservoir flooding, reducing sites for ice climbing on frozen waterfalls that were seen in the dataset. In addition, by viewing photo contents, the contrast between the popularity of kayaking and canoeing in Site C and motor/pontoon boating and cruising in Mactaquac indicates the different opportunities provided by two types of waterbodies, the turbulent river and the serene lake-like reservoir. Other activities like hunting and bird watching for particular species can be influenced or eliminated (Abreu et al., 2020), when the inundation of riparian lands disrupts crucial wildlife habitats (Rodrigues & Silva, 2012; Zhao et al., 2012). Relevant observations in this study include that clustered keywords of plant, grass, tree, flower, leaf, and botany were more related to urban green spaces such as parks and backyards than riparian areas in the dammed Mactaquac area, and more wildlife keywords (e.g., fawn, bird) were identified in Site C and Oldman. The contrast may indicate the limited or at least changing opportunities for wildlife-involved activities in dammed landscapes.

Reservoir-based landscapes also offer recreational activities that increase human interaction with waterbodies primarily in the summertime, which was also demonstrated by Mácová and Kozáková (2023) where visits to water reservoirs were sensitive to seasonality and the main recreation season was summer; a similar conclusion was made

from Kriz et al.'s (2020) survey. This was observed in Mactaquac: most of the water-related recreation happened in summertime according to viewing photo contents and no recreation-related keywords were identified within the winter topic. The adopted reservoir lifestyle, including most water-based recreational activities (e.g., swimming, sunbathing, boating), benefited people in Mactaquac more in the summertime due to accessibility and water temperature, something that was also found in previous work via manual coding of Instagram (Chen et al., 2018; Chen et al., 2019). In contrast, the pre-dam landscape of Site C also provided venues for outdoor recreational activities in the winter (keywords of outdoor recreation and winter features clustered).

The reshaping of livelihoods and lifestyles can be attributed in part to migration and rural development associated with dams. Dam construction, operation and reservoir tourism can create job opportunities and cause labor migration (Tilt et al., 2009), while development of roads and settlements near the site and reservoir can often accelerate waterfront development (Zhao et al., 2012), leading to a rural-to-urban shift in livelihood and lifestyle (Wilmsen, 2018). In Mactaquac, animals were clustered with asphalt and road, and buildings were a prominent element in the winter landscape. These topic clusters and Instagram photo contents showed a more anthropogenic and domesticated landscape, and an industrial (compared with farming) way of life that may be characteristic of some hydroelectric transitions. However, we can also see that such trajectories differ across the country. As mentioned earlier, in Oldman, despite dam construction, the reservoir's formation, and the establishment of a provincial park, high-quality agricultural land remains widespread so farming remains a vital practice in the area (Poirier & de Loë, 2011). Internationally, such differences also existed in Wang et al.'s (2013) study on cascading dams on the Upper-Mekong River in China. They found that the enhancement of public infrastructures, including roads, hospitals, and schools, could promote a more urbanized lifestyle, but regions maintaining agricultural conditions akin to those before dam construction enabled relocated farmers to sustain their farming practices.

#### 3.4.1.4 Hydropower infrastructure has limited visual impact

Compared to wind turbines, hydroelectric dam infrastructure seems to have limited visual impact or at least salience in the dammed landscape to attract the eye. Based on posts, the dam wall was more prominent in Mactaquac than Oldman, although both dams feature a road across the top and thus ready access to the infrastructure. The higher prominence in Mactaquac (0.91% versus 0.16% in Oldman) might imply more acceptance and awareness of dam infrastructure due to its longer existence, as observed by Keilty et al. (2016) in the same area, but there is also higher population density nearby and a flatter topography as discussed above. However, data bias might exist because the collection points in Oldman were not as close to the reservoir as they were in Mactaquac. In Oldman, Pincher Creek was within visible distance of the Castle River Wind Farm, which explained why wind turbines were widely seen in the dataset, a proportion (18%) much higher than that of the dam images in the Mactaquac case (1%). Additionally, wind energy facilities often have greater visual impacts than hydropower ones (Ioannidis & Koutsoyiannis, 2020). Salience analysis was used by Mohammadi et al. (2023) to understand the visual impacts of renewable energy facilities and demonstrated differences higher impact of wind than solar infrastructure, but built dam infrastructure can be very far from its reservoir impacts and thus would be lower still.

### *3.4.2 Implications for study cases*

The Mactaquac case, as an aging dam that is now approved for refurbishment, may help us to understand the future of Oldman and the in-construction Site C. Although the three cases were at different stages of a dam's lifespan, their landscapes had similar topics portrayed through Instagram use, which may indicate that a long-dammed landscape can be considered by residents as comparable to a natural one. This is consistent with Jørgensen's (2017) findings that people could develop different understanding of what 'nature' is and accept the dammed landscape as natural, which is also discussed by Ioannidis and Koutsoyiannis (2020). Jørgensen (2017) also noted that this can cause conflicts around dam removal or make recommissioning decisions difficult to resolve, which will happen sooner in Mactaquac than Oldman and Site C. The initiation and construction of a new mega hydroelectric dam can be highly controversial (Clarke et al., 2008; Sternberg, 2008), but the conflict at the end of a dam's lifespan can be just as

complex, as seen in local resistance to dam removal. Jørgensen and Renöfält (2012) studied cases in Sweden and found that proponents and opponents could have different framings of the impacts of the dam removal: ecosystem services and river fishing versus cultural services of recreation, aesthetic, and heritage. Similar findings were also discussed in Fox et al.'s (2016) research in New England, US. Such controversy happened in Mactaquac as well in the 2010s, when local and wider-scale discourses converged to resist the loss of the dam and reservoir (Sherren et al., 2017). Yet changes of landscape, livelihood, and lifestyle may benefit very different people among residents (Keilty et al., 2016), as well as between locals and visitors (Terkenli et al., 2019). Hence, decision makers involved in large-scale hydroelectric projects should not only balance the benefits against the potential losses, but consider who is impacted and who benefits to make more informed decisions about just landscape transitions.

### *3.4.3 Opportunities and limitations of AI tools to analyze social media data for SIA*

#### 3.4.3.1 Opportunities

Social media data has been shown to offer valuable insights into social science topics, including landscape uses and values, due to its extensive scale and the cost-effectiveness of data collection (Chen et al., 2023). Even though the study areas were not in densely populated cities but rather in rural Canada, over 86,000 posts were gathered. The recent advancements in AI tools, particularly pre-trained models, enable the automated analysis of large-scale social media data including text (Gone et al., 2023) and images (as shown here). These models are user-friendly, cost-efficient, and highly accurate (Vigl et al., 2021). Additionally, commercial AI platforms offer tools for training customized models, which aids researchers with limited programming skills to create models better suited to their specific data and cases. Labels generated by image detection models can be further analyzed by Natural Language Processing models such as LDA. These opportunities for data acquisition and quick analysis can significantly enhance the SIA toolkit.

#### 3.4.3.2 Limitations

The limitations of this study mainly appear in two aspects, regarding the social media data and the AI tools for analysis. The accessibility of social media data for many most

popular platforms has been significantly limited since the Cambridge Analytica Scandal in 2018 (Confessore, 2018). Many large social media platforms such as Instagram and X have removed free access for public search, especially the fine-grained location data (Freelon, 2018). Thus, in this study, we relied solely on existing location tags on Instagram to collect data, resulting in gaps in remote areas where no place names were created as tags. For example, in the Oldman River case, only one location was adjacent enough to the reservoir; the small town named Cowley was located upstream, and the largest dataset was retrieved from Pincher Creek which was 10 km from the water. By contrast, 25 location tags were identified along the reservoir in the Mactaquac area. Moreover, data collection was time-intensive and required a certain level of programming skills (Chen et al., 2024). Collecting Instagram data by geo-tags encountered other challenges like using different names for the same place (e.g., Oldman River Dam, Oldman Dam, Old Man Dam, etc.) and unpredictable geographic location for a lengthy river (e.g., the geo-tag of Oldman River was located far away downstream from the dam area). Besides, lacking demographic details of social media users in the datasets can be a problem to distinguish residents from visitors and understand how landscape changes can impact different groups of people.

AI tools leveraged in this study have limitations as well. Google Cloud Vision API has limited accuracy of identifying specific objects in the image content, such as distinguishing a reservoir from a lake, or a year-round house from a summer vacation house. It could only return a maximum of 10 labels per image at the time when analysis was conducted. Important landscape features, the energy infrastructure for instance, can be overlooked if they appear small or are unfamiliar to the trained model (e.g., an earthen dam can look like natural topography). Those are limitations commonly shared by pre-trained models (Gosal et al., 2019). Another limitation is that pre-trained models are often better at detecting high- to median-level concepts, and details and nuance can be weak (Pathak et al., 2019). In this study, high-level concept categories were identified like nature, atmosphere, ecoregion, plant, and waterbody. However, Vision API also detected very detailed categories like damselfly and apple, which caused unevenness of the results and made the interpretation work quite challenging. In addition, the pre-defined categories may contain some uninterpretable and useless labels such as ecoregion.

Regarding the LDA topic clustering model, it requires knowledge and skills in machine learning, the model tuning is time-consuming (Qomariyah et al., 2019), and noise in the datasets may influence training performance (Yoon et al., 2021).

### **3.5 Conclusion**

This study employed social media image data and artificial intelligence models including an automated image coding tool (Google Cloud Vision API) and a topic clustering model (Latent Dirichlet Allocation) to understand landscape changes and social impacts in three study areas involved hydroelectric dams in Canada. Over 19,000 valid landscape images were retrieved from Instagram, generating more than 180,000 labels detected by Vision API. The topics clustered by the LDA model were interpreted by key features such as plant, water, energy infrastructure, sunset/sunrise, winter, pet and wildlife, people in nature, road and vehicle, and recreation. The similarities among the three study areas suggest that the construction of dams and the formation of reservoirs have many similar impacts on different regions, while the differences may provide insights about landscape changes at varying stages of a dam's lifespan or by geographic characteristics, offering opportunities for cross-site learning. For example, the expansion of the waterbody by reservoir filling made water more prominent in the dammed landscape, benefiting some locals in various ways while causing losses to others. Lifestyles in the dam era will likely be reshaped due to increased water accessibility and calmer flows that may enhance the variety and popularity of reservoir-based recreational activities, particularly in the summer. Additionally, dams can contribute to suburban development and tourism, leading to more modernized and urbanized lifestyles; however, the loss of certain types of landscapes, such as agricultural land and Indigenous traditional lands, harms cultures and activities. The visual impact of hydroelectric dams may be less than that of other renewable energy facilities like wind turbines, and their long-term presence may increase the acceptance of the infrastructure, complicating decisions regarding aging dam projects. Methodologically, this study demonstrates that social media data and AI models can aid in understanding energy landscapes and efficiently processing large datasets, making them viable as a novel tool for social impact assessment. However, they also have certain

limitations, including data biases, and the unevenness of results generated by pre-trained automated image analysis tools.



## **Chapter 4 Using computer vision to assess cultural ecosystem services relating to hydropower landscapes in Canada.**

A version of this chapter is in preparation for submission to the journal *Ecosystem Services* (by Elsevier). Refer to Appendix B for the copyright agreement to reproduce this material.

Chen, Y., Smit, M., Lee, K, Y., McCay-Peet, L., Margeson, K., & Sherren, K. (In preparation). Using computer vision to assess cultural ecosystem services relating to hydropower landscapes in Canada. *Ecosystem Services*.

Statement of Authors' Contributions: YC is responsible for conducting the literature review, data collection and analysis, and writing all sections. KS and MS supervised the writing development including providing feedback and editorial revisions. KL and LMP provided feedback and editorial revisions on a final draft. KM coded the training dataset to validate the coding results.

## **Abstract**

Cultural ecosystem services (CES) assessment is challenging due to its complex and subjective nature. However, CES analysis can be a promising tool for social impact assessment in large hydroelectric dam projects, which often cause extensive landscape changes and social disruptions. Recently, social media data has demonstrated its utility in understanding CES provision. With recent advancements in artificial intelligence, particularly in computer vision, automatic analysis of social media image content has become feasible. In this research, we trained a computer vision model to identify CES-related themes from over 18,000 social media images collected from three study areas involving dam projects, achieving an average precision of 93.8%. The results reveal how damming for hydroelectricity impacts CES provision, with insights including limited effects on aesthetic values, the local nature of place identity, potential for recreational enhancement, and the need for careful planning to safeguard cultural heritage. Importantly, this study underscores the potential of using large-sized social media image data and customized computer vision models in CES assessment, discussing benefits (including compared with manual approach and pre-trained models), limitations, and future research directions.

## 4.1 Introduction

Cultural ecosystem services (CES) are one of the four main categories of ecosystem services, along with supporting, regulating, and provisioning services (Millennium Ecosystem Assessment, 2005). Ecosystem services are defined as “the benefits people obtain from ecosystems” (MEA, 2005, p. V). CES often refers to the non-material benefits arising from human-nature interactions (Tengberg et al., 2012). CES are key to human wellbeing, related in part to good social relationships, a sense of cultural identity, and a sense of security (Díaz et al., 2006; Kosanic & Petzold, 2020). Widely agreed categories of CES include recreation, aesthetic values, spiritual, education, cultural heritage, intrinsic (existence), inspiration, sense of place, knowledge, social relations, and cultural diversity (Milcu et al., 2013). A more recently updated guidance, Common International Classification of Ecosystem Services (CICES 5.1 by Haines-Young & Potschin, 2018), listed similar classes of culture services related to biotic and abiotic aspects. The CES categories investigated in specific studies vary depending on the topic, the study area, and data. Márquez (2023) found that recreation, aesthetic values, cultural heritage, and sense of place and identity were most assessed among papers published from 2010 to 2022, while knowledge systems, cultural diversity, and sports services (recently proposed in research articles) were the least touched. In general, cultural services are not tangible, and so it can be more difficult to integrate into assessments based on monetary value (Chan et al., 2012; Hirons et al., 2016).

Social impact assessment (SIA) has now been widely discussed and practiced as a tool to predict and manage the social issues of development (Esteves et al., 2012), including energy projects (Buchmayr et al., 2022). Large hydroelectric dam developments can be an immense driver for changes in landscape use and values in host areas, such as altering traditional practices (Hernández-Ruz et al., 2018) or developing a recreational lifestyle around lake-like reservoirs (Chen et al., 2018). Darvill and Lindo (2016) mentioned that the development of a hydroelectric dam can impact cultural aspects of ecosystem services including heritage resources, cultural and provisioning use of biodiversity, and recreation. Hydroelectricity landscapes can be complex since they are often an amalgam of the pre-dam landscape and subsequent alterations brought about by the new ecosystem and

residents in the dammed landscape (Calvert et al., 2019). There are relatively few studies focused on cultural services of hydroelectric-related landscapes (Davis & Kidd, 2012; Hale et al., 2019). However, understanding baseline pre-dam CES provision is important for SIA, while monitoring CES changes may help to assess the social impacts brought by such dam projects.

Lacking sufficient data and tools for assessment has been a challenge in CES analysis when integrating it into decision making and landscape management (Kosanic & Petzold, 2020). The shift toward leveraging social media data and recent advancements in computer vision (CV: a sub-field of artificial intelligence) brings opportunities (Langemeyer et al., 2023; Havinga et al., 2020). Social media data has been proven to be valuable to understand social phenomena and values (Chen et al., 2023), with some advantages over traditional survey methods: large data size, extensive spatial and temporal scales, low cost, detailed metadata (e.g., geographic tag, time stamp), various types of data (e.g., text, image, video) (Havinga et al., 2024). There is increasing research leveraging social media to assess cultural services, such as Ruiz-Frau et al.'s (2020) applied network analyses on hashtags collected from Instagram and Twitter to assess CES in coastal areas. More recently, Benati et al. (2024) retrieved Twitter data to probe the relationship between accessibility of urban green spaces and CES provision. Nevertheless, social media data has not been fully taken advantage of due to increasing constraints on data access (Chen et al., 2024) and limitations in analysis models and methods (Manley et al., 2022). Manually processing and analyzing large datasets often demands immense human effort (Chen et al., 2019). Thus, even recent CES measurements heavily relied on generic indicators like counting numbers of posts and geographic tags, without considering the contents (Havinga et al., 2024).

The recent rapid advancement of artificial intelligence facilitates the substitution of some human labor with machine power, allowing in-depth analysis when extracting meanings from textual or visual contents (Johnson et al., 2021; Vigl et al., 2021). The first wave of leveraging artificial intelligence in CES studies started with Natural Language Processing (NLP) models on analyzing textual social media data. For example, recently Gugulica and Burghardt (2023) applied NLP to annotate Flickr and Instagram textual data to map

CES indicators in urban green spaces. Gone et al. (2023) used NLP to understand human activities impacted by hydroelectric energy projects. More recently, computer vision has been used to detect image content. Richards and Lavorel (2022) used a pre-trained model to extract keywords from Flickr images to understand CES in New Zealand. Cardoso et al. (2022) trained a Convolutional Neural Network model to automatically classify natural and human elements related to CES based on Flickr and Wikiloc images. Constructing and training models for computer vision often requires more sophisticated design, tuning, and calculating power, which drives many studies to rely on these ready-to-use pre-trained models (Chen et al., 2024; Huai et al., 2022). However, such models can have characteristics such as pre-defined categories, high- to mid-level categories, useless or irrelevant categories, limited accuracy, and tend to miss details and small objects in photo contents. With more platforms providing services to train a model based on users' own data, without learning coding and algorithms, custom-trained models can be a promising opportunity for future research.

