

March 2024

# Incorporating Machine Learning with Satellite Data to Support Critical Infrastructure Measurement and Sustainable Development

Aggrey Muhebwa  
*University of Massachusetts Amherst*

Follow this and additional works at: [https://scholarworks.umass.edu/dissertations\\_2](https://scholarworks.umass.edu/dissertations_2)



Part of the [VLSI and Circuits, Embedded and Hardware Systems Commons](#)

---

## Recommended Citation

Muhebwa, Aggrey, "Incorporating Machine Learning with Satellite Data to Support Critical Infrastructure Measurement and Sustainable Development" (2024). *Doctoral Dissertations*. 3072.  
<https://doi.org/10.7275/36332652> [https://scholarworks.umass.edu/dissertations\\_2/3072](https://scholarworks.umass.edu/dissertations_2/3072)

This Open Access Dissertation is brought to you for free and open access by the Dissertations and Theses at ScholarWorks@UMass Amherst. It has been accepted for inclusion in Doctoral Dissertations by an authorized administrator of ScholarWorks@UMass Amherst. For more information, please contact [scholarworks@library.umass.edu](mailto:scholarworks@library.umass.edu).

**INCORPORATING MACHINE LEARNING WITH SATELLITE  
DATA TO SUPPORT CRITICAL INFRASTRUCTURE  
MEASUREMENT AND SUSTAINABLE DEVELOPMENT**

A Dissertation Presented

by

AGGREY MUHEBWA

Submitted to the Graduate School of the  
University of Massachusetts Amherst in partial fulfillment  
of the requirements for the degree of

DOCTOR OF PHILOSOPHY

February 2024

Electrical & Computer Engineering

© Copyright by Aggrey Muhebwa 2024

All Rights Reserved

# INCORPORATING MACHINE LEARNING WITH SATELLITE DATA TO SUPPORT CRITICAL INFRASTRUCTURE MEASUREMENT AND SUSTAINABLE DEVELOPMENT

A Dissertation Presented

by

AGGREY MUHEBWA


Approved as to style and content by:

DocuSigned by:  
  
0A0CG82F16EF41F...


Jay Taneja, Chair

DocuSigned by:  
  
06C0B566E62A450...

Colin J. Gleason, Member

DocuSigned by:  
  
1DD908920DB64AB...

Paul Siqueira, Member

DocuSigned by:  
  
423F60EC0306479...

Gabriel Cadamuro, Member

---

Christopher Hollot, Chair of the Faculty  
Electrical & Computer Engineering

## **DEDICATION**

To my parents, Alfred and Jane Munyabwera, for setting an unparalleled example of excellence. To Billy “Blanks” Kaye (RIP), for opening my eyes to the world of unlimited possibilities through computer programming. And to aunt Merinah Kateeba (RIP) for teaching me the importance of approaching every situation with a positive attitude.

## ACKNOWLEDGMENTS

*“Sometimes people appear in your life unexpectedly like a gift from the Universe. You didn’t even know you needed them, or that you had called out silently to them. They appear when you needed them most, to lift you, educate you, wake you up, or shine a light on your path. They sprout the seed that was in you, and patiently watch that seed emerge from the soil. Sometimes it wears them out to water and fertilize you every single day as you grow. This is a delicate time, you as the plant, and they as the nurturer. You as the plant need them for your growth, and they as your nurturer, have to have the energy to believe in your growth. Then, one day you blossom, and awaken to the beauty around you and rejoice. The only thing you ask from them anymore, is to celebrate the flower they have brought to life, and to accept the riches you now will give to them.” - Riitta Klint.*

Growing up in rural western Uganda, not in a million years would I have ever imagined that I would one day earn a Ph.D. from one of the world’s top universities. Although I knew from a young age that I wanted to become an engineer and put in the required work, it has been a collective effort of many people that has brought me to this moment. First, I would like to thank my Ph.D. Advisor, Prof. Jay Taneja. Thank you for taking a chance on me and for your patience, kindness, and gentle guidance over the years. You have helped me to become the researcher I am today. To the people who have mentored me through life; Sara Muhahala, Billy Kaye (RIP), and Dr. Jeff Dean, who took me on after Billy passed. Thank you for believing in me, even when I didn’t believe in myself. To the members of the STIMA Lab, especially my batchmates: Santiago, June, Zeal, and Bob Muhwezi. Thank you for your support and friendship. To my collaborators; Prof. Taneja, Dr. Cadamuro, Dr. Dongmei Feng, Dr. Lukuyu, and Prof. Colin J. Gleason, thank you for teaching me how to do research correctly. To my committee members, Prof. Taneja, Dr. Cadamuro,

Prof. Gleason, and Prof. Paul Siqueira, thank you for your guidance and feedback. To my family in Amherst: Dr. Jimi B. and Tolu Oke, Dr. Favorite Iradukunda, Antoine and Angelique Nzeyimana, and the Rwandan Students Association in Amherst area, thank you for being my family in a foreign land. I am eternally grateful to my parents, Alfred and Jane Munyabwera, for always being our north star. I am also grateful to my siblings, Judith, Abel, Albert, Juliet, Andrew, and Alvin, for setting a good example of excellence in every facet of life. Thank you to my extended family, Dr. Muramuzi, Sam Asiimwe, Sylvia, Mary, and Lisa, for opening your homes to me during my school holidays. Thank you to my best friend, Jacob “Wits” Ihunga, for constantly pushing me to excel. Finally, thank you to my auntie, Merinah Kateeba, for making a man out of me. May you continue to rest in peace.

## **ABSTRACT**

# **INCORPORATING MACHINE LEARNING WITH SATELLITE DATA TO SUPPORT CRITICAL INFRASTRUCTURE MEASUREMENT AND SUSTAINABLE DEVELOPMENT**

FEBRUARY 2024

AGGREY MUHEBWA

B.Sc., MAKERERE UNIVERSITY

M.Sc., CARNEGIE MELLON UNIVERSITY

Ph.D., UNIVERSITY OF MASSACHUSETTS AMHERST

Directed by: Professor Jay Taneja

Under the umbrella concept of Artificial Intelligence (AI) for good, recent advances in machine learning and large-scale data analysis have opened new opportunities to solve humanity's most pressing challenges. Improvements in computation complexity and advances in AI (e.g., Vision Transformers) have led to faster and more effective techniques for extracting high-dimensional patterns from large-scale heterogeneous datasets (big data). Further, as satellite data become increasingly available at varying temporal-spatial resolutions, AI tools are helping us to better understand the underlying causes of environmental and socioeconomic changes at an unprecedented scale, ushering in an era of data-driven decision-making to support sustainable and equitable development. Based on these, we propose data-driven methods and techniques for critical infrastructure measurement and sustainable development. Using machine learning and remotely sensed data, we



show that we can exploit knowledge and temporal-spatial characteristics learned from data-rich regions to improve data-driven predictions in regions with scant to no data. Specifically, we focus on three critical infrastructures: rivers, roads, and electricity access. Knowledge rivers, particularly their discharge, can help us understand how climate change is evolving, its manifestation on global water resources, and its impact on critical sectors like agriculture and renewable energy generation. On the other hand, better roads facilitate societal development, enabling access to local and global markets and socioeconomic opportunities, leading to better equality in service provision, faster socioeconomic development, and, ultimately, better human outcomes. Finally, we develop tools to support sustainable development, focusing on supporting electricity demand stimulation to improve energy access in rural communities. These methodologies and techniques can help emerging economies achieve their primary sustainable development goals (SDGs) by 2030.

# TABLE OF CONTENTS

	<b>Page</b>
<b>ACKNOWLEDGMENTS</b> .....	<b>v</b>
<b>ABSTRACT</b> .....	<b>vii</b>
<b>LIST OF TABLES</b> .....	<b>xiv</b>
<b>LIST OF FIGURES</b> .....	<b>xv</b>
 <b>CHAPTER</b>	
<b>1. INTRODUCTION</b> .....	<b>1</b>
1.1 Thesis Contributions .....	1
1.1.1 Global River Discharge Estimation .....	3
1.1.2 Explainable Machine Learning for Hydrology .....	4
1.1.3 Road Quality Prediction using Satellite Imagery .....	5
1.1.4 Tools to support electricity demand stimulation: Converting Fishing Boats for Electric Mobility to Serve as Mini-grid Anchor Loads .....	6
1.2 Proposal Outline .....	7
<b>2. RELATED WORK</b> .....	<b>9</b>
2.1 Is There a Need for Machine Learning in Hydrology? .....	9
2.2 What Role does Explainable Artificial Intelligence(XAI) Play in Hydrology? .....	11
2.3 How Can Machine Learning Transform Road Infrastructure Measurement and Analysis? .....	13
<b>3. GLOBAL RIVER DISCHARGE ESTIMATION</b> .....	<b>15</b>
3.1 Towards Improved Global River Discharge Prediction in Ungauged Basins Using Machine Learning and Satellite Observations .....	15

3.1.1	Motivation .....	15
3.1.2	Related Work .....	16
3.1.3	Methods .....	18
	3.1.3.1 Dataset and Problem Definition .....	18
	3.1.3.2 Sequential Learning .....	18
	3.1.3.3 Training and Evaluation Metrics .....	19
3.1.4	Results .....	20
3.1.5	Conclusion .....	21
3.2	Improving River Discharge Prediction with Machine Learning via Distributed Learning of Hydrologic Information in Remotely Sensed Data .....	22
3.2.1	Motivation .....	22
3.2.2	Data and Methods .....	26
	3.2.2.1 Data .....	26
	3.2.2.2 Sequential Learning via LSTMs .....	29
	3.2.2.3 Experiment design .....	31
	3.2.2.4 Evaluation Metrics .....	36
3.2.3	Results .....	36
	3.2.3.1 Predictions in ungauged basins .....	36
3.2.4	Discussion .....	40
3.2.5	Conclusion .....	47
<b>4.</b>	<b>EXPLAINABLE MACHINE LEARNING FOR RIVER DISCHARGE PREDICTION .....</b>	<b>49</b>
4.1	Explainable Machine Learning Models for River Discharge Prediction Using Remotely Sensed Data .....	49
4.1.1	Motivation .....	49
4.1.2	Data and Methods .....	53
	4.1.2.1 River Discharge Prediction .....	54
	4.1.2.2 Machine Learning in Hydrology .....	55
	4.1.2.3 Explainable Machine Learning .....	56
	4.1.2.4 Equitable Machine Learning in Hydrology .....	57
4.1.3	Experimental design .....	59
4.1.4	Evaluation Metrics .....	63
4.1.5	Results .....	64

4.1.6	Discussion .....	74
4.1.7	Conclusion .....	83
4.1.8	Future Work .....	84
<b>5.</b>	<b>ROAD QUALITY PREDICTION USING SATELLITE IMAGERY .....</b>	<b>85</b>
5.1	Road Quality Prediction using High Resolution Satellite Imagery .....	85
5.1.1	Motivation .....	85
5.1.2	Background and Related Work .....	86
5.1.3	Methodology .....	89
5.1.3.1	Datasets.....	89
5.1.3.2	Training and metrics .....	93
5.1.3.3	Convolutional neural nets and auto-encoders.....	95
5.1.3.4	Sequence learning via LSTMs .....	96
5.1.4	Results .....	97
5.1.4.1	Single tile regression .....	97
5.1.4.2	Tile sequence regression.....	98
5.1.5	Future Work and Conclusions .....	100
5.2	Using Vision Transformers to Improve Road Quality Predictions from Medium Resolution and Heterogeneous Satellite Imagery .....	101
5.2.1	Motivation .....	101
5.2.2	Background and related work .....	102
5.2.3	Methodology .....	104
5.2.3.1	Dataset .....	104
5.2.3.2	Training and metrics .....	105
5.2.3.3	From Convolutional Neural Networks to Vision Transformers .....	108
5.2.4	Results and Discussion.....	110
5.2.4.1	Impact of data size on model performance.....	112
5.2.5	Case Study 1: Correlation between road quality and household asset wealth .....	115
5.2.6	Conclusion and Future Work .....	118
5.3	Evaluating Road Quality over Time Using Satellite Imagery to Assess the Long-term Effects of Infrastructure Investments in Eastern Democratic Republic of Congo .....	119

5.3.1	Motivation .....	119
5.3.2	Background and Related Work .....	120
5.3.2.1	Road Quality Measurement .....	122
5.3.3	Methodology .....	122
5.3.3.1	Dataset .....	122
5.3.3.2	Transfer Learning .....	125
5.3.3.3	Handling Imbalanced Data .....	126
5.3.4	Results and Discussion .....	128
5.3.5	Conclusion .....	135
<b>6.</b>	<b>CONVERTING FISHING BOATS FOR ELECTRIC MOBILITY TO SERVE AS MINI-GRID ANCHOR LOADS .....</b>	<b>136</b>
6.1	Motivation .....	136
6.2	Background and Related Work .....	139
6.2.1	Minigrids – Financial and Operational Challenges .....	139
6.2.2	Why Demand Stimulation? .....	140
6.2.3	Electric Mobility as Flexible Demand .....	141
6.3	Data and Methodology .....	142
6.3.1	Data collection and description .....	142
6.3.2	Sizing of an electric outboard motor and battery system .....	149
6.3.3	Residential and commercial demand estimation .....	149
6.3.4	Modeling ice factory power demand .....	150
6.3.5	Minigrid operation model considering stochastic electric boat charging load .....	151
6.4	Analysis .....	154
6.4.1	Residential and small commercial demand profile .....	154
6.4.2	Electric outboard motor and battery sizing .....	156
6.4.3	Impact on minigrid operation .....	157
6.4.4	Impact on economics of minigrid project .....	159
6.4.5	Economic impact for boat owners .....	161
6.4.6	Demand response .....	163
6.5	Discussion and Future Work .....	165
6.6	Conclusion .....	167
<b>7.</b>	<b>CONCLUSION .....</b>	<b>168</b>

**BIBLIOGRAPHY ..... 171**

## LIST OF TABLES

Table	Page
3.1	Statistical distribution of discharge prediction results in ungauged basins. With the exception of class one, mean discharge across the remaining classes outperforms state-of-the-art process-based model predictions, which report NSE and KGE values in the range of 0.0 to 0.5. . . . . 21
3.2	Table showing the number of generated and contributed sets used for training in each Strahler river order . . . . . 34
4.1	Mean of KGE, NSE and RBias across the three models: Linear regression, random forests and LSTMs . . . . . 65
5.1	A summary of the diverse set of roads in our labeled and filtered data set recording both the length and distribution of road quality labels. For each road, the modal road quality class is in bold. The set ranges from first-class highways (e.g., A104) to rough dirt roads (e.g., C67) and includes roads with significant internal variation (e.g., C77). . . . . 91
5.2	5-class accuracy and regression R-squared results under <i>standard</i> train-test and <i>held-out</i> conditions for the single-tile regression problem. . . . . 97
5.3	Results for LSTMs in the <i>held-out</i> test scenario as a function of the length of sequence trained on. Regressing on the final tile value (last) is compared to regressing on the average tile value (mean). . . . . 99
5.4	2-class & 5-class mean AUROC for production results under standard train-test and held-out conditions for original CNN (ResNet), CNN inspired by Vision transformers (ConvNext)and, and vision transformer models trained on Planet Lab and GEP Datasets. . . . . 111

## LIST OF FIGURES

Figure	Page
1.1 Map showing global population density (2022). Over 80% of the world's population lives in the global south. On the other hand, a large percentage of real-time hydrological data (Fig 1.2) is available in the global north. ....	1
1.2 Map showing the global location of active river gauge stations from 1919 to the present. Most active gauge stations (blue) are present in the global North, while a big percentage of the inactive or decommissioned stations are in the global South. (Source: World Meteorological Organization) .....	2
3.1 A Map showing the location of gauge stations (red circles) in the Mackenzie basin used in the study. The insert shows a map of the 20 biggest basins in Canada and the Mackenzie basin (shaded).....	27
3.2 Schematic representation of a hypothetical order eight basin network. The red circle represents the location of a gauge station on the delineated basin's outlet. At each hierarchical level, a single-order basin and its lower-order basins are selected (filled) while the remaining basins on the same level or not upstream of the selected basin within that level are ignored (hatched). This topological representation (Strahler river order system) integrates the temporal-spatial variation of physical processes at different stages of a river network. ....	35



3.3	Cumulative distributions functions (CDFs) of NSE and KGE for defined experiments and selected benchmarks calculated from distributions across all Pfafstetter orders. Figures (I) and (II) compare the performance of models in the at-station and lumped experiments against the models trained with data from the distributed experiment. Figures (III) and (IV) compare the performance of models in the distributed experiment against two literature models: [81]; [150]. A shift to the right indicates an improvement in model performance. Baseline models from the literature show lower skill than the ML here when all models perform poorly ( $-\infty < \text{NSE} \& \text{KGE} \leq 0.0$ ) but better performance when all models have good predictions ( $0.5 < \text{NSE} \& \text{KGE} \leq 1.0$ ). The distributed model outperforms the at-station and lumped models across the entirety of the results. CDFs are preferred because they represent the overall model performance across the entire test dataset . . . . .	37
3.4	Top to Bottom: Distribution comparisons of selected metrics on held-out predictions for at station (I-IV), lumped (V-VII), and distributed (IX-XII) experiments. Note that distributions for seventh and eighth orders are not included due to limited gauge stations in the training set. Figure S1 shows a distribution comparison across all experiments and literature models. . . . .	38
3.5	Representative hydrographs showing randomly selected models with $0.0 < \text{NSE} \leq 0.6$ in each of the experiments; At-station (left), lumped (middle) and distributed (right) experiments across the defined orders, i.e., from order 4 (top) to order 8 (bottom). Here, we plot hydrographs for the first 2.5 years. . . . .	42
3.6	Left to right: Representative hydrographs showing the worst performing ML models in each of the experiments and the non-ML literature model; At station experiment, lumped experiment, distributed experiment, and RADR model (Feng et al., 2021) across the defined orders, i.e., from order 4 (top) to order 8 (bottom). The RADR model overestimates peak flows and underestimates base flows in lower orders. Here, we plot hydrographs for the first 2.5 years. . . . .	43
3.7	Left to right: Pairwise comparison of KGE distributions with varying lookback window sizes and corresponding statistical significance tests across the three experiments. Inter-experiment comparisons show that distributions of lookback for at-station and lumped experiments are similar, while there is an observable difference in distributions of lookback windows of the distributed experiment. . . . .	46

4.1	predicted discharge vs. observed discharge curves for Linear regression and Random Forest models. LR yields an $R^2 = 0.75$ while RF yields an $R^2 = 0.83$ . This shows that the RF model fits the data more accurately than the LR model .....	65
4.2	Comparative feature importance across three models: (a) Coefficients from a multi-linear regression model, (b) Feature Importance from the Random Forest model, (c) SHAP summary plot from the Random Forest model, and (d) SHAP summary plot from the LSTM model. Here, we display the top 14 most important features .....	66
4.3	Three plots showing the impact of the same features on the model's prediction across different instances (different dates/seasons of the year): December is typically winter in the Mackenzie basin, mid-March is early spring while mid-May is early summer .....	69
4.4	Dependence plots showing the relationship between a single feature (x-axis) and the SHAP values (or model output) for that feature (y-axis). The coloration is based on a second "interaction" feature, which captures interaction effects, i.e., how the primary feature's impact changes with varying values of another feature .....	70
4.5	Distributions of attention weights across LSTM models per river order (based on the Strahler River order system. We ignore order 1 statins due to limited data ( $n = 1$ ) .....	72
5.1	Three different roads highlighting the challenging diversity of our dataset. Left: an urban environment along the A104 highway. A104 is a major highway in Kenya, and the selected road tile is of "great" quality. Center: the C47 minor road. It passes through an arid environment, and the road segment has "poor" quality. Right: the C67 minor road. It passes through large forests and cropland, and the road segment in the image has "good" quality.....	88
5.2	Roads with labeled quality data as collected by the Kenya National Highway Authority (KenHA). Indicated are the original dataset and a dataset that has been filtered to match the availability of concurrent satellite imagery. ....	89

5.3	An example of how a road segment can be separated into tiles. The top segment shows a road divided into overlapping 64x64 squares; this generates the tiles shown in the middle segment. The tiles are always aligned in the direction of the road (red arrow). The bottom shows the same segment if divided into 224x224 tiles instead. Note that the road constitutes a much smaller proportion of each tile relative to the 64x64 case. ....	94
5.4	Figures showing the distribution of Mean square errors (y-axis, lower indicates better predictive power) of different roads using a Resnet CNN (left) and auto-encoder regression (right). The x-axis is a measure of the heterogeneity of the road, the color provides the average road quality, and the circle size indicates the relative sizes of the roads. Comparisons to VGG-11 and Resnet yield similar results. ....	99
5.5	A Series of roads from two data sources; Planet Labs(top) at 3m/pixel and Google Earth Pro (bottom) at ~1.6m/pixel. The different sizes, visibility and resolutions of the road images highlight the heterogeneity of our dataset. The red rectangle indicates the range of the road segment over which the IRI measurements are averaged. ....	105
5.6	Histogram showing the 5-class distribution of labels in the dataset. The dataset is heavily imbalanced. Labels associated with “great” road quality contributed the largest distribution percentage. ....	106
5.7	Model Accuracy and loss curves across training and validation data splits (at 50% of all roads ) for the ViT model. Performance measurement curves follow the same trend across all transformers models defined in section 5.2.3.3 and data split percentages. ....	111
5.8	Graphs showing AUROC scores for 2-class and 5-class classification results on I.I.D roads from the identical road segments. Top: Planet Lab imagery (3m/pixel) and bottom: Google Earth Engine scrapped Imagery (~1.6m/pixel). The X-axis on all graphs represents the number of roads (as %) used for training the models. ....	113
5.9	Graphs showing AUROC scores for 2-class and 5-class classification results on held-out roads from the identical road segments. Top: Planet Lab imagery (3m/pixel) and bottom: Google Earth Engine scrapped Imagery (~1.6m/pixel). The X-axis on all graphs represents the number of roads (as %) used for training the models ....	114

5.10	Comparison of change in road quality metric against change in Household Asset Wealth Index between 2016 and 2020. A line of best fit ( $R^2 = 0.14$ ) is plotted to represent the relationship between the two metrics. There is a weak positive correlation between changes in the predicted quality of good roads and changes in asset wealth within this period. In general, small changes in road quality are associated with small changes in asset wealth, whereas large changes are highly correlated with substantial changes in asset wealth. ....	116
5.11	Map of Democratic Republic of Congo showing the location of the four projects used in the study .....	124
5.12	Graphs showing the distribution of labels in the Kenyan(left) and Liberia(right) dataset. Both datasets follow the same distribution. We observed that Labels associated with “great” road quality contributed the largest percentage of the distribution. ....	126
5.13	Bar plots showing the count of images used for inference in each of the four projects over several years. The number of images fluctuates annually, stemming from the temporal-spatial variability in image availability on Google Earth Pro (our data source) .....	127
5.14	Graphs showing binary classification predictions for selected projects. The y-axis indicates the percentage of image patches classified as “good” or “bad”. The shaded area marks the project’s timeline from inception to completion. A noticeable trend reveals improvement in road quality throughout individual project durations, with higher percentage of patches predicted as “good” towards and beyond project completion.....	129
5.15	Bar plots showing the five-class classification predictions across the six projects. The distribution of predictions among the five classes is depicted on the vertical axis for one year. A consistent uptick is observed in the percentage of patches categorized as “great” over time. This trend is especially evident in projects 2 and 4. In contrast, project 3 exhibits a relatively consistent road quality throughout the project duration.....	130
5.16	The images from Project ID: P101745 depict the condition of various roads before and after construction and rehabilitation. These examples highlight situations where the model saw the most substantial changes in road quality. The model effectively highlights the transformations, including the transition from murram to asphalt and the grading and widening of the murram roads. ....	133

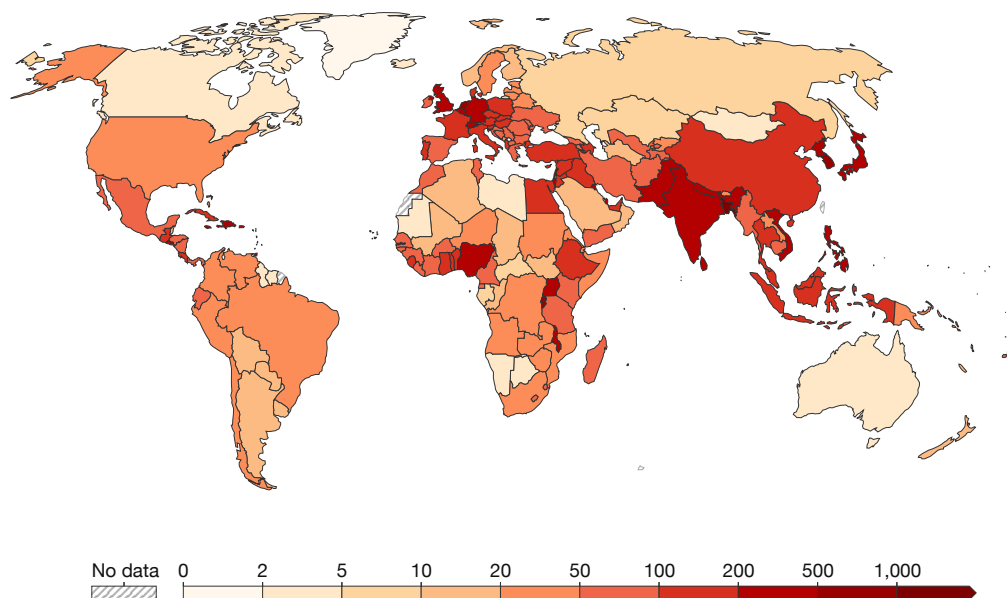
6.1	Example fishing boat on Lolwe Island. Inset: boat tracking device attached to the boat before deployment. ....	137
6.2	Flowchart of the study methodology. ....	143
6.3	Select daily traces of fishing boats. Three departure villages are each denoted with an X. ....	147
6.4	Probability distributions of the arrival and departure times of all 77 recorded fishing trips. ....	148
6.5	Average daily load profile of clusters among (a) small business, (b) residential, and (c) home business customers. ....	155
6.6	Daily load profile of residential and small business customers. ....	156
6.7	Cumulative distribution of boat battery usage from two 9.1kWh BMW i8 lithium ion battery packs connected in parallel. ....	157
6.8	Minigrid supply and demand curve on an average day of the year (a) without ice factory and electric boat charging load and (b) with ice production of 13,000 kg/day of and 102 boats charged at 15 charging stations. ....	158
6.9	Change in maximum daily electric boat charging demand a function of number of charging stations and ice produced. ....	159
6.10	Net Present Value, annual diesel consumption and annual excess PV supply as a function of number of charging stations and daily ice production. Note that the number of boats are maximized in each scenario. ....	161
6.11	Payback period on purchase of 40 HP Torqeedo Deep Blue electric outboard motor and 18.2 kWh lithium-ion battery system. ....	163
6.12	Impact of shifting electric boat charging load. ....	163
6.13	Impact of charging electric boats to 100% vs. 80%. ....	165
6.14	Distribution of signal strength as a function of distance from the island. High signal strength = 4; No signal strength = 0. ....	167

# CHAPTER 1

## INTRODUCTION

### 1.1 Thesis Contributions

Population density, 2022  
The number of people per km<sup>2</sup> of land area

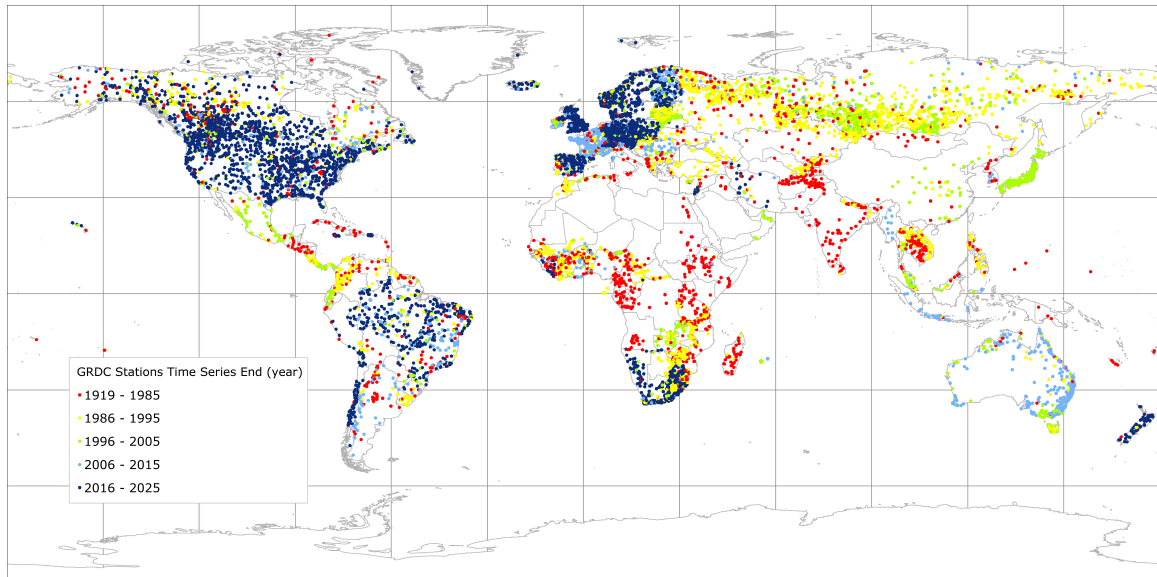


Source: Food and Agriculture Organization of the United Nations via World Bank (2021); Gapminder (v6); HYDE (v3.2); UN (2022)  
OurWorldInData.org/world-population-growth • CC BY

**Figure 1.1:** Map showing global population density (2022). Over 80% of the world’s population lives in the global south. On the other hand, a large percentage of real-time hydrological data (Fig 1.2) is available in the global north.

In this thesis, we propose data-driven methods and techniques to support critical infrastructure measurement and sustainable development. Using machine learning and remotely sensed data, we show that we can exploit knowledge and temporal-spatial characteristics learned from data-rich regions to improve data-driven predictions in regions

with little to no data. Specifically, we focus on three key infrastructures: rivers, roads,



**Figure 1.2:** Map showing the global location of active river gauge stations from 1919 to the present. Most active gauge stations (blue) are present in the global North, while a big percentage of the inactive or decommissioned stations are in the global South. (Source: World Meteorological Organization)

and electricity access. Knowledge of rivers, particularly their discharge, can help us understand how climate change is evolving and manifesting itself on global water resources and the resulting impact on key sectors such as agriculture, renewable energy generation, and the overall economy. On the other hand, better roads hasten societal development, enabling access to local and global markets and the movement of goods and ideas, leading to better equality to service provision, faster economic development, and, ultimately, better human outcomes. Finally, we develop tools to support electricity demand stimulation to improve energy access in rural communities. Here, we designed a mobile application to (a) understand the movement patterns of fishing boats to estimate the optimal times when they can be charged if converted to fishing boats and (b) understand the cellular network signal strength as fishing boats move further from the landing site in order to design best network communication and data synchronization strategies. By juxtaposing methods and techniques that effectively use highly stochastic data (e.g., climate data) with

those that use deterministic data (e.g., satellite imagery), we show that with the right data, data modeling techniques and data-driven methods, data scarce regions can benefit from reliable and consistent data generated by data-rich regions.

### **1.1.1 Global River Discharge Estimation**

The recent increase in the frequency and severity of natural disasters indicates an immediate need to address the cascading impacts of climate change. However, climate change cannot be measured directly. In a weather cycle, river discharge is the result of any hydrologic process and thus directly measures the effect of two major parameters used to measure the impacts of climate change: Temperature and Precipitation. Unlike current methods that can infer climate change patterns over a long period, river discharge is an effective proxy for measuring the effects of climate change within a short period. Unfortunately, current statistical and physics-based models neither take full advantage of hydrometeorological information encoded in over 100 years of historical hydrologic data nor are they applicable globally.