In this study, we gathered Instagram images by geo-tags from three study areas associated with hydroelectric dams in Canada: The Site C Dam (Site C) in British Columbia (construction pending in late 2024), the Oldman River Dam (Oldman) in Alberta (constructed in the 1990s), and the Mactaquac Dam (Mactaquac) in New Brunswick (built in the 1960s and approved for refurbishment in 2016). There were two main objectives. First, we aimed to explore the feasibility of training a model to categorize images according to CES coding themes (i.e., themes for content analysis) by evaluating the accuracy of the training. The complexity stems from the abstract and subjective nature of CES, which encompasses a broad spectrum of human experiences, perceptions, and cultural contexts (Mouttaki et al., 2022). The discussion will extend to the general application of social media data and CV tools in CES research, as well as future directions. Second, we aimed to use the trained model to categorize valid images collected from the three study areas, to indicate the provision of CES at these somewhat chronosequenced locations, shedding light on the impacts upon CES from hydroelectric dams. The Mactaquac site was chosen for more detailed interpretation at the geocode scale as an example to show the patterns within one site.

## 4.2 Methods

### 4.2.1 Study areas

The study areas in this research include three regions across Canada, each featuring a hydroelectric dam project at a different maturity: the Site C, scheduled to begin reservoir filling and power generation in 2024; the Oldman, constructed in the early 1990s with power generation starting in the early 2000s; and the Mactaquac, built in the late 1960s and approved for refurbishment in 2016 (see Figure 3.1). Given that social media data is only available from about 2006 onward, obtaining longitudinal data over an entire dam lifespan is challenging, particularly for rural areas. Therefore, selecting three hydroelectric dams at different stages can illuminate changes in CES provision over the lifespan of a dam, using space for time replacement.

Site C is in the Peace River Valley, an area with a longstanding history of farming and ranching. It is the third mega-hydroelectric infrastructure on the Peace River in northern British Columbia (Clarke, 2014), producing up to 1,100 megawatts, forming a reservoir 83 km long and up to 55 meters deep (Canadian Environmental Assessment Agency, 2014), and inundating a total of 5,550 hectares of riparian lands (BC Hydro, 2018). The Oldman Dam, situated at Three Rivers in southern Alberta, was originally constructed for agricultural irrigation to assist farmers frequently facing extreme droughts (Glenn, 1999). In 2003, it was augmented with a 32-megawatt hydroelectric facility (Clarion Energy Content Directors, 2003), with the total area of the Oldman River Reservoir covering 2,203 hectares (Angler's Atlas, 2024), considerably smaller than the other two. Mactaquac Dam, built in the late 1960s on the St. John River in New Brunswick, is the largest hydroelectric dam in Atlantic Canada (Bourgoin, 2013). It has an installed capacity of 670 megawatts and the reservoir has inundated approximately 5,000 hectares of riparian land. The dam's structural integrity was compromised by an alkali-aggregate reaction, leading to a premature service life end by 2030. Consequently, in 2015, a decision was made to refurbish the dam, opting against other alternatives like retaining the reservoir without power generation or river restoration (Stantec, 2016).

### 4.2.2 Data collection

Data was gathered from Instagram, one of the leading visual content-sharing platforms, launched in 2010. As of January 2024, it boasted over two billion users, ranking fourth after Facebook, YouTube, and WhatsApp (Statista, 2024). Instagram was selected for its prevalent use by individuals posting everyday photos, which are crucial for understanding CES provision through daily land use (Chen et al., 2018; Gugulica & Burghardt, 2023). Following a 2018 update, Instagram restricted searches by precise geographic coordinates (Chen et al., 2024; Kaiser et al., 2021), necessitating reliance on pre-set geo-tags for location-based post collection. We initially used Google Maps to search for cities, towns, parks, visiting sites, and landmarks around the three hydroelectric dam sites and near their existing or planned reservoirs. These locations were then checked on Instagram for: 1) if there were associated geographic location tags (geo-tags); and 2) if there were posts tagged with them. When both conditions were satisfied, location IDs from Instagram and geographical coordinates from Google Maps were documented. Our study found 5 geo-tags at Site C, 4 at Oldman, and 25 at Mactaquac, as listed in Appendix C. The variation in geo-tag numbers across study areas can be attributed to Mactaquac's higher population density and the presence of numerous towns along its reservoir, in contrast to the more rural settings of the other sites.

To collect Instagram images and metadata, we created a scraping tool using two main Python packages: Selenium and Instascrape (Chen et al., 2024). Selenium was employed to log into Instagram, search for posts by location IDs (i.e., geo-tags) from the study areas, and record post links in a log file (Titration, 2022a). This initial phase ran from May 30 to June 2, 2021, and the earliest post retrieved dated back to July 18, 2011. In this phase, we identified 49,351 posts for Site C, 8,727 for Oldman, and 28,393 for Mactaquac (Figure 4.1). Subsequently, from June 7 to July 25, 2021, the Instascrape package was utilized to scrape images (or videos) and metadata from the saved post links (Titration, 2022b). Approximately two percent of the posts were not successfully scraped, and video data were excluded, resulting in 44,875 images for Site C, 7,980 for Oldman, and 26,178 for Mactaquac. We manually excluded images where landscape features constituted less than 60% of the content, such as indoor photos and selfies that did not prominently display landscapes. Consequently, 6,817 images remained as valid data for Site C, 3,060 for Oldman, and 9,292 for Mactaquac. Although we attempted to develop a

CV model to autonomously filter the raw datasets, the complex nature of the data and the limitations of computer vision technology in 2021 hindered its accuracy and we relied on manual analysis.

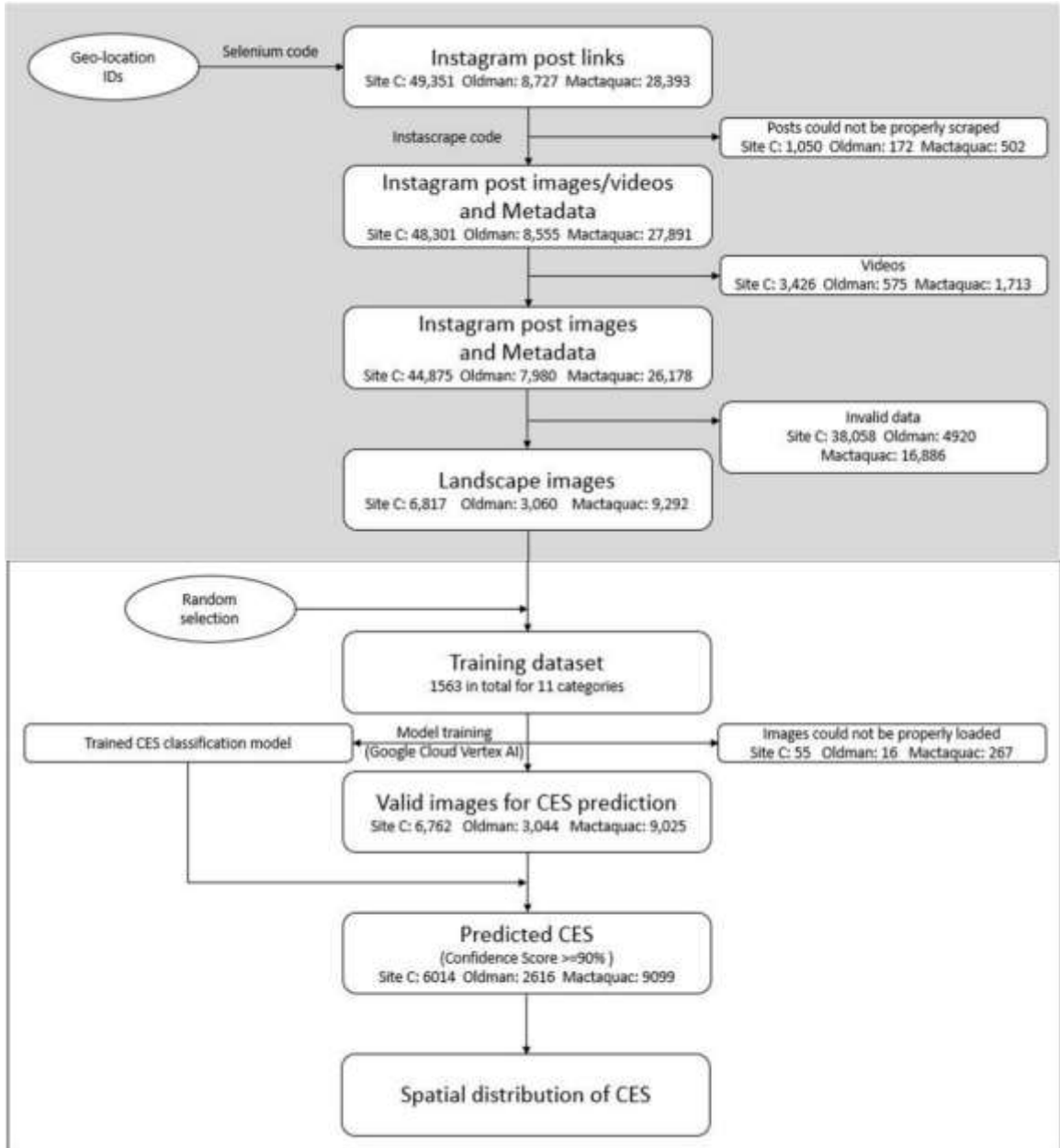


Figure 4.1 Workflow of data collection (shaded background) and analysis including model training and prediction (white background).

#### 4.2.3 CES coding themes and training dataset preparation



Modern computer vision models are typically based on Convolutional Neural Network (LeCun et al., 1998). Researchers can construct these models using open-source libraries like Keras for Python, which allows them to customize the architecture. They can also utilize structured frameworks such as ResNet-152 or opt for commercial platforms that offer computer vision models ready for customer-specific training (e.g., Google Cloud Vision Vertex AI), during which the model refines itself to produce accurate responses by learning from prepared data. Finally, they can use pre-trained models with various capabilities and no customization. We chose to use Vertex AI because it offers the most time- and cost-efficient solution for researchers who may not be proficient in algorithms and model tuning, but still allows the creation of custom-trained models using a prepared training dataset.

Media-based CES measures face a challenge in identifying indicators of people's positive experiences of nature: how does an individual piece of media convey a person's experience (Havinga et al., 2024). Existing CES research leveraging CV tools relied on annotating physical features (Cardoso et al., 2022; Mouttaki et al., 2022), often with specific and distinct foci; for example, Mouttaki et al. (2022) coded fishing recreation and Cardoso et al. (2022) recognized species of animals and plants. In this study, the provision of CES was identified with a similar approach through coding themes: images were not directly assigned to CES categories, instead we chose to assign images to feature-based coding themes based on a detailed codebook (Table 4.1 and Appendix E). CES complexity and subjectivity necessitates linkage to the physical features of ecosystems where value is attributed through conceptual bridges like place and landscape (Gee & Burkhard, 2010; Milcu et al., 2013). Given this fact, and since CV tools excel in identifying objective attributes, we crafted coding themes anchored in tangible elements, and then interpreted coding themes in the sense of CES provision (Table 4.1). This approach aids human coders in categorizing images using clear-cut and replicable standards. For instance, instead of coding the aesthetic value of a landscape on coders' subjective perceptions of beauty, we focused on coding wide-angled natural landscape views. Subsequently we trained the CV model by feeding it with manually coded data to identify such images, thereby inferring the landscape's aesthetic value as indicated.

Under this framework, eleven coding themes were identified, indicating 6 CES categories (Table 4.1, based on Milcu et al.'s [2013] work). Each coding theme also needed enough data of minimum 100 images for training as required by Vertex AI. Landscape aesthetics were primarily denoted by coding themes of 'landscape' and 'natural features'. Other themes like 'human in nature' and 'object in nature' in which the landscape serves as a background also indicate aesthetic appreciation for the landscape. The concepts of sense of place and place identity were captured through place-specific features such as signage, flags, or landmarks (e.g., bridges, railroads, statues, dams), and sense of home that was coded to identify the images showing private houses and their proxies for ownership and belongingness. Recreational value was implied by the category of general 'recreational' and 'dog walking'. We separated dog walking from other recreations because it was pervasive across various locations and might have a close connection with sense of home. Social relations were reflected directly from social events and gatherings. Cultural heritage values were identified by historical sites and buildings and the agricultural practices presenting rural cultural heritages (Swinton et al., 2007).

The training dataset was coded by two coders (YC and KM) based on the codebook (Appendix E). We randomly selected 1563 images<sup>1</sup> and both coders independently categorized them into the 11 feature-based coding themes; multi-category coding was allowed. Disagreements accounted for 8.2%, and there was an additional 3.3% where at least one of the coders was unsure. Agreements were reached after discussions with a third person (KS). We originally included spirituality and religion services that were coded by relevant activities or buildings. However, there were insufficient images to train the model for this category.

---

<sup>1</sup> Initially, 1,500 images were randomly selected. However, to meet the minimum requirement of 100 images per category for training, an additional 63 images were added. Categories were multi-coded.

Table 4.1 Coding themes and cultural ecosystem services (adapted from MEA, 2005; more details can be found in Appendix E).

<b>Coding themes</b> (# of images coded in training dataset)	<b>Main coding criteria</b>	<b>Cultural ecosystem services</b>	<b>Description (adjusted to suit research purpose)</b>
(Natural) Landscape (150)	The image mainly focuses on aesthetic appreciation of the natural landscape and there is no indicator to assign this image to any other category.	Aesthetic values	People find beauty or aesthetic value in various aspects of ecosystems, as reflected in scenic drives and the selection of taking pictures, including using nature as a backdrop for pictures of people and things.  (Cardoso et al., 2022; Mouttaki et al., 2022; Richards & Friess, 2015)
Natural features (300)	The image aims to appreciate the natural features (e.g. animal and plant) in detail or at the focus point.		
Human in nature (158)	The image shows the person(s) in the place without indicators for specific activities or purposes (except for being in the place).		
Object in nature (121)	The image shows inanimate object(s) using landscape as photo setting.		
Place-based features (213)	The image shows place-based features such as signage, flag, and landmark.	Sense of place and identity	People recognize features of their environment that represent a strong identity to them like their home place or memories attached to a place.  (Chen et al., 2020)
Sense of home (100)			
Recreational (245)	The image shows recreational activities or equipment.	Recreation	People often choose where to spend their leisure time based in part on the characteristics of the natural or cultivated landscapes in a particular area.  (Fox et al., 2021; Lingua et al., 2022)
Dog walking (170)			

Social relationship (194)	The image shows social gatherings and events, or people with intimacy (e.g. arm in arm, hug, kiss, etc.)	Social relations	Ecosystems influence the types of social relations that are established in particular cultures/places.  (Riechers et al., 2016; Xin et al., 2020)
Historical features (163)	The image shows historical site or building, or activity.	Cultural heritage values	Many societies place high value on the maintenance of historically important landscapes, such as historical sites and agricultural traditions.  (Power; 2010; Richards & Friess, 2015; Swinton et al., 2007)
Agriculture (148)	The image shows agricultural land, activity, or equipment.		

#### 4.2.4 Model training and performance, and CES prediction

Utilizing the coded images as the training dataset, we trained a model through the Vertex AI platform. The 1563 images manually labelled (Table 4.1 shows details for each coding theme) were split into: 80% for training (1247), 10% for validation (157), and 10% for testing (159). The trained model was then used to predict all valid photos which were automatically assigned to one or multiple coding themes. Only those with a confidence score higher than 90% were kept. As a result, based on this threshold, 5406 images (90.6% out of valid images) coded with at least one theme in Site C, 2371 in Oldman (89.9%), and 7738 in Mactaquac (85%).

The prediction results (shown in Table 4.2) were displayed by the average percentage of each coding theme per study case that was calculated as follows (using agriculture in Site C as an example):

$$\text{Average percentage of agriculture in Site C} =$$

$$\frac{\text{sum of images predicted with agriculture in Site C (3 valid geo-tags)}}{\text{sum of total valid images in Site C (3 valid geo-tags)}}$$

And, CES richness aims to show that one place may provide multiple CES. It was calculated as follows:

*CES richness (average coding themes per image) =*  
*sum of themes coded from valid images in Site C (3 valid geo-tags) / sum of total valid*  
*images in Site C (3 valid geo-tags)*

*or*

*CES richness for images have at least one theme coded =*  
*sum of themes coded from valid images in Site C (3 valid geo-tags) / sum of images have*  
*at least one theme coded in Site C (3 valid geo-tags)*

## **4.3 Results**

### *4.3.1 Model training results*

The average precision (calculated by the area under the precision-recall trade-off curve) for the whole training set was 93.8%, the precision (the percentage of predictions that were correct or true positive) was 92.9% with the confidence threshold at 90% (this threshold will also be used when analyzing predicted results), and the recall was 77.3% (the percentage of all ground truth items that were successfully predicted by the model). For each coding theme, the precision at 90% confidence threshold was landscape (85.7%), dog walking (100%), human in nature (91.7%), historical features (87.5%), sense of home (87.5%), agriculture (100%), place-based features (100%), object in nature (100%), natural features (85.7%), recreational (93.8%), and social relationship (94.4%). However, there is no universally agreed-upon threshold for what constitutes a sufficiently well-performing model; this can vary depending on user expectations and the baseline performance achievable by humans (Kay et al., 2015). In this case, the 8.2% disagreement rate and an additional 3.3% uncertainty in manual coding can serve as a reference. The precision of 92.9% at a 90% confidence threshold is comparable.

### *4.3.2 CES prediction results*

#### *4.3.2.1 Comparison among study areas*

In Mactaquac, there were 26 geo-tags for data collection, of which 14 had over 100 valid landscape images and 4 had over 1000 (Woodstock, 2197; Mactaquac Provincial Park,

1593; Hartland, 1288; and Kings Landing, 1187). Site C had fewer geo-tags (5 in total) with most valid images concentrated in Fort St. John (5781), followed by Hudson's Hope (907) and Bear Flat (55). Oldman had 4 with only two yielding large datasets: Pincher Creek (2526) and Cowley (470). To minimize bias from smaller geocode datasets, the analysis and discussion will focus only on the larger datasets (those geocodes with  $\geq 100$  images for Mactaquac, and  $\geq 50$  for Site C and Oldman). More detailed results for each study area are available in Appendix F.

Mactaquac had the largest dataset with 9139 images and the most themes coded at 8957, followed by Site C with 6743 images and 5951 themes, and Oldman with 2996 images and 2568 themes (Table 4.2). CES richness was highest in Mactaquac (0.98 coding themes per image for all valid photos; and 1.1 for images have at least one theme coded), followed by Site C (0.88, 1.08) and Oldman (0.85, 1.15). Mactaquac also had five coding themes with over 10% prevalence where the other two sites only had three. Aesthetic value relates to coding themes of landscape, natural features, humans in nature, and objects in nature. There were no significant patterns indicating impacts from damming on the landscape. Natural landscape images were slightly more prevalent in Oldman (19.3%), while the other three themes were more concentrated in Site C. Place identity is interpreted through themes of place-based features and sense of home, as defined in Table 4.1. Mactaquac had the highest percentage of place-based features (13.4%), with no notable difference in sense of home across the areas. Recreational activities were more frequently identified in Mactaquac (8.8%). Dog walking was slightly higher in Site C (6.6%). Social relationships, including social events and gatherings, were more prominent in Site C (13.5%). Historical features were mostly coded from Mactaquac (12.7%), while agricultural heritage was most concentrated in Oldman (18.2%).

Table 4.2 Average percentage of coding themes in study areas (top 3 bolded for each case; highest site for each coding theme shaded). Note that the columns will not sum to 1 because images could be categorized in more than one theme (or none).

	Site C, BC	Oldman, AB	Mactaquac, NB
Total valid Image#	6743	2996	9139
Image# coded with $\geq 1$ theme at 90% confidence score	5406	2371	7738
Coding theme#	5951	2568	8957
Coding themes:			
(Natural) Landscape	<b>16.5%</b>	<b>19.3%</b>	<b>18.0%</b>
Natural features	<b>19.3%</b>	<b>17.4%</b>	<b>16.5%</b>
Human in nature	5.6%	2.8%	4.5%
Object in nature	4.1%	1.9%	1.5%
Place-based features	4.8%	6.3%	<b>13.4%</b>
Sense of home	2.7%	1.6%	2.4%
Recreational	7.5%	4.6%	<b>8.8%</b>
Dog walking	6.6%	4.2%	5.0%
Social relationship	<b>13.5%</b>	6.6%	10.2%
Historical features	1.4%	2.6%	<b>12.7%</b>
Agriculture	6.3%	<b>18.2%</b>	5.0%
CES Richness for all valid images	0.88	0.85	0.98
CES Richness for images have at least 1 theme coded	1.10	1.08	1.15

#### 4.3.2.2 Results for Mactaquac

This subsection will focus only on the results in Mactaquac; additional details are available in Appendix F, which includes tables for each study area displaying results by coding themes and geo-tags. The Mactaquac case was selected for detailed examination due to its distribution across 14 geo-tags, all very proximate to the dam and reservoir, suggesting reduced bias as no single geo-tag dominates the dataset.

In Mactaquac, CES richness exceeded 1 in four geo-tags, specifically Hartland (1.164), Nackawic (1.004), Bear Island (1.031), and Kings Landing (1.073): the first two are a historical town and a town built alongside the dam over 50 years ago to house those relocated from the area of reservoir inundation, Bear Island is a rural community and campground, and Kings Landing is a tourist site and historical re-enactment site comprising historic houses. In terms of aesthetic coding themes, seven geo-tags had over 20% of images coded by natural landscape and six by natural features, with the most

prominent being Mactaquac Provincial Park, Davidson Lake (characterized by cottages and second homes on a lake near the reservoir), rural areas like Dumfries and Prince William, and Bear Island's campsite. Place-based features were primarily identified in upstream old towns such as Hartland (37.3%), Woodstock (11.8%), and Meductic (12.3%), as well as in the newer town of Nackawic (31.9%). Agriculture was most prevalent (16.7%) in Kings Landing, which simulates early settlement farming activities, which also recorded the highest percentage of historical features (49.6%). Sense of home was identified highest in rural areas like Dumfries and Keswick Ridge. Recreational activities were most frequently identified in tourism parks like Woolastook Park (26.3%) and Mactaquac Provincial Park (20.4%), as well as at Davidson Lake (23.5%).

## **4.4 Discussion**

### *4.4.1 CES provision*

#### ***Aesthetic value may not be widely impacted by long-term damming.***

Damming a natural landscape may not necessarily diminish its aesthetic value over time, especially regarding the appreciation of expansive, unobstructed views. This observation emerges from comparing three study areas: the highest frequency of natural landscape views was captured in Oldman, dammed 32 years ago, followed by Mactaquac, dammed for 56 years, and lastly, the pre-dam landscape in Site C. The numerical differences were not substantial, which can be explained by Ioannidis and Koutsoyiannis (2020) who suggested that artificial reservoir-based landscapes can eventually be perceived as 'natural.' Jørgensen (2017). Keilty et al. (2016) also discussed how 'nature' can be variously interpreted, allowing people to accept dammed landscapes as natural. A slight potential decline in the experience of aesthetic landscapes due to dam construction might be inferred from the more frequent appearance of 'human in nature' and 'object in nature' themes in the pre-dam landscape of Site C compared to the dammed landscapes in Oldman and Mactaquac. In our study, these coding themes often depicted individuals or man-made items within the natural scenery, likely chosen for their perceived beauty as backgrounds, as also suggested by Kaiser et al. (2021). This is consistent with interpretations of users' behaviors and expectations when sharing content on social media, particularly on Instagram. Lee et al. (2019) suggested that anticipation of



encountering picturesque natural locales, as might be expected for Site C in a pre-dammed state, leads to a higher prevalence of nature-centric images on social platforms. However, as noted by Art et al. (2021), a common trend of emphasizing appealing landscapes in social media posts can result in the prevalence of such images across study areas, whether there was a dam or not. Gugulica & Burghardt (2023) also observed that aesthetic appreciation was more prominently identified in Instagram posts than on Flickr. A 3-point chronosequenced approach like the one we used is not temporally granular enough to detect the aesthetic impacts immediately post-dam, but this will soon be possible for Site C.

While the presence of an energy facility often reduces the aesthetic appreciation of a place, the impact is generally less significant for hydroelectric dams compared to more visible energy infrastructure like wind turbines (Gee & Burkhard, 2010; Ioannidis & Koutsoyiannis, 2020). In Mactaquac, for instance, the natural landscape coding theme was most frequently identified in the natural setting of Davidson Lake (not on the reservoir) and the theme of natural features at Bear Island, a campsite well upriver from the dam infrastructure. However, landscape aesthetics near the dam did not show a marked decrease compared to points upriver, as open landscape views were still commonly captured (details in Appendix F). This suggests that the impact of hydroelectric dams on aesthetic values might be subtle and limited to areas within visible range of the infrastructure itself, although even that can be difficult to identify up close from the reservoir side (Chen et al., in review).

***Sense of place and place identity impacts can be quite place-specific.***

The findings of our study may indicate a trend where locations dammed for longer periods are more likely to feature place-based elements, suggesting an enhanced sense of place and place identity. However, the key determinant appears to be the settlement history of the location rather than the presence of the dam, as evidenced by the prevalence of place-based features in older towns upstream in Mactaquac. Some elements contributing to the symbolic value of a place and its identity may be compromised by hydroelectric developments. A notable example, Kings Landing near Mactaquac, involves historical buildings that were rescued to avoid reservoir flooding, subsequently

transforming the area into a heritage site for visitors in the dammed era (Lawson et al., 1985).

The concept of 'sense of home' might remain stable despite the damming of a river if it is strictly linked to residential settlements (as per our coding guide), since no notable differences existed across study areas. Yet, a broader interpretation of sense of home might include individuals' farms, workplaces, and family connections (Stephenson, 2008). Damming may preserve sense of home by promoting community development and enhancing population density. Sherren et al. (2016) spoke to locals in the Mactaquac area and found people perceived the long-dammed place as home as long as they were still living there. A sense of home was not observed to be high in Oldman where remains primarily an agricultural community without expansive industrial development or urbanization (Chen et al., 2024). It was not possible to have a longitudinal comparison with the data before damming, however, relocation due to reservoir flooding can be a reason for losing sense of home (Million, 1992). Furthermore, our results suggest that sense of home could be more prominent in rural areas in Mactaquac, where other CES aspects like recreational opportunities and specific landmarks may be less prevalent.

***Recreation may thrive if venues are provided.***

Introducing a hydroelectric dam and reservoir into a landscape typically creates new construction and opportunities for sports and recreational activities (World Commission on Dams, 2020), as seen in the Mactaquac dataset where such activities were more prevalent compared to other study areas. However, the mere presence of a dam and reservoir does not automatically enhance recreational value. For example, in Oldman, recreational activities remained sparse due to the continued dominance of agricultural practices, despite the construction of the dam and the establishment of a provincial park. Additionally, the availability of recreational values may not be directly tied to proximity to dam infrastructure but is more often associated with locations that provide suitable venues or facilities. In Mactaquac, notable recreational spots include Davidson Lake, favored for water-based activities like boating, and the two provincial parks near the dam. Although agricultural lands can present significant opportunities for agritourism (Power,

2010), it was not evident in our dataset except for the simulated farming at Kings Landing near Mactaquac.

***Settler-focused cultural heritage values can survive.***

It was not surprising that settler-focused cultural heritage values were most evident in Mactaquac, particularly due to the presence of Kings Landing, a popular historical tourism site, and Hartland, home to the world's longest covered bridge since 1901. The agricultural heritage of an area can be compromised by the introduction of a hydroelectric dam and its reservoir, especially when fertile lowlands are submerged. In Mactaquac, agricultural lands and scenes were least represented, with the most pertinent images coming from Kings Landing, where farming scenes are recreated for tourism. In contrast, the extensive history of farming in Site C and Oldman positions agriculture as a fundamental aspect defining both the place and its people. Although agriculture primarily provides food, fuel, and fiber, it also offers diverse cultural services such as aesthetic rural landscapes, the cultural heritage of farming practices, and opportunities for agritourism (Power, 2010; Swinton et al., 2007). Apart from images from Kings Landing in Mactaquac, categorizing other images definitively under agricultural heritage or recreational values when coding training dataset proved challenging. For instance, an image of a man riding a horse could ambiguously represent either a recreational activity or a working scene, depending on the context. In addition, cultural services and values related to Indigenous groups were rarely represented in this study due to the data nature, though three study areas included Indigenous lands and people.

***4.4.2 Implications for hydroelectric landscapes***

Damming a river can adversely affect certain CES, yet it may also enhance its overall CES diversity and richness, as evidenced by the high CES richness in the long-dammed landscape of Mactaquac. Areas with a long history of settlement, diverse communities, and various industries tend to offer a wider and more varied range of CES and thus deserve special consideration during the assessment of hydroelectric dam projects. For example, the old towns identified in Mactaquac exhibited higher CES richness. Stephenson (2008) emphasized the importance of "time-thickness" in landscapes, which enhances people's perception and appreciation of a place by connecting the present to the

past; inappropriate development can disrupt these historical connections, but advance planning (as for Kings Landing) can reduce such disruption. However, it is important to note that Indigenous history and practices can be underrepresented when using social media data to assess CES, as evidenced in this study where only a few images may indicate relevant content.

Another key point is that evaluating the effects of landscape changes such as development on CES provision needs to be tailored to individual cases; while overarching insights and comparisons between different cases can aid in understanding and predicting changes, the specifics can vary from one case to another and obfuscate the impacts of development. Establishing a baseline understanding of the area's existing CES is crucial such as using conventional media to capture longitudinal pictures (Pimentel da Silva et al., 2021). Subsequent planning and management after dam construction are vital for preserving and enhancing CES provision, advocating for the establishment of sustainable tourism that harmonizes with local natural resources and cultural heritage (Albrecht et al., 2024; Stamatiadou et al., 2023). Provincial parks are examples of the former, and Kings Landing in Mactaquac exemplifies the latter. Indigenous CES outcomes require much more in-depth and fine-grained attention than is possible using CV.

#### *4.4.3 CES assessment via social media images and custom-trained CV tools*

The innovative methodology is a key focus of this study, offering valuable insights for the broader research community interested in social media data and CV tools. To address the objective outlined earlier—evaluating the feasibility of training an CV model to assess CES in hydropower-related landscapes—we will discuss the methodology with respect to the following questions: 1) Is it feasible and worthwhile to train a CV model to analyze social media images? 2) Does this model effectively capture CES based on assumptions about connections to physical landscape features? 3) What are the directions for future research in this area?

Firstly, it is feasible and beneficial to explore such a method. The primary reason is the robust results from model training and the meaningful interpretation of the prediction results by the trained model as discussed above. Utilizing the Vertex AI platform to assess

CES provision through feature-based coding themes achieved a high average accuracy (93.8%). Labor and time investments included coding a training dataset of 1563 images by two coders over one month, with the training process taking 3 hours and prediction for 18,831 images taking 10 hours. The monetary cost other than data gathering and human labour was 95.50 USD for training and 95.99 USD for deploying (cost for data gathering can be found at Chen et al., 2024). Overall, such CV tools are time and cost-effective for large datasets.

CES evaluation has developed over decades, incorporating a variety of assessment methods, some of which have been established through the introduction of new data sources and technologies. Nevertheless, neither social media data nor artificial intelligence technologies can fully replace traditional, established approaches. Social media data can serve as a supplementary resource, helping to address gaps or mitigate some of the drawbacks inherent in conventional methods such as interviews, surveys, and public participation GIS (PPGIS)—these include the underrepresentation of younger demographics, low response rates, and biases introduced from preset questions (Chen et al., 2018), and the extensive time and resources required for data collection (Langemeyer et al., 2023). Platforms like Instagram offer fresh perspectives on everyday life and local lifestyles that traditional data sources like maps, aerial photographs, and satellite images cannot provide (Gugulica & Burghardt, 2023; Skokanová et al., 2021). Kaiser et al. (2021) utilized interview results to affirm that social media predominantly showcases daily activities in the landscape, especially in non-touristic areas. While traditional methods such as surveys and interviews can be designed to provide valuable longitudinal data, typically they capture only temporal and spatial snapshots (Gee & Burkhard, 2010; Tajima et al., 2023). The dataset in this study, spanning from July 18, 2011, to May 30, 2021, offers a decade-long perspective that could be pivotal for monitoring shifts in CES. Future research might explore these chronological changes more comprehensively; a particular opportunity exists to continue to monitor Site C following the filling of its reservoir later this year. That said, changes to platform rules or demographic use can also affect longitudinal datasets.

The limitations of using social media data in CES assessment should be carefully considered, relating to the second question regarding the use of a feature-based coding strategy based on selected CES categories instead of directly coding services. It's crucial to acknowledge that social media data may not effectively represent all CES categories. Intangible CES categories such as spiritual values, inspiration, education, arts, life-sustaining, therapeutic/health, intrinsic/existence, and wilderness/pristine values are less feasible to identify, not only through social media data but also through some conventional methods like PPGIS (Daymond et al., 2023; Vieira et al., 2021). Additionally Indigenous cultural heritage is less identifiable in such datasets than built settler-based heritage. This limitation was evident in our research, where spiritual value, education, and arts were underrepresented in the dataset, as also seen in previous research using Instagram data (Chen et al., 2020). Other CES categories suggested in social media images can only be imperfectly inferred based on tangible features, such as a wide-open landscape photo for aesthetic value (Mouttaki et al., 2022). Otherwise, interpretations can be overly subjective if depending on the researcher's judgment of what constitutes beauty in a landscape.