Our contribution is twofold. First, we train Long Short Term Memory (LSTM) Recurrent Neural Network models on satellite observations and daily discharge from gauged basins to predict discharge in ungauged basins. Our models show Kling-Gupta and Nash-Sutcliffe Efficiency scores of 85% and 81%, respectively, in ungauged basins with limited to no existing data, while the latest state-of-the-art process-based hydrology models show performance between 0% and 50% in similar circumstances. However, categorizing basin-wide rivers into classes is less efficient because of varying hydrometeorology characteristics across river basins. To overcome this, we propose leveraging the projected increase in the availability of fine-grained hydrometeorological data from the recently launched surface water and ocean topography (SWOT) mission. Data from the SWOT mission will help improve our knowledge of local rivers on a global scale. We hypothesize that integrating spatiotemporal hydrologic knowledge into the data modeling process (distribu-



tion/disaggregated modeling) will improve the performance of discharge prediction models. To test this hypothesis, we design experiments to compare the performance of identical Long Short Term Recurrent Neural Network (LSTM-RNN) models based on two data modeling approaches, i.e., lumped vs. distributed/disaggregated modeling. We expect the distributed modeling approach to outperform the latest state-of-the-art data assimilation and lumped machine learning models in ungauged basins. Our proposed approach can potentially improve methods for predicting river discharge on a global scale and advance our understanding of the cascading impacts of anthropogenic climate change on global water resources. Additionally, this work sets the stage to examine the constraints of process-based modeling approaches and better characterize how machine learning-based models can be used to model physical processes in hydrology and other physical sciences.

### **1.1.2 Explainable Machine Learning for Hydrology**

Little or no information exists about the majority of global water resources. Modern machine learning (ML) techniques can facilitate the transfer of hydrologic information across regions with varying amounts of historical and current data. However, the black-box nature of ML algorithms makes it difficult to understand how they arrive at accurate predictions, resulting in poor adaptation in traditionally first-principles-based hydrologic sciences. In contrast, process-based hydrology models, partial differential equations (PDEs) that characterize the properties of a particular system and its underlying processes, are easy to explain and comprehend, making them ideal for discharge prediction. However, process-based models are a simplification of reality due to epistemic limitations and are not rapidly scalable in the presence of new heterogeneous data. The optimal solution would be to use explainable machine learning to make accurate and timely predictions. Our thesis contribution is as follows. First, using heterogeneous data, we train a series of ML models to predict discharge in ungauged basins. Then, based on these models, we use cooperative game theory tools to deconstruct the contribution of each feature towards an

ML model prediction. By replicating input features as coalition members, the marginal contribution of each feature can be estimated as the sum of its contributions (Shapley values) across all possible combinations (coalitions). We hypothesize that these experiments and results can demonstrate whether hydrology can benefit from more accurate machine learning predictions without sacrificing the traditional physics-based empiricism.

### **1.1.3 Road Quality Prediction using Satellite Imagery**

Critical infrastructure, such as roads and electricity, are core systems that enable economic development. These crucial systems are frequently under-monitored in developing regions, resulting in lost growth opportunities. Recent advances in remote sensing and machine learning have made it possible to monitor and measure infrastructure faster and more frequently than traditional methods. However, ground data is often unavailable, resulting in a disconnect between labels and remotely sensed data. Although data from industrialized regions can be used to transfer innate characteristics to regions with sparse data, there exist differences in the concept of quality between regions. Additionally, consistency in data and the complexity of ML models can introduce bias due to learned characteristics across diverse regions, leading to inaccurate predictions and recommendations for action.

In this part of the thesis, our contribution is threefold. In the first part, we develop a set of convolutional neural network (CNN) models for monitoring the quality of road infrastructure using satellite imagery, enabling much larger scale and much lower costs than are achievable with current methods. For this task, we harness two trends: the increasing availability of high-resolution, often-updated satellite imagery. We train models for intercity road quality prediction using a unique dataset of road quality measurement labels (57 roads, total length is 7000km) throughout the Republic of Kenya combined with corresponding 50cm resolution satellite imagery. Using a variety of neural network architectures, we create and evaluate regression models for predicting road quality. Our

results show a best-case  $R^2$  value of 0.79 for the regression problem using a standard train-test split and an  $R^2$  value of 0.35 for the substantially harder heldout regression problem, which has the added potential to generalize more readily to other contexts. In the second part, we train traditional neural networks and cutting-edge vision transformers to predict road quality from medium-resolution satellite imagery and apply them to realistic data conditions: heterogeneous temporal-spatial resolutions. These models achieve AUROC scores of 0.934 and 0.685 for binary and five-class classification tasks, respectively, exhibiting results appealing for inference in otherwise unmeasured areas. In addition, these experiments and results show that accurate models can be derived from limited and low-resolution data. Finally, we combine the best techniques from our previous experiments to predict road quality in regions without ground observations, such as the Democratic Republic of the Congo. We share these results with the World Bank, a major investor in infrastructure development in the global south, to help them understand the impact of large-scale infrastructure on peace and socioeconomic development. This is a key step towards developing data-driven socioeconomic policies.

#### **1.1.4 Tools to support electricity demand stimulation: Converting Fishing Boats for Electric Mobility to Serve as Mini-grid Anchor Loads**

<sup>1</sup> Though electricity access remains out of reach for roughly one billion, primarily rural and low-income people, crucial strides have been made in developing new pathways for connecting households and businesses to electricity supplies. Among these, decentralized mini-grids – typically comprised of generation, storage, and a medium- and low-voltage distribution network – have considerable technical promise for balancing recent advances in decentralized generation and grid sensing and communication systems with the overwhelming economies-of-scale enjoyed by electricity grids. However, low revenues and, in

---

<sup>1</sup>Aggrey Muhebwa was a second author on this work. This work has been included in this thesis with permission from the project lead and Main Author (June Lukuyu [jlukuyu@uw.edu]). The contributions by Aggrey Muhebwa have been outlined at the start of chapter 6.

response, high tariffs necessary for cost recovery stifle the widespread development of this promising pathway for electrification. In this thesis contribution, we study techniques for addressing the principal challenge for sustainable mini-grids: demand stimulation among rural customers. Specifically, we evaluate the potential for converting diesel-based fishing boats in Lake Victoria to electric motor and battery-based systems that can provide a crucial anchor load for a nascent  $650 \text{ kW h}^{-1}$  hybrid solar-battery-diesel mini-grid. We surveyed fishing boat operators ( $n = 69$ ) to characterize the target population and deploy a custom tracking system to measure fishing boat movement patterns. Using these primary and secondary data on customer consumption, we select a candidate electric mobility system, create synthetic loads of residential and business customers, and construct technical and financial models of the complete mini-grid system. We then use these models to evaluate the excess capacity on the mini-grid for electric boats, evaluate the tradeoffs among electric mobility and manufacturing on the mini-grid, and assess the impacts of demand response capabilities for charging the boats. We find that electric boat charging contributes to at least 17% more daily consumption, resulting in substantial technical and financial value to the mini-grid system, though perhaps at the cost of additional use of the system's backup diesel generator. On the other hand, adding shifting capabilities to electric boat charging can save up to 6% of diesel expenditures at little to no impact on the system's Net Present Value. Finally, we combine these mini-grid-scale evaluations with design considerations for a future boat tracking system, providing guidance for mini-grid designers and operators to incorporate the potentially attractive load class of electric mobility systems.

## 1.2 Proposal Outline

The rest of this thesis proposal is as follows: Chapter 2 reviews traditional and emerging methods that harness artificial intelligence and big data to improve global river discharge predictions, machine learning explainability, and the necessity of regular road quality

measurement in achieving sustainable development goals. Chapter 3 reviews the current river discharge prediction techniques, highlighting their limitations. We then introduce new approaches combining machine learning with temporal and spatial hydrological data to improve global river discharge predictions, focusing on ungauged basins. Chapter 4 outlines foundational techniques for adopting machine learning in hydrology. It centers on statistical methods designed to improve the explainability of machine learning models in hydrology, which can be extended to most applications in physical sciences. In Chapter 5, we present an innovative method for road quality prediction using satellite imagery and machine learning. This approach makes use of both high and medium-resolution satellite imagery. We conclude this chapter by assessing the World Bank's road infrastructure investment in the Democratic Republic of Congo (formerly Zaire) to showcase the methodology's effectiveness. In Chapter 6, we focus on tools to promote sustainable development initiatives. Here, we detail our contributions to creating software tools to stimulate electricity demand in off-grid communities, with a case study on Lolwe Island, Lake Victoria (East Africa). Finally, Chapter 7 summarizes the thesis contributions, suggests possible applications in other areas, and discusses future work.

## CHAPTER 2

### RELATED WORK

In this chapter, we provide a background on the transformative role of machine learning (ML) in hydrology and physical sciences in general, highlighting the current applications and the growing need for explainable machine learning. We then discuss ML's contributions to sustainable development, focusing on existing literature about the role of good quality roads in the socio-economic and political development of emerging economies. Finally, we outline our vision for the future of ML applications across various facets of hydrology and critical infrastructure assessment.

#### **2.1 Is There a Need for Machine Learning in Hydrology?**

Managing water resources and responding to hydro-meteorological extremes has become increasingly important as mankind becomes aware of the cascading impacts of climate change. There is, however, a lack of real-time data, especially in the global south, where no infrastructure exists to collect hydrology data. The lack of data makes it difficult to manage water resources effectively and respond to extreme events such as floods and droughts.

Machine learning (ML) techniques such as transfer learning, zero-shot learning, data augmentation, and self-supervised learning can be used to address data gaps throughout the world by transferring learned knowledge from regions with large amounts of data to those with insufficient data. Currently, ML models can be trained on large hydrological and hydrometeorological data datasets to learn complex relationships between different variables. For example, ML models can predict river discharge in ungauged basins using only

remotely sensed data and basin physiography data. As a result, machine learning can be an integral part of water resource management in developing countries and regions with limited access to data.

Process-based models are another tool that can be used to manage water resources. Process-based models simulate the hydrological and hydrometeorological physical processes that drive the hydrologic cycle. This allows them to provide insights into the underlying mechanisms that control water movement and to predict how these systems will respond to changes in climate and other factors.

Several studies have demonstrated the superior performance of ML models for river discharge prediction over traditional process-based models, especially in ungauged basins. For example, Kratzert et al. (2018) showed that ML models outperform traditional hydrological models in predicting river discharge in ungauged basins in the United States. Similarly, Feng et al. (2021) found that ML models outperformed traditional hydrological models in predicting river discharge in ungauged basins in China.

ML and process-based models can be used together to improve the accuracy and reliability of hydrologic predictions. For example, ML models can be used to calibrate process-based models or to develop new process-based models that are more efficient and computationally feasible. Additionally, ML models can post-process the output of process-based models to improve their accuracy.

Machine learning (ML) is rapidly gaining traction in hydrology, potentially revolutionizing water resources management. ML models can simulate climate change impacts on the hydrologic cycle, assess flooding and drought risks, and develop adaptation strategies. Additionally, integrating ML with other advanced technologies, such as the Internet of Things (IoT) and blockchain, could facilitate real-time data acquisition, ensure data integrity, and foster collaborative decision-making among stakeholders.

To achieve sustainable water management outcomes, embracing a multidisciplinary approach, engaging with stakeholders, and tailoring solutions to local contexts will be es-

sential. Furthermore, developing open-source ML platforms, frameworks, and tools will likely catalyze innovation and foster a more inclusive and informed community dedicated to tackling hydrological challenges. Finally, the continual exchange of knowledge and best practices between the ML and hydrology communities will be pivotal in driving the frontier of ML applications in hydrology, thus contributing to a more resilient and sustainable water future.

## **2.2 What Role does Explainable Artificial Intelligence(XAI) Play in Hydrology?**

Machine Learning (ML) application in hydrology and the physical sciences in general has gained increasing attention in recent years. Researchers have shown that machine learning models can improve the predictions of many hydrologic processes, such as river discharge, by leveraging more than one hundred years of historical data. However, the hydrologic cycle is complex, dealing with multiple stochastic and non-linear processes governed by natural laws. Hydrologists use process-based models to simulate the hydrological cycle and predict river discharge. However, process-based models can be computationally intensive and challenging to parameterize for complex scenarios. Fortunately, ML models can overcome these challenges by rapidly learning from large amounts of data, offering more accurate predictions, and revealing many non-linear and dynamic interactions between physical processes within the hydrologic cycle.

However, hydrologists accustomed to process-based modeling frequently exhibit skepticism towards ML techniques, driven by their “black-box” nature and inherent lack of transparency, which complicates interpretation and communication of results to peers and the public. Additionally, ML techniques often overlook established physical laws governing the hydrologic system, which diminishes the understanding of the predictive accuracy of these models in varied hydrological contexts.



As such, Explainable AI (XAI) is increasingly important in the physical sciences, including hydrology. XAI techniques can help hydrologists understand how ML models work, identify their limitations, and build trust in their predictions. These techniques can provide insights into the inner workings of AI models and identify the factors most important for making accurate predictions. Thus, XAI can serve as a diagnostic tool, ensuring these AI models are congruent with established physical principles or providing insight when deviations arise. This information can then be used to improve the design and use of AI models in hydrology. Another advantage of XAI's insights is the refinement of AI models. Discerning the salient features and their influence on model decisions allows for model enhancement, data processing alterations, or integration of novel features.

While the application of XAI in hydrology is still emerging, there have been several studies and efforts to improve the explainability and adaptability of ML models in hydrology. Kratzert et al. (2019) developed a method for interpreting deep-learning predictions for rainfall-runoff forecasting. This method utilizes a technique called attention to identify the input features that are most relevant for making a given prediction. Liu et al. (2021) developed a technique for explaining deep-learning flood forecasting predictions. The method uses gradient-weighted class activation mapping (Grad-CAM) to identify the regions of an input image most important for making a given prediction.

The integration of ML in hydrology offers promising advancements but faces acceptance challenges due to its "black-box" nature. However, Explainable AI (XAI) provides the much needed transparency, enhancing trust and adaptability in ML models. As the field evolves, XAI is set to be crucial in the future, merging accurate predictions with clarity in hydrological and physical science modeling.

## **2.3 How Can Machine Learning Transform Road Infrastructure Measurement and Analysis?**

The United Nations Sustainable Development Goals (SDG) Report (2022) emphasizes the importance of built infrastructure in achieving the Sustainable Development Goals by 2030. The report notes that “investments in infrastructure are essential for boosting economic growth, creating jobs, and reducing poverty.” Machine Learning (ML) has shown significant promise in facilitating the improvement of infrastructure measurement and analysis. One of the most promising applications of Machine Learning in the context of infrastructure measurement and analysis is using remote sensing data to predict road quality. Road quality is essential for socio-economic and political development. However, traditional methods for measuring road quality are time-consuming, expensive, and impractical for large road networks. Several studies, e.g., Cadamuro et al. (2019) and Muhebwa et al. (2023), have demonstrated the efficacy of ML in measuring road quality from satellite imagery. Satellite imagery is a widely available and affordable data source. ML algorithms can extract informative features from satellite images, such as road texture, color, and geometry. These features can then be used to train ML models to predict road quality indicators, such as roughness, cracking, and potholes.

ML-based methods for measuring road quality from satellite imagery have several potential implications for sustainable development. First, these methods can help to improve the efficiency and effectiveness of road maintenance and rehabilitation programs. By providing timely and accurate information on road quality, ML-based methods can help to identify and address road problems early before they become more serious and costly to repair. Second, good quality roads, especially in rural areas, can help to reduce poverty and inequality by improving access to markets and services. Additionally, they can also help improve agriculture production and food security by making it easier for farmers to transport their produce to market and acquire farm inputs such as fertilizers. Finally, good roads can be seen as a good measure of good governance, peace, security, and po-

litical stability. Indeed, a growing body of research suggests a correlation between good roads and peace, especially in regions like the eastern Democratic Republic of Congo (DRC)[232, 6, 243, 77, 12]. For example, a study by the World Bank in 2019 found that a 10% increase in road density is associated with a 2% decrease in the likelihood of civil conflict. In the eastern DRC, road quality is a particularly important issue. The region has been plagued by conflict for decades, and much of the infrastructure has been damaged. As a result, many communities are isolated and have difficulty accessing essential services. However, there is some evidence that improved road quality is helping to promote peace and rehabilitation in the region.

Overall, ML has the potential to play a significant role in measuring infrastructure and achieving SDGs. ML-based methods for measuring road quality are particularly promising, as they can help to improve the efficiency and effectiveness of road maintenance and rehabilitation programs and promote peace and rehabilitation in regions affected by conflict. Additionally, ML for road quality measurement could be used to develop new methods for assessing other types of infrastructure, such as bridges, railways, and power grids.

Although the research into using ML to measure road quality from satellite imagery (and fostering sustainable development) is still in its early stages, the results are promising. ML-based methods have the potential to provide a more efficient and affordable way to measure and monitor infrastructure, which can support sustainable development, especially in the global south.

## CHAPTER 3

### GLOBAL RIVER DISCHARGE ESTIMATION

#### 3.1 Towards Improved Global River Discharge Prediction in Ungauged Basins Using Machine Learning and Satellite Observations

##### 3.1.1 Motivation

Anthropogenic climate change and explosive population growth are straining already scarce water resources, and the resulting impact is borne in many crucial sectors: Agriculture, renewable energy, and manufacturing, among others [110, 287, 288]. Therefore, there is a need for near real-time and accurate systems to measure the direct impact of climate change on water resources. River discharge is the result of all hydrologic processes within a river basin and, as such, can be used as a proxy for measuring increased surface melting and runoff, temporary injection of meltwater to the bed of grounded glaciers, and hydrofracturing, i.e., melt water-induced ice shelf collapse [66], all of which are key indicators of an increase in global temperature. However, there is limited measurement of river discharge on a global scale, which has hampered the ability to measure the true depth, scale, and pace of climate change.

Traditionally, river discharge has been measured *in situ* using water gauges strategically placed along the river. However, this approach does not scale well to the global level. In a weather cycle, hydrometeorological variables combine to produce the flow of water in rivers (discharge) and, as such, can be used to estimate the amount of river discharge in a hydrologic cycle. Fortunately, these variables are recorded globally using numerous satellite constellations that rotate the Earth at regular intervals. Machine Learning

approaches can encode domain knowledge and leverage the spatial-temporal relationship between hydrometeorology variables (satellite data) and *in situ* discharge data. This opens up the opportunity for more accurate river discharge predictions on a global scale, especially for the majority of the global rivers, which have no *in situ* data. In this work, we demonstrate the improved performance of machine learning methods that leverage both spatial and temporal information existing in hydrometeorological data to improve daily discharge prediction. We demonstrate that using a Long short-term memory (LSTM) Recurrent Neural Network, we can achieve Kling-Gupta Efficiency (KGE) and Nash-Sutcliffe Efficiency (NSE) <sup>1</sup> scores of 85% and 81% respectively on held-out discharge data drawn from a different distribution, outperforming the latest state-of-the-art process-based hydrology models in ungauged basins with limited to non-existing data.

These experiments and results demonstrate the impact of integrating spatial and temporal information in improving prediction of daily river discharge using modern machine learning algorithms in a physical sciences field that relies heavily on both conventional time-series and process-based models for analysis

### 3.1.2 Related Work

**Discharge measurement:** *In situ* measurements are the standard approach for measuring daily river discharge where water gauges are strategically placed at gauge stations along a river network. In places where gauge stations do not exist, process-based models, for example, the Manning Equation (Eq. 3.1) for daily discharge, is used if the geomorphological characteristics of the river are known.

$$Q_t = \frac{1}{n} A_{it}^{5/3} W_{it}^{-2/3} S_{it}^{1/2} \quad (3.1)$$

---

<sup>1</sup> **KGE** and **NSE** are the common performance metrics for measuring accuracy of river discharge predictions in hydrology

Where  $Q$  is discharge ( $m^3S^{-1}$ ),  $A$  is the cross-sectional area ( $m^2$ ),  $W$  is width ( $m$ ),  $S$  is slope (unitless), the index  $i$  specifies the cross section and  $t$  specifies the day. However, process-based models tend to degrade when trained on non-independent and identically distributed data (i.i.d), i.e., data drawn from varying geographical regions. This means that it is difficult to transfer hydrological information learned about one river basin to another river basin, making it difficult to predict discharge for basins with little to no data.

**Machine Learning in Hydrology:** The success of machine learning has largely been due to its ability to extract complex spatial and temporal patterns existing in the training data, thus overcoming the drawbacks of conventional time-series models. Long-Short Term Memory (LSTM) Recurrent Neural networks [124] have demonstrated exceptional performance in predicting discharge in gauged basins [152, 151, 79] at both local- and continental-scale. Models trained on over 100 years' worth of historical data have demonstrated the ability to extract inherent patterns in large hydrological datasets whose dynamics are dependent on various direct and indirect interconnected phenomena, thus opening up the possibility of solving a longstanding problem of regional modeling via transfer learning [212]. However, machine learning models are stochastic and non-deterministic in that they tend to encode correlation in the training data instead of causation. Furthermore, machine learning models require large training data to make better predictions, which do not exist for a majority of the basins in the world. Finally, unlike process-based models, ML models provide black box predictions, which are not easily explainable or interpretable. These make them less useful for modeling physics-driven processes in which the interactions between the underlying variables must be interpretable to enhance broader understanding.

### 3.1.3 Methods

#### 3.1.3.1 Dataset and Problem Definition

Our ultimate goal is to predict the average amount of water flowing through a particular gauge station per day. Our data is from 1980 to 2010. To achieve this goal, we leverage *in situ* discharge values obtained from the Government of Canada [102], climate forcing variables from Google Earth Engine [101], simulated discharge from the Princeton discharge database [164], river reach widths obtained from Landsat images [82], and river classes originally defined by C. B. Brinkerhoff *et. al* [34].

Although 17 classes were initially defined in [34], we focus on the five largest classes as a proof-of-concept for our proposed approach. We make the following data selection decisions. First, although previous studies [96] have shown that width is a strong predictor of daily river discharge, Landsat4-8 have repeat cycles of 16 days, with some overhead days being too cloudy to pick out river width outlines. As such, we use other features to train an intermediate model to impute widths for the missing days. Secondly, we only consider gauge stations with more than two years of *in situ* discharge data and at least five upstream reaches. This ensures sufficient data to quantify the impact of upstream hydrometeorological factors on daily discharge at a given gauge station.

#### 3.1.3.2 Sequential Learning

The standard approach in machine learning is to train, validate, and test models on data drawn from the same distribution (i.i.d); applications of these techniques for river discharge predictions are common in the literature [79, 151, 152]. However, we focus on training models that can perform well on previously unseen data (i.e., ungauged river basins), which is needed for most basins where *in situ* data are unavailable. Section 5.1.4 reports results obtained via transfer learning. By modeling daily discharge prediction as a sequential problem, we can utilize the full power of LSTMs and the historical context of related physics of the hydrologic systems to improve predictions across time and

space, both in gauged and ungauged basins. Our preliminary analysis led us to use a Bi-directional LSTM model with 4 layers because additional layers showed no substantial improvement in performance. Furthermore, we choose Swish [225] as the activation function after comparison with existing state-of-the-art activation functions. Finally, we train our Bi-directional LSTM model with L2 regularization to prevent over-fitting and present the results in Section 5.1.4. In practice, we train  $n$  models where  $n$  corresponds to the number of classes selected.

### 3.1.3.3 Training and Evaluation Metrics

Both single and ensemble models [151, 79] trained on basin-wide datasets have demonstrated remarkable results in predicting daily discharge. However, the Mackenzie River basin (where we perform our analyses) has extreme variations in the average discharge across its tributaries, and as such, a single model performed relatively similar to the current state-of-the-art process-based models [111].

As stated in 3.1.3.1, we train five models, one for each class of rivers considered. Whereas we designed multiple experiments with varying volumes of observations and meteorological variables to quantify the impact of data quantity and quality on the model performance, we only report results for one experiment that combines dynamic and static features at a particular gauge station and one upstream reach.

Consider a class with  $n$  stations; we can create all possible combinations of classes using Equation (3.2) that vary the type and volume of data available to the model.

$${}^nC_k = \frac{n!}{k!(n-k)!}; k = 1, 2, \dots, n-1 \quad (3.2)$$

Then, we train a model on each of the selected sets and test on  $(n-k)$  held-out stations. For large sets, we randomly select 20 sets at most. Our results consist of distributions across these sets to reduce bias towards a single set of high-performing gauge station datasets. Finally, we choose to report our results based on three major metrics used in hydrology to



evaluate river discharge prediction performance: Nash-Sutcliffe Efficiency (NSE) [191], Kling-Gupta Efficiency (KGE) [108], and Relative Bias (RBIAS).

NSE is a normalized statistic that determines the relative magnitude of residual variance compared to the measured data variance. NSE ranges between  $(-\infty, 1]$  with  $NSE = 1$  being the optimal value. Values between 0.0 and 1.0 are generally acceptable, while values  $\leq 0.0$  indicate that the mean of observed values is a better predictor than the predicted value.

KGE is based on decomposing NSE into its constituent components (correlation, variability bias, and mean bias). Like NSE, KGE ranges between  $(-\infty, 1]$  with  $KGE = 1$  being the desired value that indicates perfect agreement between observed and simulated values while values  $\leq 0.0$  indicate that the mean of observed values is a better predictor than the predicted value.. Finally, RBIAS quantifies the relative systematic bias in the predicted discharge values. A positive or negative value indicates a corresponding bias in predicted values, respectively, while 0.0 shows no bias in the predicted values.

Overall, a stable performance should always have KGE values higher than NSE, although it should be noted that NSE and KGE values cannot be directly compared [145].

### 3.1.4 Results

In Table 3.1, we report statistics of set combinations based on equation (3.2) of predicted discharge across the five selected classes in ungauged basins (previously unseen data). We compare our results to the existing state-of-the-art process-based models [34] with average scores of NSE and KGE in the range of 0.0 to 0.5. Class one performs poorly as compared to other classes. This is mainly attributed to the smaller widths for rivers in this class compared to others. River width is a stronger predictor of discharge relative to other features [96]. Overall, models across the remaining classes can generalize well across ungauged basins, as indicated by high values of NSE and KGE, and values of RBIAS close to 0.0, indicating less deviation of models' predictions from the actual observations. These

results strongly suggest that machine learning models are better at generalizing hydrological information across ungauged basins than state-of-the-art process-based models.

**Table 3.1:** Statistical distribution of discharge prediction results in ungauged basins. With the exception of class one, mean discharge across the remaining classes outperforms state-of-the-art process-based model predictions, which report NSE and KGE values in the range of 0.0 to 0.5.

River class		1	2	3	4	5
KGE	Mean	0.17	0.60	0.71	0.47	0.54
	Median	0.26	0.61	0.72	0.47	0.58
	Max	0.73	0.88	0.86	0.81	0.86
	Min	-1.05	0.41	0.31	-0.04	0.07
NSE	Mean	-0.28	0.58	0.72	0.27	0.47
	Median	0.10	0.62	0.74	0.41	0.50
	Max	0.62	0.84	0.87	0.84	0.81
	Min	-4.77	0.26	0.35	-0.72	-0.54
RBIAS	Mean	0.23	-0.03	-0.06	0.01	0.09
	Median	0.17	-0.03	-0.07	-0.01	0.07
	Max	1.95	0.30	0.47	0.79	0.71
	Min	-0.57	-0.29	-0.42	0.77	-0.44

### 3.1.5 Conclusion

In this part of the thesis, we demonstrated the improved performance of machine learning approaches over process-based models for predicting discharge in ungauged basins. However, it should be noted that categorizing basin-wide rivers into classes is a less efficient method because of varying hydrometeorology characteristics across basins. Future work will improve river classification by adopting stream orders or Pfafstetter units since these are more hydrologically-informed approaches for grouping rivers based on climatic regions, geomorphological, and tributary characteristics. Furthermore, we hope to statistically quantify the impact of additional training data, both qualitatively and quantitatively, on model performance. Finally, this work sets the stage to enable the examination

of constraints of process-based modeling approaches for predicting river discharge and better characterizing how machine learning-based models can be used to model physical processes, not only in hydrology but also in other physical sciences.

## **3.2 Improving River Discharge Prediction with Machine Learning via Distributed Learning of Hydrologic Information in Remotely Sensed Data**

### **3.2.1 Motivation**

Knowledge of rivers, in particular their discharge, can help us understand how climate change is evolving and its manifestation on global water resources, agriculture, renewable energy generation, and the overall global economy [215, 147, 53, 201]. The hydrologic cycles that generate river discharge are stochastic, complex, and non-deterministic systems characterized by processes and events whose dynamics depend on various direct (e.g., meteorological and environmental factors) and indirect (e.g., human interactions) inter-connected phenomena [65, 310]. This complexity ensures that *in situ* monitoring via gauges is the best way to understand rivers: a direct measurement is best. However, continuous *in situ* monitoring of global rivers is a difficult challenge due to logistical difficulties, expense, and politics [230, 95].

As a result of these challenges, process-based hydrology models are often deployed to estimate river discharge. Models are rapidly scalable in response to changing hydro-meteorological characteristics and can explain and interpret underlying model performance to describe how they arrive at predictions. However, process-based hydrology models are highly dependent on their calibrated parameters and degrade significantly when calibrated on rivers of different average discharges, seasonal variations, river widths, and geographical characteristics [280, 10, 220, 185, 18, 172], which is especially crucial with the recent increase in intensity and frequency of hydrologic extremes. This is important for modeling discharge in remote and developing regions where many assumptions must

be made to achieve accurate predictions [177, 270, 51, 218]. The needs and benefits of process-based models are an especially circular problem in ungauged basins between the need for robust models to replace gauges and the lack of gauged data to calibrate them. Watershed regionalization techniques such as spatial calibration, interpolation, and regression of basin and hydro-meteorological characteristics are often used to adopt these models and their parameters to ungauged basins [126, 210, 20]. Finally, models can simulate future projections based on physically realistic processes, i.e., “what if” scenarios [185, 18, 172], which is especially crucial with the recent increase in intensity and frequency of hydrologic extremes.

Although process-based models are widely adopted and trusted in hydrology, they have several limitations. First, dominant physical processes depend on different fluvial and hydro-geomorphological characteristics [143, 252], and as such, it is difficult to model complete interaction among all processes. Second, equifinality and model parsimony often obscure true process-based understanding for models with many parameters. That is, if a model represents dozens of processes with dozens of physical parameters, the model likely cannot be calibrated to accurately represent those physical parameters without a large volume of a priori knowledge. Finally, most process-based models are made for specific regions and conditions, making them less robust to fast-evolving temporal-spatial variability in physical processes across scales [301, 49, 51]. To overcome these limitations, heterogeneity and temporal-spatial variability of physical processes must be integrated into modeling [204, 294, 275]. Remote sensing has emerged as a way to provide these data to models.

Remote sensing has demonstrated a huge potential to improve discharge predictions on a global scale [96, 271, 54, 29, 247, 81]. Previous studies have shown that river discharge can be predicted purely from remotely sensed data [35, 205, 159, 9] or largely improved by combining satellite observations with *in-situ* discharge data [120, 164, 59, 129, 134]. In addition, models improve when remotely sensed data is included to replicate complex

hydrologic processes [137, 130]. The recently launched NASA SWOT mission will provide surface water measurements (width, height, and slope) for all global rivers and lakes greater than 50m in width [27], and a major component of its design is its expected river discharge product. Although remote sensing is a fast and efficient method for collecting hydro-meteorological data, it still faces several limitations. Satellite orbits reduce the chances of detecting high-frequency streamflow dynamics, especially with optical sensors obscured by clouds. Secondly, both optical and active sensors are prone to climate interference (e.g., storms and thick clouds), layover, and terrain interference, which reduces data quality captured in each snapshot [99, 302, 70, 283]. Finally, although purely remote sensing solutions produce excellent discharge dynamics [71, 88], the most accurate remote sensing methods are calibrated to gauges. Thus, as with models, high-accuracy remote discharge sensing is limited by *in situ* data [94].