Furthermore, intrinsic biases in social media data include the lack of socio-demographic information, which can restrict its utility for policymaking, such as demonstrating the disproportionate impacts of landscape changes on various groups (Kaiser et al., 2021; Kosanic & Petzold, 2020). This issue might be mitigated through a data donation approach, where users voluntarily provide their data from social media platforms for research purposes, but this is difficult to achieve at scale (Chen et al., 2024). Additionally, the instability of social media platforms' policies and APIs poses challenges for longitudinal research (Kaiser et al., 2021). Before 2018, precise locations of Instagram photos were accessible; now they are not, and collecting these photos has become increasingly difficult. Consequently, this leads to an imbalance in data collection, favoring platforms like Flickr that are easier to utilize for data gathering (Vieira et al., 2021).

Artificial intelligence technologies are particularly effective when handling large datasets, which explains their widespread use with social media data (Mouttaki et al.,

2022; Vieira et al., 2021). The advanced capabilities of CV tools allow for an analysis based on photo content, providing insights into CES assessment beyond mere counts of posts, a method prevalent in earlier research (Langemeyer et al., 2023; Sherren et al., 2017). However, when we initiated this research in 2021, computer vision tools faced significant limitations, particularly in distinguishing relevant landscape photos from irrelevant ones like selfies or indoor photos, necessitating labor-intensive manual work. Computer vision models are primarily feature- and statistic-based, making them adept at recognizing or categorizing images based on tangible distinctions. A crucial principle is that if the training dataset's codebook is confusing for human coders, it will likely confuse the CV model even more. This limitation also explains why CES categories were often coded based on physical features. Yet, challenges arise in nuances, such as differentiating between a year-round house and a summer vacation home using these models. There is also a risk of self-evidence in the interpretation of results: for instance, if a historical building (or in our case, historic covered bridge) is coded as an indicator of historical features, it's expected that this location will show a high concentration of such features in the predictive results. Therefore, researchers must interpret results with caution.

In our longitudinal research on hydroelectric landscapes, we utilized various methods to analyze social media image data, including manual coding (Chen et al., 2018; 2019; 2020), pre-trained CV models (Chen et al., submitted), and custom-trained CV models in this study. Due to the typically large size of social media datasets and the advancements in CV technologies, we generally do not recommend manual coding unless the dataset is small and requires detailed analysis. The use of a pre-trained model to detect an image is largely limited by its preset labels, which are mostly based on physical features. Another study analyzing the same dataset with a pre-trained CV model identified 1704 unique labels, ranging from high-level concepts like 'landscape' and 'nature' to detailed ones like 'damselfly' (Chen et al., submitted). Categorizing these labels into CES categories would be challenging and time-consuming. Custom-trained CV models show more promise for future applications, particularly as commercial platforms develop user-friendly training tools. However, manual coding for training data remains a burden, though its higher quality will enhance training results. Another potential direction for future research is to

incorporate different types of data, such as textual captions and comments, which were not utilized in this study. Textual data can help address some limitations of passive crowdsourcing of imagery data, providing insights into CES categories like spirituality and education that are less easily captured through images alone (Chen et al., 2020; Gugulica & Burghardt, 2023).

#### **4.5 Conclusions**

This paper illustrates the use of large-scale social media image data and a custom-trained computer vision model to evaluate CES provision as a tool for SIA across three study areas at different stages in the lifecycle of hydroelectric dams: pre-dam, 32-year, and 56-year damming landscapes. The results demonstrate the model's ability to categorize visual content based on landscape feature-related themes, achieving an average precision 93.8%, with some themes reaching 100%. The findings from the hydroelectric dam landscapes show that meaningful patterns can be discerned using this method regarding aesthetic value, place identity, recreation, and cultural heritage values. However, it is important to note that this approach cannot replace traditional first-hand methods like interviews and surveys, such as for the purposes of SIA. Indigenous impacts are particularly invisible. Social media data should be seen as providing a supplementary perspective, especially given their demographic and temporal biases. CV models currently have limitations, such as a reduced capability to identify subjective concepts and a tendency to overlook small details. A key challenge for researchers is integrating social media data and computer vision technology into the CES assessment framework to achieve a more comprehensive understanding of CES provision. To achieve this, future studies should explore diverse tools and models, combine various types of online data (such as text, images, videos, and metadata), and develop reliable CES indicators that can be analyzed through social media data.



## **Chapter 5 Conclusion**

This dissertation has two main objectives:

- (1) To explore methodological innovation for SIA by retrieving alternative data sources and applying computer vision-based analysis models;
- (2) To understand social impacts caused by large hydroelectric dams and their reservoirs, providing insights for decision-making processes in similar projects.

This chapter will conclude this dissertation following the two objectives, respectively, integrating key findings, contributions and recommendations for further work.

### **5.1 SIA methodological innovation**

This dissertation research project began in 2018 and quickly encountered challenges in the post-API era when the previously used data collection tool, Netlytic, discontinued its Instagram service. Several alternative approaches were tested and failed until self-developed web scraping scripts successfully retrieved data from Instagram by geotags. Although progress has been made with the recent introduction of research APIs from platforms like TikTok, Meta, and others, researchers still struggle to obtain social media data in a suitable format without legal and privacy concerns. The discussion of the eight approaches tested in this dissertation provided insight into the advantages and limitations of each. It became clear that collaboration among government, technology companies, and academia is essential to improve access to social media data for research purposes. Government leadership is particularly crucial in providing a safe and regulated environment. For example, the Digital Services Act, approved by the European Parliament in 2022, establishes rules for social media and search engine data access for research purposes (Joint Research Centre, 2023). However, it is important to note the limitations and biases in social media data, such as demographic biases (more young users), spatial biases (less data from rural areas) and positive biases (lifestyle posts tend to be positive). Therefore, social media data should be seen as supplementary rather than a substitute for conventional data sources like census data, surveys, and interviews.

Two computer vision approaches were used to understand the social impacts of hydroelectricity, demonstrating the potential of applying these state-of-the-art technologies in real-world cases, a relatively unexplored area in social impact studies. Using different approaches for the same dataset reveals how they can be used more efficiently. The meaningful patterns identified through the pre-trained model indicate its suitability for detecting direct and physical feature-based elements in visual content. Our success in custom-training a computer vision model highlights its potential to understand complex and sophisticated concepts like CES provision in landscapes through large-sized social media images. However, there is no simple answer to choosing between a pre-trained or a custom-trained computer vision model when generalizing to other research. The decision depends on research goals, timeline, budget, and data. Pre-trained models may seem easy to use initially because they do not require the effort of preparing a training dataset or knowledge of computer vision algorithms and programming languages. However, the workload increases when dealing with numerous labels and the noise from irrelevant and incorrect labels which can obscure key information. Conversely, a custom-trained model is labor-intensive at the beginning, requiring careful preparation of the training dataset and tuning of the model. Once successfully trained, it can be directly used on similar datasets to obtain categorical results.

Another insight lies in the choice between manual and automatic approaches. Manual approaches require significant effort but offer precise control over the meaning-extraction process, allowing for the identification of multiple layers of meaning. Manual processes can also more easily identify cryptic elements such as dam structures, which can be overlooked or mistaken for other infrastructure by computer vision (e.g., a reservoir versus a lake). Automatic approaches can handle larger datasets more efficiently but may be more suitable for seeking specific features. The cost of using commercial tools like Google Cloud Platform is reasonable: \$3,000 USD in research credits were received over three years, but only 15% of the credits were utilized for data analysis in this research. While current artificial technology can detect subjective and subtle concepts to some degree, it is not as agile as humans in distinguishing them. However, considering how rapid the recent development has been in the artificial intelligence area, it will be interesting to see how far it can go toward matching human analysts.

This research also brings insights for methodological innovation in SIA. First, it demonstrates the validity of image-based impact assessments, contrasting with traditional language-based methods like surveys and interviews, or metadata-based social media data usage, such as simply counting posts without analyzing content (e.g., value is assumed by the appearance of a post without investigating post contents, as observed in Chen et al., 2023). Visual content reveals patterns that cannot be fully captured by other data types. Second, the space-for-time substitution approach, also known as comparative SIA, can be employed when longitudinal data is difficult or expensive to obtain from conventional sources (Asselin & Parkins, 2019; Pickett, 1989). Comparing cases with similar situations and features can indicate how a place may be impacted by development by learning from how it has played out elsewhere. Third, the models tested in this research can serve as successful exemplars for other applications of computer vision technology and social media data in SIA research and practice, such as other renewable energy projects like wind which also bring landscape changes. Additionally, the innovative approaches in this study may contribute to two further layers of social impact assessment in the future: 1) monitoring real-time impacts (social media data provides real-time updates and computer vision analysis is faster than human analysis); and 2) longitudinal research: social media could offer consistent data for longitudinal studies for events occurring after 2006 (the advent of social media, or 2010 when it became more prevalent).

Future research could consider combining different types of data, encompassing various sources (e.g., conventional and big data; social media data from different platforms) and formats (e.g., text, images, videos, and metadata). Each data type has its limitations: conventional methods can be difficult to engage young participants in the digital era, while social media data may lack perspectives, such as the Indigenous communities affected by the dams in this study. To integrate social media data and computer vision models into the existing SIA framework, future researchers need to identify gaps these methods can fill, additional insights and implications they can bring, and how they can make SIA a more reliable and rigorous process for planning and managing large projects. Another direction is to investigate how the public perceives and accepts the use of their data in public-good research. The future direction of custom-trained models lies in their

potential to identify more subjective themes, which was not fully achieved in this study where CES was indicated through feature-related coding criteria. Specific questions may arise, such as whether future computer vision models can identify very subjective concepts without relying on physical indicators (e.g., determining if a landscape in a photo is beautiful) and whether such models can identify key information while filtering out noise (e.g., identifying a small hydroelectric dam facility in the background, which was a hurdle in this study). Future AI tools should facilitate human-machine communication in natural language and operate on user-friendly interfaces. The evolution from deploying computer vision generative models through coding scripts to using the Midjourney Discord interface and now Dall-E 3 in ChatGPT illustrates this trend.

## **5.2 Hydroelectric social impacts**

This study identified a broad trend of social acceptance of hydroelectric landscapes in long-dammed areas, but the results also underscore the importance of case-specific characteristics in understanding the cultural and social value of these landscapes. The results highlight the significance of reservoirs, despite their artificial nature, and the recreational lifestyles they foster among locals. These findings align with the concepts of social acceptance of energy discussed by Wüstenhagen et al. (2007), where acceptance can increase after deployment. The Mactaquac project, once controversial during its construction (Lawson et al., 1985), now benefits locals by reshaping lifestyles around increased water accessibility and recreational activities. However, place-specific characteristics play an important role in this process, as demonstrated by the different outcomes in Oldman, where ongoing agricultural practices create a contrast with Mactaquac.

Despite these differences, this research holds pragmatic value for the recently constructed Site C dam by providing insights into how people might perceive and use the new landscape, drawing from the findings of older dams like Oldman and Mactaquac. Given Site C's history of agricultural practices, the current situation at Oldman may offer a more relevant perspective on its future. If the future of Mactaquac is locally desirable, tourism and industry development could help to enhance such a goal. These implications can be generalized to other hydroelectric projects with similar developments and features, but

caution is needed to account for the unique characteristics of each case that can influence outcomes.

However, as mentioned earlier in this section it is important to consider the limitations of social media data, which may present a positive bias toward hydroelectric landscapes, especially data retrieved from Instagram where users tend to share scenic views. We do not see posts from those who have left the region, for instance, perhaps because of displacement or trauma, and there is comparatively little from those who are unhappy with the landscape. Therefore, social media insights should be triangulated with other data sources, such as news media, public opinion surveys and Indigenous engagement, to provide a more balanced perspective (Pimentel da Silva et al., 2021). This understanding can guide future research towards mixed methods studies.

### **5.3 Summary**

In conclusion, this dissertation compared three different hydroelectric landscapes in Canada – pre-dam, 32-year, and 56-year dammed – and indicated changes in social and cultural values, confirming that reservoirs and long-dammed landscapes can be positively perceived by people; thus, hydroelectric projects need further investigations in decision-making processes in the future. It also explored two computer vision models (pre-trained and custom-trained) to analyze social media images for SIA. These models showed high accuracy in identifying meaningful patterns, though limitations still exist. This work sets the foundations for many possible research directions in SIA innovation, hydroelectricity landscape change, and applied computer vision. It also provides examples of uncovering patterns using large-scale passive crowdsourced image data to assess social impacts and inform decision-makers.

## References

- Aaen, S. B., Lyhne, I., & Nielsen, H. (2018). The use of social media in impact assessment: experiences among national infrastructure developers in Denmark. *Impact Assessment and Project Appraisal*, 36(6), 456-466. <http://doi.org/10.1080/14615517.2018.1500091>
- Abreu, T. L. S., Berg, S. B., Faria, I. P. De, & Gomes, L. P. (2020). River dams and the stability of bird communities: A hierarchical Bayesian analysis in a tropical hydroelectric power plant. *Journal of Applied Ecology*, 2020(00), 1-13. <https://doi.org/10.1111/1365-2664.13607>
- Acker, A., & Kreisberg, A. (2020). Social media data archives in an API-driven world. *Archival Science*, 20, 105-123.
- Acquisti, A., Brandimarte, L., & Loewenstein, G. (2015). Privacy and human behavior in the age of information. *Science*, 347(6221), 509-514.
- Aiken, S. R. & Leigh, C. H. (2015). Dams and indigenous peoples in Malaysia: development, displacement and resettlement. *Geografiska Annaler: Series B, Human Geography* 97(1): 69–93.
- Albrecht, E., Isaac, R., & Räsänen, A. (2024). Legal and political arguments on aquatic ecosystem services and hydropower development – A case study on Kemi River basin, Finland. *Ecosystem Services*, 67, 101623. <https://doi.org/10.1016/j.ecoser.2024.101623>
- Angler's Atlas. (2024). *Oldman River Reservoir*. <https://www.anglersatlas.com/place/726229/oldman-river-reservoir>
- Arts, I., Fischer, A., Duckett, D., & Wal, R. Van Der. (2021). The Instagrammable outdoors – Investigating the sharing of nature experiences through visual social media. *People and Nature*, 3, 1244–1256. <https://doi.org/10.1002/pan3.10239>
- Asselin, J., & Parkins, J. R. (2009). Comparative Case Study as Social Impact Assessment: Possibilities and Limitations for Anticipating Social Change in the Far North. *Social Indicators Research*, 94(3), 483–97. <https://doi.org/10.1007/s11205-009-9444-7>

- Arnold, L. M., Hanna, K., Noble, B., Gergel, S. E., & Nikolakis, W. (2022). Assessing the Cumulative Social Effects of Projects: Lessons from Canadian Hydroelectric Development. *Environmental Management*, *69*, 1035-1048.  
<https://doi.org/10.1007/s00267-022-01622-x>
- Azevedo, A. K., Vieira, F. A. S., Guedes-Santos, J., Gaia, J. A., Pinheiro, B. R., Bragagnolo, C., Correia, R. A., Ladle, R. J., & Malhado, A. C. M. (2022). A big data approach to identify the loss of coastal cultural ecosystem services caused by the 2019 Brazilian oil spill disaster. *Ecosystems An. Acad. Bras. Ciênc.* *94* (suppl 2). <https://doi.org/10.1590/0001-3765202220210397>
- Barnhart, B. (2023, April 28). *Social media demographics to inform your brand's strategy in 2023*. Accessed on January 31, 2024, at Sproutsocial.  
<https://sproutsocial.com/insights/new-social-media-demographics/>
- BC Hydro. (2018, February). *Site C clean energy project: Site C reservoir*.  
[https://www.sitecproject.com/sites/default/files/info-sheet-site-c-reservoir-feb-2018\\_0.pdf](https://www.sitecproject.com/sites/default/files/info-sheet-site-c-reservoir-feb-2018_0.pdf)
- BC Hydro. (2023). *Explore construction progress at Site C*.  
<https://www.sitecproject.com/>
- Benati, G., Calcagni, F., Martellozzo, F., Ghermandi, A., & Langemeyer, J. (2024). Unequal access to cultural ecosystem services of green spaces within the city of Rome – A spatial social media-based analysis. *Ecosystem Services*, *66*, 101594.  
<https://doi.org/10.1016/j.ecoser.2023.101594>
- Blei, D. M., Ng, A. Y., Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, *3*, 993-1022.
- Borgman, C. L. (2019). The lives and after lives of data. *Harvard Data Science Review* (1.1), 1-9.
- Bourgoin, S. (2013). *Disregarded sentiments: Discovering the voices of opposition to the Mactaquac Dam* [Master's thesis]. Saint Mary's University.
- Byrne, J., Kienzle, S., Johnson, D., Duke, G., Gannon, V., Selinger, B., & Thomas, J. (2006). Current and future water issues in the Oldman River Basin of Alberta, Canada. *Water Science and Technology*, *53*(10), 327–334.  
<https://doi.org/10.2166/wst.2006.328>
- Breuer, J., Bishop, L., & Kinder-Kurlanda, K. (2020). The practical and ethical challenges in acquiring and sharing digital trace data: Negotiating public-private partnerships. *New Media & Society*, *22*(11), 2058-2080.