Therefore, gauges are the best means of monitoring rivers, but these are impractical globally. Hydrologic models and remote sensing, whether used separately or in combination, are excellent tools but have unique challenges in ungauged basins. How, then, do we best combine the richness of primary satellite data with process-based hydrologic knowledge and sparse *in situ* data? We argue the answer can be found in machine learning. The earliest machine learning applications for discharge prediction were demonstrated by training a feed-forward network to predict discharge across flow regimes, which outperformed a calibrated process-based model [127]. Recent studies [207, 79, 81, 171, 150] have demonstrated the ability of Long short-term memory (LSTM) artificial neural networks to outperform process-based models on improving predictions at continental scales and in ungauged basins. Additionally, transfer learning [309, 264, 166, 304, 171], which is analogous to regionalization [144, 298, 206, 284], shows promise in tuning ML to well-measured basins and applying it to ungauged basins. At its core, ML for hydrology involves automatically discovering inherent temporal-spatial patterns in historical hydrologic data. Although current machine-learning approaches have demonstrated improved streamflow

predictions, they still have several limitations. First, ML models are still relatively non-interpretable, i.e., while we can produce accurate hydrographs of streamflow, we do not learn how or why they were produced or which combinations of hydrologic processes improved the model's learning process. Second, ML models are complex and require access to specialized computing, particularly GPU clusters. Third, ML models typically require much more training data with stricter consistency requirements than hydrologists are used to working with, and the amount of data needed for quality training far outstrips the amount of data needed to calibrate a model or remote sensing technique [178, 246, 57]. ML is moving toward interpretability [175, 168, 169, 286], but for now, it remains a powerful predictive tool that often divides opinions in the traditionally process-based discipline of hydrology.

Current ML for hydrology retrofits ML techniques to hydrologic data. However, we argue that aspects of hydrologic modeling and remote sensing for hydrology can easily be implemented in an ML-driven hydrology (hereafter known as hydroML) framework to move toward a type of ML that is more hydrologically aware and purpose-built for the discipline. For instance, hydrologists have long known that so-called distributed modeling - where inputs are spatiotemporally heterogeneous, outperforms lumped modeling - where inputs are spatiotemporally homogenous [17, 199, 89, 274, 59], yet almost all previous ML in hydrology has been lumped modeling. Moving to distributed hydroML would allow known correlations between altitude and temporal-spatial variation in isotopic signatures of snow melt, glacier melt, and rainwater to express themselves in the data [133, 219, 244, 90, 193, 214]. This shift would require changes to the input structure of ML models but should improve them considerably. Further, since ML requires huge quantities of training data, remotely sensed inputs are the best way of obtaining primary data in ungauged basins [94] in conjunction with globally available climate model output currently used in ML-driven hydrology modeling [159, 171, 79, 150, 207].

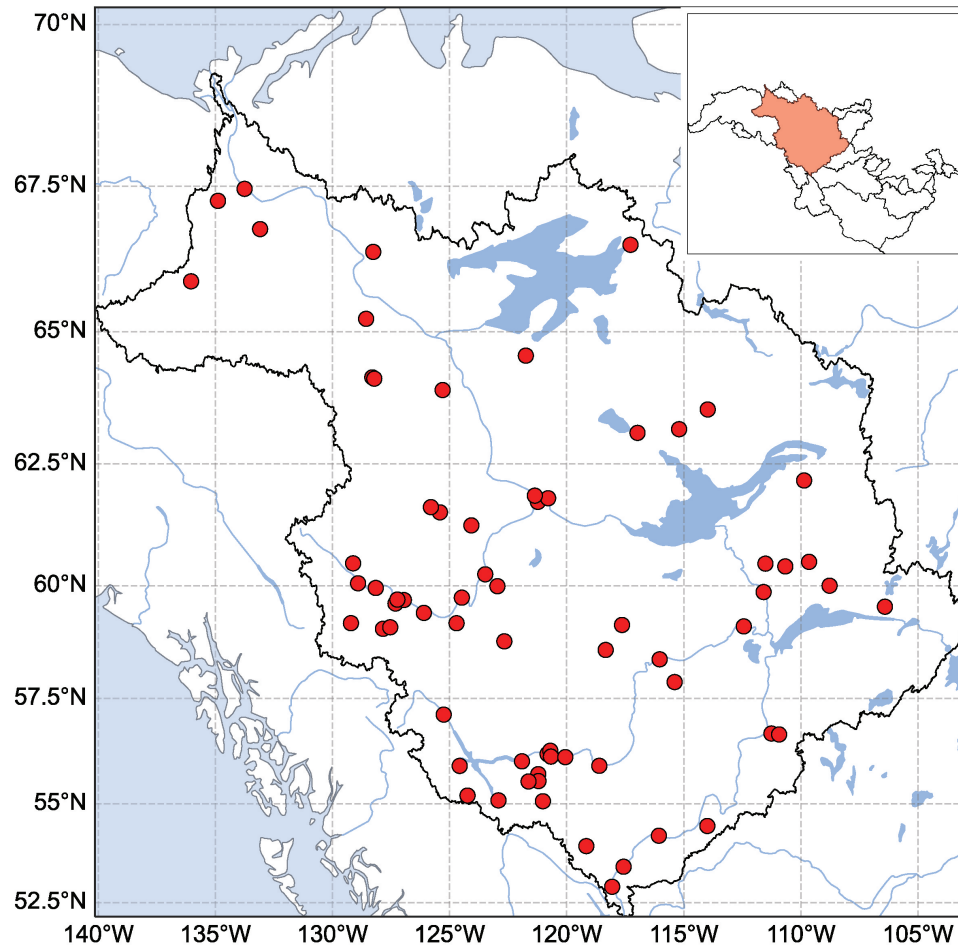
Therefore, the goal of this work is twofold: first, we hypothesize that creating a distributed LSTM model based on topologically organized geomorphologic and hydrologic information should improve ML discharge estimation performance. To test this, we compare the impact of aggregating LSTM training data over the entire upstream basin (lumped modeling) with separating upstream basin information based on the Pfafstetter Coding System (distributed modeling) while holding the LSTM architecture and input data constant. Second, we demonstrate this comparison in ungauged basins by training generalizable machine learning models in hydrologically similar basins to validation zones in ungauged basins (transfer learning). We also compare to previous LSTM architectures. Ultimately, we aim to show how tenets of hydrologic modeling and remotely sensed data improve ML in ungauged basins as we move toward a future integrated hydroML framework for predicting water resources.

### **3.2.2 Data and Methods**

#### **3.2.2.1 Data**

We test our proposed ML approach on the Mackenzie basin (Figure 3.1). The basin covers an area of approximately  $1.8 \times 10^6 \text{ km}^2$  and encompasses a wide variety of climatic conditions that include mountainous, cold temperate, sub-Arctic, and Arctic zones. The Mackenzie River drains approximately one-fifth of the total land area of Canada (Rocky and Mackenzie Mountains and the Canadian Shield) and contains over 39,000 river reaches in the MERIT Basin river network [164] developed on the MERIT HYDRO topography data [11, 296]. We select a subset of the gauge stations ( $n = 69$ ) with at least 10 years of consistent daily gauge data. This data is publicly available, courtesy of the Environmental and Climate Change Canada (ECCC). These gauge data form the basis of our training and validation work

We include both static and dynamic variables in our training data. Static variables do not change over timescales of a few decades, e.g., bed slope, sinuosity, and stream length. In



**Figure 3.1:** A Map showing the location of gauge stations (red circles) in the Mackenzie basin used in the study. The insert shows a map of the 20 biggest basins in Canada and the Mackenzie basin (shaded)



contrast, dynamic features reflect changing hydrologic processes. We gathered daily data from 1981 to 2010 that include simulated discharge and runoff from the GRADES database [164], reach averaged widths obtained from the Remotely-sensed Arctic Discharge Re-analysis (RADR) database [81], and climate model data. Climate data are from the Global Land Data Assimilation System (GLDAS)-2.1 model [235, 19] and include 3 hourly climate data gridded at 0.25 x 0.25 degrees resolution which was downsampled to daily data. These data were downloaded from the Google Earth Engine platform [101]. This mixture of modeled climate data and remotely sensed data gives us both primary and secondary data for the basin. Previous studies have shown that stationary data are relatively easy to model with ML [125, 63]. Appendix A gives all the variables used in this study.

Previous studies have shown that river width is a strong predictor of river discharge [96, 111, 35, 83, 80]. However, Landsat-derived river widths have a frequency of 16 days before considering cloud cover and seasonality. This is not a problem for hydrology approaches, but LSTMs require training data without gaps [45, 162]. Therefore, we “impute” a complete width record from the Landsat observations in the RADR dataset. Imputation is a statistical process of determining and assigning replacement values for missing or invalid data points in a multivariate dataset by leveraging possible correlation between covariates [37, 135]. Thus, we estimated missing width values using a regression model fitted with the remaining covariates in the dataset. We have chosen this imputation approach to retain river widths as a strong predictor of discharge as important primary data. Since our goal is to compare lumped and distributed ML, we only trained/tested at gauges with at least five upstream reaches to ensure that there is sufficient data to quantify the impact of upstream climatology factors toward daily discharge at a given gauge station. Further, we limited our dataset to gauges with at least 10 years of daily discharge data as preliminary tests indicated that this reflected the scale of data that was needed to accurately train an LSTM model without overfitting [240]. Finally, we select Pfafstetter orders with at least 4 gauge stations, i.e., order 4 (25 gauge stations), order 5 (23 gauge stations),

order 6 (13 gauge stations), order 7 (4 gauge stations), and order 8 (4 gauge stations) for a total of 69 gauge stations.

### 3.2.2.2 Sequential Learning via LSTMs

Our machine learning models are based on Long-Short Term Memory Recurrent Neural Network (LSTM-RNN) model architecture. This is a type of artificial neural network originally proposed by [124] that is capable of processing sequential data. LSTMs have been successfully applied to language modeling, video understanding, music transcription, discharge prediction for hydrology, and other applications [73, 259, 93, 207, 79, 150]. Unlike standard neural networks that only understand the spatial context of data, LSTMs can extract both the temporal and spatial context encoded in the training data [300, 292]. At a structural level, an LSTM network consists of a series of identical recurrent neural networks where the previous neural network ( $t_{i-1}$ ) passes information to the current network ( $t_i$ ). This cascading architecture allows LSTMs to handle the sequential context encoded in historical data, e.g., hydrologic data. Unlike traditional RNNs, LSTMs can maintain information in memory over long periods of time, thereby overcoming the problem of vanishing gradients [46, 128]. This allows LSTMs to learn long-term temporal dependencies, i.e., where the desired output depends on inputs presented at times far in the past (lookback window), which is important when modeling physical processes that occur at different spatial resolutions. Consequently, the size of the lookback window determines how much information a model learns about a particular physical process at any time. The LSTM network architecture can be either unidirectional or bidirectional [104, 251, 87]. Unidirectional LSTMs learn encoded features in a time-increasing manner (forward chain) i.e., information from each feature is derived at every timestep  $t = t[0], t[1], t[2], \dots, t[n]$ , but only information from previous timesteps ( $t_{i-1}$ ) is used to improve prediction at the current timestep ( $t_i$ ). On the other hand, bidirectional LSTMs concatenate two unidirectional LSTMs in opposite directions such that the model learns encoded features in both

time-increasing (forward chain) and time-decreasing (backward chain) manners. This is important where information encoded at the next timestep( $t_{i+1}$ ) can further improve prediction at the current timestep ( $t_i$ ). Knowledge of river discharge at the next timestep can improve our prediction at the current timestep, and as such, we adopted the bidirectional LSTM network model architecture for our experiments.

Designing an optimal machine learning model requires a rigorous hyperparameter search [22, 47, 303]. Hyperparameters are model configurations that cannot be learned from the training data and, therefore, tunable to a specific predictive modeling problem. However, finding optimal hyperparameters is a complex and computationally expensive task. On the other hand, a poor choice of hyperparameters can cause a model to overfit, i.e., the model memorizes patterns in training data and as such, fails to generalize to previously unseen data. Therefore, we use regularization techniques and deeper layers to prevent our LSTM models from overfitting. Regularization [28, 92] constrains the model's coefficient estimates (learned parameters), making it generalizable to new data. Layers are topological structures of neurons that make up a neural network and are defined by weights and biases. Weights are sharable in recurrent neural networks and define how important a given input is to the next neuron, while biases define how easily a given neuron can get fired. Layers ensure easy sharing of parameters and statistical strengths across different parts (temporal positions) of an LSTM model, making it more generalizable to sequences not seen in training data [100]. Finally, the choice of an activation function [181, 3] impacts how fast and efficiently a neural network can extract contextual information in training data. Thus, we adopted Swish [225] as the output layer activation function based on its superior performance over existing state-of-the-art activation functions on benchmark datasets. Ultimately, our bidirectional LSTM network model had 4 layers, as additional layers showed no substantial improvement in performance. Between each LSTM layer, we added a dropout layer [119, 281] that randomly dropped 20% of the connections.

### 3.2.2.3 Experiment design

We hypothesize that an ML model trained with topologically organized distributed geomorphologic and hydrologic information should outperform other discharge prediction models that lump the same training data. To this end, we designed three experiments with identical ML models per section 3.2.2.2 but different organizations of the training data. We also include a comparison to a state-of-the-art hydrologic modeling approach for the basin that assimilates remotely sensed river widths (RADR- [81]) and a recently published LSTM model (PUB-LSTM [153]).

#### 3.2.2.3.1 Experiments and literature comparisons

1. At station experiment: We used both dynamic and geomorphological static variables in addition to climate data in a 25 km buffer around a given gauge station as input features to an ML model. This is the least possible data we can use to train any ML model that leverages temporal-spatial information encoded in historical data around a gauge station.
2. Lumped experiment: Besides leveraging local information around the river outlet (at station experiment), we included aggregated climate data from the largest possible upstream basin. Therefore, this experiment has static and dynamic variables from the prediction reach and averaged upstream climatology. This represents the approach taken by [207, 79, 80, 171, 150] among others
3. Distributed Experiment: We took the static and dynamic variables at the prediction reach as in the last two experiments but disaggregated the upstream climate data by Pfafstetter level. Therefore, given a river with  $n$  orders of upstream sub-basins, we generated  $(n * k)$  additional input features, where  $k$  = several modeled hydrometeorological processes. The data were averaged across all upstream basins per order

4. Comparison datasets: We compare our approach against off-shelf results from RADR and PUB-LSTM models. The RADR [81] dataset was calibrated on data from 1984 to 1998 and assimilated with remotely sensed discharge data from 1984 to 2018 for the entire Arctic region (including the Mackenzie basin). Data assimilation in process-based modeling provides time-dependent distributed estimates that are updated whenever new data becomes available, i.e., the model's states are updated in response to how it performs at a given time [179, 50]. We also implemented the PUB-LSTM model defined in [153] and trained it with data defined in the lumped experiment but with two major changes to the training process; we have fewer catchment (geomorphological) characteristics per our inputs described above and use 7 stations (as compared to 12 stations in Kratzert's work) for held-out data for k-fold cross-validation

Our approach requires us to develop order-specific ML models given the rigid requirements for LSTM training. That is, all three of our ML experiments each have five different LSTMs - one for each order from 8 to 4, as these orders contain sufficient training data. In order to apply our model to an ungauged basin, we would need first to identify the order of the river reach of interest and then select the appropriate order model to deploy. This means that our methods are unable to predict flows in orders other than 4-8, but in return for this compromise, we can estimate flows quickly, efficiently, and accurately in ungauged basins, as proven below. Further, global datasets like those used to build our models already identify all global rivers' order, so there is no additional computational burden on future users of these methods.

**3.2.2.3.2 Validation design and applicability to ungauged basins** Our goal is to train ML models that can accurately predict daily river discharge in ungauged basins. A standard approach in machine learning is to split the model's input data into training and validation sets by a particular ratio[293, 223, 249]. This means that models are trained and

validated on data drawn from the same distribution and are referred to as independent and identically distributed data (*i.i.d*), whereby each random variable has the same probability distribution as the others, and all variables are mutually independent. As such, it is easy to train models that perform well on both the train and test data but cannot generalize well to previously unseen data (overfitting). However, our goal is to transfer hydrologic knowledge to ungauged basins. For this reason, we use cross-validation to evaluate the performance of our ML models. Cross-validation is a machine learning technique [261, 228, 233, 23] where several ML models are trained on subsets of available input data and evaluated on complementary subsets of the same data. This introduces heterogeneity in the training data by repeated resampling, thereby improving the ability of models to generalize to previously unseen data.

Since we use stream order as a unifying concept to do our distributed modeling, we must build, train, and validate models that function per order. Previous studies [80, 150, 263] have either treated training data as a single entity, thereby making it easier to implement out-of-sample testing using k-fold validation (dividing data into groups of approximately equal sizes) or splitting training data by a given percentage (e.g., 70/30 split) for models trained and tested on *i.i.d* data. Conversely, different Strahler orders in our training data have unequal gauge stations (Table 1), making it difficult to implement an identical k-fold validation strategy. The imbalance in data across different orders can result in model uncertainties (e.g., inverse relationship between NSE and increase in the number of sub-basins). We try to mitigate this by a combinatorial selection of training data for individual models in each order and by maintaining an equal number of stations ( $k$ ) in each training and validation subset. This strategy of organizing training data maintains a relatively consistent volume of training data across the entire data strata. Consider a stream order with  $n$  stations; we can create sets of all possible combinations of stations in that order where each set contains  $k$  stations where  $k$  is any arbitrary number less than  $n$ . We chose  $k = 3$  for our experiment as a tradeoff between the minimum number of stations in each

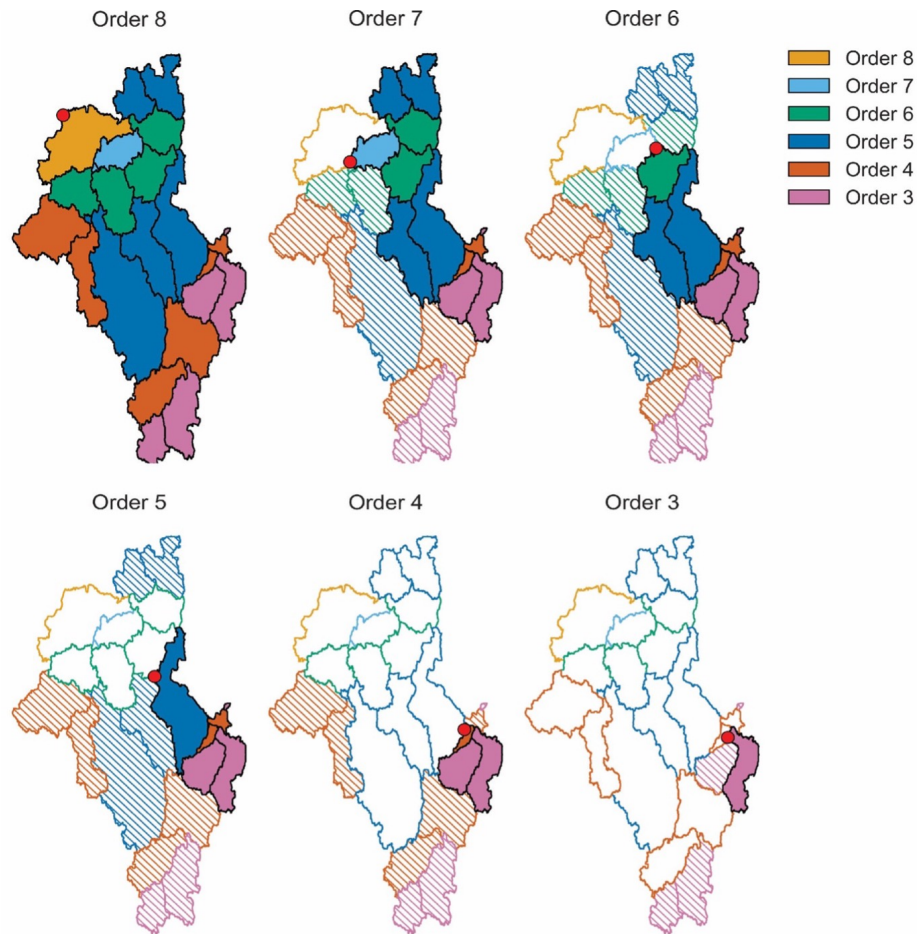
order (orders 7 and 8 each have 4 stations) and the computation time to train models for all subsets in each order. We then train a model on each subset and evaluate it on the complementary subsets of the same order. Therefore, in a basin with  $n = 25$  gauge stations, we try all combinations of  $k = 3$  training and  $(n - k) = 22$  validation stations. For stations with a large number of subsets, i.e., orders 4 to 6 (Table 1), we randomly select 24 sets from all possible set combinations in order to balance model compute time with statistical representativeness. Preliminary experiments to increase the size of the sets from 24 to 50 and 100 had no substantial improvement/degradation in model performance. Our results are presented as distributions of predictions across the complementary (validation) sets, as opposed to reporting the results of individual or selected ML models that may perform particularly well or particularly poorly at a gauge. The width of these distributions, therefore, corresponds to the sensitivity of our three experiments to a particular combination of training/validation data.

**Table 3.2:** Table showing the number of generated and contributed sets used for training in each Strahler river order

Strahler order	Number of gauge stations ( $n$ )	Number of training stations per set ( $k$ )	Number of ungauged validation stations per set ( $n-k$ )	Possible training/validation combination sets ( $nCk$ )	Number of selected sets used to report results
4	25	3	22	2300	24
5	23	3	21	1771	24
6	13	3	10	286	24
7	4	3	1	4	4
8	4	3	1	4	4

Note that orders 7 and 8 have sufficient data to train and test but insufficient data to cross-validate. Remember also that we build per-order ML models, and thus, the performances here reflect only rivers of that order, and we cannot predict in orders below 4 and above 8 given the available data in the Mackenzie.

Ultimately, and importantly, all results represent an ungauged case where validation is only done on the  $n - k$  stations not used in training and then tested in combination per Table 3.2. This represents a common hydrologic situation where there are some gauge data in a basin but not in areas where you'd like them to be. Our methods would use the gauge data in hand, per order, to estimate all ungauged reaches of the basin of the same order. Here, we withhold gauge data to make that test, and each validation set is completely independent of the others for a true ungauged case.



**Figure 3.2:** Schematic representation of a hypothetical order eight basin network. The red circle represents the location of a gauge station on the delineated basin’s outlet. At each hierarchical level, a single-order basin and its lower-order basins are selected (filled) while the remaining basins on the same level or not upstream of the selected basin within that level are ignored (hatched). This topological representation (Strahler river order system) integrates the temporal-spatial variation of physical processes at different stages of a river network.



### 3.2.2.4 Evaluation Metrics

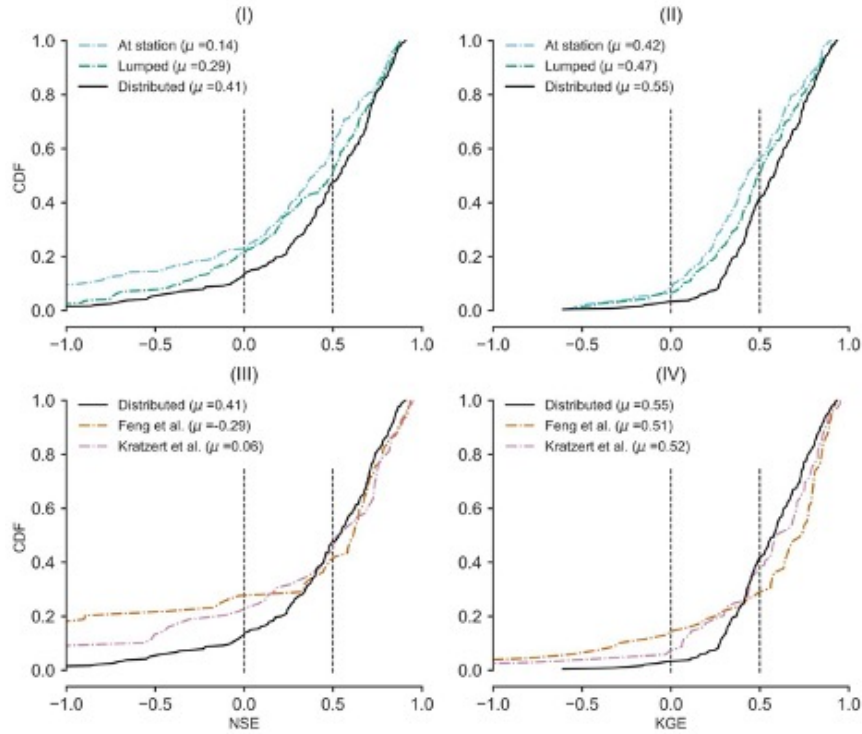
We report our results based on four major metrics used to evaluate the performance of discharge prediction models: Kling-Gupta Efficiency (KGE) (Gupta et al., 2009), Nash-Sutcliffe Efficiency (NSE) [191], Relative Bias, and Normalized Root Mean Squared Error (NRMSE). These standard hydrology metrics assess different aspects of the hydrograph and errors in both timing and volume of water [164, 111].

### 3.2.3 Results

Our experiments show that a distributed data modeling approach produces more accurate models than at station and lumped approaches when training ML models for predicting discharge in ungauged basins. Figure 3.3 shows key results of these experiments as cumulative distribution functions (CDFs) for KGE and NSE across each of the experiments defined in section 3.2.2.3.1 As a reminder, all results represent an ungauged case where validation is only done on the  $n-k$  stations not used in training and then tested in combination per Table 3.2.

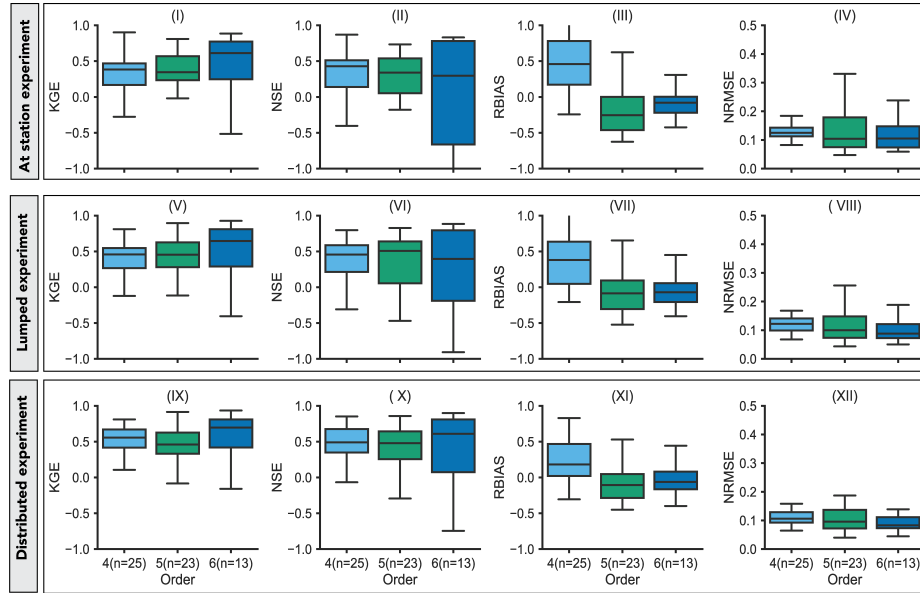
#### 3.2.3.1 Predictions in ungauged basins

First, we compare results from at-station, lumped, and distributed experiments. Figures 3.3(I) and (II) show that increasing quantities of upstream basin gradually improves discharge estimation. Here, we define performance improvement as the *CDF* curve of the distributed experiment results shifting to the right of both at-station and lumped experiment curves. Order level-specific models trained with the least possible data (at station experiment) have 77% positive NSE predictions and 92% positive KGE predictions. Both KGE and NSE values range between  $(-\infty, 1]$ ; in general, positive values are desirable; for instance, a negative NSE value indicates that the mean of observed values is a better predictor than the predicted value. When aggregated upstream basin information (lumped experiment) was included in the model training process, there was no significant improvement in performance ( $P$ -value  $> 0.05$ ) although training the same models with



**Figure 3.3:** Cumulative distributions functions (CDFs) of NSE and KGE for defined experiments and selected benchmarks calculated from distributions across all Pfafstetter orders. Figures (I) and (II) compare the performance of models in the at-station and lumped experiments against the models trained with data from the distributed experiment. Figures (III) and (IV) compare the performance of models in the distributed experiment against two literature models: [81]; [150]. A shift to the right indicates an improvement in model performance. Baseline models from the literature show lower skill than the ML here when all models perform poorly ( $-\infty < \text{NSE} \& \text{KGE} \leq 0.0$ ) but better performance when all models have good predictions ( $0.5 < \text{NSE} \& \text{KGE} \leq 1.0$ ). The distributed model outperforms the at-station and lumped models across the entirety of the results. CDFs are preferred because they represent the overall model performance across the entire test dataset

topologically organized data (distributed modeling) improved NSE by 9.6% and KGE by 4.6%, respectively.



**Figure 3.4:** Top to Bottom: Distribution comparisons of selected metrics on held-out predictions for at station (I-IV), lumped (V-VII), and distributed (IX-XII) experiments. Note that distributions for seventh and eighth orders are not included due to limited gauge stations in the training set. Figure S1 shows a distribution comparison across all experiments and literature models.

Figure 3.3 summarizes performance across individual experiments (inter-experiment differences) but ignores order level intrinsic differences (intra-experiment differences). Figure 3.4 shows the model performance of order-level delineated models in each defined experiment for orders with at least 20 validation %sets. First, we compared predicted discharge metrics across individual experiments (rows in Figure 3.4) as follows: When ML models were trained with the least possible data (at station experiment) i.e., Figure 3.4(I)-(IV), we observed a significant improvement in median KGE from 0.38 to 0.61 as basin size increased from order 4 to order 6. When we included aggregated upstream information (lumped modeling) in the training data, i.e., Figure 3.4(V)-(VIII), median KGE improved linearly with an increase in the number of sub-basins (from 0.45 in the fourth order to 0.64 in the sixth order). Finally, when we trained ML models with hierarchically organized data (distributed modeling), i.e., Figures 3.4 (IX)-(XII), median KGE improved from

0.56 to 0.69 from the fourth order to the sixth order. NSE, however, was relatively constant across orders, with a noticeable increase in the interquartile range (IQR), the median, for the largest order with 10 stations. When we compared similar spatial resolutions (orders) across the three experiments (columns) at-station, lumped, and distributed experiments, we observed an improvement in both NSE and KGE scores as orders increased and more information was added to the data modeling process. Consider Figures 3.4(I), (V), and (IX), KGE improved from 0.38 to 0.56 in the fourth order, 0.34 to 0.46 in the fifth order and 0.61 to 0.69 in the sixth order, from at station to distributed experiments respectively. Likewise, we observed an equivalent improvement in NSE, i.e., Figures 3.4(II), (VI) and (X) from 0.42 to 0.48 in the fourth order, 0.34 to 0.47 in the fifth order and 0.29 to 0.60 in the sixth order.

When we compared the performance of literature models on an order level basis, we observed a much more substantial improvement in performance as the number of sub-basins increased. The RADR model [81] had the most noticeable improvement in skill scores, with median KGE improving from 0.63 in the fourth order to 0.77 in the sixth order, while median NSE improved from 0.47 to 0.58 in the corresponding orders. On the other hand, the [150] model demonstrated an improvement in KGE from 0.68 in the fourth order to 0.72 in the sixth order but a decline in NSE scores from 0.72 in the fourth order to 0.56 in the sixth order. The key takeaway is that more hydrological information improves model certainty for both process-based and data-driven models but that the ML introduced here has a more consistent performance.