- Bruns, A. (2019). After the 'APIcalypse': Social media platforms and their fight against critical scholarly research. *Information, Communication & Society*, 22(11), 1544-1566.
- Buchmayr, A., Verhofstadt, E., Van Ootegem, L., Thomassen, G., Taelman, S. E., & Dewulf, J. (2022). Exploring the global and local social sustainability of wind energy technologies: An application of a social impact assessment framework. *Applied Energy*, 312, 118808. <https://doi.org/10.1016/j.apenergy.2022.118808>
- Calvert, K., Greer, K., & Maddison-MacFadyen, M. (2019). Theorizing energy landscapes for energy transition management: Insights from a socioecological history of energy transitions in Bermuda. *Geoforum*, 102, 191-201. <https://doi.org/10.1016/j.geoforum.2019.04.005>
- Canadian Environmental Assessment Agency. (2014). *Report of the Joint Review Panel - Site C Clean Energy Project, BC Hydro and Power Authority, British Columbia*. Canada Department of Canadian Environmental Assessment Agency. <https://publications.gc.ca/site/eng/9.652806/publication.html>
- Cardoso, A. S., Renna, F., Moreno-Llorca, R., Alcaraz-Segura, D., Tabik, S., Ladle, R. J., & Vaz, A. S. (2022). Classifying the content of social media images to support cultural ecosystem service assessments using deep learning models. *Ecosystem Services*, 54, 101410. <https://doi.org/10.1016/j.ecoser.2022.101410>
- Chan, K. M. A., Satterfield, T., & Goldstein, J. (2012). Rethinking ecosystem services to better address and navigate cultural values. *Ecological Economics*, 74, 8–18. <https://doi.org/10.1016/j.ecolecon.2011.11.011>
- Chen, Y., Caesemaeker, C., Rahman, T. H.M., & Sherren, K. (2020). Comparing cultural ecosystem service delivery in dykelands and marshes using Instagram: A case of the Cornwallis (Jijuktu'kwejk) River, Nova Scotia, Canada. *Ocean & Coastal Management*, 193, 105254. <https://doi.org/10.1016/j.ocecoaman.2020.105254>
- Chen, Y., Parkins, J. R., & Sherren, K. (2019). Leveraging social media to understand younger people's perceptions and use of hydroelectric energy landscapes. *Society & Natural Resources*, 32(10), 1114-1122. <https://doi.org/10.1080/08941920.2019.1587128>
- Chen, Y., Parkins, J. R., & Sherren, K. (2018). Using geo-tagged Instagram posts to reveal landscape values around current and proposed hydroelectric dams and their reservoirs. *Landscape and Urban Planning*, 170, 283-292. <https://doi.org/10.1016/j.landurbplan.2017.07.004>



- Chen, Y., Smit, M., Sherren, K., Lee, K. Y., McCay-Peet, L., & Xue, S. (2024). From theory to practice: Insights and hurdles in collecting social media data for social science research. *Frontiers in Big Data*, 7. <https://doi.org/10.3389/fdata.2024.1379921>
- Chen, Y., Smit, M., Lee, K. Y., McCay-Peet, L., & Sherren, K. (In review). Image auto-coding tools for social impact assessment: Leveraging social media data to understand human dimensions of hydroelectricity landscape changes in Canada. *Landscape and Urban Planning*.
- Chen, Y., Sherren, K., Smit, M., & Lee, K. Y. (2023). Using social media images as data in social science research. *New Media & Society*, 25(4), 849–871. <https://doi.org/10.1177/14614448211038761>
- Clarke, M. (2014, December 16). Site C dam: How we got here and what you need to know. *CBC*. <https://www.cbc.ca/news/canada/british-columbia/site-c-dam-how-we-got-here-and-what-you-need-to-know-1.2874998>
- Clarion Energy Content Directors. (2003, October 3). *ATCO Power opens Oldman River plant*. Power Engineering. <https://www.power-eng.com/renewables/atco-power-opens-oldman-river-plant/>
- Clarke, K. D., Pratt, T. C., Randall, R. G., Scruton, D. A., & Smokorowski, K. E. (2008). *Validation of the flow management pathway: Effects of altered flow on fish habitat and fishes downstream from a hydropower dam*. Canadian Technical Report of Fisheries and Aquatic Sciences 2784.
- Confessore, N. (2018). *Cambridge Analytica and Facebook: The scandal and the fallout so far*. The New York Times. Retrieved on April 5, 2022, and May 16, 2024, from <https://www.nytimes.com/2018/04/04/us/politics/cambridge-analytica-scandal-fallout.html>
- Cortese, D. K., Szczyпка, G., Emery, S., Wang, S., Hair, E., & Vallone, D. (2018). Smoking selfies: Using Instagram to explore young women’s smoking behaviors. *Social Media + Society*, 4(3), 205630511879076. <https://doi.org/10.1177/2056305118790762>
- Cox, S. (2018). *Breaching the Peace: The Site C Dam and a valley’s stand against big hydro*. On Point Press, Vancouver.
- Dangi, D., Dixit, D. K., & Bhagat, A. (2022). Sentiment analysis of COVID-19 social media data through machine learning. *Multimedia Tools and Applications*, 81, 42261–42283.

- Darvill, R., & Lindo, Z. (2016). The inclusion of stakeholders and cultural ecosystem services in land management trade-off decisions using an ecosystem services approach. *Landscape Ecology*, *31*, 533-545. <https://doi.org/10.1007/s10980-015-0260-y>
- Daschuk, J., & Marchildon, G. P. (1993). *Historical chronology of the Oldman River Dam conflict*. <https://www.parc.ca/mcri/pdfs/HistoricalChronologyoftheOldmanRiverDamConflict.pdf>
- Davis, J., & Kidd, I. M. (2012). Identifying major stressors: The essential precursor to restoring cultural ecosystem services in a degraded estuary. *Estuaries and Coasts*, *35*, 1007-1017. <https://doi.org/10.1007/s12237-012-9498-7>
- Daymond, T., Andrew, M. E., & Kobryn, H. (2023). Crowdsourcing social values data: Flickr and public participation GIS provide different perspectives of ecosystem services in a remote coastal region. *Ecosystem Services*, *64*, 101566. <https://doi.org/10.1016/j.ecoser.2023.101566>
- Dehkordi, F., A. & Nakagoshi, N. (2004). Impact evaluation of Haizukai Dam on its up stream: A case study in Hiroshima Prefecture, Japan. *Chinese Geographical Science*, *14*(4), 350-354.
- Díaz, S., Fargione, J., Chapin III, F.S., & Tilman, D. (2006). Biodiversity loss threatens human well-being. *PLoS Biol.* *4*(8), e277.
- Dixon, S. J. (2023a). Instagram – Statistics & Facts. *Statista*. <https://www.statista.com/topics/1882/instagram/#topicOverview>
- Dixon, S. J. (2023b). Distribution of Instagram users worldwide as of January 2023, by age group. *Statista*. <https://www.statista.com/statistics/325587/instagram-global-age-group/>
- Edwards, A., Housley, W., Williams, M., Sloan, L., & Williams, M. (2013). Digital social research, social media and the sociological imagination: Surrogacy, augmentation and re-orientation. *International Journal of Social Research Methodology*, *16*(3), 245-260.
- Esteves, A. M., Franks, D., & Vanclay, F. (2012). Social impact assessment: The state of the art. *Impact Assessment and Project Appraisal*, *30*(1), 34-42. <https://doi.org/10.1080/14615517.2012.660356>
- European Commission. (December 18, 2023). *Commission opens formal proceedings against X under the Digital Services Act*. Retrieved April 6, 2024, from [https://ec.europa.eu/commission/presscorner/detail/en/ip\\_23\\_6709](https://ec.europa.eu/commission/presscorner/detail/en/ip_23_6709)

- Fabris, M. (2023). Review process. *Annals of the American Association of Geographers*, 113(7), 1664–1673. <https://doi.org/10.1080/24694452.2022.2142087>
- Fox, C. A., Magilligan, F. J., & Sneddon, C. S. (2016). “You kill the dam, you are killing a part of me”: Dam removal and the environmental politics of river restoration. *Geoforum*, 70, 93-104.
- Fox, N., Graham, L. J., Eigenbrod, F., Bullock, J. M., & Parks, K. E. (2021). Reddit: A novel data source for cultural ecosystem service studies. *Ecosystem Services*, 50, 101331. <https://doi.org/10.1016/j.ecoser.2021.101331>
- Freelon, D. (2018). Computational research in the post-API age. *Political Communication*, 35, 665-668. <https://doi.org/10.1080/10584609.2018.1477506>
- Fukushima, K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, 36, 193–202.
- Galway, L., Beery, T., Jones-Casey, K., & Tasala, K. (2019). Mapping the solastalgia literature: A scoping review study. *International Journal of Environmental Research and Public Health*, 16(15), 2662. <https://doi.org/10.3390/ijerph16152662>
- Gee, K., & Burkhard, B. (2010). Cultural ecosystem services in the context of offshore wind farming: A case study from the west coast of Schleswig-Holstein. *Ecological Complexity*, 7(3), 349-358. <https://doi.org/10.1016/j.ecocom.2010.02.008>
- Gensim. (2022). *Gensim: topic modelling for humans*. <https://radimrehurek.com/gensim/index.html>
- Ghermandi, A., & Sinclair, M. (2019). Passive crowdsourcing of social media in environmental research: A systematic map. *Global Environmental Change*, 55, 36-47.
- GitHub. (2022). *arc298/instagram-scraper*. Accessed on February 28, 2021, from <https://github.com/arc298/instagram-scraper>
- Glenn, J. (1999). *Once upon an Oldman: Special interest politics and the Oldman River Dam*. UBC Press, Vancouver.
- Gone, K., P., Chen, Y., & Smit, M. (2023). Natural Language Processing to Understand Human Activities Impacted by Hydroelectric Energy Projects. *2023 IEEE International Conference on Big Data (BigData)*, Sorrento, Italy, pp. 3770-3778. doi: 10.1109/BigData59044.2023.10386212

- Google Cloud. (2023a). *Cloud Vision documentation*.  
<https://cloud.google.com/vision/docs>
- Google Cloud. (2023b). *Batch image annotation offline*.  
<https://cloud.google.com/vision/docs/batch>
- Gosal, A. S., Geijzendorffer, I. R., Václavík, T., Poulin, B., & Ziv, G. (2019). Using social media, machine learning and natural language processing to map multiple recreational beneficiaries. *Ecosystem Services*, 38(June), 100958.  
<https://doi.org/10.1016/j.ecoser.2019.100958>
- Gramling, R., & Freudenburg, W. R. (1992). Opportunity-threat, development, and adaptation: Toward a comprehensive framework for Social Impact Assessment. *Rural Sociology*, 57(2), 216-234. <https://www.proquest.com/scholarly-journals/opportunity-threat-development-adaptation-toward/docview/1291037911/se-2?accountid=10406>
- Grieco, C. (2018). What do social entrepreneurs need to walk their talk? Understanding the attitude-behavior gap in social impact assessment practice. *Nonprofit Management and Leadership*, 29, 105-122.
- Gugulica, M., & Burghardt, D. (2023). Mapping indicators of cultural ecosystem services use in urban green spaces based on text classification of geosocial media data. *Ecosystem Services*, 60, 101508. <https://doi.org/10.1016/j.ecoser.2022.101508>
- Hale, R. L., Cook, E. M., & Beltrán, B. J. (2019). Cultural ecosystem services provided by rivers across diverse social-ecological landscapes: A social media analysis. *Ecological Indicators*, 107, 105580.  
<https://doi.org/10.1016/j.ecolind.2019.105580>
- Haines-Young, R., & Potschin, M. (2018). *Common International Classification of Ecosystem Services (CICES) V5.1: Guidance on the application of the revised structure*. Fabis Consulting Ltd. UK.  
<https://cices.eu/content/uploads/sites/8/2018/01/Guidance-V51-01012018.pdf>
- Havinga, I., Bogaart, P. W., Hein, L., & Tuia, D. (2020). Defining and spatially modelling cultural ecosystem services using crowdsourced data. *Ecosystem Services*, 43, 101091. <https://doi.org/10.1016/j.ecoser.2020.101091>
- Havinga, I., Marcos, D., Bogaart, P. W., Tuia, D., & Hein, L. (2024). Understanding the sentiment associated with cultural ecosystem services using images and text from social media. *Ecosystem Services*, 65, 101581.  
<https://doi.org/10.1016/j.ecoser.2023.101581>

- Heaton, J. (2015). *Artificial Intelligence for Humans, Volume 3: Deep Learning and Neural Networks*. CreateSpace Independent Publishing Platform.
- Hernández-Ruz, E. J., Silva, R. D. O., & do Nascimento, G. A. (2018). Impacts of the construction of the Belo Monte Hydroelectric Power Plant on traditional knowledge of riverine communities in Xingu River, Pará, Brazil. *International Journal of Research Studies in Biosciences (IJRSB)*, 6(6), 13-20. <http://dx.doi.org/10.20431/2349-0365.0606003>
- Hirons, M., Combetti, C., & Dunford, R. (2016). Valuing cultural ecosystem services. *Annual Review of Environment and Resources*, 41, 545-574. <https://doi.org/10.1146/annurev-environ-110615-085831>
- Hough, M. (1990). *Out of place: restoring identity to the regional landscape*. New Haven and London, CO: Yale University Press.
- Huai, S., Chen, F., Liu, S., Canters, F., & de Voorde, T. V. (2022). Using social media photos and computer vision to assess cultural ecosystem services and landscape features in urban parks. *Ecosystem Services*, 57, 101475.
- Imperiale, A. J., & Vanclay, F. (2023). From project-based to community-based social impact assessment: New social impact assessment pathways to build community resilience and enhance disaster risk reduction and climate action. *Current Sociology*. <https://doi.org/10.1177/00113921231203168>
- Imrie, A. S. (1991). Stress-induced response from both natural and construction-related processes in the deepening of the Peace River valley, B.C. *Canadian Geotechnical Journal*, 28, 719-728.
- International Energy Agency. (2022). *World Energy Outlook 2022*. IEA, Paris. <https://iea.blob.core.windows.net/assets/830fe099-5530-48f2-a7c1-11f35d510983/WorldEnergyOutlook2022.pdf>, License: CC BY 4.0 (report); CC BY NC SA 4.0 (Annex A).
- International Hydropower Association. (2020). *2020 Hydropower Status Report*. International Hydropower Association. Retrieved on May 16, 2024, from <https://www.hydropower.org/publications/2020-hydropower-status-report>
- Ioannidis, R., & Koutsoyiannis, D. (2020). A review of land use, visibility and public perception of renewable energy in the context of landscape impact. *Applied Energy*, 276(August), 115367. <https://doi.org/10.1016/j.apenergy.2020.115367>
- Jacobsen, J. K. S. (2007). Use of landscape perception methods in tourism studies: A review of photo-based research approaches. *Tourism Geographies*, 9(3), 234-253.