Finally, we compare results of the distributed experiment against model predictions of both an off-the-shelf re-implementation of an ML model proposed by [150] with minor modification and off-the-shelf results of a remote sensing data assimilation over the same basin and time period from [81], i.e., Figure 3.3(III)-(IV). For all positive predictions, the distributed experiment outperformed both literature models; 86.7% of NSE and 96.7% of all KGE are positive as compared to 78.3% of all NSE and 92.8% of all KGE predictions for the

Kratzert et al. (2019) model and 72.5% of all NSE and 86.7% of all KGE predictions from the Feng et al. (2021) model. However, both literature models outperformed the distributed experiment in areas where all models performed extremely well, i.e., KGE and NSE > 0.5. Here, the Feng et al. model (59% of all NSE values and 71% of all KGE values) outperformed both our proposed approach (53% of all NSE values and 58% of all KGE values) and the re-implemented Kratzert et al. model (57% of all NSE values and 62% of all KGE values). Interestingly, our methods outperform the literature models when all models perform poorly (Figures 7 and S1): The distributed modeling approach has 13% of all NSE values and 3% of all KGE values as negative predictions across the entire experiment, the Kratzert et al. model has 22% of all NSE values and 7% of KGE values as negative predictions across all orders, and the Feng et al. model has 28% of all NSE values and 13% of all KGE values as negative predictions across all Pfafstetter orders.

The overarching result from Figures 3, 4, and 5 is that distributed models outperform other defined discharge prediction models, which confirms our hypothesis that integrating topologically organized geomorphological and hydrologic information improves the performance of ML models for discharge prediction. Our results against literature models are mixed, and as such, we explore the skill of all models next.

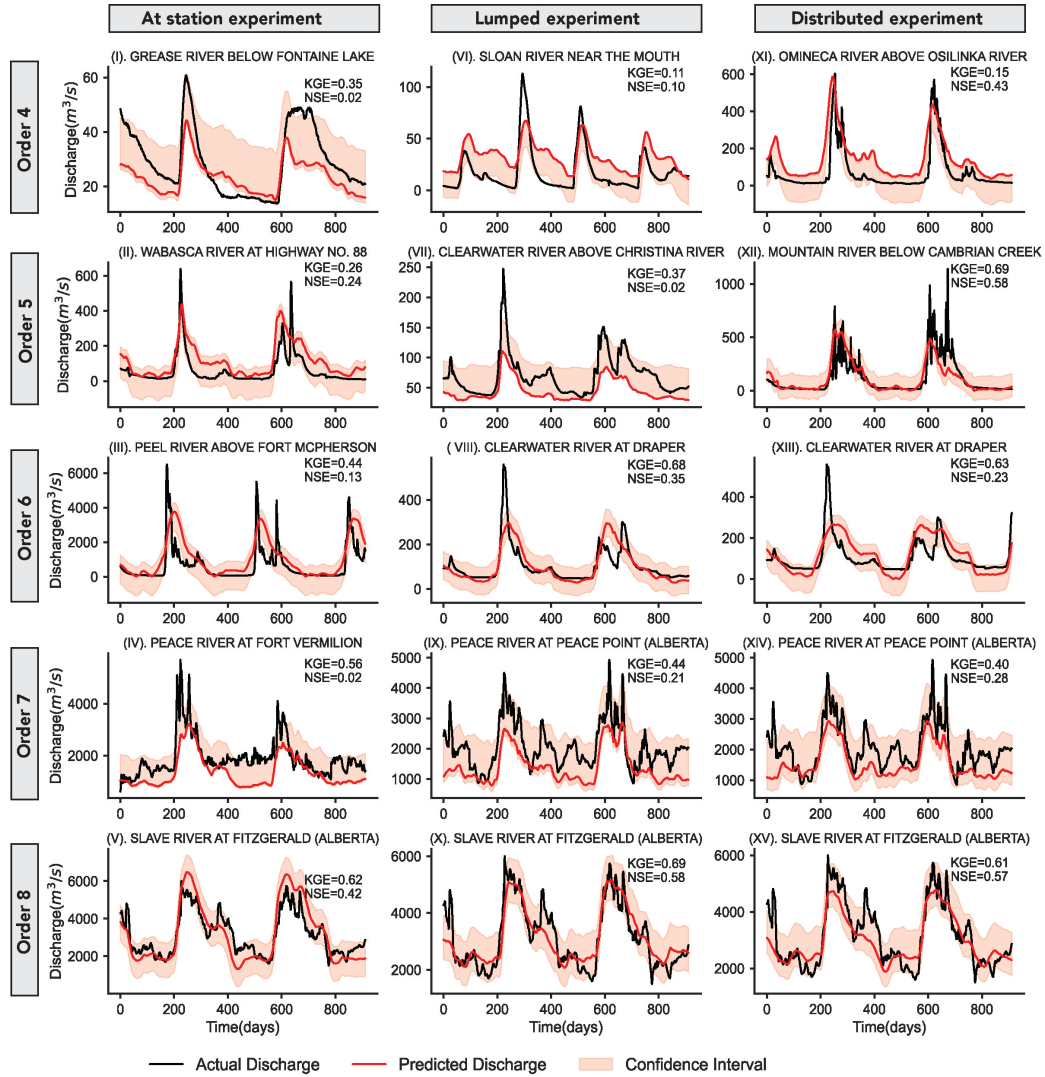
### **3.2.4 Discussion**

We hypothesized that a distributed data modeling approach improves the performance of ML models for discharge prediction. To this, we compared three experiments: At-station, lumped, and distributed modeling approaches (section 3.2.2.3.1). In theory, comparison between at-station and lumped experiments tells us the importance of including additional hydrologic information in the model training data while comparison between at-station or lumped experiment with the distributed experiments highlight the importance of integrating information about varying temporal-spatial scales of physical processes (topologically guided training data) into the data modeling process i.e., Figure 3.3(I)-(II) and Figure

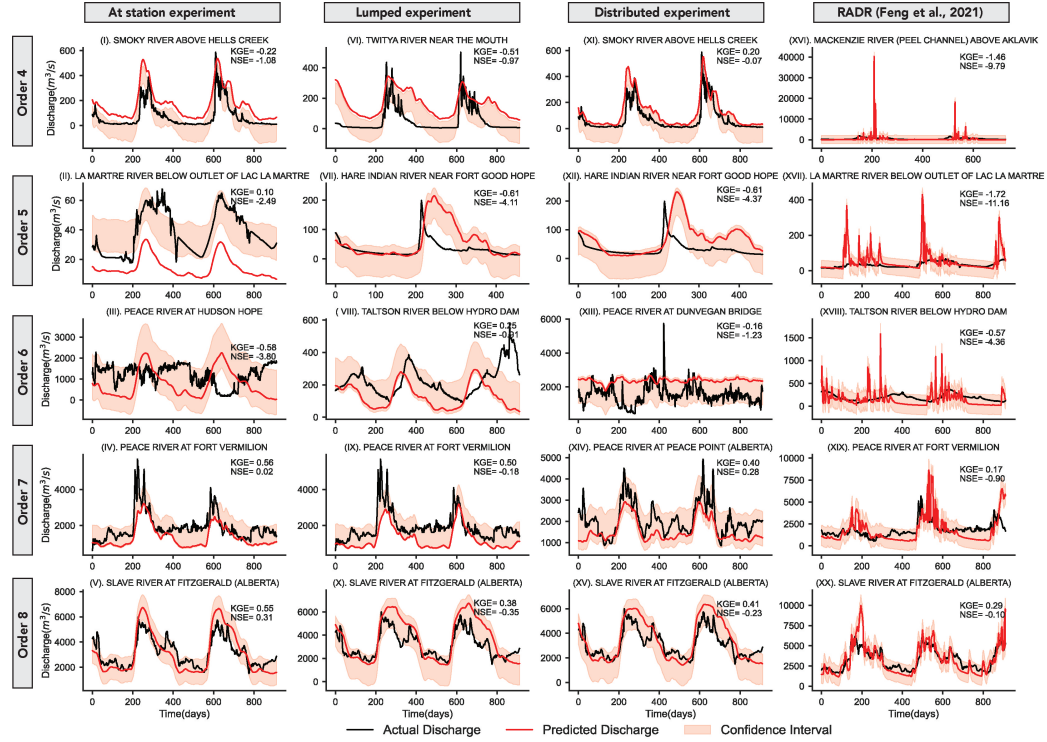
3.4. Indeed, previous studies, e.g., [17, 89, 274, 59, 187] showed that both process-based and data-driven models performed well when trained with spatially distributed data. This is because the distributed modeling approach assumes a non-uniform spatial influence of both physical processes and anthropogenic drives in the upstream basin toward discharge at the basin outlet (spatial-temporal heterogeneity), while lumped modeling assumes that discharge at the basin outlet is a true representation of the integral response of all hydrologic process in the upstream (spatial-temporal homogeneity). This is why, although we observed performance improvement in both the lumped and disturbed experiments as we added information by increasing the order (a result of added hydrologic information about physical processes in the upstream basin), model performance was more pronounced and consistent when training data was topologically organized (distributed modeling) than when it was aggregated (lumped modeling).

When we compared results of the distributed experiment against model predictions of both an off-the-shelf re-implementation of an ML model proposed by [150] and the process-based RADR model assimilation proposed by [81], we found both strengths and weaknesses of our approach. Our model outperformed the literature models when all produced poor hydrographs 3.6, and our skills scores have a much higher “floor” than the literature models. However, we have a lower “ceiling” as well- the literature models performance exceeds ours when all models perform well, although the difference between this study and the literature is much more pronounced at a lower skill (where our results improve skill). We attribute the superior performance of Feng et al. RADR product at the high skill areas to two factors: First, RADR was calibrated on remotely sensed data drawn from the same distribution (independent and identically distributed data), and second, the model was assimilated on heterogeneous data from the entire arctic region (as compared to our models trained on data from only the Mackenzie basin). We attribute the Kratzert[153] model’s better performance to a different training strategy than the distributed experiment. Whereas models in the distributed experiment were trained and val-

idated on order-specific training data, the Kratzert model used a k-fold validation strategy and trained on the entire spectrum of data (all 69 gauge stations). This strategy ensured that the model was trained with more heterogeneous data, which improved its generalization to previously unseen data. This also has the advantage of predicting flows at all river basins.



**Figure 3.5:** Representative hydrographs showing randomly selected models with  $0.0 < NSE \leq 0.6$  in each of the experiments; At-station (left), lumped (middle) and distributed (right) experiments across 7 the defined orders, i.e., from order 4 (top) to order 8 (bottom). Here, we plot hydrographs for the first 2.5 years.



**Figure 3.6:** Left to right: Representative hydrographs showing the worst performing ML models in each of the experiments and the non-ML literature model; At station experiment, lumped experiment, distributed experiment, and RADR model (Feng et al., 2021) across the defined orders, i.e., from order 4 (top) to order 8 (bottom). The RADR model overestimates peak flows and underestimates base flows in lower orders. Here, we plot hydrographs for the first 2.5 years

Our distributed experiment, on the other hand, has two advantages. First, when all models performed poorly (Figure 3.6), models in this experiment still performed better than literature models. In general, we attribute poor performance (poor generalization) to limited training data, a reality for much of the world where training data are rare, nonexistent, or proprietary [96]. Second, acknowledging the influence of physical processes on the hydrologic cycle, the existence of these processes at different spatial resolutions, and their varying dominance across different geographical regions, order-specific models in the distributed experiment firmly integrate this hydrologic knowledge in the data modeling process as compared to the literature models.

One possible explanation of why models in the distributed experiment perform better when all models have low skill scores is that despite limited training data, these models



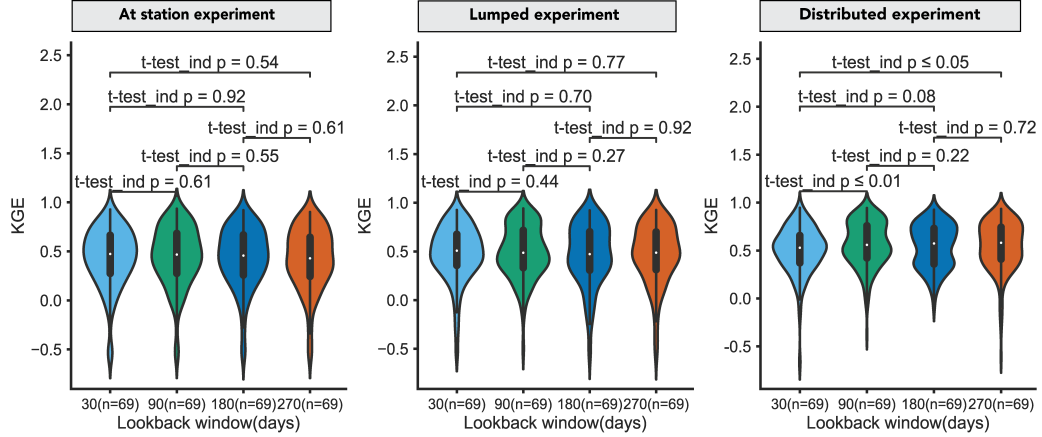
are better than literature models at leveraging the high correlation between temporal-spatial variability and physical processes to extract meaningful patterns in the training data. Therefore, as we move towards understanding discharge on a global scale (as compared to a few well-hydrologically mapped regions), it is sufficient to say that despite the unique advantages that each of these models possess, a distributed data modeling approach would be more applicable on a global scale, if not always more accurate.

Figure 3.5 shows hydrographs of randomly selected ML models in each of the orders 4 to 8 whose NSE scores lie between 0.0 and 0.60. Here, we use  $0.0 < \text{NSE} \leq 0.6$  as a representative average performance range across the prediction distribution. Across individual experiments, the models' confidence to re-create discharge increases as the number of sub-basins increases. For example, absolute relative bias ( $|\text{RBias}|$ ) improves from 0.24 to 0.007 in the station experiment, 0.80 to 0.002 in the lumped experiment, and 0.82 to 0.06 in the distributed experiments, as the number of sub-basins increases (i.e., from fourth to eight order). However, there is a distinct difference in hydrographs across the defined experiments. Consider the fourth order across the three experiments, normalized root mean squared error (NRMSE) reduces from 0.17 in the at-station experiment to 0.09 in the distributed experiment, indicating an improvement in model performance in response to additional hydrologic information in the training data. Further, we observed that even the best-performing models in the at-station experiment fail to recreate medium to high peak discharges by a big margin in the lower orders. This is not surprising, given that peak discharges are a function of events in the upstream basin, e.g., after maximum rain intensity or melting of accumulated snow [278, 138, 91, 139], information that is not included in the training data. Indeed, the impact of the knowledge of events in the upstream basin becomes more prevalent as more information is added to the training data. This is visible in the hydrographs of both the lumped and distributed experiments in Figure 3.5 (average-performing), in which models recreate most of the peak discharges (or miss them by a small margin). To verify this, we aggregated the top 10 peak flows of each station

and observed that the mean error of the best-performing models across each experiment (defined as the average of the top 10 peaks in each order) reduced from  $2901.58 \text{ m}^3\text{s}^{-1}$  in the lumped experiment to  $2518.74 \text{ m}^3\text{s}^{-1}$  in the distributed experiment and observed a similar pattern between the same orders across the two experiments.

Different geographical and climatic regions have different dominant physical processes that occur at different temporal-spatial scales. Results in section 3.2.3.1 showed that integrating this knowledge of temporal-spatial variations (distributed modeling) improved the discharge prediction of ML models. Earlier studies [140, 62] showed that longer lookback windows with a longer “memory” of past hydrologic conditions improve model performance. However, this performance improvement comes with increased computational power and time. To further evaluate the impact of the lookback window on model performance, we repeat experiments defined in section 3.2.2.3.1 with varying lookback window sizes of 30, 90, 180 and 270 days. We hypothesize that longer lookback windows improve model performance. Pairwise comparisons of distributions for both at-station and lumped experiments indicate that the size of the lookback window has no impact on model performance (P-value  $> .05$ ). However, there is a significant difference between distributions of results for lookback pairs (30, 90), (30, 270) days of the distributed experiment (P-value  $\leq .05$ ).

We attribute the high correlation between pairs of lookback windows for both the at-station and lumped experiments to the fact that both experiments ignore spatial variations of events in the upstream basin (physical processes). On the other hand, we attributed the differences across the lookback window pairs of the distributed experiment to the integration of knowledge of both temporal and spatial variations of physical processes in the data modeling process, indicating that the impact of dominant physical processes on model performance is prevalent at different temporal and spatial scales. We found that at various temporal scales (with similar spatial scales), a lookback of as little as 90 days was enough to capture temporal information encoded in the training data and as such,



**Figure 3.7:** Left to right: Pairwise comparison of KGE distributions with varying lookback window sizes and corresponding statistical significance tests across the three experiments. Inter-experiment comparisons show that distributions of lookback for at-station and lumped experiments are similar, while there is an observable difference in distributions of lookback windows of the distributed experiment

we saw no additional value in longer lookback windows, although this could be different for different geographical regions and/or data.

We do not report individual skill scores of the seventh and eighth orders (Figure 3.4) due to a limited number of gauge stations (Table 1). Further, data availability limits the minimum number of gauge stations ( $k$ ) to include in each subset, which reduces data heterogeneity for each order-specific model. For instance, on order 8,  $k = 3$  represents 75% of the data as training, while on order 4,  $k = 3$  is only 12% (Table 3.2). We chose to keep  $k$  constant instead of choosing a constant train/test ratio because this allows sharing model hyper-parameters (and structure) and is easier to compare results of models trained on the same number of gauge stations ( $k$ ) across different orders of the same experiment. Finally, randomly selecting 24 subsets from all possible combinations for spatial resolutions with many gauge stations (Table 3.2) is not the best representation of complete data heterogeneity. Although we experimented with up to 100 validation sets and observed no substantial change in model performance. Future work could explore all possible combinations of training and testing and/or vary  $k$  to learn the effect of increasing the training sample.

Machine learning has demonstrated encouraging results in global river discharge predictions and a potential to solve many of existing problems in hydrology [248, 192]. However, up until now, these results have all been based on lumped data modeling techniques, ignoring temporal-spatial variations of physical processes that drive the hydrologic cycle. We have demonstrated that integrating this knowledge into training data modeling (distributed experiment) can further improve the performance of ML models, more especially for prediction in ungauged basins. Further, we have shown that even with limited data, there is a possibility that a distributed modeling strategy could provide improved predictions (especially in ungauged basins) than any of the benchmarked models. We acknowledge that literature models from ML and hydrologic modeling represented by [150] and [81] have unique advantages that can improve our understanding of global discharge as a proxy for understanding cascading impacts of climate change on water resources and as such, leveraging distributed modeling could further improve their performance and applicability.

### **3.2.5 Conclusion**

In this work, we have demonstrated the importance of distributed data modeling in improving the performance of ML models for discharge prediction in ungauged basins. Further, we leveraged topologically guided river hierarchies as a proxy for understanding the contributions of different physical processes at varying spatial resolutions and showed that as spatial resolution increases, model performance improves in response to more fine-grained hydrologic information. Finally, we compared our distributed approach against two literature models, i.e., [150] and [81], and showed that when all models performed poorly on previously unseen data, the models trained with the distributed modeling approach demonstrated better performance. This makes our proposed approach more applicable for predicting discharge for most global river basins with limited to no data.

Our experiments and results demonstrate the importance of integrating hydrologic and geographical differences in the data modeling process, a notion that has, up until now been largely ignored when building data-driven hydrology models. With the upcoming launch of the SWOT mission that will provide more consistent and fine-grained hydrologic information on global rivers, our proposed approach can improve methods for predicting river discharge on a global scale, and as a result, our understanding of the cascading impacts of anthropogenic climate change on global water resources. However, we did not specifically discover which physical processes are dominant at varying spatial scales, but this opens up questions in future work in which we will aim at quantifying the temporal-spatial contribution of distinct features towards model performance and overall interpretability and explainability of ML models in hydrology and physical sciences in general.

## CHAPTER 4

# EXPLAINABLE MACHINE LEARNING FOR RIVER DISCHARGE PREDICTION

Machine learning has emerged as a powerful tool for predicting river discharge, but its black-box nature raises concerns about the underlying mechanisms driving these predictions. To bridge the gap between data-driven models and traditional physics-based hydrology, it is crucial to develop explainable ML approaches that can demystify the reasoning behind their predictions. Chapter 3 demonstrated the effectiveness of ML in predicting river discharge, particularly in ungauged basins. However, the black-box nature of these algorithms raises concerns among hydrologists who rely on explainable models to derive physical insights. Addressing these concerns, this chapter introduces statistical techniques to explain the internal workings of these ML models (explainable ML). By unraveling the decision-making process of these complex algorithms, this approach can foster trust in ML models and pave the way for their seamless integration into hydrological practice. The ability to statistically explain ML predictions presents an opportunity to bridge the divide between data-driven and physics-based hydrology, paving the path towards more holistic and reliable river discharge prediction models.

### 4.1 Explainable Machine Learning Models for River Discharge Prediction Using Remotely Sensed Data

#### 4.1.1 Motivation

Reliable water supply is the most vulnerable natural resource to anthropogenic climate change. Changes in water cycles due to global warming and other anthropogenic cli-

mate change factors have a wide range of environmental and socio-economic impacts: increasing sea levels, shifts in precipitation patterns, changes in groundwater recharge, salinization of freshwater sources and ocean acidification, changes in bio-eco systems, as well as impact on agriculture and clean energy generation (e.g., hydropower), all of which are a resulting of rising global temperatures[148]. As such, continuous monitoring of the cascading effects of anthropogenic climate change on water resources is critical to address the intricate interplay of environmental, political, and socio-economic challenges arising from shifting water dynamics- critical to ensuring the long-term survival of the global population, a majority of whom live in the global south. However, such a real or near-real-time monitoring system is unavailable for most regions worldwide. It is, therefore, difficult to obtain accurate data on the quantity and quality of water available for human consumption, agricultural and industrial use, and clean energy generation, among others[197, 215, 147]. This means that a major disaster, such as a hurricane, storm, or drought, can devastate an economy, including the loss of millions of lives and billions of dollars worth of property[113]. As a result, there is an urgent need for continuous monitoring of global water resources for which there are limited historical and current data.

Recent advances in machine learning (ML) such as deep learning, transfer learning, remote sensing integration, feature learning, and AutoML and neural architecture search, among others, have made it possible to transfer hydrologic information from well-mapped regions to those with little to no historical data with superior performance compared to centuries-old process-based techniques [248, 152, 79, 67]. Machine learning methods can “learn” intrinsic patterns in hydro-meteorological data and exploit this hydrologic knowledge to generate new predictions with greater precision than existing process-based models. However, ML models are still black-box, without a comprehensive knowledge of how they arrive at predictions. In other words, it is difficult to fully understand an ML model’s processes, decisions, and predictions. This has led to skepticism and slow adoption in hy-

drologic sciences: the discipline emerged from civil water infrastructure management in previous centuries and has a strong tradition of the classical physics-based “proof” from conservation of mass, momentum, and energy. Hydrology does have a strong empirical tradition (e.g., Manning’s Equation and Darcy’s Law), but hydrologic empiricism always assigns physical meaning to empirical parameters and expects any solution to be traceable, repeatable, and have a physical meaning. Further, process-based models are inflexible to novel conditions (such as novel climate change scenarios) and struggle to adapt to conditions different from those under which they were calibrated. This results in an empirical abstraction of complex and unknown physics, resulting in lower-skill predictions and forecasts as an empirical representation of a physical system. This also makes it difficult to identify dominant physical processes automatically because hydrologic responses frequently exhibit threshold behaviors due to sensitivity to varying physical phenomena happening at distinct temporal-spatial scales, leading to regions of state space (multi-dimensional space that encompasses all possible combinations) where some processes and their parameters are negligible (Ogden 2021). Consequently, physics-based models are frequently highly tailored for specific applications (e.g., specific watersheds and climatic regions). In addition, the number of model parameters increases with the number of different physical processes that are represented, making process-based models complex and computationally costly and ultimately equifinal- the empirical parameters abstracting the physics end up absorbing uncertainty in non-physical ways (Wagener and Montanari 2011; Arsenault and Brissette 2014). Because of this simplified depiction of reality due to epistemic limitations, process-based models often fail to capture non-linear interactions and emergent phenomena, especially in the wake of cascading climate change, making them neither easily transferrable to new regions with unique hydro-meteorological characteristics nor rapidly scalable in the presence of new and heterogenous data, rendering them less useful for real-time learning and insightful information sources.



Process-based models in hydrology [78, 185] are a set of mathematical equations that model natural laws that govern physical processes (e.g., mass conservation, energy, and momentum in hydrologic cycles). This is especially significant considering that dominating physical processes are specific to geographical or climatic locations [25]. As a result, the ability of process-based models to adjust their parameters in response to the changing dynamics of the prevailing physical processes makes them more useful, particularly in situations with limited to no training data. The simplicity with which process-based models represent and mimic physical processes allows users to understand how the underlying physical processes influence the model's prediction consistent with the established understanding of hydrologic processes. However, not all physical processes are known in hydrology, which makes them difficult to model at scale and in different environmental and climatic conditions [280, 10].

Fortunately, with their data-driven nature, ML models [160, 182] hold the promise to bridge these gaps through adaptive modeling capabilities and provide timelier and context-specific insights into hydrologic systems. These models assume that their input-output relationship can be explained by a set of mathematical equations, eliminating the need to understand the underlying physical processes and how they interact. Therefore, this begs the question for the hydrologic sciences: Is it possible to leverage the predictive prowess of machine learning models while maintaining the transparency of their internal working mechanisms? This would allow for higher-skill predictions of ML that match the tradition of physically based empiricism of hydrology.

In this work, we propose an explainable artificial intelligence (explainable AI) methodology to improve the interpretability of machine learning models for river discharge prediction. Our goal is to bridge the gap between traditional process-based models and the emerging domain of machine learning. Starting with traditional machine learning methods, we use linear regression to explain the global relationships between the input and output variables, focusing on how to use the model's coefficients to interpret the sen-

sitivity of the output in response to each input variable. To address the limitations of linear regression, we use a combination of cooperative game theory and ensemble learning methods to learn local relationships between inputs and output variables. Consequently, we can identify the most predominant variables across different temporal-spatial contexts. Due to the sequential nature of hydrometeorological data, we incorporate attention mechanisms [307, 198, 44] in Long-Short Term Recurrent Neural Networks (LSTMs) to learn long-term temporal dependencies in the data. As a result, our models become adept at focusing on critical segments (specific temporal points) of the input sequences, allowing us to understand the physical processes that drive the hydrologic cycle in time and space. Our proposed methodology had several advantages over process-based modeling methods. First, it is inherently data-driven, meaning we can garner insights from various data sources. Secondly, explainability emphasizes the internal workings of machine learning models by building on simple and easy-to-explain principles like model coefficients and Shapely values. Machine learning can overcome some of the limitations of process-based modeling. Thus, we believe that this shift towards explainable machine learning can improve the adaptability of machine learning models in hydrology, leading to the development of more accurate, reliable, and fair models that can help us manage and understand the impact of anthropogenic climate change on water resources.

#### **4.1.2 Data and Methods**

Building upon the foundational work presented in Chapter 3, this chapter employs the same dataset and methods outlined in Section 3.1.3.1 to train machine learning models for river discharge prediction, as a basis for explainable AI in hydrology. By applying advanced analytical techniques to uncover deeper insights into the data, this chapter provides critical evidence for the thesis arguments in the subsequent chapters.

#### 4.1.2.1 River Discharge Prediction

River discharge is the volume of water per unit of time passing through a given cross-section in a river. This is the most important measurement for human and ecosystem services for any river[97]. The most accurate method for measuring river discharge is *in situ*, which is impractical at scale. Therefore, most river discharge records are derived from continuous measurements of river elevations or stages transformed empirically to discharge via a ‘rating curve’ rather than direct discharge measurements[122]. A rating curve is a mathematical model that estimates the relationship between discharge (flow rate) and river stage (water level) by fitting a curve through a set of measurements at a single place in a river[180]. This is the most accurate method for estimating water flow in a river or stream at a specific point in time based on the water level or stage measurement. This process must be repeated for each point of interest along the river, which is expensive, time-consuming, and not feasible globally. Thus, other means of estimating discharge are needed.

In the hydrologic cycle, river discharge is an integral function (outcome) of various interactions and physical processes (e.g., precipitation evapotranspiration) that occur within a watershed (or water catchment basin). As such, these processes and interactions can be used as proxies to estimate discharge emanating from a given land area (watershed) by utilizing process-based hydrology models such as VIC, SAC-SMA, and HyMOD[161, 38, 31], which are deployed in static and operational global assessments like GRADES, GLDAS, and GloFAS[164, 235, 121]. Process models can update their internal working mechanisms (empirical parameters) in response to the underlying changing phenomenon of the forcing variables, offering insights into how different components of the hydrologic systems interact and influence each other. However, these models have varying parameters that introduce equifinality into the final solutions, degrading their physical meaning. Further, many of these parameters are not directly measurable and must be estimated. This can introduce uncertainty and the potential for overfitting: the model performs well on cal-

ibration data but fails to reproduce the same patterns on new data. Finally, they do not easily transfer hydrologic information across watersheds, making them less applicable for prediction in ungauged basins (PUB) with limited to no historical data.

#### **4.1.2.2 Machine Learning in Hydrology**

Machine learning has been applied to various problems in hydrology for several decades. Early ML applications in hydrology (e.g., [112, 141, 48, 58]) focused on developing models for predicting various hydrological variables such as streamflow. ML algorithms, particularly Long-Short Term Memory neural networks [152, 79, 294], have demonstrated better discharge prediction performance than process-based models in recent years. This is because neural networks are universal approximators that can simulate linear and non-linear systems without any major underlying assumptions, making them better suited for solving huge and complicated issues with sufficient and insufficient data[32, 285].

However, ML models assume a direct causal relationship between hydrological processes and discharge, which is not necessarily correct. This is because numerous factors, including geographical and climatic characteristics of a given watershed (water catchment basin), contribute to a high degree of temporal and spatial variation in hydrologic processes. In other words, dominant physical processes differ by region, which might bring uncertainty into the assessment of model parameters. Additionally, hydrologic processes are governed by physical laws, which are challenging to integrate into ML models unless explicitly designed to do so (e.g., physics-driven ML models). Without incorporating these constraints, ML models can produce physically inconsistent or unrealistic predictions. Further, machine learning models are computationally costly and require voluminous training data, which does not exist for most of the world's basins or is expensive and difficult to collect over a long period of time across different geographic regions. Finally, unlike process-based models, ML models produce black-box predictions that are difficult to explain or comprehend. The lack of transparency(explainability) is a massive concern in

hydrology, where understanding the underlying processes is extremely important. These drawbacks reduce their utility for modeling physics-driven processes, where the relationships between underlying factors need to be interpretable to improve global acceptance and comprehension.

#### **4.1.2.3 Explainable Machine Learning**

In machine learning, explainability (explainable AI) is the extent to which machine learning models' internal working mechanisms (decision-making processes) can be made interpretable, transparent, and understandable to non-domain experts. Hydrologic systems are inherently complex, with multiple interacting processes and non-linear relationships. Although ML models can capture these complex interactions, understanding how they do so is very important for experts and policymakers to trust, validate, and use the predictions and recommendations provided by these models. Additionally, hydrologic systems are often region-specific, varying widely across regions and scales. Thus, by understanding how ML models work in one context, hydrologists can better understand their suitability and adaptability to other contexts. Further, hydrologic observations and predictions often come with uncertainty due to measurement errors, incomplete data, or inherent variability. Hydrologists can understand, quantify, and address these uncertainties using explainable AI when ML models are used for discharge prediction. Finally, critical decisions like flood prediction, drought monitoring, and water resource management have significant socio-economic and political(hydro-politics) implications. As such, explainable and transparent ML models can give policymakers and decision-makers the confidence to act on AI-driven insights and recommendations. While there have been attempts to explain the performance of LSTMs (e.g., [149]), explainability in hydrologic ML is still largely unexplored. Model explainability can be achieved through various methods, such as counterfactual explanations, feature importance, Local Interpretable Model-agnostic Explanations (LIME), Shapley Additive explanations (SHAP), Attention Mechanisms for

sequential models such as LSTMs and transformers, feature visualization and activation maximization, and probing influential training data, among others. [26, 107].

In this work, we use model-specific and model-agnostic methods to improve explainability of ML models [114, 157, 158, 170]. Model-specific methods are designed for specific types of ML models (usually neural networks) and require access to the internal details or architecture of the model. As such, they can often provide deeper insights into the overall functioning of these models at a deeper scale. On the other hand, model-agnostic methods apply to any type of model, ranging from traditional ML models such as linear regression to the latest state-of-the-art models like time-series transformers. Model-agnostic methods don't rely on the internal works of the model but focus on the model's inputs and outputs. As such, they are mostly useful in scenarios where the model's internal details may be unknown or too complex to analyze directly. Although model-agnostic methods are more broadly applicable, they do not provide as deep or more detailed insights into the model's internal workings as model-specific methods might for that particular model. As such, exploring the applicability of both explainability methods provides a wide set of options that hydrologists can choose from, depending on the task.

Therefore, As ML becomes more integrated into data-driven decision-making in hydrologic sciences, it is of utmost importance to have explainable and interpretable models that can justify their predictions simply and concisely.

#### **4.1.2.4 Equitable Machine Learning in Hydrology**

Equity in machine learning refers to the just and fair distribution of benefits, risks, and responsibilities associated with developing and deploying ML models, thereby ensuring that no group is unfairly disadvantaged or marginalized. In the context of hydrology, equitable ML refers to the development of ML models in a manner that ensures fairness, inclusivity, and justice across different regions and time periods while recognizing the diverse hydrologic challenges faced by different geographic and climatic regions and seeking to ensure

that these challenges are addressed without bias. In physical science and hydrology, fairness is an important issue that needs to be addressed, as it can affect the accuracy and reliability of decisions and recommendations made by these ML models. Understanding the extent to which models are biased or affected by external factors is essential to ensure equitable outcomes are realized. This work focuses on temporal and spatial equity concerning ML models for predicting river discharge.

Temporal equity emphasizes consistent performance (prediction accuracies) of ML models across different time frames (e.g., seasons and years). Since hydrometeorological data is inherently time-series, models trained on this data should be robust to temporal variables. Over time, river systems change due to climate variations, anthropogenic impacts, and land use changes. As such, an ML model that offers accurate predictions for a certain timeframe might have lower accuracies in the next if it doesn't account for these temporal dynamics. Conversely, spatial equity emphasizes that if data is collected from multiple locations, it is essential to ensure that the model trained on this data performs equally well (or within acceptable margins) across all these locations. Historically, data-rich regions (e.g., the conterminous United States), which have robust infrastructure, have dominated the datasets used for model training (e.g., CAMELS dataset). This imbalance in data diversity can result in models that produce biased predictions when validated in regions with limited data (predictions in ungauged basins), usually regions in the global south. ML techniques such as transfer learning, where models are trained on one region and finetuned for another region, can bridge data gaps and ensure region-specific accuracy. Additionally, tailoring models to local conditions using a hybrid of ML and physics-based models (physics-driven ML) can enhance regional relevance, ensuring that models perform well across diverse geographic regions.

Implications of neglecting temporal and spatial equity can be significant. If left unchecked, temporal-spatial inconsistent predictions can lead to poor long-term water management strategies if the models fail to recognize emerging temporal patterns, resulting in inad-

equate preparedness for extreme events like floods and droughts, serious environmental and socio-economic consequences, including increased water pollution, damage to infrastructure, decline in agricultural productivity and cleaning energy generation as well as loss of lives.

### 4.1.3 Experimental design

To demonstrate the internal decision-making process of ML models and how they arrive at given predictions, we train a series of ML models with increasing complexity: Linear regression, random forests, and Long-Short Term Memory (LSTM) recurrent neural networks with an attention mechanism [262, 245, 124]. This section focuses on making the outcomes of these models more understandable to non-domain experts by breaking down the “black box” nature of these models, thereby making them transparent and their predictions more interpretable. Linear regression is a traditional ML model that assumes a linear relationship between the inputs (independent variables) and the outputs (dependent variables). This relationship between the inputs and outputs can be represented by equation (4.1), for multiple linear regression (multiple independent variables).

$$y = \beta_0 + \beta_1x_1 + \beta_2x_2 + \dots + \beta_nx_n + \varepsilon \quad (4.1)$$

where  $y$  is the dependent variable;  $x_1, x_2, \dots, x_n$  are the independent variables;  $\beta_0$  is the y-intercept (constant term);  $\beta_1, \beta_2, \dots, \beta_n$  are the coefficients of the independent variables;  $\varepsilon$  represents the error term.

Linear regression operates on several assumptions, which makes it one of the simplest and most interpretable ML models. These assumptions include linearity, The relationship between the independent and dependent variables is linear; Independence: The observations (model instances) are independent of each other; homoscedasticity, The variance of errors is consistent across independent variables, and the absence of multicollinearity, i.e., the independent variables are not highly correlated with each other.



Equation 4.1 is a straightforward representation of how each feature affects the prediction  $y$ . The coefficients of the equations (Beta) indicate the strength and direction of the relationship between the independent variables  $x_1, x_2, \dots, x_n$  and the dependent variable  $y$ . A positive coefficient indicates a direct relationship: as the independent variable increases, the dependent variable increases, while a negative coefficient indicates an inverse relationship. The magnitude of the coefficients indicates the importance of the corresponding features: large values suggest a stronger impact on the dependent variable (predicted value). To explain the linear regression, randomly select a subset ( $k = 4$ ) of stations from the entire dataset ( $n = 89$ ) and concatenate them into a single dataset. We then randomly shuffle the data (since each observation is independent of the other) and split it into train and test sets in the ratio of 80% to 20%, respectively. We repeat this procedure several times (k-fold cross-validation) to ensure consistency in our results and report the average of the coefficients across these experiments. Finally, we examine the direction and magnitude of the model's coefficients, indicative of the interaction between the inputs and outputs.

However, hydrological processes are often complex, non-linear, and interdependent, making it difficult to model them based on linear regression assumptions accurately. This is because hydrometeorological interactions involve many variables, such as the effects of climate change, land use, and population dynamics. These factors can all have a significant impact on the dynamics of river discharge. As a result, more complex models are required to capture the interactions between hydrological processes accurately. Thus, while linear regression is inherently interpretable, its assumptions of linearity, independence, and homoscedasticity can limit its real-world applicability, potentially leading to oversimplified or misleading explanations.

Random Forests (RF) is an ensemble algorithm that combines multiple decision trees to improve prediction. This enables the model to capture complex and non-linear relationships between the input(independent) and output(dependent) variables, resulting in higher ac-

curacy and, subsequently, more trustworthiness in the predictions the model provides. RF is especially useful in applications where the data is high-dimensional and the relationships between the variables are complex. Although RF models are powerful predictors, they are considered black-box models, especially considering the many trees (decision steps) involved in arriving at the correct decision. Therefore, interpreting RF models can improve both the model’s robustness and build trust among non-technical users and stakeholders. In this work, we use SHAP (Shapley Additive exPlanations) to provide a unified feature importance measure by distributing the prediction value across the input variables (features) by leveraging the SHAP framework.

$$\phi_j = \sum_{S \subseteq N \setminus \{j\}} \frac{|S|! \cdot (|N| - |S| - 1)!}{|N|!} (f(S \cup \{j\}) - f(S)) \quad (4.2)$$

where  $N$  is the set of all features;  $S$  is a subset of  $N$  that does not include feature  $j$ ;  $f(S)$  is the prediction of the model for the instance with only the features in  $S$  active.

This unified model-agnostic Python framework uses cooperative game theory [279, 208, 202] to calculate each feature’s marginal contribution (Shapley value) toward overall model accuracy. The Shapley method is a statistical method for allocating payouts to players based on their marginal contribution to the overall payout. Shapley values conform to the natural axiomatic attributes of a fair allocation: efficiency, anonymity, proportionality, and dumminess.

Regarding explainability, the SHAP framework considers each predictive task as a single game, input features as players, and the model’s prediction as the overall payout. Thus, the Shapley value of a given input feature can be defined as its contribution to the total payout (prediction), weighted and summed over all possible combinations (coalitions). However, the computational cost of calculating Shapley values across all possible feature combinations grows exponentially with the number of features. Tree SHAP is a faster and more efficient method for estimating SHAP values for tree-based models, which reduces computational demands without sacrificing accuracy.

Although Random Forest models are powerful predictors, they are not naturally designed to handle the inherent temporal dependencies in time series data like river discharge. This limitation makes them less adept at capturing the sequential order and relationships between observations over time. Additionally, their use requires extensive feature engineering to address seasonality and non-stationarity. Finally, the model's complexity can lead to computational challenges and a risk of overfitting specific timeframes or anomalies.

Long Short Term Recurrent Neural Networks (LSTMs) are a type of artificial neural network (ANN) originally proposed by Hochreiter & Schmidhuber (1997), capable of processing sequential data. LSTMs have been successfully applied to language modeling, video understanding, music transcription, discharge prediction for hydrology, and other applications. Unlike standard neural networks that only understand the spatial context of data, LSTMs can extract both the temporal and spatial context encoded in the training data. At a structural level, an LSTM network consists of a series of identical recurrent neural networks where the previous neural network  $t_{i-1}$  passes information to the current network  $t_i$ . This cascading architecture allows LSTMs to handle the sequential context encoded in historical data, e.g., hydrologic data. Unlike traditional RNNs, LSTMs can maintain information in memory over long periods, thereby overcoming the problem of vanishing gradients. This allows LSTMs to learn long-term temporal dependencies, i.e., where the desired output depends on inputs presented at times far in the past (lookback window), which is important when modeling physical processes that occur at different spatial resolutions. Consequently, the lookback window size determines how much information a model learns about a particular physical process.

In this work, we use a bi-directional LSTM with four layers and an attention mechanism. Bidirectional LSTMs concatenate two unidirectional LSTMs in opposite directions such that the model learns encoded features in both time-increasing (forward chain) and time-decreasing (backward chain) manners. This is especially important where information encoded at the next timestep( $t_{i+1}$ ) can improve prediction at the current time ( $t_i$ ). Knowl-

edge of river discharge at the next timestep can improve our prediction at the current timestep, and as such, we adopted the bidirectional LSTM network model architecture for our experiments. The attention mechanism is a computation technique that enables the network to selectively focus on segments of the input sequence when generating a prediction. For each output time step  $t$ , it computes scores for every input time step  $t'$  based on the model's hidden states,  $\text{score}(h_{t'}, h_t)$ . These scores are then transformed into attention weights using a softmax function,  $\alpha_{t,t'} = \frac{\exp(\text{score}(h_{t'}, h_t))}{\sum_{t''} \exp(\text{score}(h_{t''}, h_t))}$ . The weighted sum of the input hidden states, based on these attention weights, produces a context vector  $c_t = \sum_{t'} \alpha_{t,t'} h_{t'}$  for the decoder. This context vector, combined with the current hidden state of the decoder, determines the final output for that time step, allowing the LSTM to focus adaptively on different parts of the input for each output. By allowing the model to focus selectively on significant past events or patterns, attention mechanisms enhance the LSTM's ability to capture long-term dependencies and offer increased interpretability. Hydrologists can then visualize which past timesteps in each sequence were deemed most influential by the model for a given forecast.

#### 4.1.4 Evaluation Metrics

To compare explainability across different models, we use three different metrics: coefficients of linear regression, feature importance scores, and shapely values. Linear regression coefficients represent the change in the dependent variable for a one-unit change in the independent variable, *ceteris paribus*. On the other hand, feature importance is represented as a relative value between 0 and 1, indicating the proportion of predictive power attributed to each independent variable. Finally, Shapely values provide a contribution score for each feature (independent variable), indicating its marginal contribution to the model's prediction for a specific instance relative to the average prediction for all dataset instances. Further, we report prediction results based on four major metrics used to evaluate discharge prediction models: Kling-Gupta Efficiency (KGE) (Gupta et al., 2009),

Nash-Sutcliffe Efficiency (NSE)[191], Relative Bias, and Normalized Root Mean Squared Error (NRMSE). These standard hydrology metrics assess different aspects of the hydrograph and errors in both timing and volume of water[164, 111].

#### 4.1.5 Results

In this section, we present the results of our comparison of three data-driven (ML) models for predicting river discharge: linear regression, random forests, and long-short-term memory networks. We evaluated the models using standard hydrology performance metrics: NSE, KGE, and RBIAS, and explainable ML techniques: regression coefficients, feature importance, Shapley value, and attention mechanisms. Overall, our results suggest that linear regression, random forests, and LSTMS can all be utilized to improve the explainability of ML models in hydrology and the physical sciences in general.

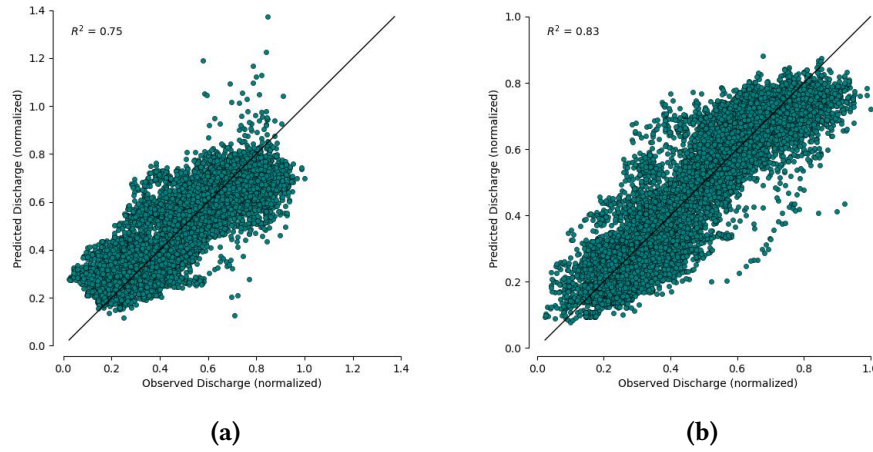
Table 4.1 shows the results of three ML models evaluated for river discharge prediction: linear regression, random forest, and LSTM. Linear regression, a simple and interpretable model, achieved an NSE of 0.78 and KGE of 0.77, suggesting a well-balanced representation of the observed river discharge dynamics. Random forest, a more complex model, achieved the same NSE of 0.78 but a slightly superior KGE of 0.86, indicating its potential to capture intricate statistical characteristics of the observed data accurately. Both models exhibited a near-negligible underestimation bias. On the other hand, the LSTM model achieved an NSE of 0.68 and a KGE of 0.8, suggesting a satisfactory fit. Although random forest and linear regression seem to outperform LSTMs in terms of mean metrics, it is important to understand the performance of the respective models beyond mean metrics. For example, it is important to consider the models' ability to capture extreme events and their performance on different sub-basins of a river network.

Ultimately, the choice of explainability technique depends on different factors, including the complexity of the data and the desired level of explainability. The remaining part of

**Table 4.1:** Mean of KGE, NSE and RBias across the three models: Linear regression, random forests and LSTMs

Model Name	NSE	KGE	Rbias
Linear Regression	0.78	0.77	-0.008
Random Forests	0.78	0.86	-0.008
LSTM	0.68	0.80	0.050

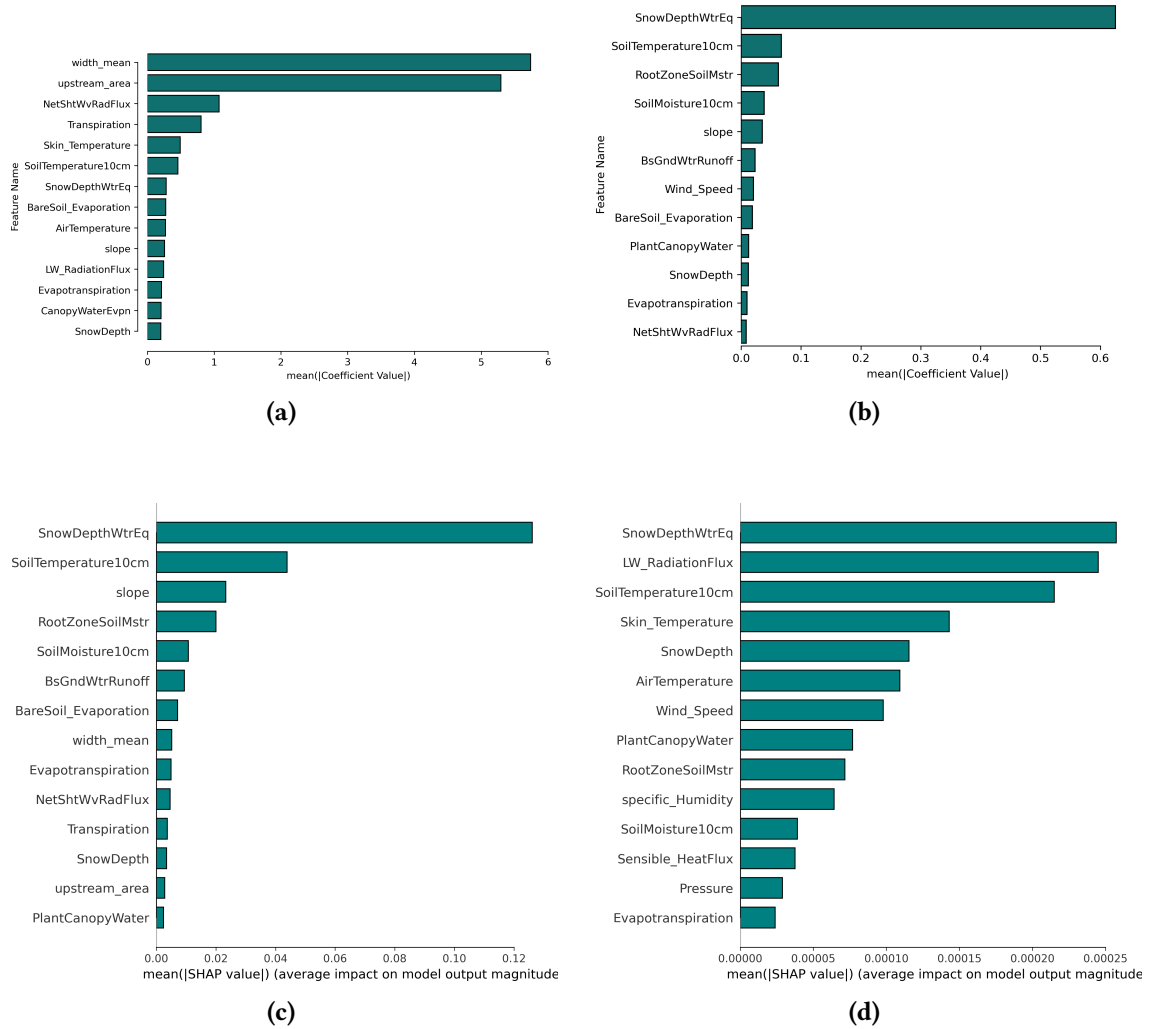
this section presents a detailed description of our findings, highlighting the comparison of performance and potential shortcomings of each method.



**Figure 4.1:** predicted discharge vs. observed discharge curves for Linear regression and Random Forest models. LR yields an  $R^2 = 0.75$  while RF yields an  $R^2 = 0.83$ . This shows that the RF model fits the data more accurately than the LR model

First, we compare predicted discharge vs. observed discharge curves for Linear regression and Random Forest models (Figure 4.1). As a reminder, we randomly shuffle training and testing data ( $k=4$ ). This is because both LR and RF models assume that individual observations (data points) are independent of each other, and as such, unlike LSTMs, there is no need to maintain temporal dependence in the dataset. The LR model (Figure 4.1a) yields an  $R^2$  of 0.75, indicating that the model can explain 75% of the observed river discharge variability.  $R^2$ (coefficient of determination) is the proportion of the variation in the dependent(output) variable explained by the independent(input) variables.  $R^2$  can take on

any value between 0 and 1, with a higher value indicating a better fit of the model: an  $R^2$  of 1 indicates that the model perfectly fits the data, while a value close to 0 indicates that the model fails to fit the data. On the other hand, the random forest model (Figure 4.1b) shows a higher  $R^2$  value of 0.83, indicating that the RF model can explain 83% of the variability in the observed data on average.



**Figure 4.2:** Comparative feature importance across three models: (a) Coefficients from a multi-linear regression model, (b) Feature Importance from the Random Forest model, (c) SHAP summary plot from the Random Forest model, and (d) SHAP summary plot from the LSTM model. Here, we display the top 14 most important features

Next, we compare the influence of the input (independent) variables on the model prediction as captured by the two models across all dataset instances (Figure 4.2). The features are ranked in their order of influence, starting with the most influential ones. This is also known as global explainability since the influence is represented as an average across the entire dataset. Global explainability allows us to identify which input variables have the largest influence on model predictions. The length of each bar represents the magnitude (importance) of the corresponding feature. Linear regression coefficients can be negative or positive, indicating the direction of the relationship between the feature and the target variable. A positive coefficient value means that as the predicted value increases, the input feature also increases. Conversely, a negative coefficient means that the input feature value increases as the predicted variable decreases. However, to compare feature importance across the three models, we opt to represent the regression coefficients as absolute values, thereby focusing on the magnitude of the influence rather than the direction.

Based on this, the LR model (Figure 4.2a ) ranks the mean width of the river reach as the most important feature, followed by upstream basin area and net short-wave radiation flux, while traditional hydrometeorological processes like precipitation and soil moisture are ranked least important. However, it should be noted that LR assumes a linear relationship between the input and output variables and, as such, may fail to capture complex non-linear relationships that are inherent in hydrologic processes. The RF model (Figure 4.2b) ranks the Snow depth water equivalent as the most influential feature, followed by soil temperature at the depth of 10cm and Root Zone soil moisture. Soil depth water equivalent and root zone soil moisture are the amount of water stored in the snowpack and soil layer where plant roots extract water, respectively. In the Mackenzie basin, we expect both factors to have a major influence on river discharge. The random forest model captures this influence better than the Linear regression model. This demonstrates the RF model's ability to capture non-linear interactions between features and suggests that

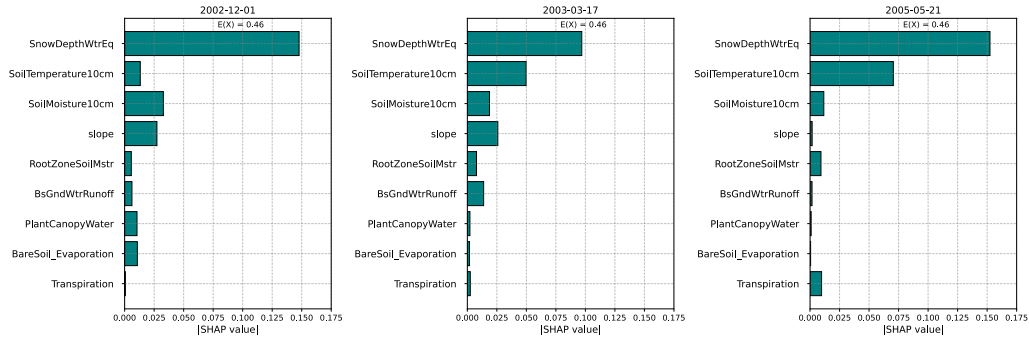


the influence of a given input feature on the prediction may be multi-faceted, potentially interacting with other features in ways that a linear model fails to capture.

Figure 4.2c shows the SHAP values of the features as represented by the RF model. SHAP values provide a more granular value decomposition of all predictions in the dataset. SHAP values, which are an extension of Shapley values, achieve this by attributing portions of the interaction to each feature while simultaneously considering all interactions. Given that there is high consensus between feature importance and SHAP values of the random forest model, i.e, the top 6 most important features (snow depth water equivalent, soil temperature at 10cm, root zone soil moisture, slope, and baseflow ground water runoff), it can be said that RF inherently captures non-linear and complex interactions between the features better than regression.

The SHAP values calculated using an LSTM model ( Figure 4.2c) show the order of importance of the input features on river discharge prediction. The LSTM model considers snow depth water equivalent, longwave radiation flux, soil and skin temperatures, and snow depth as the most influential features. Interestingly, the model captures both snow depth and snow depth water equivalent, which are critical drivers of river discharge in regions like the Mackenzie basin, which are covered by snow for most of the year. Despite having lower mean metrics (KGE and NSE) than the linear regression (LR) and random forest (RF) models, we observe that the LSTM model can capture the complex non-linear interactions between features and the most influential features that impact river discharge better than the LR and RF models.

Figure 4.2 - global explainability provides insights into the overall behavior of the RF model in predicting river discharge as an average of individual data instances across the entire dataset. However, this representation can limit the ability to explain why a model makes specific predictions at specific periods in specific climatic regions. On the other hand, local explainability (Figure 4.3) focuses on the model's prediction for specific in-



**Figure 4.3:** Three plots showing the impact of the same features on the model’s prediction across different instances (different dates/seasons of the year): December is typically winter in the Mackenzie basin, mid-March is early spring while mid-May is early summer

stances (days). This makes local explainability ideal for understanding why the model makes specific predictions at specific time periods, river basins, or geographic regions.

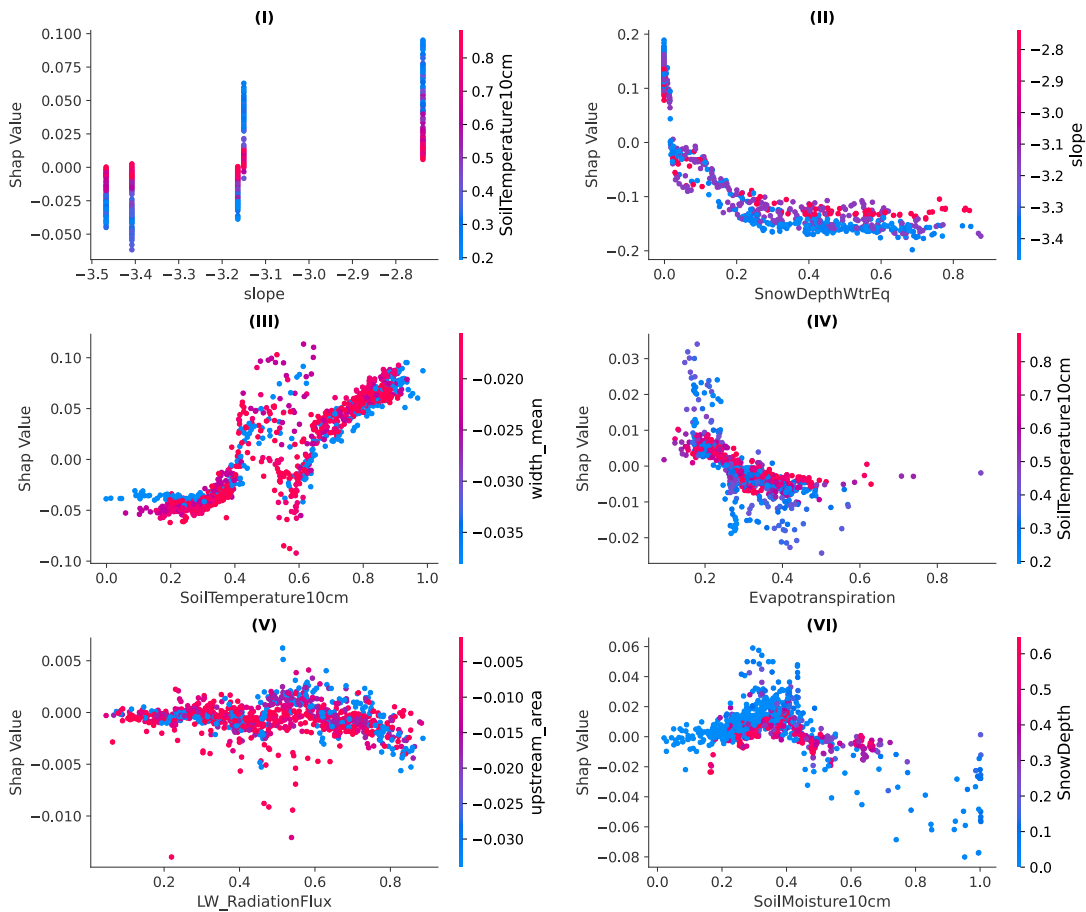
Starting with the base value,  $E(X)$ , which denotes the average discharge prediction across all instances, each feature’s contribution is highlighted with corresponding SHAP values.

To compare the impact of each feature across different seasons, we represent the SHAP values as magnitude, thereby focusing on how much impact each feature has on model prediction rather than the direction of the impact, i.e., negative or positive impact. The magnitude and order of features, represented by the length and position of bars, respectively, help prioritize the most influential factors. Finally, by summing all these SHAP values with the base value, we can compute the model’s precise prediction for that instance, thereby ensuring a comprehensive and transparent interpretation.

From Figure 4.3, we observe that Snow depth water equivalent (SWE) consistently exerts the most significant influence on the model’s predictions throughout the three seasons.

This influence is particularly pronounced during winter and summer. As the seasons transition from winter to summer, the model’s sensitivity to temperature, slope, root zone soil moisture, and transpiration increases. Conversely, parameters like soil moisture at a depth of 10 cm, bare soil evaporation, and canopy water evaporation (PlantCanopyWater) display an inverse relationship to the model’s performance. The effects of these features are most pronounced during winter and less pronounced during summer. Overall, we

observe the multi-facet and complex interactions between the physical processes, resulting in the interpretation we observe in Figure 4.3: during the summer, evaporation, and transpiration are higher, resulting in less soil moisture, bare soil evaporation, and canopy water evaporation. As such, these parameters become less important in the model’s performance during the summer.



**Figure 4.4:** Dependence plots showing the relationship between a single feature (x-axis) and the SHAP values (or model output) for that feature (y-axis). The coloration is based on a second “interaction” feature, which captures interaction effects, i.e., how the primary feature’s impact changes with varying values of another feature

Figure 4.4 shows dependence plots for six features in the dataset. Dependence plots visually represent how a specific feature affects a model’s predictions by plotting feature values against SHAP values. Additionally, they explain both directly and indirectly influenced effects. Direct effects are observed through the general trend of the plots: as the

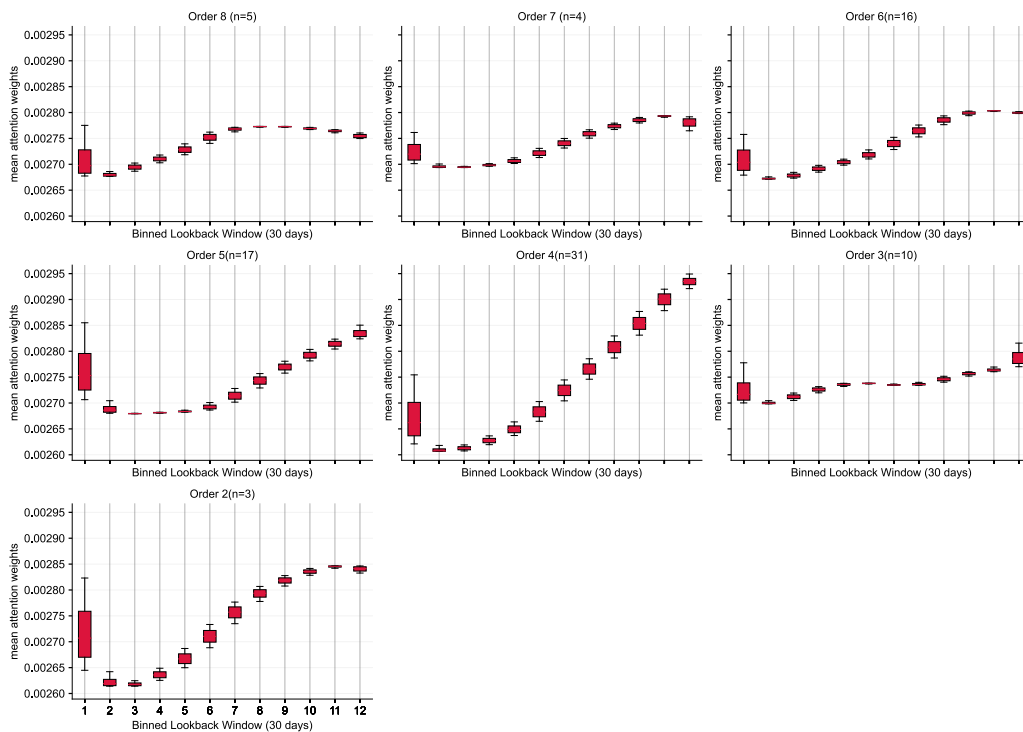
feature value changes, how does its impact (SHAP) value on prediction change? On the other hand, interaction effects are highlighted in two ways: by coloring data points based on the value of the second (interacting) feature, which shows how the primary feature's influence varies with another feature, and by observing the vertical dispersion of SHAP values at a single feature value. The interpretation is that if data points for a specific feature value are dispersed vertically, it suggests that other features also play a significant role in the prediction, indicating significant interactions.