- Jaichand, V. & Sampaio, A. A. (2013). Dam and be damned: The adverse impacts of Belo Monte on indigenous peoples in Brazil. *Human Rights Quarterly*, 35(2), 408-447.
- John, N. A., & Nissenbaum, A. (2019). An agnotological analysis of APIs: or, disconnectivity and the ideological limits of our knowledge of social media. *The Information Society*, 35(1), 1-12.
- Johnson, T. F., Kent, H., Hill, B. M., Dunn, G., Dommett, L., Penwill, N., Francis, T., & González-Suárez, M. (2021). classecol: Classifiers to understand public opinions of nature. *Methods in Ecology and Evolution*, 12(7), 1329-1334. <https://doi.org/10.1111/2041-210X.13596>
- Joint Research Centre. (December 13, 2023). *FAQs: DSA data access for researchers*. European Centre for Algorithmic Transparency. Retrieved on April 7 and May 22, 2024, from [https://algorithmic-transparency.ec.europa.eu/news/faqs-dsa-data-access-researchers-2023-12-13\\_en#:~:text=Article%20of%20the%20DSA,systemic%20risks%20in%20the%20EU.](https://algorithmic-transparency.ec.europa.eu/news/faqs-dsa-data-access-researchers-2023-12-13_en#:~:text=Article%20of%20the%20DSA,systemic%20risks%20in%20the%20EU.)
- Jørgensen, D. (2017). Competing ideas of 'natural' in a dam removal controversy. *Water Alternatives* 10(3), 840-852.
- Jørgensen, D. & Renöfält, B. M. (2012). Damned if you do, dammed if you don't: debates on dam removal in the Swedish media. *Ecology and Society*, 18(1), 18. <http://dx.doi.org/10.5751/ES-05364-180118>
- Kaiser, N. N., Ghermandi, A., Feld, C. K., Hershkovitz, Y., Palt, M., & Stoll, S. (2021). Societal benefits of river restoration – Implications from social media analysis. *Ecosystem Services*, 50, 101317. <https://doi.org/10.1016/j.ecoser.2021.101317>
- Kapadia, S. (2019, April 14). *Topic modeling in Python: Latent Dirichlet Allocation (LDA)*. Medium. <https://towardsdatascience.com/end-to-end-topic-modeling-in-python-latent-dirichlet-allocation-lda-35ce4ed6b3e0>
- Karpathy, A., & Li, F. (2017). Deep visual-semantic alignments for generating image descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(4), 664-676.
- Kay, M., Patel, S. N., & Kientz, J. A. (2015). How Good is 85%?: A Survey Tool to Connect Classifier Evaluation to Acceptability of Accuracy. *CHI '15: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, April 2015, 347 – 356. <https://doi.org/10.1145/2702123.2702603>

- Keilty, K., Beckley, T. M., & Sherren, K. (2016). Landscape acceptability and generational change on the Mactaquac hydroelectric dam headpond, New Brunswick, Canada. *Geoforum*, 75, 234-248.
- Kinder-Kurlanda, K., & Weller, K. (2014). "I always feel it must be great to be a hacker!" The role of interdisciplinary work in social media research. *Bloomington, IN, USA: WebSci'14*.
- Kinder-Kurlanda, K., & Weller, K. (2020). Perspective: Acknowledging data work in the social media research lifecycle. *Frontiers in Big Data*, 3, Article 509954.
- King, G., & Persily, N. (2020). A new model for industry - academic partnerships. *Political Science & Politics*, 53(4), 703-709.
- Kirchherr, J., & Charles, K. J. (2016). The social impacts of dams: A new framework for scholarly analysis. *Environmental Impact Assessment Review*, 60, 99-114. <https://doi.org/10.1016/j.eiar.2016.02.005>.
- Kosanic, A., & Petzold, J. (2020). A systematic review of cultural ecosystem services and human wellbeing. *Ecosystem Services*, 45, 101168. <https://doi.org/10.1016/j.ecoser.2020.101168>
- Kriz, K. A., Bland, J. T., & Payne, L. L. (2020). *Springfield reservoir study of aquatic recreation Supply and demand*. Springfield City Water, Light, and Power. <http://supplementalwater.cwlp.com/Springfield%20Reservoir%20Project%20-%20Final%20Report-FINAL.pdf>
- Kulanthaivel, A., Fogel, R., Jones, J., & Lammert, C. (2017). Digital cohorts within the social mediome: An approach to circumvent conventional research challenges? *Clinical Gastroenterology and Hepatology*, 15(5), 614–618. <https://doi.org/10.1016/j.cgh.2017.02.015>
- Langemeyer, J., Ghermandi, A., Keeler, B., & van Berkel, D. (2023). The future of crowd-sourced cultural ecosystem services assessments. *Ecosystem Services*, 60, 101518. <https://doi.org/10.1016/j.ecoser.2023.101518>
- Lawson, P. M., Farnsworth, G., & Hartley, M. A. (1985). *The Nackawic Bend: 200 years of history*. Town of Nackawic, New Brunswick.
- Lazer et al., D. (2009). Life in the network: The coming age of computational social science. *Science*, 323(5915), 721-723.
- Lee, H., Seo, B., Koellner, T., Lautenbach, S., 2019. Mapping cultural ecosystem services 2.0 – Potential and shortcomings from unlabeled crowd sourced images. *Ecol. Indic.* 96, 505–515. <https://doi.org/10.1016/j.ecolind.2018.08.035>

- LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *PROC. OF THE IEEE*, 1-46.
- Li, R., Crowe, J., Leifer, D., Zou, L., & Schoof, J. (2019). Beyond big data: Social media challenges and opportunities for understanding social perception of energy. *Energy Research & Social Science*, 56, 101217. <https://doi.org/10.1016/j.erss.2019.101217>
- Lingua, F., Goops, N. G., & Griess, V. G. (2022). Valuing cultural ecosystem services combining deep learning and benefit transfer approach. *Ecosystem Services*, 58, 101487. <https://doi.org/10.1016/j.ecoser.2022.101487>
- Liu, Y., Wu, G., & Guo, R. (2016). Changing landscapes by damming: the Three Gorges Dam causes downstream lake shrinkage and severe droughts. *Landscape Ecology*, 31(8), 1883–1890. <https://doi.org/10.1007/s10980-016-0391-9>
- Mácová, K. & Kozáková, Z. (2023). How important for society is recreation provided by multi-purpose water reservoirs? Welfare analysis of the Vltava River Reservoir system. *Water*, 15, 1966. <https://doi.org/10.3390/w15101966>
- Manley, K., Nyelele, C., & Egoh, B. N. (2022). A review of machine learning and big data applications in addressing ecosystem service research gaps. *Ecosystem Services*, 57, 101478. <https://doi.org/10.1016/j.ecoser.2022.101478>
- Márquez, L. A. M., Rezende, E. C. N., Machado, K. B., do Nascimento, E. L. M., Castro, J. D. B., & Nabout, J. C. (2023). Trends in valuation approaches for cultural ecosystem services: A systematic literature review. *Ecosystem Services*, 64, 101572. <https://doi.org/10.1016/j.ecoser.2023.101572>
- McElroy, J. (2016, September 11). Why B.C.'s Site C dam could become a national issue. *CBC*. <https://www.cbc.ca/news/canada/british-columbia/site-c-primer-amnesty-trudeau-1.3754463>
- Meta. (2023). *Meta Content Library and API*. Retrieved January 23, 2024, from <https://transparency.fb.com/researchtools/meta-content-library>
- Meta for Developers. (2023). *Instagram Platform*. Retrieved August 29, 2023, from <https://developers.facebook.com/docs/instagram>
- M.E.A. (2005). *A Report of the Millennium Ecosystem Assessment. Ecosystems and Human Well-Being*. Island Press, Washington DC.
- Milcu, A. I., Hanspach, J., Abson, D., & Fischer, J. (2013). Cultural ecosystem services: A literature review and prospects for future research. *Ecology and Society*, 18(3), 44. <http://dx.doi.org/10.5751/ES-05790-180344>



- Million, M. L. (1992). *"It was home": A phenomenology of place and involuntary displacement as illustrated by the forced dislocation of five southern Alberta families in the Oldman River Dam flood area* [Doctoral thesis]. Saybrook Institute Graduate School and Research Centre.  
<https://www.proquest.com/docview/304030314?pq-origsite=gscholar&fromopenview=true&sourcetype=Dissertations%20&%20Theses>
- Milusheva, S., Marty, R., Bedoya, G., Id, S. W., Resor, E., & Legovini, A. (2021). Applying machine learning and geolocation techniques to social media data (Twitter) to develop a resource for urban planning. *PLoS ONE* 16(2): e0244317. 1–12. <https://doi.org/10.1371/journal.pone.0244317>
- Mioc, D., Nkhwanana, N., Moreri, K. K.... Santos, M. (2015). Natural and man-made flood risk mapping and warning for socially vulnerable populations. *International Journal of Safety and Security Engineering*, 5(3), 183-202.  
<https://doi.org/10.2495/SAFE-V5-N3-183-202>
- Mohammadi, M., Chen, Y., Rahman, H. M. T., & Sherren, K. (2023). A saliency mapping approach to understanding the visual impact of wind and solar infrastructure in amenity landscapes. *Impact Assessment and Project Appraisal*, 41(2), 154–161. <https://doi.org/10.1080/14615517.2023.2169460>
- Mouttaki, I., Bagdanavičiūtė, I., Maanan, M., Erraiss, M., Rhinane, H., & Maanan M. (2022). Classifying and mapping cultural ecosystem services using artificial intelligence and social media data. *Wetlands*, 42, 86.  
<https://doi.org/10.1007/s13157-022-01616-9>
- Muir, B. R. (2018). Effectiveness of the EIA for the Site C Hydroelectric Dam reconsidered: Nature of Indigenous cultures, rights, and engagement. *Journal of Environmental Assessment Policy and Management*, 20(4), 1-3.
- Muldoon, P., Williams, J., Lucas, A., Gibson, R. B., & Pickfield, P. (2020). An introduction to environmental law and policy in Canada (3rd edition). Emond, Toronto.
- Natural Resources Canada. (2001). *An outline map without names which shows only Canada's coastline and boundaries*. Retrieved August 27, 2016, from <http://www.nrcan.gc.ca/earth-sciences/geography/atlas-canada/reference-maps/16846>.

- Qomariyah, S., Iriawan, N., & Fithriasari, K. (2019). Topic modeling Twitter data using Latent Dirichlet Allocation and Latent Semantic Analysis. *The 2nd International Conference on Science, Mathematics, Environment, and Education AIP Conf. Proc.* 2194, 020093-1–020093-7. <https://doi.org/10.1063/1.5139825> Published.
- Pathak, A., Ruhela, A., Saroha, A. K., & Bhardwaj, A. (2019). Examining robustness of Google Vision API based on the performance on noisy images. *International Journal of Computer Sciences and Engineering*, 7(3), 2347-2693. <https://doi.org/10.26438/ijcse/v7i3.8993>
- Pickett, S. T. A. (1989). *Space-for-time substitution as an alternative to long-term studies*. In: Likens, G.E. (eds) Long-Term Studies in Ecology. Springer, New York, NY. [https://doi.org/10.1007/978-1-4615-7358-6\\_5](https://doi.org/10.1007/978-1-4615-7358-6_5)
- Pimentel da Silva, G. D., Parkins, J. R., & Sherren, K. (2021). Do methods used in social impact assessment adequately capture impacts? An exploration of the research-practice gap using hydroelectricity in Canada. *Energy Research & Social Science*, 79, 102188. <https://doi.org/10.1016/j.erss.2021.102188>
- Pimentel da Silva, G. D., Sherren, K., & Parkins, J. R. (2021). Using news coverage and community-based impact assessments to understand and track social effects using the perspectives of affected people and decisionmakers. *Journal of Environmental Management*, 298, 113467. <https://doi.org/10.1016/j.jenvman.2021.113467>
- Poletti, C., & Gray, D. (2019). Good data is critical data: An appeal for critical digital studies. In A. Daly, S. Devitt, & M. Mann (Eds.), *Good Data* (pp. 260-276). Amsterdam: The Institute of Network Cultures.
- Poirier, B. A., & De Loë, R. C. (2011). Protecting aquatic ecosystems in heavily allocated river systems: the case of the Oldman River Basin. *The Canadian Geographer*, 55(2), 243–261. <https://doi.org/10.1111/j.1541-0064.2010.00322.x>
- Power, A. G. (2010). Ecosystem services and agriculture: tradeoffs and synergies. *Philosophical Transactions of The Royal Society B*, 365(1554), 2959-2971. <https://doi.org/10.1098/rstb.2010.0143>
- Procter, R., Vis, F., & Voss, A. (2013). Reading the riots on Twitter: Methodological innovation for the analysis of big data. *International Journal of Social Research Methodology*, 16(3), 197-214.
- Rainbow, S. L. (1992). Why did New Zealand and Tasmania spawn the world's first green parties? *Environmental Politics*, 1(3), 321-346.

- Reilly, K. H., & Adamowski, J. F. (2017). Spatial and temporal scale framing of a decision on the future of the Mactaquac Dam in New Brunswick, Canada. *Ecology and Society* 22(3), 21. <https://doi.org/10.5751/ES-09535-220321>
- Richards, D. R., & Friess, D. A. (2015). A rapid indicator of cultural ecosystem service usage at a fine spatial scale: Content analysis of social media photographs. *Ecological Indicators*, 53, 187-195. <https://doi.org/10.1016/j.ecolind.2015.01.034>
- Richards, D. R., & Lavorel, S. (2022). Integrating social media data and machine learning to analyse scenarios of landscape appreciation. *Ecosystem Services*, 55, 101422. <https://doi.org/10.1016/j.ecoser.2022.101422>
- Riechers, M., Barkmann, J., & Tschardt, T. (2016). Perceptions of cultural ecosystem services from urban green. *Ecosystem Services*, 17, 33-39. <https://doi.org/10.1016/j.ecoser.2015.11.007>
- Rieder, B., & Hofmann, J. (2020). Towards platform observability. *Internet Policy Review*, 9(4), 1-28. <https://doi.org/10.14763/2020.4.1535>
- Rodrigues, S. C., & Silva, T. I. (2012). Dam construction and loss of geodiversity in the Araguari River basin, Brazil. *Land Degradation & Development*, 23(4), 419–426. <https://doi-org.ezproxy.library.dal.ca/10.1002/ldr.2157>
- Rojas W. A., Magzul, L. Marchildon, G., & Reyes, B. (2009). The Oldman River Dam Conflict: Adaptation and Institutional Learning. *Prairie Forum*, 34(1), 235-260.
- Ruiz-Frau, A., Ospina-Alvarez, A., Villasante, S., Pita, P., Maya-Jariego, I., & de Juan, S. (2020). Using graph theory and social media data to assess cultural ecosystem services in coastal areas: Method development and application. *Ecosystem Services*, 45, 101176. <https://doi.org/10.1016/j.ecoser.2020.101176>
- Sæþórsdóttir, A. D. & Ólafsson, R. (2010). Nature tourism assessment in the Icelandic Master Plan for geothermal and hydropower development. Part II: assessing the impact of proposed power plants on tourism and recreation. *Journal of Heritage Tourism*, 5(4), 333-349. DOI: 10.1080/1743873X.2010.517840
- Sandvig v. Barr, 451 F. Supp. 3d 73 (D.D.C. 2020). Retrieved on April 7, 2024, from <https://globalfreedomofexpression.columbia.edu/wp-content/uploads/2020/06/Sandvig-v-Barr.pdf>

- Sargentis, G. F., Hadjibiros, K., & Christofides, A. (2005). Plastiras Lake: The impact of water level on the aesthetic value of the landscape. *Proceedings of the 9<sup>th</sup> International Conference on Environmental Science and Technology*, Rhodes Island, Greece, 1-3 September 2005.
- Savage, M., & Burrows, R. (2007). The coming crisis of empirical sociology. *Sociology*, *41*(5), 885-899.
- Schapper, A., Urban, F. (2021). Large dams, norms and Indigenous Peoples. *Development Policy Review*, *39*(S1), 61–80. <https://doi.org/10.1111/dpr.12467>
- Sherren, K., Beckley, T. M., Greenland-Smith, S., & Comeau, L. (2017). How Provincial and Local Discourses Aligned Against the Prospect of Dam Removal in New Brunswick, Canada. *Water alternatives*, *10*(3), 697-723.
- Sherren, K., Beckley, T. M., Parkins, J. R., Stedman, R. C., Keilty, K., & Morin, I. (2016). Learning (or living) to love the landscapes of hydroelectricity in Canada: Eliciting local perspectives on the Mactaquac Dam via headpond boat tours. *Energy Research & Social Science*, *14*, 102–110. <https://doi.org/10.1016/j.erss.2016.02.003>
- Sherren, K., Chen, Y., Mohammadi, M., Zhao, Q., Gone, K. P., Rahman, H. M. T., & Smit, M. (2023). Social media and social impact assessment: Evolving methods in a shifting context. *Current Sociology*. <https://doi.org/10.1177/00113921231203179>
- Shpyth, A. A. (1991). An ex-post evaluation of environmental impact assessment in Alberta: A case study of the Oldman River Dam. *Canadian Water Resources Journal*, *16*(4), 367-379. <https://doi.org/10.4296/cwrj1604367>
- Shtern, M., Simmons, B., Smit, M., & Litoiu, M. (2013). Toward an ecosystem for precision sharing of segmented Big Data. In *2013 IEEE Sixth International Conference on Cloud Computing* (pp. 335-342). IEEE.
- Skokanová, H., Slach, T., Havlíček, M., Halas, P., Divíšek, J., Špinlerová, Z., Koutecký, T., Šebesta, J., & Kallabová, E. (2021). Landscape Painting in the Research of Landscape Changes. *Journal of Landscape Ecology*, *14*(3), 110-127. <https://doi.org/10.2478/jlecol-2021-0019>
- Social Science One. (2022). Retrieved from <https://socialscience.one/>

- Sottini, V. A., Barbierato, E., Bernetti, I., Capecchi, I., Fabbrizzi, S., & Menghini, S. (2019). The use of crowdsourced geographic information for spatial evaluation of cultural ecosystem services in the agricultural landscape: the case of Chianti Classico (Italy). *New Medit*, 18(2), 105-118. <http://dx.doi.org/10.30682/nm1902g>
- Stamatiadou, V., Mazaris, A., Mallios, Z., & Katsanevakis, S. (2023). Valuation and mapping of the recreational diving ecosystem service of the Aegean Sea. *Ecosystem Services*, 64, 101569. <https://doi.org/10.1016/j.ecoser.2023.101569>
- Stantec. (2016, August). *Mactaquac Project: Final comparative environmental review (CER) report – summary document*. Retrieved from [https://www.nbpower.com/media/689743/cer\\_mactaquac\\_project\\_summary\\_document\\_aug2016.pdf](https://www.nbpower.com/media/689743/cer_mactaquac_project_summary_document_aug2016.pdf)
- Statista. (January 2024). *Most popular social networks worldwide as of January 2024, ranked by number of monthly active users*. Retrieved on April 17, 2024, from <https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/>
- Stedman, R. C., Connelly, N. A., Heberlein, T. A., Decker, D. J., & Allred, S. B. (2019). The end of the (research) world as we know it? Understanding and coping with declining response rates to mail surveys. *Society & Natural Resources*, 32(10), 1139-1154.
- Steen-Johnsen, K., & Enjolras, B. (2015). Social research and Big Data – the tension between opportunities and realities. In H. Fossheim, & H. C. Ingierd (Eds.), *Internet Research Ethics* (pp. 122-140). Cappelen Damm Akademisk.
- Stephenson, J. (2008). The cultural values model: An integrated approach to values in landscapes. *Landscape and Urban Planning*, 84, 127-139. <https://doi.org/10.1016/j.landurbplan.2007.07.003>
- Sternberg, R. (2008). Hydropower: Dimensions of social and environmental coexistence. *Renewable and Sustainable Energy Review*, 12, 1588–1621. <https://doi.org/10.1016/j.rser.2007.01.027>
- Swe, K. N., Funakawa, S., Okamoto, Y., & Chan, N. (2023). Assessment on land use changes and livelihood transition under the hydropower dam construction in Paunglaung Township, Southern Shan Highlands, Myanmar. *Land Degradation & Development*, 34(17), 5647-5661.

- Swinton, S. M., Lupi, F., Robertson, G. P., & Hamilton, S. K. (2007). Ecosystem services and agriculture: Cultivating agricultural ecosystems for diverse benefits. *Ecological Economics*, *64*(2), 245-252.  
<https://doi.org/10.1016/j.ecolecon.2007.09.020>
- Tajima, Y., Hashimoto, S., Dasgupta, R., & Takahashi, Y. (2023). Spatial characterization of cultural ecosystem services in the Ishigaki Island of Japan: A comparison between residents and tourists. *Ecosystem Services*, *60*, 101520.  
<https://doi.org/10.1016/j.ecoser.2023.101520>
- Taylor, J. G., Zube, E. H., & Sell, J. L. (1987). Landscape assessment and perception research methods. In R. Bechtel, R. Marans, & W. Michaelson (Eds.), *Methods in environment and behavioral research* (pp. 361-393). New York, NY: Van Nostrans Reinhold.
- Taylor, J., & Pagliari, C. (2018). Mining social media data: How are research sponsors and researchers addressing the ethical challenges? *Research Ethics*, *14*(2), 1-39.
- Tengberg, A., Fredholm, S., Eliasson, I., Knez, I., Saltzman, K., & Wetterberg, O. (2012). Cultural ecosystem services provided by landscapes: Assessment of heritage values and identity. *Ecosystem Services*, *2*, 14-26.  
<https://doi.org/10.1016/j.ecoser.2012.07.006>
- Terkenli, T. S., Skowronek, E., Tucki, A., & Kounellis, N. (2019). Toward understanding tourist landscape. a comparative study of locals' and visitors' perception in selected destinations in Poland and Greece. *Quaestiones Geographicae*, *38*(3), 81-93.
- The World Commission on Dams. (2001). Dams and development: A new framework for decision making. Retrieved on May 16, 2024, from  
<https://www.ied.org/sites/default/files/pdfs/migrate/9126IIED.pdf>
- TikTok for Developers. (2023). *Research API*. Retrieved August 29, 2023, from  
<https://developers.tiktok.com/products/research-api/>
- Tilt, B., Braun, Y., & He, D. (2009). Social impacts of large dam projects: A comparison of international case studies and implications for best practice. *Journal of Environmental Management*, *90*, S249–S257.  
<https://doi.org/10.1016/j.jenvman.2008.07.030>
- Titration. (2022a). Ins-Scraping. GitHub. [https://github.com/Titration/Ins-Scraping/blob/main/code/ins\\_location.py](https://github.com/Titration/Ins-Scraping/blob/main/code/ins_location.py)
- Titration. (2022b). Ins-Scraping. GitHub. [https://github.com/Titration/Ins-Scraping/blob/main/code/ins\\_profile.py](https://github.com/Titration/Ins-Scraping/blob/main/code/ins_profile.py)

- United Nations. (2011). *The guiding principles on business and human rights: Implementing the UN “respect, protect and remedy” framework*. New York: United Nations Human Rights Office of the High Commissioner. Accessed 2024 May 5.  
[https://www.ohchr.org/sites/default/files/documents/publications/guidingprinciple\\_sbusinesshr\\_en.pdf](https://www.ohchr.org/sites/default/files/documents/publications/guidingprinciple_sbusinesshr_en.pdf)
- Van Driel, I. I., Giachanou, A., Pouwels, J. L., Boeschoten, L., Beyens, I., & Valkenburg, P. M. (2022). Promises and pitfalls of social media data donations. *Communication Methods and Measures*, 16(4), 266-282.
- Vanclay, F. (2020). Reflections on Social Impact Assessment in the 21<sup>st</sup> century. *Impact Assessment and Project Appraisal*, 38(2), 126-131.  
<https://doi.org/10.1080/14615517.2019.1685807>
- Vanclay, F., & Esteves, A. M. (2011). Current issues and trends in social impact assessment. In Vanclay, F., & Esteves, A. M. (Eds.), *New Directions in Social Impact Assessment: Conceptual and Methodological Advances* (pp. 3-19). Cheltenham UK and Northampton, MA, USA: Edward Elgar.
- Vieira, F. A. S., Santos, D. T. V., Bragagnolo, C., Campos-Silva, J. V., Correia, R. A. H., Jepson, P., Malhado, A. C. M., & Ladle, R. J. (2021). Social media data reveals multiple cultural services along the 8.500 kilometers of Brazilian coastline. *Ocean & Coastal Management*, 214, 105918.  
<https://doi.org/10.1016/j.ocecoaman.2021.105918>
- Vigl, L. E., Marsoner, T., Giombini, V., Pecher, C., Simion, H., Stemle, E., Tasser, E., & Depellegrin, D. (2021). Harnessing artificial intelligence technology and social media data to support Cultural Ecosystem Service assessments. *People and Nature*, 3(3), 673-685. <https://doi.org/10.1002/pan3.10199>
- Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2016). Show and Tell: Lessons learned from the 2015 MSCOCO Image Captioning Challenge. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PP(99).
- Vogus, C. (2022, August 16). *Improving researcher access to digital data: A workshop report*. Center for Democracy & Technology. Accessed on January 31, 2024, from <https://cdt.org/insights/improving-researcher-access-to-digital-data-a-workshop-report/>
- Walker, S. (2017). *The complexity of collecting digital and social media data in ephemeral contexts*. Retrieved April 5, 2022, from <https://digital.lib.washington.edu/researchworks/handle/1773/40612>

- Wang, P., Lassoie, J. P., Dong, S., & Morreale, S. J. (2013). A framework for social impact analysis of large dams: A case study of cascading dams on the Upper-Mekong River, China. *Journal of Environmental Management*, *117*, 131-140.
- Weller, K., & Kinder-Kurlanda, K. E. (2015). Uncovering the challenges in collection, sharing and documentation: The hidden data of social media research? 2015 ICWSM Workshop.
- Wilmsen, B. (2018). Damming China's rivers to expand its cities: The urban livelihoods of rural people displaced by the Three Gorges Dam. *Urban Geography*, *39*(3), 345-366. <https://doi.org/10.1080/02723638.2017.1328578>
- World Commission on Dams. (2000). *Dams and development: A new framework for decision-making*. Earthscan, London.
- Wüstenhagen, R., Wolsink, M., & Bürer, M. J. (2007). Social acceptance of renewable energy innovation: An introduction to the concept. *Energy Policy*, *35*(5), 2683-2691. <https://doi.org/10.1016/j.enpol.2006.12.001>
- X Developer Platform. (2023). *About the Twitter API*. Retrieved August 29, 2023, from <https://developer.twitter.com/en/docs/twitter-api/getting-started/about-twitter-api>
- X Developer Platform. (2024). *Research under EU Digital Services Act*. Retrieved April 6, 2024, from <https://developer.twitter.com/en/use-cases/do-research>
- Yoon, J., Jeong, B., Kim, M., & Lee, C. (2021). An information entropy and Latent Dirichlet Allocation approach to noise patent filtering. *Advanced Engineering Informatics*, *47*, 101243. <https://doi.org/10.1016/j.aei.2020.101243>
- Zhang, J., Liu, Z., & Sun, X. (2009). Changing landscape in the Three Gorges Reservoir Area of Yangtze River from 1977 to 2005: Land use/land cover, vegetation cover changes estimated using multi-source satellite data. *International Journal of Applied Earth Observation and Geoinformation*, *11*, 403-412.
- Zhao, D., Xiao, M., Huang, C., Liang, Y., & An, Z. (2021). Landscape dynamics improved recreation service of the Three Gorges Reservoir area, China. *International Journal of Environmental Research and Public Health*, *18*, 8356.
- Zhao, Q., Liu, S., Deng, L., Dong, S., Yang, Z., & Yang, J. (2012). Landscape change and hydrologic alteration associated with dam construction. *International Journal of Applied Earth Observation and Geoinformation*, *16*, 17-26. <https://doi.org/10.1016/j.jag.2011.11.009>



Zou, J., & Schiebinger, L. (2018). AI can be sexist and racist — it's time to make it fair. *Nature*, 559(7714), 324-326.

Zuckerman, E. (2023, November 1). *When the internet becomes unknowable*. Prospect. Accessed on January 31, 2024, from <https://www.prospectmagazine.co.uk/ideas/technology/63752/when-internet-becomes-unknowable-social-media-tools>

## **Appendices List**

Appendix A: Copyright Permission – Frontiers

Appendix B: Copyright Permission – Elsevier

Appendix C: Location name and ID for data collection (X: longitude, Y: latitude)

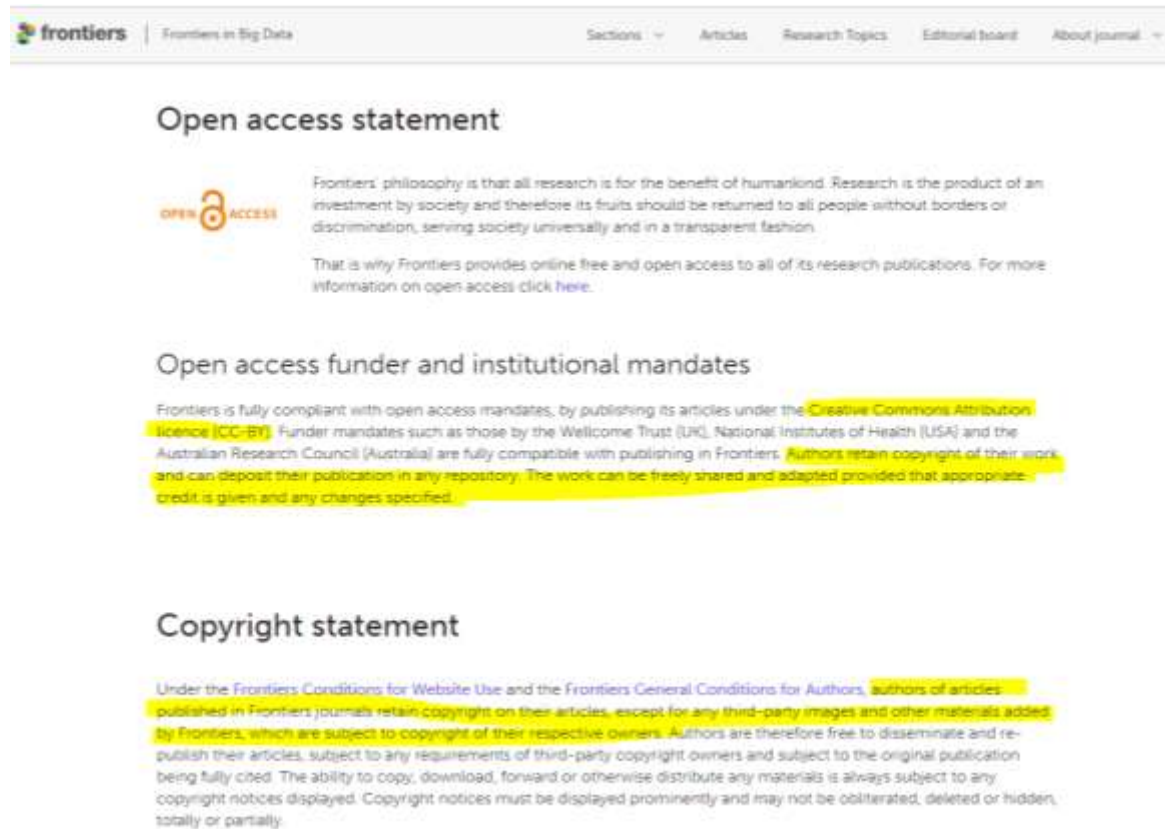
Appendix D: Latent Dirichlet Allocation model for label analysis

Appendix E: Codebook for Coding Themes

Appendix F: Prediction results for each study areas by geo-tags

## Appendix A: Copyright Permission – Frontiers

<https://www.frontiersin.org/journals/big-data/about#copyright-statement>



The screenshot displays the 'Open access statement' section of the Frontiers website. At the top, the 'frontiers' logo and 'Frontiers in Big Data' are visible on the left, and navigation links for 'Sections', 'Articles', 'Research Topics', 'Editorial board', and 'About journal' are on the right. The main heading is 'Open access statement'. Below it is the 'OPEN ACCESS' logo and a paragraph explaining Frontiers' philosophy: 'Frontiers' philosophy is that all research is for the benefit of humankind. Research is the product of an investment by society and therefore its fruits should be returned to all people without borders or discrimination, serving society universally and in a transparent fashion.' A second paragraph states: 'That is why Frontiers provides online free and open access to all of its research publications. For more information on open access click here.'

The next section is 'Open access funder and institutional mandates'. It contains a paragraph: 'Frontiers is fully compliant with open access mandates, by publishing its articles under the Creative Commons Attribution licence (CC-BY). Funder mandates such as those by the Wellcome Trust (UK), National Institutes of Health (USA) and the Australian Research Council (Australia) are fully compatible with publishing in Frontiers. Authors retain copyright of their work and can deposit their publication in any repository. The work can be freely shared and adapted provided that appropriate credit is given and any changes specified.'

The final section is 'Copyright statement'. It contains a paragraph: 'Under the Frontiers Conditions for Website Use and the Frontiers General Conditions for Authors, authors of articles published in Frontiers journals retain copyright on their articles, except for any third-party images and other materials added by Frontiers, which are subject to copyright of their respective owners. Authors are therefore free to disseminate and republish their articles, subject to any requirements of third-party copyright owners and subject to the original publication being fully cited. The ability to copy, download, forward or otherwise distribute any materials is always subject to any copyright notices displayed. Copyright notices must be displayed prominently and may not be obliterated, deleted or hidden, totally or partially.'

<https://creativecommons.org/licenses/by/4.0/>

# ATTRIBUTION 4.0 INTERNATIONAL

## Deed

Canonical URL: <https://creativecommons.org/licenses/by/4.0/>

[See the legal code](#)


### You are free to:

**Share** — copy and redistribute the material in any medium or format for any purpose, even commercially.

**Adapt** — remix, transform, and build upon the material for any purpose, even commercially.

The licensor cannot revoke these freedoms as long as you follow the license terms.

### Under the following terms:

 **Attribution** — You must give [appropriate credit](#), provide a link to the license, and [indicate if changes were made](#). You may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use.

**No additional restrictions** — You may not apply legal terms or [technological measures](#) that legally restrict others from doing anything the license permits.

### Notices:

You do not have to comply with the license for elements of the material in the public domain or where your use is permitted by an applicable [exception or limitation](#).

No warranties are given. The license may not give you all of the permissions necessary for your intended use. For example, other rights such as [publicity, privacy, or moral rights](#) may limit how you use the material.



Frontiers in Big Data - Production <bigdata.production.office@frontiersin.org>

To: Yan Chen

Reply Reply all Forward

Wed 2024-06-12 10:58 PM

Some content in this message has been blocked because the sender isn't in your Safe senders list.

Trust sender Show blocked content

**CAUTION:** The Sender of this email is not from within the business.

Dear Yan,

Thank you for your message.

That is correct, as long as you add a line that the material has been taken from your published article, which is published by Frontiers, that's fine.

All the best with the work!

Kind regards,

Joyce

Production Team

Production Manager: Nikolaos Anagnostis

Frontiers | Production Office

London Office

The Yamwiche 2nd Floor 119-121 Cannon Street London EC4N 5AF

United Kingdom

Do you need help? Visit our [Production Help Center](#) page for more information. For technical issues, please contact our Application Support Team [support@frontiersin.org](mailto:support@frontiersin.org) or visit our [Customer Help Center](#).

Wonderful, thank you!

Great, thanks for confirming!

Great, thank you so much!

Reply

Forward

## Appendix B: Copyright Permission – Elsevier

<https://www.elsevier.com/about/policies-and-standards/copyright/permissions>

[Permission guidelines](#)

[ScienceDirect content](#)

[ClinicalKey content](#)

[Tutorial videos](#)

[Help and support](#)

---

Do I need to request permission to re-use work from another STM publisher?

+

---

Do I need to request permission to text mine Elsevier content?

+

---

Can I include/use my article in my thesis/dissertation?

-

Yes. Authors can include their articles in full or in part in a thesis or dissertation for non-commercial purposes.

For any further clarifications, you can submit your query via our [online form](#).