Figure 4.4 (I) shows the dependence plot for the slope (river slope) and its interaction with soil temperature at a depth of 10cm. Five clearly defined vertical segments represent different segments of piecewise linear relationships. Each segment represents a different range of river slope values. Within each segment, the spread of values is relatively binary, implying that slope has relatively little impact on prediction across segments. The color of dots in each of the clusters suggests that for data instances in which the river slope segment has low values (most likely in higher orders: orders 6, 7, and 8 in the Mackenzie basin), the soil temperature is more significant. The model's prediction is least affected by soil temperature in regions with steep slopes.

Figure 4.4 (II) shows the dependence plot for snow depth water equivalent (SWE) and its interaction with the river slope bed slope. We find that the higher the snow water equivalent (SWE), the lower the impact of SWE on the model's prediction. This means that the more water stored in the snowpack, the less water runs into rivers. Additionally, we find that a combination of high SWE and low slope has a lower impact on the model's performance. This means the river discharge is low in low-lying areas with high SWE. We observe the same relationship in Figure 4.4 (iv), which demonstrates the impact of evapotranspiration on the model's prediction and its interaction with the soil temperature. We observe that, on average, as evapotranspiration increases, its impact on the model's prediction reduces. Furthermore, moderate soil temperature and relatively low evapotranspiration positively impact the model's prediction.

Figure 4.4 (III) shows the dependence plot for the soil temperature at a depth of 10cm and its interaction with the river width. We observe that on average, there is a positive relationship between soil temperature and its impact on the model’s performance. In other words, as soil temperature increases, its impact on the model’s performance also increases. Further, we observe that, on average, wider rivers and low soil temperatures have a negative impact on the model’s prediction while wider rivers and high temperatures have a positive impact on the model’s prediction.

In Figures 4.4 (IV), (V), and (VI), evapotranspiration, longwave radiation flux, and soil moisture all seem to have a constant effect on the model’s prediction.



**Figure 4.5:** Distributions of attention weights across LSTM models per river order (based on the Strahler River order system. We ignore order 1 status due to limited data ( $n = 1$ ))

Attention weights visualize which part of the time-series sequence (e.g., first, or last  $k$  timesteps) the model is paying attention to. Figure 4.5 illustrates the average attention weights for orders 2 through 8 of a Strahler River order system. Streams of order 1 are small streams without tributaries, usually found in high mountains. When two streams

of the same order meet, they form the next higher stream. Therefore, rivers in order 1 are the headwaters (sources), while rivers in order 8 drain their water into large water bodies, such as the ocean.

Accordingly, we observe that attention weights for higher orders (orders 8 and 7) are evenly distributed across a wide range of time lags. This indicates the effect of multiple hydrologic processes over short (immediate effects like rainfall) and long (snowmelt in high mountains) time spans. It is possible that the distinct peak in interest at the beginning of the study is due to seasonality effects, which may be a reflection of snowmelt in the high mountains. The relative decline in attention early in the input sequence (roughly between the first and third months) suggests a buffered response to precipitation, typical of large catchment areas, where runoff processes are more distributed over time.

We observe a noticeable shift in average attention weights towards more immediate lags in intermediate orders (6 and 5). This indicates that rivers in these order systems are responding more rapidly to recent precipitation events. The non-negligible weights across a broad range of time lags suggest the influences of past hydrologic events, which indicates that rivers in these older systems often retain the memory of past rainfall, perhaps a result of processes such as delayed surface runoff, soil moisture dynamics, and slower groundwater contributions.

We observe an overwhelming focus on the most recent lags in the lower orders (orders 4, 3, and 2), indicating that they respond rapidly to local hydrologic processes such as precipitation. The attention mechanism's focus on immediate time lags indicates that these rivers experience a more direct and immediate hydrologic response with minimal influence from upstream processes. This is due to the fact that rivers in these orders are characterized by small catchments and reduced buffering capacities.

#### 4.1.6 Discussion

Our results suggest that explainable AI methods can provide valuable insights into the “black box” nature of machine learning models for predicting river discharge. By comparing the plots of predicted vs. actual discharge curves of LR and RF models, we observe that RFs are more adept at capturing the complex interplay of hydrometeorological factors that drive the hydrologic cycle. Further, we observed that LR coefficients are more useful for explaining simple linear relationships between input features and output predictions. However, they fail to capture the potentially complex interactions between inputs and outputs. RF feature importance and Shapley values provided complementary information about the relative importance of different features for model predictions. LSTMs, in general, are more difficult to explain. However, we were able to understand how LSTM models were learning temporal dependencies in the dataset using a combination of Shapley values and attention mechanisms.

The predicted vs. observed plot for the LR model (Figure 4.1a) exhibits a reasonably good fit, with most data points clustered around the 1:1 line. The 1:1 line represents the perfect fit between the predicted and observed values of the dependent variables. The closer the data points (pair of observed vs. actual river discharge for each day) are to the 1:1 line, the better the predictions, and the further the spread of points away from the 1:1 line, the more discrepancies in the data. However, we observe deviations, especially for high discharge values. These deviations could be attributed to various hydrometeorological phenomena inherent in the data, which the linear model fails to capture. On the other hand, the predicted vs. observed plots for the RF model (figure 4.1b) display a tighter fit, with points densely clustered around the 1:1 line across diverse pairs of discharge values. Such improved performance underscores the ability of RFs to capture intricate, complex, and non-linear relationships in hydrometeorological data. Such interactions may include precipitation, evapotranspiration, soil moisture, and other river discharge variables.

the linear regression coefficient i.e, Figure 4.2a underscore the importance of the mean width of the river reach and the size of the upstream catchment area by ranking them as the most influential features towards river discharge. In both the continuity equation for open channel ( $Q = A \times V$ ) flow and the manning equation for discharge in open channel ( $Q = \frac{1}{n}AR^{\frac{2}{3}}S^{\frac{1}{2}}$ ),  $A$  is cross-sectional area, which is a product of the river width and slope. As such, when river width is combined with flow velocity, it has a direct impact on river discharge. Likewise, The size of the upstream catchment area has a significant impact on river discharge. In general, larger catchment areas produce higher river discharges than smaller ones. This is because larger catchment areas have more surface area to collect precipitation and more time for rainfall to runoff into the river system. However, traditional hydrologic processes like precipitation and evapotranspiration are considered less relevant, although this might not be the case. We attribute this to the fact that the LR model fails to understand the non-linear interaction between the discharge and these processes, deeming them less critical.

The RF model, i.e., 4.2b can capture non-linearities and interactions between features. The model considers Snow Depth water equivalent, soil temperature at the depth of 10cm, and root zone soil moisture as the most influential features towards discharge prediction. Snow Depth Water Equivalent represents the water content within snow. This water contributes directly to river runoff, especially if the ground is frozen or saturated. The temperature of the soil at a depth of 10 cm influences river discharge by affecting processes like freezing/thawing (which prevents infiltration) and evaporation/transpiration rates. Finally, Root Zone Soil Moisture determines how much rainfall or snowmelt is absorbed by the soil versus how much runs off into rivers: saturated soils lead to higher runoff, while unsaturated soils may allow water to percolate down, replenishing groundwater that slowly feeds into waterways. Together, these factors shape river discharge patterns. SHAP values: figures 4.2c and 4.2d for both RF and LSTM models concur with the RF model feature importance, reiterating the dominant influence of SWE and soil tempera-



ture on hydrological outcomes. This consistency amplifies the role of snow depth water equivalent (SWE) and soil temperature at a depth of 10cm in shaping water movement and distribution within the hydrometeorological system. These effects on water movement and distribution are further magnified by the fact that both SWE and temperature can evaporation and evapotranspiration, which can have an even more profound impact on water movement. Indeed, we observe that slope, specific humidity, and air temperature influence model prediction most, suggesting their consistent influence across individual predictions. In the hydrological cycle, these two factors can dictate moisture available phase changes, affecting various processes ranging from cloud formation to evapotranspiration.

Global Explainability (Figure 4.2) can provide insights into the overall importance of different features in predicting river discharge. Through this, we can identify important features common to all predictions. However, this holistic view provided by global explainability does not provide information on how models make decisions for individual data points. This can be a limitation in hydrology (and physical sciences in general), where river discharge is influenced by hydrometeorological processes that are dependent on a specific season (e.g., winter/summer vs. dry/wet season), climatic region, geographic region, or a combination of all three. Local explainability can address this limitation by providing insights into how the model makes predictions for individual data points across different temporal-spatial conditions. This can be useful for understanding why the model predicted a specific river discharge value and identifying the most influential features in making that decision.

Thus, from figure 4.3, we observe that Snow depth water equivalent (SWE) is the most important feature for predicting river discharge in the Mackenzie basin across all three seasons. However, its importance is highest in winter (and summer- reverse influence). The Mackenzie basin is a large, cold region with a long winter season. As a result, snowmelt is the primary source of runoff. Thus, SWE is most important in winter when snowmelt is

the only source of runoff. In spring, SWE is still important, but its importance decreases as other sources of runoff, such as rainfall, become more critical. In summer, SWE is still important: it negatively impacts the model's performance, as evapotranspiration becomes the dominant loss term.

The influence of soil temperature increases from winter to summer. This is because temperature is a major driver of snowmelt and evapotranspiration. In winter, temperature is less important because snowmelt rates are low. However, snowmelt rates increase as temperatures increase in spring and summer, and temperature becomes a more important factor in predicting river discharge.

Slope is most important in winter, while root zone soil moisture and transpiration are most important in spring and summer. Slope affects the rate of runoff by influencing the flow velocity of water. In winter, when the ground is frozen, and infiltration rates are low, slope significantly impacts runoff. In spring and summer, when the ground is thawed, and infiltration rates are higher, slope has a lesser impact on runoff. Root zone soil moisture affects the availability of water for plant transpiration. In spring, when plants are beginning to grow, the moisture in the root zone is more important because it limits the amount of water available for transpiration. Additionally, transpiration, the process by which plants release water into the atmosphere, is a major loss term in the summer months. In winter and spring, transpiration rates are low, so transpiration is a less important factor in predicting river discharge ads compared to summer.

Finally, Soil moisture at a depth of 10 cm, bare soil evaporation, and canopy water evaporation are most influential in winter, decreasing their influence from winter to summer.

In summary, local explainability provides insights into how input features for each data instance (day) influence discharge under specific conditions (e.g., winter, summer, anthropogenic climate change). As a result, local explainability can be used to identify hydrometeorological anomalies, such as flash floods, extreme weather events, such as landslides, and areas and times of potential drought. Greenland's sudden ice melt in 2019 (July 30

- Aug 3) is an example of such an anomaly where ice melting occurred over five days across 90% of the continent's surface. Finally, Local explainability can help incorporate local knowledge to improve river discharge prediction: Many local communities deeply understand river behavior through traditions or experiments. This knowledge can be used to align model predictions with local experience - ensuring scientifically rigorous predictions match the local reality.

Local explainability methods (Figure 4.2) are useful for understanding how an ML model makes predictions for individual data points. However, they do not show how the interaction of different features affects the model's prediction. Dependence plots can visualize these relationships and interactions between individual input features while marginalizing the other input features. In river discharge predictions, dependence plots can be used to understand how the model's predictions change in response to changes in specific input features such as precipitation, air temperature, and specific humidity, among others. Furthermore, dependence plots can identify non-linear relationships between inputs and the model's output. For example, a dependence plot might show that the model's prediction of river discharge increases rapidly with an increase in air temperature up to a certain point and then decreases more slowly at high-temperature levels. This information can then be used to develop strategies for mitigating the impacts of climate change on water resources as well as data-driven decisions for water resources management.

Figure 4.4 (I) shows the impact of the river slope and its interaction with soil temperature at a depth of 10cm on the model prediction. In the context of the Mackenzie basin, the river slope and soil temperature play intertwined roles in influencing river discharge. River slope is the measure of the steepness of a river and dictates the speed and volume of water flow: steeper slopes lead to faster movement of water and potentially higher discharges, especially after precipitation events. On the other hand, soil temperature can affect the timing and amount of water that reaches rivers. For example, in the spring, warmer soils can cause snow to melt faster, leading to increased river discharge. Addi-

tionally, warmer soils can increase evapotranspiration, reducing river discharge. Thus, the exact relationship will vary depending on several factors, including soil type, vegetation, climate, and the time of year. In tandem, areas with steeper river slopes, often mountainous regions, experiencing low soil temperatures can significantly impact the model's prediction, potentially forecasting rapid and significant decreases in river discharge due to the compounded effects of terrain, cold temperatures, and atmospheric moisture. However, it should be noted that the Mackenzie basin is very large and covers a wide range of climates and geomorphologies. Thus, the impact of the interaction between slope and soil temperature on the model's prediction might differ depending on the basin's gauge station's location.

Figure 4.4 (II) shows the impact of snow depth water equivalent (SWE) and its interaction with the river slope on model prediction. SWE refers to the amount of water contained within a snowpack, expressed as the depth of water produced if all the snow melted. The value of SWE determines how much water enters rivers and streams when snow melts. A high SWE indicates significant potential for runoff, leading to increased river discharge when temperatures rise. Conversely, a low SWE suggests limited water availability from snowmelt, potentially leading to reduced river flows during the melt season. River slope, meanwhile, modulates how water flows, with steeper slopes usually promoting quicker water movement and increased river discharge. When these factors converge, areas with steep slopes experiencing low SWE might see accelerated snowmelt and runoff, amplifying river discharge. In predictive modeling, such combined effects can manifest as heightened sensitivities in discharge forecasts, especially in regions where rapid snowmelt events are expected.

4.4 (III) shows the impact of soil temperature and its interaction with the upstream catchment basin area on model prediction. Soil temperature influences river discharge through various mechanisms. Warm soil can hasten snowmelt, leading to quicker runoff, while frozen soil can inhibit infiltration, causing increased surface runoff. Soil temperature

also impacts evapotranspiration rates, groundwater flow, and vegetation growth patterns. These patterns can modify the amount and timing of water contributing to rivers. In colder climates, the temperature of inflowing water, affected by soil conditions, can further influence river ice dynamics and potential flooding. These soil temperature-driven hydrologic processes collectively shape river discharge dynamics and the broader hydrological cycle. In some cases, large catchment areas might not contribute as much to river discharge as expected because a significant portion of the water is returned to the atmosphere before it reaches the river. However, it should be noted that the specific relationship between soil temperature and basin area can vary depending on several factors, including climate, land cover, and topography, which is especially true of the Mackenzie basin.

The consistent effect of longwave radiation flux and soil moisture i.e., Figures 4.4 V and VI on the model's prediction suggests that these variables have a stable and predictable influence on river discharge across various conditions. Longwave radiation flux affects surface energy balance and evaporation rates, while soil moisture determines water availability for runoff and evapotranspiration. The uniform impact of these variables implies that the hydrological processes they drive operate consistently within the observed range of the dataset. In addition, the Mackenzie River basin's characteristics might inherently dampen variability in response to changes in these factors, resulting in a more predictable discharge outcome.

Figure 4.5 shows the average attention weights of LSTM models trained across different Strahler River order systems. Visualizing the attention weights of LSTM models trained on data from each Strahler order can help determine the specific input features most relevant to the model's predictions. Additionally, this can help identify patterns in the data that are not immediately obvious to the human eye. For example, it may be possible to identify patterns in how different Strahler orders respond to different environmental factors.

Order 8 rivers are typically large in the Mackenzie basin, found in valleys, and flowing into the ocean. The plot shows relatively widespread attention weights across the entire sequence. This suggests that models trained on order 8 data consider a broad range of past and present events when predicting discharge from these large rivers. The widespread attention is consistent with the idea that large rivers accumulate water over vast areas and long timescales, thus having a memory of past events. Additionally, processes such as groundwater baseflow, channel storage, and the cumulative effects of multiple tributaries significantly influence river discharge in order 8 rivers. Groundwater, for instance, can have a long lag time before influencing river discharge. This is especially true for large basins. Furthermore, large-scale weather systems such as prolonged rainfall or snowmelt events in the watershed's upper reaches have downstream effects that manifest days or weeks later (lagged influence). The widespread attention suggests that the models recognize delayed contributions from these processes.

The attention weights for order 7 rivers are also relatively spread out but seem to peak in the first month and last 3 months. This suggests that while past and recent events matter, events from the intermediate past (a few months back) might be more influential for this code. This could be due to the accumulation of flows from multiple tributaries, each with its own response time. This is because  $n-1$  orders (order 7 in the Mackenzie basin) still have significant channel storage and receive water from lakes or wetlands upstream, which can introduce delayed response to precipitation events.

The attention mechanism in intermediate orders (orders 6 and 5) seems to focus more on recent events, but past events are still weighted, especially at the start of the sequence. The influence of direct runoff becomes more pronounced in these orders (transitions between high mountains and lowlands), but there is still an element of delayed response due to storage areas like small wetlands, ponds, or localized groundwater contributions. However, the models pay some attention to historical conditions, which reflect the system's memory of past events such as rainfall or snowmelt.

The attention mechanisms in the lower orders (4, 3, and 2) seem skewed towards the most recent timesteps. This is consistent with the rapid response of smaller streams to precipitation events. These streams can react quickly to rainfall or snowmelt in mountainous areas, leading the models to focus on recent events when predicting discharge. Often found in steeper terrains, these streams have a rapid hydrological response. Here, the influence of direct surface runoff is more pronounced. Furthermore, groundwater influence is minimal, and any groundwater contributing to discharge is usually from shallow aquifers with quick response times. From a hydrologic perspective, short-duration and high-intensity rainfall events often lead to flash floods in these streams. These explain why models heavily skew attention towards recent timesteps.

Local explainability in river discharge prediction is a powerful tool for understanding and managing rivers. While global explainability shows the general importance of features towards prediction, local explainability provides insights into how these features influence discharge under specific conditions, such as winter, summer, anthropogenic climate change, and extreme weather events. Local explainability can be used to identify anomalies, such as flash floods, potential risks associated with landslides, areas and periods of potential drought, and human activities contributing to climate change. For example, local explainability could have been used to identify the sudden ice melt in Greenland in 2019, where melting occurred across 90% of the continent's surface – dumping 55 billion tons of water over 5 days. Further, local explainability can show which factors are the most influential at specific times of the year, allowing targeted and more effective interventions. For example, local explainability could be used to identify the most important factors contributing to flooding in a particular region during the monsoon season. This information could then be used to develop targeted interventions to reduce flooding risk, such as building flood defenses or improving drainage systems. Finally, local explainability can help to incorporate local knowledge into river discharge prediction models. Often, local communities possess traditional or experimental knowledge of the behaviors of the

rivers. Local explainability can help to align model predictions with this knowledge – ensuring that predictions are scientifically rigorous and resonate with local experiences.

#### **4.1.7 Conclusion**

Recent advances in machine learning (ML) have revolutionized the field of hydrology, enabling accurate and timely predictions of river discharge even under complex and changing hydrometeorological conditions. However, the black-box nature of ML models has limited their adoption and usage, particularly in the hydrologic sciences, where transparency and interpretability are crucial. In this part of our thesis, we addressed this challenge by leveraging explainable AI (XAI) techniques to investigate ML models' inner workings for river discharge prediction. Our comparative analysis revealed that different ML algorithms have varying strengths and weaknesses in capturing different aspects of hydrological dynamics. For example, Linear Regression (LR) effectively highlights energy balance dynamics, while Random Forests (RF) adeptly capture intricate hydrometeorological interactions, and Long Short-Term Memory (LSTM) models reflect varied hydrological responses across different spatial scales.

Notably, XAI techniques helped us identify the specific features and interactions that drive the predictions of each ML model. This enabled us to discern specific seasonal and regional influences on river discharge. This helped us identify hydrological anomalies and shed the potential to align ML models' recommendations with traditional local knowledge. Additionally, we uncovered nuanced interactions between different hydrometeorological variables, such as the synergy between river slope and specific humidity, underscoring their combined impact on discharge predictions.

Our findings underscore the importance of explainable AI in elucidating intricate hydrological dynamics, offering invaluable insights for adaptive water resource management. In the face of climate change, the need for accurate and interpretable discharge projections is more critical than ever. Thus, by leveraging the power of ML and XAI, we can develop



robust and reliable water management strategies in response to evolving anthropogenic climate change.

#### **4.1.8 Future Work**

In this chapter, we have demonstrated statistical techniques to improve the explainability and interpretability of different ML algorithms for river discharge prediction. Going forward, we will explore incorporating explainable XAI techniques directly into the model design process. This approach, known as physics-driven machine learning, lies at the intersection of ML and physics-based modeling and incorporates a feedback component that ensures that ML models learn the physics dynamics of the underlying hydrologic processes. Additionally, we plan to conduct explainability assessments across diverse river basins in varying geographical and climate regions, crucial for elucidating climate change's local and global drivers. Through this, we hope to identify common patterns and trends contributing to climate change by comparing the explainability results from different regions. This valuable information can then be utilized to develop more effective climate change mitigation and adaptation strategies. As we face the increasing challenges of climate variability, the demand for explainable ML models in hydrology will only grow more pressing. By fostering a deeper understanding of ML model behavior, we can enhance the reliability and trustworthiness of these powerful tools, ultimately enabling more effective and sustainable water resource management practices.

## CHAPTER 5

### ROAD QUALITY PREDICTION USING SATELLITE IMAGERY

#### 5.1 Road Quality Prediction using High Resolution Satellite Imagery

##### 5.1.1 Motivation

High-quality roads are among the foremost infrastructure for hastening societal development. Roads enable goods, people, and ideas to travel easily, leading to better equity in service provision, faster economic development, and, ultimately, better human outcomes. Though enormous sums are spent on roads – for example, in sub-Saharan Africa, 1.5% of total GDP is spent on roads [267] – funds for road maintenance consistently fall short, a problem arising from an inability to prioritize investments [24]. This is partially due to limited road quality measurement, which requires large amounts of labor, time, and expensive equipment. As Lord Kelvin famously said: “If you cannot measure it, you cannot improve it.”

In urban developing settings, where road usage is heavier and increasingly more people travel with sensor-laden smartphones, crowdsourcing data on road quality is possible [282]. However, rural settings are not as conducive to smartphone-based solutions. In this work, we present a viable alternative for measuring road quality in rural, resource-constrained settings: models for predicting road quality from remote sensing imagery. Our models leverage recent advances in two areas: satellite technology and computer vision. A proliferation of satellite companies has resulted in increasingly higher resolution images collected more frequently; in developing regions, this imagery is as high as 30-50cm resolution and some urban areas are imaged near-daily. Meanwhile, advances in

computer vision have produced techniques for creating and applying sophisticated neural network-based models with thousands to millions of training examples. Further, we can explore the potential for domain adaptation, where we can apply our models trained in one setting to another, a capability that has huge potential benefits in cost-constrained contexts.

Our training dataset consists of road roughness measurements collected by specialized equipment over 7000 km of interurban roadways through diverse terrain in Kenya. We employ this unique dataset to train a regression engine to produce estimates of road quality based solely on observing satellite imagery. This learning task is well-suited to developing regions, where sensing approaches using fixed infrastructure, expensive mobile equipment, and even smartphone-based systems may be infeasible. Our models are built upon convolutional neural network architectures [131, 154, 253] with modifications to accommodate our regression task. We focus on the particular domain adaptation challenge of prediction for held-out roads, which have been explicitly excluded from the training set to evaluate whether our model can accurately predict road quality using imagery at places or times that it has never seen before. To exhibit the potential of our approach to road quality measurement, we present a case study on the positive correlation between road quality and town prosperity, as measured by satellite nighttime illumination.

### **5.1.2 Background and Related Work**

*Road quality measurement:* The road quality measure used throughout our study is called the International Roughness Index (IRI), developed by the World Bank in 1986 [241]. IRI measures the cumulative vertical displacement of a vehicle along a stretch of road due to the roughness of the road surface, is typically provided in units of  $m/km$ , and is commonly collected using a specialized vehicle with a mounted laser. IRI values can be any positive real number, where higher IRI values imply worse road quality and typical values fall

between 0 and 30. While not explicitly a measurement of road quality, in practice, IRI has been found to have a very high correlation with user perception of road smoothness. The equipment used for measuring IRI is high precision, complex, and expensive. Governments in developing countries with oversubscribed budgets for infrastructure can seldom afford to pay for these equipment and carry out this procedure on a regular basis [109]. The result is that developing countries either conduct road quality surveys as infrequently as once every few years or even do not conduct full, accurate road quality measurements. Some researchers are leveraging cheap accelerometers and gyroscopes that are fitted with mobile phones to measure road quality [86], but these cheap sensors are incapable of handling continuous acceleration and vibration intensity for more than a few minutes without losing calibration. In addition, smartphones typically use Assisted GPS, which relies on nearby cell towers for better GPS accuracy, but due to poor cellular network systems in developing regions, this process is quite expensive in terms of data and battery consumption that are required to maintain the accuracy of GPS.

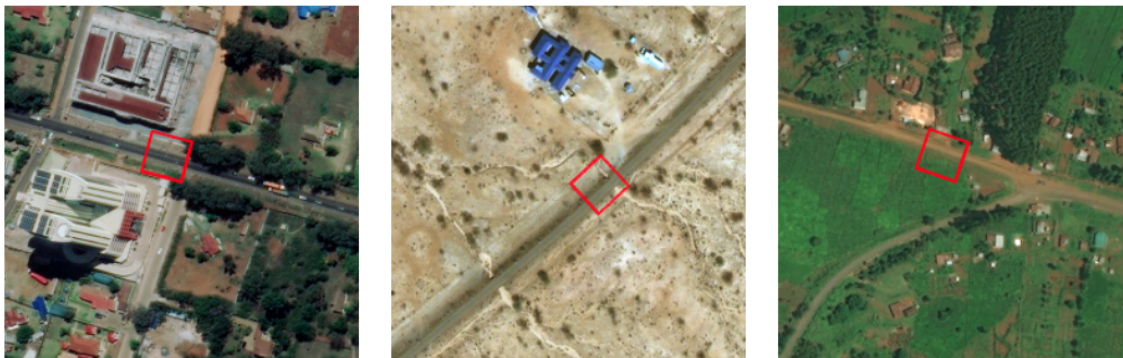
*CNNs and LSTMs on satellite imagery.* The increasing application of convolutional neural networks (CNNs) and long short-term memory/recurrent neural networks (LSTM-RNNs) to satellite imagery has been partly due to faster and cheaper computational power [242], efficient and less computationally intensive algorithms, increasing availability of satellite-gathered image datasets in the public domain, and transfer learning.

As a result, these algorithms are being applied to satellite imagery for more complicated tasks such as large-scale damage detection after calamities [106], land use classifications [5], and generating human-like descriptions of satellite images [250]. This means that highly accurate algorithms trained on traditional images can be used to evaluate satellite images via transfer learning.

Another related and popular area of work is road detection using satellite imagery, which has spawned substantial research [183, 258], competitions [60], and even companies. The road detection problem seeks to identify the locations of road infrastructure, and while

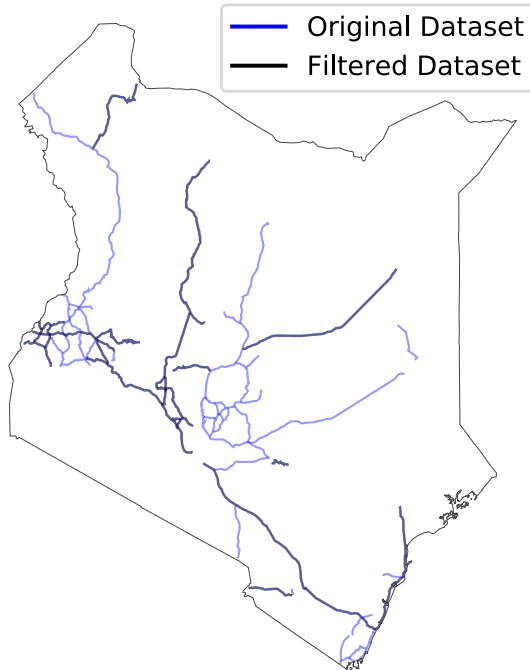
most work focuses on industrialized regions, there are examples of work that target more challenging unpaved roads [16]. Nonetheless, while the two problems – road detection and road quality estimation – have some complementarity (for example, the latter problem can leverage outputs of road detection algorithms), in most cases, systems solving these problems are independent and employ entirely different algorithms and metrics.

Traditionally, LSTMs have been applied to recognize patterns in sequential data such as speech, text, and video [239, 306, 276]. Previous work on using LSTMs on satellite imagery includes temporal vegetation modeling for crop identification [238]. However, we take a different approach by leveraging the spatial characteristics of individual image patches and the sequential nature of a single stretch of a road (road patches in a sequence) to harness the advantages of LSTMs. Essentially, by adapting sequences of patches to imitate a time series of images, we create image frames in succession that act as inputs to an LSTM.



**Figure 5.1:** Three different roads highlighting the challenging diversity of our dataset. Left: an urban environment along the A104 highway. A104 is a major highway in Kenya, and the selected road tile is of “great” quality. Center: the C47 minor road. It passes through an arid environment, and the road segment has “poor” quality. Right: the C67 minor road. It passes through large forests and cropland, and the road segment in the image has “good” quality.

This paper builds upon previous work by the authors on the topic of measuring road quality using satellite imagery, recently published in a workshop [39]. While the input datasets for both pieces of work are the same, in this work, we explore many additional methods to improve prediction performance (recurrent neural networks and auto-encoders), perform regression analysis instead of classification, and conduct a novel and significant case



**Figure 5.2:** Roads with labeled quality data as collected by the Kenya National Highway Authority (KenHA). Indicated are the original dataset and a dataset that has been filtered to match the availability of concurrent satellite imagery.

study to demonstrate the unique value of our road quality measurement techniques for studying economic activity in a developing region.

### 5.1.3 Methodology

#### 5.1.3.1 Datasets

Our intent is to ultimately predict the quality of a road seen in satellite imagery. Towards this goal, we employ two main datasets: one set of road quality measurements and a corresponding set of satellite imagery, both for Kenya.

The dataset of road quality measurements used to train our models consists of IRI measurements conducted at a resolution of 10m along a diverse set of 57 roads throughout Kenya, resulting in samples over a total length of 7000km. A map of the dataset is available in Figure 5.2. This dataset was collected as the result of a partnership between the Kenya National Highway Authority (KenHA) and the Japanese International Cooperation

Agency (JICA). Each measurement is tagged with a latitude and longitude (“lat-lon”) and a date of survey (during 2013-2015). The roads can vary from tens to several hundred kilometers in length and, as we show in Figure 5.1, span a wide variety of road sizes, terrain types, and land usage. Additionally, IRI measurements are often bucketized into 5 *road quality classes*: great (0-7), good (7-12), fair (12-15), poor (15-20) and bad (20+). Figure 5.1 also shows examples of roads falling into these categories. Roads in our data can also be split into three administrative classes: Class A, linking centers of international importance; Class B, linking national centers within the country; and Class C, linking provincially-important centers. These roads comprise the fabric of Kenya’s road transport system, serving as the primary interlinkage between major towns throughout the country.

The satellite imagery we use is the DigitalGlobe Basemap+Vivid product [64] and the coverage is the entirety of Kenya. We employ two iterations of this imagery product, each of which is a mosaic of roughly 6300 tiles that forms the illusion of a continuous map by stitching together several images collected at different points in time by multiple different satellites. The +Vivid product is post-processed to account for orthorectification, color correction, and cloud cover, though the latter is still a problem in some remote areas. The first mosaic, compiled in November 2014, consists of imagery from the QuickBird-02 and WorldView-02 satellites, and is composed of tiles with collection dates ranging from 2002 to 2014. The second mosaic, compiled in September, 2017, consists of imagery from the QuickBird-02, GeoEye-1, WorldView-02, and WorldView-03 satellites, and is composed of tiles from 2002 to 2017. The typical resolution of the tiles is roughly 50 cm per pixel, and each of the two image mosaic datasets is 7 – 8TB.

While the satellite imagery and road quality datasets are both impressively large, the wide range of dates covered by the tiles of each mosaic coupled with the range of dates of the IRI measurements creates a mismatch. This issue often appears when learning on satellite

Road	Length (km)	Bad (%)	Poor (%)	Fair (%)	Good (%)	Great (%)
A104	269.2	2.2	1.1	1.5	6.3	<b>88.9</b>
A109	86.54	0	0	1.9	10.5	<b>86.6</b>
A23	10.31	<b>44.9</b>	17.9	12.6	24.2	0.3
B8	47.56	3.4	9.8	18.3	<b>50.5</b>	17.9
B9	16.64	13.6	<b>31.9</b>	21.3	31.1	2
C31	39.33	0	0	0	1.5	<b>98.5</b>
C32	30.03	<b>61.7</b>	20.5	8.8	7.9	1.2
C33	44.79	0.1	1.0	2.4	19.3	<b>76.2</b>
C36	23.05	11.2	9.5	8.3	21.5	<b>49.6</b>
C42	40.88	31.9	13.5	6.7	9.9	<b>38.0</b>
C47	104.92	<b>40.3</b>	35.8	14.8	8.9	2.4
C51	40.89	2.1	3.8	4.9	16.3	<b>72.9</b>
C54	27.62	4.4	1	2.8	13.9	<b>77.8</b>
C67	37.86	<b>90.3</b>	4.4	3.1	2.3	0
C68	16.94	<b>56.3</b>	20	15.1	8.6	0
C69	95.07	0	0.1	1.7	26.1	<b>72.0</b>
C76	41.89	<b>79.5</b>	16	4.4	0	0
C77	110.40	22.2	16.6	21.8	<b>36.1</b>	3.2
C78	28.71	16	20	17	<b>39.3</b>	8
C83	21.38	<b>100</b>	0	0	0	0
C96	19.07	5	23	29	<b>39</b>	4
All	1153	19.2	9.5	7.6	16	<b>47.7</b>

**Table 5.1:** A summary of the diverse set of roads in our labeled and filtered data set recording both the length and distribution of road quality labels. For each road, the modal road quality class is in bold. The set ranges from first-class highways (e.g., A104) to rough dirt roads (e.g., C67) and includes roads with significant internal variation (e.g., C77).



imagery [136], but is particularly acute in our scenario since road quality can experience sudden and potentially substantial changes (*i.e.*, due to weather or construction) in a way that only a serious emergency may impact other attributes commonly predicted via satellite imagery (like wealth). In an ideal data-collection scenario, the maximum time period requirement would be a month or even a week, though given the reduced frequency of data collection in developing regions, this is untenable. Ultimately, to ensure that imagery reasonably matches the condition on the ground when the IRI sample was collected, we decided to restrict our label dataset to only those samples where the difference between the two dates was 12 months or less. Selecting any period of time shorter than 1 year for the maximal time discrepancy would have significantly decreased the amount of labeled data. This left us with a tradeoff between not having enough data to properly train a deep net and possibly having some incorrect labels. Our ideal scenario of using data no more than one month out of date would have left us with only 340 kilometers of road and 40% of the unique roads in our 1-year set. Given that anything more than three months already entails a possible shift between seasons it was decided that the extra data provided by a maximal discrepancy of 1 year was tolerable. This design decision results in a subset of the samples from the larger IRI dataset used for training; this subset consists of samples covering 1153km over 21 roads, as detailed in Table 5.1, which also includes a breakdown of the classes of labels for each road. Additionally, a map of the filtered dataset is available in Figure 5.2. This set of roads includes paved and dirt roads, consistently high-quality and consistently low-quality roads, and roads with high variability in quality.

One nuance of the data set is deciding what the fundamental unit of training data will be. We define this unit as a *patch* and define it as a quadrilateral such that the length of the patch is parallel to the course of the road and the width is perpendicular as seen in Figure 5.3. Our IRI measurements are at intervals of approximately 10 meters (20 pixels in our imagery data), so possibilities for length are bounded below by 20 pixels. Argu-

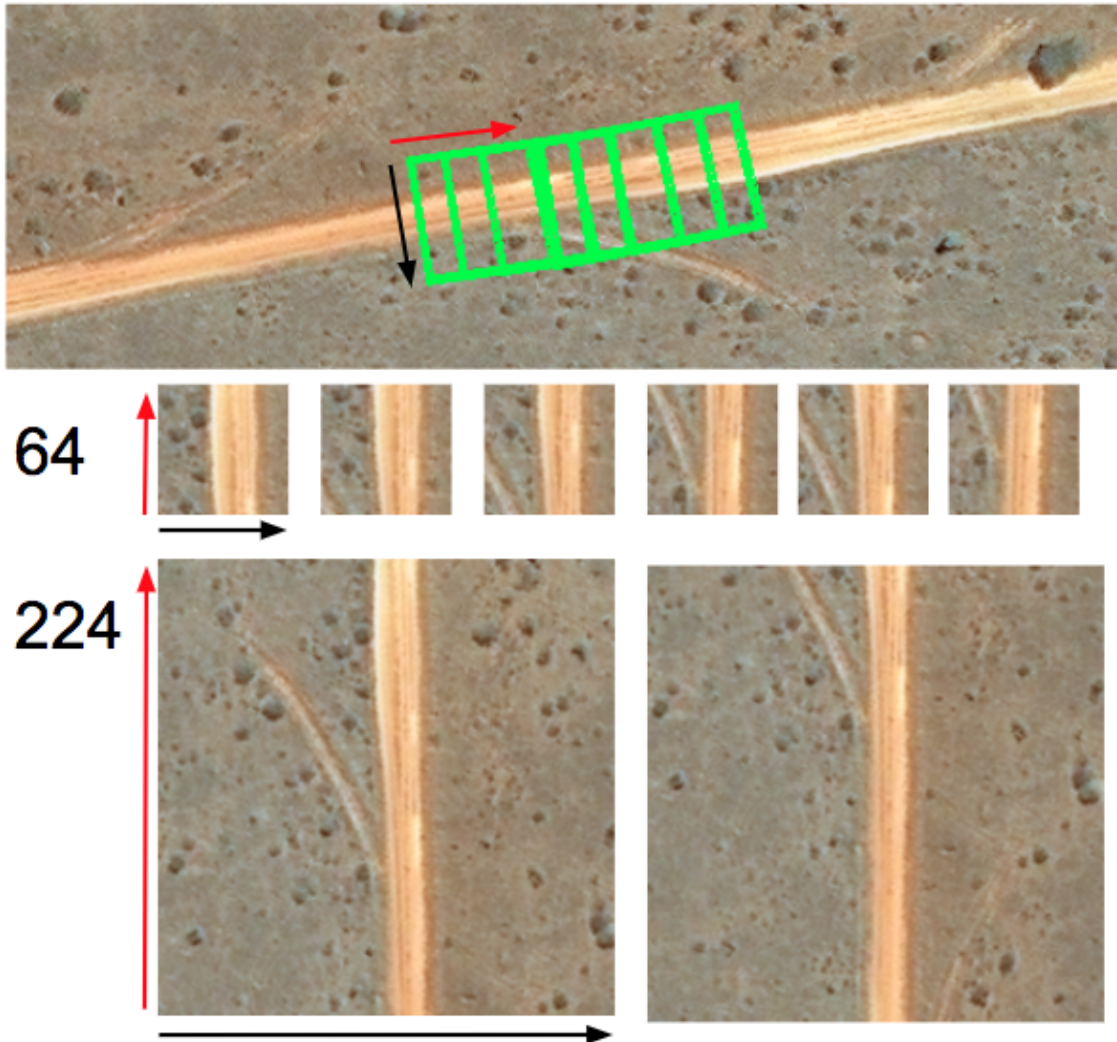
ments for a smaller patch size include greater granularity and the road forming a greater proportion of the patch's area. However, patches of lesser dimensions may sometimes not include some or all of the road due to random noise in the latitude-longitude pairs associated with IRI values. We settled on a compromise of 64x64 pixels, which was robust enough to account for this noise and also neatly covers 3 IRI measurements per patch.

### 5.1.3.2 Training and metrics

The IRI data set is sufficiently fine-grained as to allow several different choices for what exactly to predict. In the first case, IRI is a numeric measure (again, note that lower values imply higher quality) but is often broken down into the 5 *road quality classes*. One could either predict the underlying IRI number of a length of road directly or instead attempt to classify which of the five aforementioned classes it falls into. This work pursues the former approach for the reason that this is both informative and avoids corner cases with stretches of roads falling near the threshold between two different labels.

As such, our work focuses on a regression problem and will record results in terms of the mean square error (MSE) and the  $R^2$  coefficient. MSE gives us an absolute averaged error while  $R^2$  explains how much of the total variance in the IRI is explained by our prediction. Instead of directly using the IRI values  $y_i^*$ , we establish a maximum threshold  $T = 30$  and train on the labels  $y_i = \frac{\min(y_i^*, T)}{T}$ . We feel this is justified since anything above an IRI of 20 is already bad and the visual/practical difference between a tile of IRI 30 and another with IRI 40 is minimal. Note that since any predicted IRI value can easily be mapped to a prediction of one of the five *road quality classes*, we can also measure the accuracy of our predictions if they were used for a classification task instead of a regression. We report these accuracies throughout our results for comparative purposes even though we do not at any time train for classification accuracy.

A final consideration is defining the train and test sets in this scenario. Since the training data has a sequential nature, randomly splitting the data into train and test sets (as is



**Figure 5.3:** An example of how a road segment can be separated into tiles. The top segment shows a road divided into overlapping 64x64 squares; this generates the tiles shown in the middle segment. The tiles are always aligned in the direction of the road (red arrow). The bottom shows the same segment if divided into 224x224 tiles instead. Note that the road constitutes a much smaller proportion of each tile relative to the 64x64 case.

usually done) would result in cases where patches that appear next to each other in the satellite imagery might be in both the training and test sets. In addition to the problem of test data contamination, this testing scenario would be very different from the use case that we are targeting, where one would seek to predict on an entirely unseen road. As such, we devise two more appropriate methods of generating training and test sets. The first is done by splitting the entire set into 1-kilometer long “runs” which are then randomly assigned to the train or test set with proportion 70%-30% – we call this the *standard* method since it more closely resembles the random train-test split. The second method is to assign an entire road to the test set and the remaining 20 roads to the train set and average the result over the 21 possible splits (one with each road held out): we call this the *held-out* split procedure. Though this very closely approximates a real application, we note that this method breaks an often central assumption of machine learning methods: that the train and test sets are drawn from the same distribution. As the held-out problem is much harder to predict, results reported using the *held-out* methodology are significantly worse than those reported using the *standard* methodology. However, *held-out* predictions are potentially more impactful, as results can generalize to unseen contexts.

### 5.1.3.3 Convolutional neural nets and auto-encoders

Convolutional Neural Networks (CNNs) are a class of machine learning models that have shown excellent promise in several visual processing tasks. Though initially focused on classifying images into many categories (such as ImageNet [237]), these models have also been applied successfully on satellite imagery. Sometimes, complex pre-trained models can be re-purposed on another task with less data through a process known as *transfer learning*. However, with enough training data, the structure of successful nets can be re-used while all the parameters are re-learned. We experimented with both approaches but went with the latter after noting that we had a sufficiently large data set to train networks from scratch. Thus, we began with Resnet, AlexNet and VGG-11 [116, 155, 254] as initial

network structures and then simply replaced the last layer of fully connected layer nodes with a single sigmoid function instead. We then trained using our 64x64 tiles scaled to 224x224 pixels.

While our labeled dataset is already fairly large, we discussed in Section 5.1.3.1 that this only represents 15% of the total of the roads in our dataset; the remainder had to be discarded since the labels might be out of date. However, auto-encoders provide an alternative to supervised CNNs that allows us to leverage that large set without relying on the labels. Convolutional auto-encoders consist of two parts: an encoder, which compresses the images down to  $k$  features, and a decoder, which attempts to reverse the encoding back to the original image. Training this network to attempt to recreate the original image as closely as possible should ideally lead to a  $k$ -dimensional representation of the image that preserves as much information as possible. We can leverage this by training the auto-encoder over the larger, complete set of roads to learn a very efficient representation of any given tile and then doing an L2-regularized regression of these features on our training set. We perform this with a 2-convolution auto-encoder with  $k = 1000$  alongside retraining the aforementioned CNNs.

#### 5.1.3.4 Sequence learning via LSTMs

Another avenue for exploration is how much the sequential structure of roads can be leveraged to more accurately predict road quality. In the simplest sense, we can keep the same fundamental aim of predicting  $y_i$ , but instead of using only the tile image  $\mathbf{x}_i$ , one can use the last  $s$   $\{\mathbf{x}_j \mid i - s < j \leq i\}$ . Slightly more complex would be the case where we use the same segments of satellite imagery but attempt to instead predict the average IRI of the entire segment  $\bar{y}_i = \sum_{j=i-s}^i y_j$ .

We handle both of these cases by first using the auto-encoder to featurize all the roads and grouping them into contiguous sequences of length  $s$ . We will use the same simple 1-layer LSTM with 500 internal nodes, changing only the objective to optimize between

5-class accuracy		Regression $R^2$		
Model	Standard	Held-out	Standard	Held-out
ResNet	0.69	0.44	0.79	0.24
VGG-11	0.71	0.47	0.78	0.26
AlexNet	0.73	0.49	0.66	0.21
Autoencoder	0.65	0.41	0.78	0.31

**Table 5.2:** 5-class accuracy and regression R-squared results under *standard* train-test and *held-out* conditions for the single-tile regression problem.

the two. The LSTM is then trained with L2 regularization to prevent overfitting, and we record the *held-out* results in Section 5.1.4.2.

## 5.1.4 Results

### 5.1.4.1 Single tile regression

We first compare how different network structures and the auto-encoder regression perform under the aforementioned *standard* and *held-out* testing methodologies in Table 5.4. Resnet and AlexNet were retrained from scratch, while VGG-11 was transfer-learned, with only its classifier component being retrained. All the CNNs were trained over 10 epochs of the data, augmented by random horizontal and vertical flips, and completed within a few hours when trained on a GPU cluster. We found that after around 10 epochs, the training loss was roughly flat, and continuing to train would likely only result in overfitting. The auto-encoder was trained overnight on the unlabelled dataset for 20 epochs and then simply regressed with an L2 penalty. Though training the auto-encoder itself is time-consuming, this is only a one-time task and can be later used to quickly featurize any road.

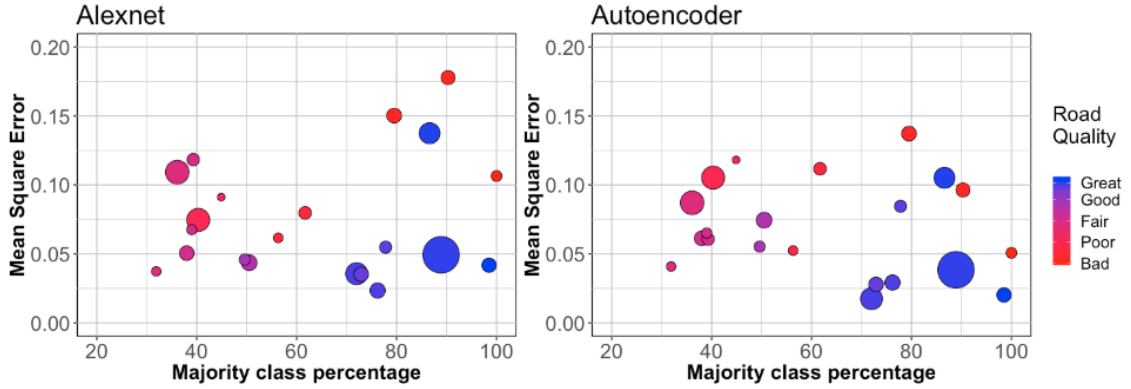
One immediate observation that can be made is that the more realistic *held-out* test case is significantly harder than the *standard* scenario. However, the results are encouraging, given that achieving even the baseline result (accuracy of 0.20 and an  $R^2$  of 0.0) is not guaranteed when the train and test sets are entirely different. This proves the problem is indeed approachable using standard machine learning techniques. The second observa-

tion we wish to highlight is that accuracy is fairly good for a 5-class problem even though we do not directly optimize for accuracy. This is a consequence that estimating the IRI value fairly accurately will translate to a correct estimate of the road quality class and seems to bolster the idea that directly regressing the IRI and then transforming it to less granular measures as the application calls for it is a viable idea. We note that measuring accuracy does not distinguish between one error mistaking a “fair” tile for a “good” tile and another error mistaking a “poor” tile for an “excellent” tile, thus potentially understating the predictive quality of the CNNs.

We find little to separate between the different CNN classes regarding performance. Similar initial learning rates and dropout rates were used in all, though VGG-11 had more problems with overfitting compared to the other two and did not move beyond transfer learning. In terms of the key regression metric, we found that the auto-encoder regression outperformed the other CNN methods. This is likely due to its superior generalization performance on unseen roads. In this we find a notable advantage of leveraging the entire set of roads; the feature representation from the auto-encoder was much more stable in the *held-out* scenario than it was on the other CNNs. This was likely because the auto-encoder could look at a much more diverse set of roads to determine how to featurize a patch, as opposed to the only 21 roads that a CNN could use. Figure 5.4 illustrates this by comparing the results on individual roads for Alexnet to those of the auto-encoder regression. In addition to better overall performance, the auto-encoder has fewer roads with very high error: this is important for real-world application as we would like to be confident of some guarantee of predictive power.

#### **5.1.4.2 Tile sequence regression**

After the single-tile regression results, we wished to explore whether incorporating the sequence of tiles leading up to the final one would improve our predictive quality. As discussed in Section 5.1.3.4, we wished to explore this for both the average and last tile



**Figure 5.4:** Figures showing the distribution of Mean square errors (y-axis, lower indicates better predictive power) of different roads using a Resnet CNN (left) and auto-encoder regression (right). The x-axis is a measure of the heterogeneity of the road, the color provides the average road quality, and the circle size indicates the relative sizes of the roads. Comparisons to VGG-11 and Resnet yield similar results.

IRI. Based on the earlier results in Section 5.1.4.1, we focused our efforts only on the *held-out* test scenario since the *standard* method already had strong results for single-tile regression, and the former is in any case a more representative approximation of real-world applications. We also decided to focus on the autoencoder features instead of using the CNN features since the two-step featurization/sequence training makes the strong generalization performance of the former an important asset.

Sequence length	5-class accuracy		Regression $R^2$	
	Last	Mean	Last	Mean
1	0.41	0.41	0.31	0.31
10	0.42	0.43	0.34	0.35
25	0.43	0.43	0.35	0.32

**Table 5.3:** Results for LSTMs in the *held-out* test scenario as a function of the length of sequence trained on. Regressing on the final tile value (last) is compared to regressing on the average tile value (mean).

Our results are summarized in Table 5.3. These show a modest improvement over the non-sequential LSTM, though these initial experiments do not suggest much improvement when increasing the sequence length over 10. However, this is likely influenced by the



well-known difficulties of training on longer sequence LSTMs and may not reflect the actual limit of this technique. Further investigation into both the structure of the LSTM as well as its training will be important.

### **5.1.5 Future Work and Conclusions**

In this work, we describe a methodology to infer the quality of intercity roads in developing regions, with the primary goal of enabling useful and practical applications. To do this, we trained models using satellite data and road roughness data from Kenya and demonstrated that the models performed well in some cases in locations previously unseen while remaining cognizant of remaining challenges. We saw that while the normal train-test paradigm can be approached readily, achieving reliable results on the held-out case is significantly harder. We also demonstrated a novel use case of our road quality measurement at a larger scale than traditional methods would feasibly allow.

There are several machine-learning aspects that could be explored further. Though we have shown that auto-encoders are greatly beneficial in this scenario where there is a plethora of unlabelled data, understanding the ideal setup will take further investigation. What sort of convolutional structure, whether it should have any regularization, and how far to compress each tile will all need to be rigorously investigated. Likewise, for the recurrent neural network structure. In this vein, further investigation of the LSTM structure is possible, as is the idea of using an entirely different type of RNN to model the continuous sequential nature of this data. This would be an interesting novelty compared to the discrete sequential nature of data, such as words in a sentence or frames in a video.

More germane to potential use cases, we can further explore the ability of different modeling decisions and methods on the ability to generalize to different contexts, especially those further afield. We can also consider the effects of improved or degraded satellite imagery quality on results and measure performance improvements from including other

features available from further remote sensing data (e.g., land use data or multispectral imagery).

In general, more accurate and granular measurement of road quality can lead to reduced road maintenance costs, allowing expensive rehabilitation efforts to be replaced by targeted repairs. Further, these capabilities can empower governments, donors, and policymakers to identify particularly hazardous roads and monitor the short- and long-term performance of construction firms and contractors, improving public safety and enabling more efficient public investments. Additionally, these models can enhance the work of economists and others researching public policy in a variety of domains, ideally leading to a clearer understanding of the levers of societal development in a diverse array of contexts.

## **5.2 Using Vision Transformers to Improve Road Quality Predictions from Medium Resolution and Heterogeneous Satellite Imagery**

### **5.2.1 Motivation**

In section 5.1, we discussed the fundamental importance of high-quality roads for societal development. Furthermore, we discussed the challenges associated with traditional methods for measuring road quality, especially when resources are limited. Building upon this work, this section examines the opportunities for using more advanced machine learning techniques, notably, Vision Transformers (ViTs) to improve road quality measurement using medium-resolution satellite imagery. Vision Transformers have marked a breakthrough in machine learning, particularly in the analysis of low-resolution images. Unlike traditional convolutional neural networks (CNNs), which rely on local connections between pixels, ViTs employ a global attention mechanism that allows them to effectively capture patterns across the entire image, even if the patterns are only a small fraction of the image. Consequently, ViTs are extremely valuable for tasks like road quality assess-

ment using low-resolution satellite images when roads are only a small part of the image patch. Here, we make a trade-off between resolution, accessibility frequency, and acquisition cost. Although both Google Earth Pro and Planet Data have relatively low resolution, they are frequently accessible and available in the public domain either at a cheaper price or for free. Furthermore, Planet Data has daily imagery for most economies (geographic locations), including the global south. We show that when our models are validated with previously unseen (held-out) medium-resolution and sparse data, significant results can be achieved without degrading region-specific characteristics or introducing bias due to inconsistencies in the data quality. Finally, to illustrate the practical applicability of our approach to road quality measurement, we present a case study examining the relationship between infrastructure quality and the average household asset wealth index. The results of the case study demonstrate that a positive correlation exists between improvement in road quality and an increase in asset wealth.

### 5.2.2 Background and related work

**Impact Evaluation of Infrastructure:** Despite the importance of measuring road quality (built infrastructure) regularly and the role that infrastructure will play if we are to achieve sustainable development goals by 2030 [4, 174, 55, 1], there is a paucity of literature available on the topic. This is because road quality is frequently overlooked when discussing sustainable development goals since more emphasis is placed on topics such as poverty eradication, health, and education. However, infrastructure plays a crucial role in improving productivity, allowing access to services such as education, healthcare, and the market for goods and services, thus reducing socioeconomic barriers and stimulating economic growth, all of which are critical components of sustainable development. By regularly measuring road quality, governments can assess the efficacy of their previous investments in infrastructure and determine whether these investments are having the desired effect (positive socioeconomic impact). This information can inform future

investments and ensure they are made most efficiently and effectively. Furthermore, understanding the quality of infrastructure (roads and beyond) allows government and the private sector to generate lock-in patterns of social and economic development as well as understand the impact of investing in quality infrastructure [269, 1], allowing development opportunities to be maximized, effectively improving regional planning (e.g., where to construct commercial airports and industrial parks) and allocating resources that power economic developments and, subsequently, creating a domino effect of economic growth.

**Vision Transformers:** Vision Transformers (ViTs) [69, 142] are a type of neural network architecture designed for image recognition tasks. ViTs are based on the transformer architecture [277], which was originally developed for natural language processing (NLP) tasks [189, 123]. ViTs adapt to image tasks by breaking up an image into smaller patches that are then treated as being analogous to the words making up a natural language input. In the context of image recognition, Vision Transformer (ViT) models have been shown to be competitive with state-of-the-art convolutional neural network (CNN) models on various benchmark datasets. ViTs can perform vision-related tasks directly from raw pixel data, without the need for manual feature engineering. Unlike traditional CNNs that rely on convolutional layers, ViT models use an attention mechanism [14], a computational technique for relating different parts (small patches) of a given image to improve understanding of the visual contents of the image. The attention mechanism allows the model to focus on the most relevant parts of the input when making predictions. This can be especially useful for image classification tasks, such as road quality prediction from low-resolution satellite imagery, where the road may make up only a small part of the satellite imagery, where the object of interest may only occupy a fairly small portion of the total image. Transformers have also been shown to have strong generalization capabilities, which means that they can perform well on a wide range of tasks without the need for task-specific architectural modifications. This is important for image classification, where it is often difficult to design a model that can handle a diverse set of images. Regarding

training capabilities, ViTs are more efficient to train and can be easily scaled to larger models simply by increasing the number of transformer layers. However, ViTs are generally more computationally expensive to run than CNNs and require more training data and computation resources, which is attributed to convolutional inductive bias. In contrast to CNNs, which are translation invariant, ViTs lack translation invariance (i.e., the model’s output is unaffected by the object’s location in the image). As a result, the output of the model can be influenced by the location of the object in the image. Additionally, ViT models are difficult to interpret because they do not separate the different “layers” of the model. Finally, transformers tend to perform poorly on tasks that require fine-grained localization, such as object detection and segmentation [167, 190, 224].

### **5.2.3 Methodology**

#### **5.2.3.1 Dataset**

Our ultimate goal is to predict road quality based on medium-resolution satellite imagery. We use two sources of satellite imagery: Planet Labs imagery (PLD) and satellite imagery scraped from Google Earth Pro (GEP). PLD [266] has a resolution of 3 meters per pixel, while GEP [72] has a resolution of approximately 1.6 meters per pixel. We use the same label data defined in section 5.1.3.1.

We consider only the earliest available 3m/pixel resolution imagery from Planet Labs, taken between July 2015 and January 2016, in order to align the satellite imagery with our label data. As a result, the number of road segments is reduced from 57 to 36. Additionally, because we match the labels to relatively lower spatial resolution satellite imagery, we down-sample the measurements to 50m and then calculate the average coordinates (latitude/longitude) and IRI within the sampling window. Downsampling IRI measurements lowers the probability of matching measurements to overlapping patches. As a result, the correlation between different but adjacent classes is reduced when regressive IRI measurements are converted to class labels based on a predefined threshold. Finally, we run



**Figure 5.5:** A Series of roads from two data sources; Planet Labs(top) at 3m/pixel and Google Earth Pro (bottom) at  $\sim 1.6\text{m/pixel}$ . The different sizes, visibility and resolutions of the road images highlight the heterogeneity of our dataset. The red rectangle indicates the range of the road segment over which the IRI measurements are averaged.

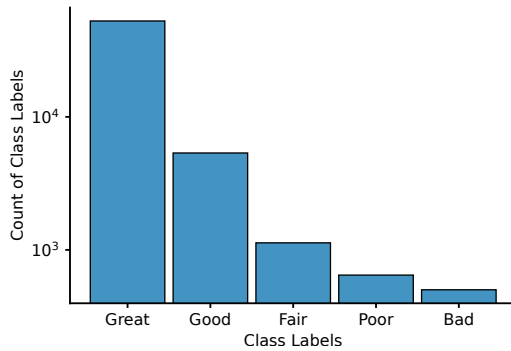
experiments on binary and multiclass classification problems. We follow the same approach defined in Section 5.1.3.1 to discretize IRI measurements into two and five classes. Considering the geographical location of our study area, we selected the earliest available cloud-free images obtained during the period of July 2015 to January 2016. This choice of image selection is based on the assumption that changes in road quality during this period are insufficient to change quality predictions from one class to another (e.g., from good to bad).

Finally, we scrape publicly available satellite imagery from GEP using custom data scraping tools. We strive to keep all images at a resolution of  $\sim 1.6\text{m/pixel}$ . It should be noted, however, that Google Earth Pro aggregates the highest quality images from various data sources (Maxar, Airbus, etc), and as such, the quality and pre-processing strategy of the images is variable.

### 5.2.3.2 Training and metrics

Measurements of the International roughness index (IRI) are recorded as continuous values. As a result, it is difficult to train models that can be deployed across a wide range of geographical regions where the measure of goodness is relative. It is possible, for ex-

ample, for an “excellent” quality road in developing regions to be visually equivalent to a “good” quality road in industrialized regions. This combination of IRI measurements and satellite imagery can introduce bias when training models in one economic region and deploying them in another.



**Figure 5.6:** Histogram showing the 5-class distribution of labels in the dataset. The dataset is heavily imbalanced. Labels associated with “great” road quality contributed the largest distribution percentage.

Thus, we focus on a classification problem whereby we define classification categories (binary, multi-class) in accordance with predefined criteria. This is mainly because it is difficult to infer exact IRI values as compared to predefined classes. With this reconceptualization of road quality measurement, we bound the IRI values within the same range to be categorized under similar categories. By doing so, we can predict roads with similar characteristics with more precision. Furthermore, the class thresholds can be adjusted according

to the user’s needs.

In light of the availability of IRI measurements and the limitation of the measurements to measurements collected in 2015, our dataset is highly imbalanced. We observed labels from 0 to 7 being the most frequent (Fig 5.12). There are various ways to handle imbalanced data in machine learning, including reweighting class labels, oversampling the smaller distribution and undersampling the larger distribution, and data enhancement, among others. Considering the various models, we used various strategies to increase the label size of minor classes without overfitting. When training the baseline model (CNN), we explored two approaches: sampling distribution of the majority classes in relation to the size of the minority classes (over/under sampling) and assigning significant weight to

minority classes and small weights to majority classes (class weighting). Ultimately, we adopted the former approach, as it resulted in better prediction results. For all models, we perform data augmentation, artificially increasing the amount of data by generating new data points from existing data or making small modifications to data (such as flipping images or reducing their contrast) to increase the size of minor classes in the distribution. In terms of evaluating model performance, we record the results in terms of both area under the receiver operating characteristic (AUROC) [76] and accuracy. The AUROC curve is a metric for measuring the performance of binary classification problems. It can also be used to measure the performance of multiclass classification problems by applying statistical techniques such as 1-vs-the-rest. Given the optimal threshold, AUROC can calibrate the trade-off between sensitivity and specificity of the prediction. The AUROC is calculated by plotting the true positive rate (TPR) against the false positive rate (FPR) for different threshold settings of a binary classifier and calculating the area under the curve. In contrast to accuracy, AUROC is not sensitive to the choice of probability thresholds, making it a useful metric when evaluating the performance of models trained on imbalanced data. In particular, any AUROC score above 0.5 represents a predictive value added, whereas a model that trivially predicts the majority class would always obtain an accuracy score greater than 50%.