---

**Appendix C: Location name and ID for data collection (X: longitude, Y: latitude)**

Site Name	Site ID	Y	X
Mactaquac Provincial Park	4146896	45.96117	-66.8948
Mactaquac	965634720	45.9788	-66.6276
Woolastook Park	2072336633008070	45.86425	-66.8986
Keswick Ridge	1032136740	46.00255	-66.8746
Kingsclear	315405255	45.85424	-66.9285
Upper Queensbury, New Brunswick	139911169373043	45.99558	-67.2076
Queensbury Parish, New Brunswick	409880646	45.96757	-67.0146
Kings Landing, New Brunswick	16565482	45.87976	-66.9774
Prince William, New Brunswick	430746127	45.91998	-67.0514
Bear Island (New Brunswick)	108158722546469	45.92232	-67.0263
Davidson Lake (sjö i Kanada, New Brunswick)	1956126391279140	45.93943	-67.1549
Dumfries, New Brunswick	809660864	45.9618	-67.1418
Nackawic, New Brunswick	258975485	46.00754	-67.2382
Hawkshaw, New Brunswick	135620713138737	45.97425	-67.2327
Pokiok, New Brunswick	129194380455217	45.95155	-67.249
Meductic, New Brunswick	285194361	45.99279	-67.4797
Northampton Parish, New Brunswick	297048430	46.06671	-67.5478
Woodstock, New Brunswick	116657058	46.14942	-67.5767
Woodstock First Nation	289475363	46.11143	-67.5714
Lower Woodstock, Nouveau Brunswick	253004261379353	46.11726	-67.5835
Grafton, New Brunswick	139643799399519	46.14778	-67.5621
Pembroke, New Brunswick	138323019531085	46.18393	-67.5332
Wakefield, New Brunswick	630774388	46.2341	-67.5164
Victoria Corner, New Brunswick	1005548118	46.27382	-67.5106
Hartland, New Brunswick	258264458	46.30179	-67.5223
Somerville, New Brunswick	339720957	46.30408	-67.54
Fort Saint John, British Columbia	268647134	56.25327	-120.84
Attachie, British Columbia	107349732633924	56.21933	-121.433
Hudson's Hope, British Columbia	267521741	56.03595	-121.899
Farrell Creek, British Columbia	465407726	56.17737	-121.576
Bear Flat, British Columbia	417374340	56.27324	-121.227
Old Man Dam	379665350	49.58311	-113.92
Pincher Creek	225208439	49.48556	-113.948
Pincher Station, Alberta	571002191	49.52115	-113.949
Cowley, Alberta	334100342	49.56811	-114.074

## **Appendix D: Latent Dirichlet Allocation model for label analysis**

All labels detected by Vision API were exported and saved into an Excel spreadsheet with columns titled Image (Instagram shortcode), Description (label), and Score (confidence of the label detected). Later, we transformed the data format to comma-separated values (csv) files. Data collected from different locations were compiled by study areas (Site C, Oldman, and Mactaquac). We then utilized a topic clustering model – Latent Dirichlet Allocation (LDA) – to identify patterns within the 188,301 labels. LDA is a Bayesian model that automatically identifies topics based on the co-presence of certain words in textual corpora (Blei et al., 2003), and we use it to identify topics based on the co-presence of labels for specific images. The analysis was conducted in Jupyter Notebook using Python. The script was developed based on Kapadia’s (2019a) work and revised to meet the needs of this study. The details of each phase will be presented in the following subsections.

### Label data loading and cleaning

First, comma-separated values files containing labels from each study area were loaded as a dataframe. Each image had up to 10 labels detected by Vision API. Labels from the same image were grouped together and treated as a single document in subsequent LDA model training. There were 3031 documents (labels for each image) in the AB case, 6752 in BC, and 8980 in NB. A minor portion (2%) of images was excluded due to unrecognizable shortcodes (tokenized words were grouped as documents by shortcodes and the index).

### Data normalization and tokenization

Second, data normalization was conducted to make the data more regular and reduce the complexity, and thus comparable. Often, this process is used to process natural language (e.g., articles), including removing punctuation and stop words (those such as ‘and’ and ‘the’ that do not add meaning), lower casing, and stemming (finding the root or base form of words so that terms such as like, liking, liked can all be clustered together) or lemmatization (similar process but based on the meaning of the word). In this case, labels were pre-identified categories, and stored as separate words or phrases. We thus skipped



the steps of removing punctuation and stemming/lemmatization, because it was not necessary to further simplify the textual materials. The documents were then tokenized into list data as separate words instead of phrases.

#### LDA modeling tuning and training

LDA model tuning (find the parameters that can guide its learning process to optimize the performance) and training (identify the topic clusters) used the gensim library—a free open source of topic models including Latent Semantic Indexing, Latent Dirichlet Allocation, etc.—unsupervised models that automatically discover statistical co-occurrence patterns within a corpus (Gensim, 2022). The hyperparameters in the LDA model which are used to guide the learning process were tuned by measuring the coherence score (Kapadia, 2019b). There are three pre-determined hyperparameters that will shape the distribution: the number of topics, alpha, and beta. Alpha and beta define the prior distributions of the document-topic and word-topic assignment matrices, respectively. The results show the coherence score of the LDA model given different combinations of  $k$  (number of topics), alpha, and beta. The range of  $k$  is 1 to 20, 25, and 50. The potential values of alpha are 0.01, 0.05, 0.1, 0.2, 0.5, and 1; while that of beta are 0.01, 0.05, 0.1, 0.2, 0.5, 1, and  $1/k$  (Maier et al., 2018). Also, 10% and 25% of corpus was randomly removed and the rest was tested for coherence because it might improve the score by deleting noise in the dataset.

By choosing the highest coherence scores, the hyperparameters were determined for the model training process (see Table A1). In the Mactaquac case, the highest coherence score was 0.514 achieved by 90% of the corpus. Because the other two cases had the highest coherence scores with the complete corpus; to be comparable, we decided to use the highest hyperparameters achieved with the complete corpus as well for the Mactaquac case: 8 topic numbers while alpha and beta both equal to 1. The coherence score is 0.490 which is close to the highest of 75% corpus (0.514). The parameters were then used to train the model and cluster the landscape topics by analyzing the processed labels.

#### Table A1: Model tuning results

Study case	Best coherence score	Corpus percentage	K (number of topics)	Alpha	Beta
The Oldman Dam, AB	0.461	100%	10	0.1	1
The Site C Dam, BC	0.548	100%	11	0.5	1
The Mactaquac Dam, NB	0.490	100%	8	1	1

## Reference

- Blei, D. M., Ng, A. Y., Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3, 993-1022.
- Gensim. (2022). *Gensim: topic modelling for humans*. <https://radimrehurek.com/gensim/index.html>
- Kapadia, S. (2019a, April 14). *Topic Modeling in Python: Latent Dirichlet Allocation (LDA)*. Medium. <https://towardsdatascience.com/end-to-end-topic-modeling-in-python-latent-dirichlet-allocation-lda-35ce4ed6b3e0>
- Kapadia, S. (2019b, August 19). *Evaluate Topic Models: Latent Dirichlet Allocation (LDA)*. Medium. <https://towardsdatascience.com/evaluate-topic-model-in-python-latent-dirichlet-allocation-lda-7d57484bb5d0>
- Maier, D., Waldherr, A., Miltner, P., Wiedemann, G., Niekler, A., Keinert, A., Pfetsch, B., Heyer, G., Reber, U., Häussler, T., Schmid-Petri, H., & Adam, S. (2018). Applying LDA topic modeling in communication research: Toward a valid and reliable methodology. *Communication Methods and Measures*, 12(2-3), 93-118.

## Appendix E: Codebook for Coding Themes

Coding Theme (# of images coded)	Description (focusing on what is valued or beneficial to human well-being)	Criteria
(Natural) Landscape (150)	The image mainly focuses on aesthetic appreciation of the natural landscape and there is no indicator to assign this image to any other category. This is a category exclusive to others.	<ul style="list-style-type: none"> <li>• Wide scale and open landscape</li> <li>• Focus of image is view of landscape</li> <li>• No humans or pets in image</li> <li>• Wild animals are not the focus of the image</li> <li>• Artificial objects (e.g., roads, energy facilities which are not indicators for other coding themes) do not occupy over 40% of the frame</li> </ul>
Natural features (294)	The image aims to appreciate the natural features which are captured in detail (close-up) or at the focus point.	<ul style="list-style-type: none"> <li>• Images showing details of or focusing on wild animals, plants, or living organisms</li> <li>• (or) images showing details of or focusing on natural things (rock, ice, sky, sun, water, tree, etc.)</li> </ul>
Human in nature (158)	The image shows the person(s) in the place without indicators for recreational activities or social events and relations. The main value comes solely from being in the place.	<ul style="list-style-type: none"> <li>• People in the landscape without signs of specific activities or social events and relations</li> <li>• (if there are more than one person, they do not meet the criteria for social events or relationships)</li> </ul>
Place-based features (213)	The image shows place-based features.	<ul style="list-style-type: none"> <li>• Images showing place-based signage, flag, or landmark (e.g., bridges, railroad, statues, dams, etc.)</li> </ul>
Object in nature (121)	The image shows the inanimate object(s) in the place which is an indispensable part of the image.	<ul style="list-style-type: none"> <li>• Inanimate manmade objects (e.g., cars, buildings) in the landscape which is the focus of the image (highlighting to the objects)</li> </ul>
<i>Sense of home</i> (100)		<ul style="list-style-type: none"> <li>• Images showing private house(s)</li> </ul>
Recreational	The image shows the person(s) doing recreational activities or equipment in the landscape.	<ul style="list-style-type: none"> <li>• Images showing people doing recreational activities</li> <li>• (or) images showing activity equipment, except cars, motorcycles or boats</li> </ul>
<i>Dog walking</i>		<ul style="list-style-type: none"> <li>• Images showing dog(s)</li> </ul>
Social relationship (194)	The images show social gatherings/events or indicators of social relations.	<ul style="list-style-type: none"> <li>• Images showing events (e.g., graduation, prom, wedding, funeral), or group of people</li> <li>• (or) images showing family together (adults and kids) or people showing intimacy (arm in arm, hug, kiss, look at each other, etc.)</li> </ul>
Historical features (162)	The images show indicators of historical values.	<ul style="list-style-type: none"> <li>• Images showing historical site or building</li> </ul>
Agriculture (147)	The images show agricultural landscapes, activity, or equipment.	<ul style="list-style-type: none"> <li>• Images showing agricultural activities (including farming, logging, and grazing), facilities, tools, or livestock</li> <li>• (or) images showing farmland or meadow (must be focus of image or occupy over 40% of frame)</li> </ul>
N/A Spirituality and religion	The image shows spiritual and religious facilities or activities.	<ul style="list-style-type: none"> <li>• Images showing religion activities and buildings (e.g., church)</li> </ul>

		<ul style="list-style-type: none"> <li>• (or) Images showing spiritual monuments and activities</li> </ul>
N/A Energy infrastructure	The image shows energy infrastructure (i.e., dams, wind turbines, and solar panels).	<ul style="list-style-type: none"> <li>• The image shows dams, wind turbines, or solar panels.</li> </ul>

\*N/A means the number of valid images coded was less than the model training threshold of 100.

**Appendix F: Prediction results for each study areas by geo-tags**



107

Table A2: Percentage of coding themes in each geo-tag in Mactaquac (=images assigned to the coding theme/total valid image# of the geo-tag); Only dataset >=100; CES richness = coding theme#/Image#; top 1 shaded for each coding theme; numbers for each coding theme are percentage)

	Hartland	Woodstock	Meductic	Nackawic	Davidson Lake	Dumfries	Prince William	Bear Island	Kings Landing	Kingsclear	Woolastook Park	Mactaquac Provincial	Keswick Ridge	Mactaquac
Characteristics	Old town	Old town	Small old town	New town	Cottage	Rural	Rural	Camp site	Historical tourism	Indigenous	Tourism park	Tourism park	Rural (farm)	Tourism park (wrong location)
Image#	1288	2197	154	787	102	159	391	128	1187	694	114	1593	194	151
Coding theme#	1499	1955	137	790	97	130	358	132	1274	634	102	1544	176	129

Landscape	0.112	0.197	0.279	0.146	0.294	0.220	0.240	0.13	0.072	0.183	0.202	0.272	0.180	0.212
Natural features	0.095	0.200	0.149	0.154	0.108	0.201	0.220	0.46	0.098	0.210	0.114	0.176	0.201	0.113
Human in	0.032	0.056	0.032	0.042	0.029	0.013	0.056	0.03	0.052	0.042	0.009	0.045	0.031	0.086
Object in nature	0.026	0.021	0.045	0.014		0.006	0.010		0.001	0.026	0.009	0.008	0.010	0.013
Place-based	0.373	0.118	0.123	0.319	0.049	0.050	0.028	0.00	0.074	0.053	0.044	0.026	0.041	0.046
Sense of home	0.012	0.039	0.006	0.024	0.029	0.057	0.026	0.00	0.019	0.035		0.008	0.046	0.033
Agriculture	0.038	0.035	0.019	0.017		0.019	0.043	0.05	0.167	0.056		0.010	0.144	0.033
Recreational	0.045	0.056	0.058	0.079	0.235	0.031	0.087	0.14	0.009	0.112	0.263	0.204	0.072	0.093
Dog walking	0.021	0.034	0.058	0.055	0.049	0.132	0.069	0.09	0.003	0.082	0.132	0.087	0.067	0.066
Social	0.068	0.103	0.097	0.142	0.157	0.057	0.107	0.08	0.083	0.095	0.105	0.129	0.072	0.132
Historical	0.342	0.031	0.019	0.013		0.031	0.028		0.496	0.019	0.018	0.004	0.041	0.026
CES richness	1.164	0.890	0.890	1.004	0.951	0.818	0.916	1.03	1.073	0.914	0.895	0.969	0.907	0.854

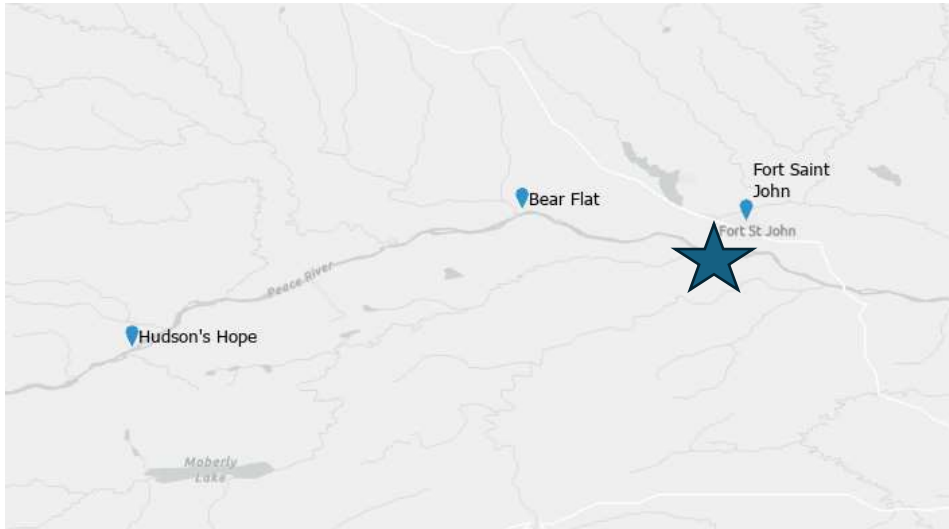


Table A3: Percentage of coding themes in each geo-tag in Site C (=images assigned to the coding theme/total valid image# of the geo-tag); Only dataset  $\geq 50$ ; CES richness = coding theme#/Image#; top 1 shaded for each coding theme; numbers for each coding theme are percentage)

Site Name	Hudson's	Bear Flat	Fort Saint John
Characteristics	Historic town	Rural	Largest town nearby
Image#	907	55	5781
Coding theme#	778	59	5114
Landscape	0.314	0.218	0.141
Natural features	0.149	0.182	0.200
Human in nature	0.053	0.018	0.057
Object in nature	0.031	0.073	0.042
Place-based features	0.049	0.018	0.048
Sense of home	0.010	0.018	0.030

Agriculture	0.041	0.182	0.066
Recreational	0.072		0.076
Dog walking	0.040	0.109	0.069
Social relationship	0.077	0.255	0.143
Historical features	0.023		0.013
CES Richness	0.858	1.073	0.885



110

Table A4: Percentage of coding themes in each geo-tag in Oldman (=images assigned to the coding theme/total valid image# of the geo-tag); Only dataset >=50; CES richness = coding theme#/Image#; top 1 shaded for each coding theme; numbers for each coding theme are percentage)

Site Name	Cowley	Pincher Creek
Characteristics	Rural (agriculture)	Old town (agriculture)
Image#	470	2526
Coding theme#	411	2157



Landscape	0.217	0.188
Natural features	0.179	0.173
Human in nature	0.013	0.031
Object in nature	0.021	0.019
Place-based features	0.055	0.065
Sense of home	0.011	0.017
Agriculture	0.232	0.172
Recreational	0.023	0.050
Dog walking	0.040	0.043
Social relationship	0.043	0.071
Historical features	0.040	0.024
CES Richness	0.874	0.854