In traditional machine learning, data are divided into train and test sets according to a predefined ratio (e.g., 70: 30) (hereafter referred to as the standard split). However, this approach has two disadvantages. First, since the road patches are spatially sequential, inter-class differences are not absolute, which means that train and test patches might be close to each other. This could contribute to overfitting of the model [236], a process in which a machine learning algorithm memorizes patterns in training data and fails to generalize the learned information to new data from a different distribution. Our goal is to evaluate how these models would perform if deployed in economic regions in which roads may differ from those with available data. This is a common scenario, given that



developing regions rarely have reliable or available data. Therefore, we use an additional train-test strategy. Consider our dataset with  $n$  roads ( $n=36$ ) and  $k$  as the number of roads in each group, where  $k$  is the number of roads used to train the prediction models. For each percentage split, we randomly sample subsets of size  $k$  from the entire set of  $n$  roads. We refer to this data split strategy as “held-out.” We then train a model on each of the subsets and test it on the  $(n - k)$  roads that were excluded from the training set. We report our results in table 5.4 and Figure 5.9 as the average of the distributions in these sets. This training strategy ensures that the trained models are not biased towards a single set of roads. Note that while this sampling strategy is much more realistic and useful for real-world applications, it significantly increases the difficulty from a machine learning perspective, as it means that the train and test set are no longer drawn from the same distribution.

### 5.2.3.3 From Convolutional Neural Networks to Vision Transformers

Convolutional neural networks (CNNs) [105, 203] have achieved great success in solving computer vision and image classification tasks, including remote sensing [136, 213, 217]. This is largely attributed to spatial feature preservation through a series of convolution functions. Vision transformers [69, 105, 142] have recently shown promise in solving several visual challenges. Vision transformers use self-attention mechanisms to process an entire image as a sequence of image patches, as compared to convolutional operations synonymous with CNNs. This enables visual transformers to focus on the most pertinent features of an image for a given task and might be of particular use in this application given the relatively small proportion of a satellite image that a road might cover. Despite their superior performance over CNNs, vision transformers require a significant amount of training data, are computationally expensive, and are often large in size [224, 15]. Consequently, they are not always suitable for solving tasks related to sustainable development in the global South, where data can be scarce or completely unavailable. There are,

however, several disadvantages associated with CNNs, including difficulty capturing relationships in sequenced input data, overfitting, sensitivity to hyperparameters, and fixed input sizes, all of which can result in information loss when working with low-resolution satellite images.

In order to experimentally determine which model be best for our task, we train a mixture of CNN and Vision transformer models; ResNet[117], ViT[69], Data-Efficient Transformers (DeiT)[273] and ConvNext[165] on identical data and road classification tasks. First, we train a ResNet model as our baseline model. This baseline builds on previous work by [40], who demonstrated that CNNs could perform well on high-resolution satellite imagery. Next, we use a pre-trained vision transformer available through the hugging faces model repository[291]. This vision transformer was pre-trained on ImageNet-21k, a dataset of one million natural images across 21,843 classes, and finetuned on ImageNet 2012, a dataset of one million natural images across 1,000 classes[61]. Further fine-tuning (using lower-/higher-resolution satellite imagery) did not result in significant performance gains. A major issue when applying machine learning to solve problems related to developing regions is data availability (both on the ground and remotely sensed data). As such, vision transformers may not have the necessary data. Therefore, we train and evaluate Data Efficient Transformers (DeiT) based on different train/test data splits. Data efficient transformers are a variation of the vision transformer that allows computer vision tasks to be performed on smaller datasets.

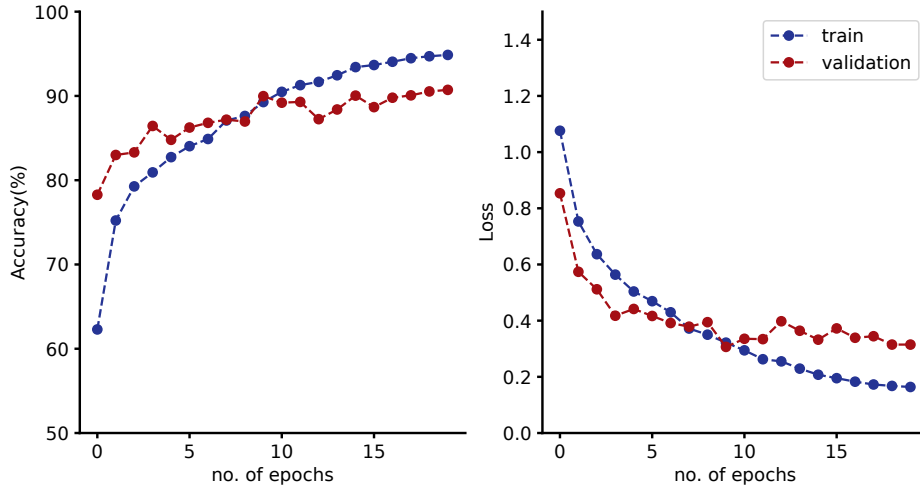
Consequently, they are suitable for computer vision-related tasks that target developing regions. In this work, we use a DeiT model pre-trained and finetuned on ImageNet-1K, one million natural images across 1000 classes, available on the hugging face model hub. Further fine-tuning using satellite imagery did not yield a significant performance improvement. Finally, we use the new Convnext method that combines elements of both CNNs and vision transformers. Thus, we evaluate the performance of ConvNext models against ViT, DeiT, and Resnet models. The initial layers are effective at extracting local

features from input data (inspired by CNNs), while the last layers are designed to capture global dependencies and relationships in the data, leading to improved performance, particularly on low-resolution and heterogeneous training data. Here, we use a ConvNext model pre-trained on ImageNet-224, 14 million natural images across 21841 classes. Finally, we report our findings in Section 5.2.4.

#### 5.2.4 Results and Discussion

The primary objective of this study is to predict the quality of roads based on medium-resolution satellite imagery. To achieve this, we trained a series of machine learning models defined in section 5.2.3.3 and reported the results in table 5.4. The data included in table 5.4 includes average AUROC scores for both models trained using both completely held-out data (non-IID) and standard train/test split data (IID). All models were initially trained on ImageNet data [61] and then fine-tuned using Planet or GEP satellite imagery. We wanted to compare the performance of both models when trained on non-IID and IID data. We trained all models on a single NVIDIA Tesla A100 GPU with 512 GiB of RAM. The vision transformer-related models were trained for 15 epochs, as we found that, on average, additional epochs did not produce any notable performance improvement (Figure 5.7). On the other hand, we trained the baseline CNN for 50 epochs as additional epochs yielded no further improvements in model performance.

The data in table 5.4 indicate little appreciable difference between the three vision transformer models. We see similar results for all three when we compare results for the same classification tasks: 2-class classification and 5-class classification; data split strategies; standard and held-out; and data sources; GEP and Planet Lab. In contrast, the baseline CNN model (ResNet) performs worse on binary and five-class classification tasks. Given that previous research (e.g., [40]) demonstrated that CNNs could perform well when trained on high-resolution satellite imagery (50 cm), we attribute the poor performance to the low resolution of our current dataset, rather than the computational technique or the



**Figure 5.7:** Model Accuracy and loss curves across training and validation data splits (at 50% of all roads ) for the ViT model. Performance measurement curves follow the same trend across all transformers models defined in section 5.2.3.3 and data split percentages.

amount of data. Our current data is, on average, three times (GEP) - six times (PLD) lower in resolution than the data used in [40]. Consequently, we are confident that increasing the resolution of the dataset would significantly improve the performance of all models, and particularly the base CNN model.

Model	2-class		5-class		2-class		5-class	
	Planet Lab		Planet Lab		GEP		GEP	
	I.I.D	Held-out	I.I.D	Held-out	I.I.D	Held-out	I.I.D	Held-out
ResNet (CNN)	0.626	0.574	0.502	0.500	0.608	0.503	0.500	0.499
ViT	<b>0.872</b>	<b>0.741</b>	0.661	0.525	<b>0.934</b>	<b>0.680</b>	0.685	<b>0.505</b>
ConvNext	0.869	0.732	0.630	<b>0.529</b>	0.932	0.670	0.687	0.501
DeiT	0.870	0.739	<b>0.685</b>	0.527	0.925	0.660	<b>0.702</b>	0.504

**Table 5.4:** 2-class & 5-class mean AUROC for production results under standard train-test and held-out conditions for original CNN (ResNet), CNN inspired by Vision transformers (ConvNext)and, and vision transformer models trained on Planet Lab and GEP Datasets.

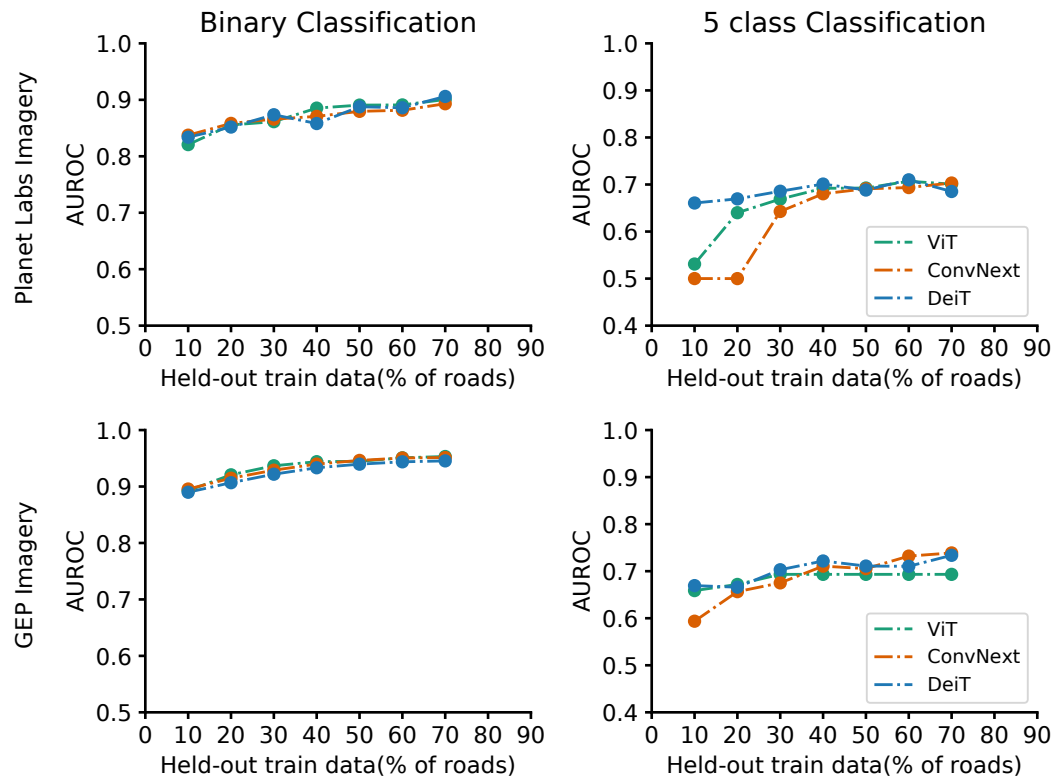
Focusing on transformer models, we see another few trends emerging. The results for i.i.d data splits are superior to those of the held-out case since, in the latter, the distribution of the train and test set are no longer the same. However, given how fundamental this

assumption is to machine learning methods, the results (particularly for binary classification) are very strong. Another interesting trend is that while the i.i.d results for GEP data is stronger than Planet, the opposite is true when looking at the case with fully held-out roads. This can be explained by considering the fact that GEP has, on average, better resolution, but the distribution of image sources is location-dependent. As such, when an equal amount of training data is available from all places, the transformers can take advantage of this higher resolution, but in the held-out scenario, there may be cases where one imagery source is only available in the train or test set. This would generalize much harder and probably result in the lower scores we see. Conversely, models trained on Planet imagery have a much more modest reduction in quality when going from i.i.d. to held-out scenarios.

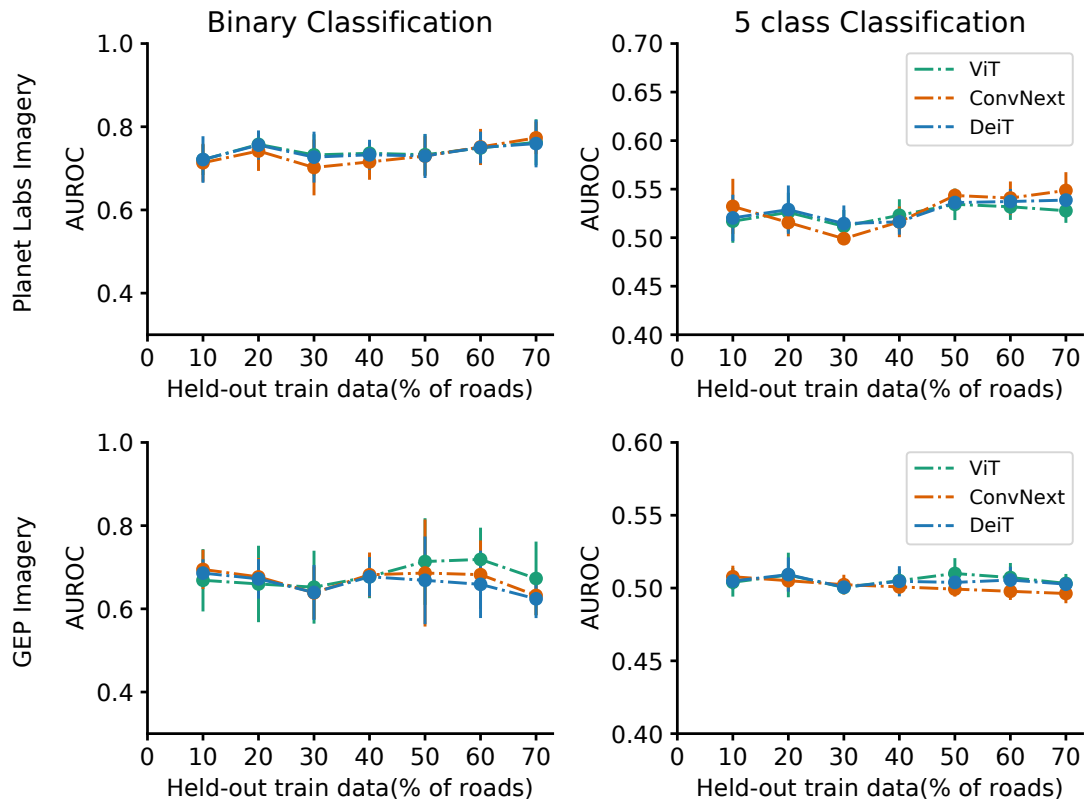
#### **5.2.4.1 Impact of data size on model performance**

Next, we set out to empirically determine the impact of data scarcity on the performance of different models. Knowing the sensitivity of models to data constraints is particularly important for developing region applications where large sets of label data can be hard to come by. We compared the performance of different models trained on the same amount of data from our two data sets to see if different resolutions ( $\sim 1.6\text{m}/\text{pixel}$  vs  $3\text{m}/\text{pixel}$ ) impacted the degradation in performance. We hypothesized that the lower the resolution, the more significant the degradation in performance caused by data scarcity. We also wanted to determine the amount of data necessary to create satisfactory predictions, both when the data was IID (identically and independently distributed) and non-IID.

To do this, we compared the performance of models with different train/test ratio splits. According to Figures 5.8 and 5.9, both standard (IID) and held out (non-IID) data splits yielded similar AUROC scores. Based on both figures, we observed that performance improved linearly as training data percentages increased from 10% to 70%. As would be expected, this is particularly evident in the standard data split (Fig 5.8), where there is an



**Figure 5.8:** Graphs showing AUROC scores for 2-class and 5-class classification results on I.I.D roads from the identical road segments. Top: Planet Lab imagery (3m/pixel) and bottom: Google Earth Engine scrapped Imagery (~1.6m/pixel). The X-axis on all graphs represents the number of roads (as %) used for training the models.



**Figure 5.9:** Graphs showing AUROC scores for 2-class and 5-class classification results on held-out roads from the identical road segments. Top: Planet Lab imagery (3m/pixel) and bottom: Google Earth Engine scrapped Imagery (~1.6m/pixel). The X-axis on all graphs represents the number of roads (as %) used for training the models

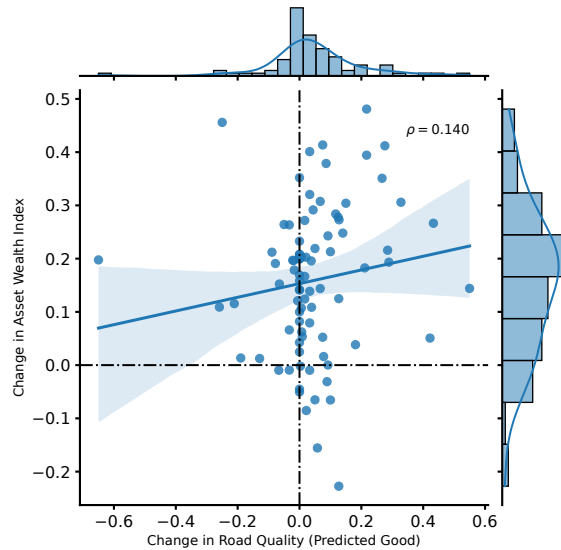
identical distribution of locations and labels across both the training and testing data. The same positive trend can still be seen in the held-out data but is less recognized since the fact that entire roads are held out in the test set remains true regardless of the amount of data.

One interesting pattern that emerges is the differences between the models as a function of data size: we see some of the transformer models performing much worse when the data size decreases. This is particularly noticeable for the 5-class classification problem (which is significantly harder than the two-class problem), where convNext and ViT are a little better than random guessing until we use more than 30 or 40% of the data for training purposes. Notably, we observe that the DeiT model (explicitly designed to operate with low levels of data) does not suffer from this problem. This validates the hypothesis that DeiT might be more robust in a data-poor environment and that the relative difficulty of the problem (e.g., 2 class vs 5 class) mediates the impact of data scarcity. Based on these observations, we conclude that large amounts of low- to medium-resolution satellite imagery can produce more accurate vision models than small amounts of high-resolution and heterogeneous satellite imagery.

### **5.2.5 Case Study 1: Correlation between road quality and household asset wealth**

In Section 5.2.4, we demonstrate that ML models trained on medium-quality satellite imagery could produce high prediction accuracies when applied in previously unseen settings. To demonstrate the potential applications of these findings on a scale, we performed a further analysis to examine the relationship between infrastructure quality and household wealth in a developing region. Specifically, we are interested to see how changes in infrastructure quality correlate with changes in household asset wealth. A household asset wealth measure consists of a household's ownership of items such as televisions and bicycles, building materials, types of water access, and sanitation systems, among others. We recognize that road quality is only one of the potentially many factors that





**Figure 5.10:** Comparison of change in road quality metric against change in Household Asset Wealth Index between 2016 and 2020. A line of best fit ( $R^2 = 0.14$ ) is plotted to represent the relationship between the two metrics. There is a weak positive correlation between changes in the predicted quality of good roads and changes in asset wealth within this period. In general, small changes in road quality are associated with small changes in asset wealth, whereas large changes are highly correlated with substantial changes in asset wealth.

may correlate with or even influence asset wealth, so we do not expect perfect agreement between the two measures. Furthermore, we acknowledge that asset wealth is one of many different potential measures of income, consumption, and wealth of communities that can indicate economic activity; we have selected this particular measure because of the availability of data sources documented below.

For the case study, we randomly sample a subset of all towns throughout the Republic of Kenya ( $n=85$ ) having road location data available in OpenStreetMaps as of 2016. A significant amount of infrastructure development is carried out in major cities and towns regularly; consequently, the results cannot be easily quantified. Therefore, we exclude major towns and cities from our analysis. To obtain household asset wealth information, we use an Asset Wealth Index (AWI) layer from Atlas AI. The dataset consists of annual estimates of individual household asset wealth based on asset ownership at a resolution

of 2 km/pixel and is produced from a deep learning model that predicts survey-based estimates from satellite imagery [299, 136]. This dataset is crucially available at both a high spatial resolution and periodic temporal resolution, making it appropriate for considering longitudinal changes. For satellite imagery, we select all roads within a radius of 2km of these towns and sample  $k$  image patches ( $k = 30$ ) for each road segment within this radius. We then select the highest quality cloud-free images, at 3m/pixel resolution, collected between the months of July and August across two years: 2016 and 2020. This provides a before-and-after view of roads needed for our model's inference.

In order to perform inference, we select the model with the highest AUROC score for binary class classification predictions on Planet Imagery (3m/pixel resolution) from Table 5.4. Following that, we calculate the difference in the percentage of roads classified as “good” in each town between 2016 and 2020. In our opinion, this metric score can indicate either improvement or deterioration in road quality. Finally, we compare the change in the percentage of roads predicted as “good” with the change in the asset wealth index over the same period of time.

Figure 5.10 illustrates the relationship between the change in road quality and the asset wealth index for this set of towns. Our analysis reveals that there is a small but statistically significant positive correlation between changes in road quality and changes in asset wealth between 2016 and 2020. Generally, a small change in road quality is associated with a small change in asset wealth, while a large change is highly correlated with a substantial change in asset wealth. Despite this, it should be noted that we cannot claim the existence of a causal relationship between changes in road quality and changes in asset wealth. Several factors may contribute more to asset wealth than roads, such as education, access to markets and financial services, the quality of the workforce in the selected towns, other nearby economic activities, and the presence of other infrastructure. Nonetheless, even though this relationship is not perfect, it can still serve as a useful indicator of the extent

to which high-quality roads may contribute to improving the livelihoods of citizens living in low-income regions since they provide access to markets, jobs, and other services.

### **5.2.6 Conclusion and Future Work**

We have demonstrated that under realistic data conditions, such as heterogeneous data sources and intermediate spatial resolution, traditional convolutional neural networks and cutting-edge vision transformers can predict road quality from medium-resolution satellite imagery that is becoming broadly available. Furthermore, we found that with as little as 3 roads (10% of the data set), these models can achieve substantial accuracies, demonstrating that it is possible to train accurate models with limited and relatively low spatial resolution satellite imagery. These experiments and results suggest that by using the right combination of models and data sources, it is possible to predict road quality from remote sensing imagery accurately.

There are several aspects we hope to explore in the future. These include assessing infrastructure damages after natural disasters, facilitating logistical support for activities like last-mile vaccine delivery, and evaluating the performance of vision transformers when trained on satellite imagery at low resolution, Sentinel and LandSat imagery at 10m/pixel, and 30m/pixel, respectively, and exploring alternative machine learning approaches that perform better when trained on data from limited and heterogeneous data sources. This is because, while satellite imagery is increasingly available in the public domain, it remains difficult to acquire high-resolution imagery of developing regions. As a result, we can assume that these will be the only true temporal-spatial datasets available for many years to come, which makes them extremely valuable for longitudinal studies. Consequently, improving the ability to gain insight from low-resolution imagery could provide governments with much-needed evidence-based input in the short term to assist them in achieving sustainable development objectives.

Machine learning and remotely sensed data can enable us to monitor, track, and address complex environmental, economic, and social issues related to sustainable development. By utilizing machine learning and remotely sensed data, governments and policymakers can accurately map and measure key infrastructure, such as roads and bridges, to assess and understand their location, condition, and connectivity, as well as the roles they play in the socio-economic development of certain regions. These data can then be used to inform decisions about investments in infrastructure, as well as track their impact on the local economy and environment. By investing in infrastructure improvements, governments can create jobs, reduce poverty, and improve access to health care, education, and other services.

### **5.3 Evaluating Road Quality over Time Using Satellite Imagery to Assess the Long-term Effects of Infrastructure Investments in Eastern Democratic Republic of Congo**

#### **5.3.1 Motivation**

Based on the lessons learned in sections 5.1 and 5.2, we select the best of the ML techniques and apply them to a real-world case study; predicting road quality from satellite imagery in regions where we do not have ground observations. Specifically, we train the latest state-of-the-art vision transformers on data from different economies (the Republic of Kenya and the Republic of Liberia) within the global south that share significant commonalities in their road infrastructure. We then use these trained models to predict road quality (using the scale from the International Roughness Index) in regions without ground labels, focusing on four multi-year road construction and rehabilitation projects in the eastern provinces of the Democratic Republic of Congo (formally Zaire). Our experiments and results demonstrate the efficacy of combining ML and remote sensing to assess progress towards the Sustainable Development Goals in a low-cost manner. This insight can guide future research and applications that use satellite imagery, especially in the

Global South, where datasets are typically scarce or available only at poor spatiotemporal resolutions.

### **5.3.2 Background and Related Work**

The Democratic Republic of Congo (DRC) has been subject to longstanding and devastating conflicts, triggering a humanitarian crisis of staggering proportions [52, 260, 173]. The Eastern regions of the country (defined as the provinces of North and South Kivu, Ituri, Katanga, Maniema, and Tanganyika), are some of the most mineral-rich regions [188] of sub-Saharan Africa and have been the site of conflicts between government forces and various armed groups since DRC (formally Zaire) gained independence from Belgium in 1960. The area's strife is deeply rooted in complex socio-political factors, including ethnic tensions, competition over access to rich mineral resources, and a power vacuum from weak state control. Armed groups, both local and foreign, leverage these factors to create a state of constant instability, resulting in widespread violence, forced displacement of communities, and gross human rights abuses. The impact on the civilian population is profound and multifaceted, leading to acute food insecurity, limited access to essential services such as education and healthcare, and disrupted economic activities. Thus, the persistent conflict in Eastern DRC presents a significant challenge to regional peace and stability and critically impedes the nation's progress towards sustainable development and attaining human security for its population.

The World Bank has invested billions of dollars in infrastructure in the Eastern part of the Democratic Republic of Congo (DRC) to boost the region's socioeconomic recovery substantially [6, 56, 118]. Improved infrastructure: roads, bridges, ports, and utilities have enhanced connectivity within the region and the neighboring countries, fostering trade, facilitating the movement of people and goods, and opening up isolated communities to new opportunities. These developments have significantly boosted local economies by stimulating sectors such as agriculture and small-scale (artisanal) mining, making them

more productive and efficient and creating job opportunities crucial for poverty reduction. Consequently, these investments have upgraded essential services such as water, sanitation, and electricity, profoundly impacting public health and quality of life, reducing disease, and raising living standards. Finally, building and investing in good quality infrastructure has provided immediate benefits, offering employment opportunities, spurring demand in related industries, and injecting money into local economies. Given the backdrop of the Eastern DRC's conflict-impacted landscape, these investments continue to have the potential to be not just an economic lifeline but also a cornerstone of peace-building and social stability.

However, large-scale infrastructure investments in Eastern DRC pose considerable barriers and disadvantages, mainly tied to the region's ongoing conflict and instability. First, with proper infrastructure, militias, and armed groups can smuggle weapons across borders and set up roadblocks to extort traders and the local population, thereby setting up a steady source of income to sustain these armed groups. Second, with improved infrastructure, there are significant environmental considerations, with the risk of large-scale projects causing or exacerbating environmental degradation or loss of biodiversity as large companies now have access to the hinterland. Third, without proper planning and community engagement, these projects can lead to forced displacement, local opposition, and exacerbation of social inequalities, as benefits may not be equitably distributed. Finally, without strengthening the local capacity for operation and maintenance, the sustainability of these investments could be compromised, leading to a cycle of disrepair and requiring further investments for refurbishment. Therefore, the inherent challenges make large-scale infrastructure investment a complex proposition for the Eastern DRC.

### **5.3.2.1 Road Quality Measurement**

Given the scale and depth of investment that the World Bank undertakes in the global south, it is paramount to regularly measure and monitor these investments' performance and their overall socio-economic impact across time and space.

In the context of the Eastern Democratic Republic of Congo, regular monitoring and measurement of infrastructure becomes even more vital. The region, marred by years of conflict and instability, must ensure its infrastructure investments yield maximum benefits. Regular monitoring and measurement can aid in maintaining the functionality and safety of the roads, which is critical in a region where transportation networks can be lifelines for communities cut off by conflict or geography. This process can help prevent the deterioration of roads due to weather conditions or heavy use, which is crucial for uninterrupted access to markets, healthcare, and other essential services. In a region where resources are scarce and the need for development is high, data from regular monitoring can guide policy-makers and donors in making informed decisions on where to channel their investments. It can also foster transparency and accountability in using funds, a significant factor in a region where corruption can be challenging. Finally, regular monitoring and measurement of infrastructure can contribute to peace-building efforts, as well-managed and equitable infrastructure development can help alleviate some of the socio-economic tensions that feed into the region's conflict.

### **5.3.3 Methodology**

#### **5.3.3.1 Dataset**

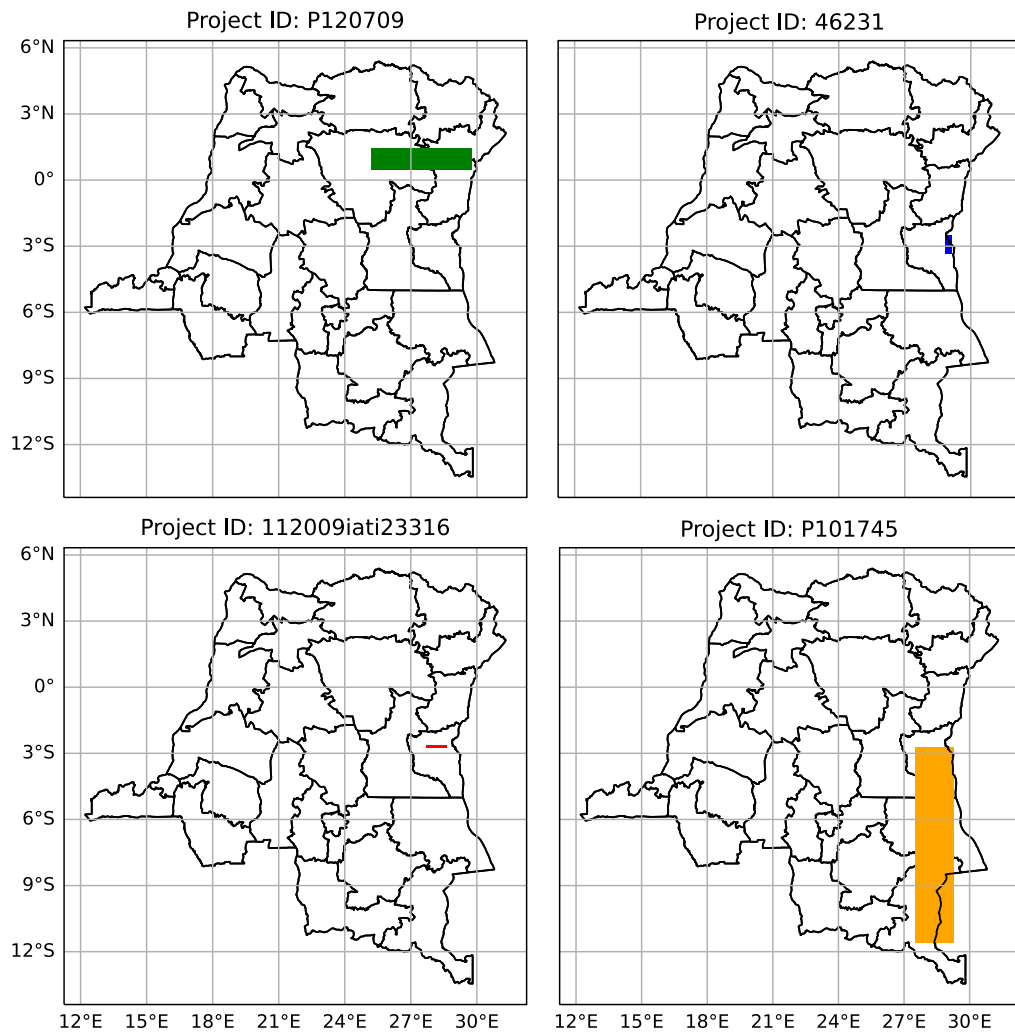
We aim to predict road quality across four major road construction and rehabilitation projects in the Eastern Democratic Republic of Congo (see Figure 5.11): Project 112009iati23316 (hereafter referred to as Project 1), Project P101745 (hereafter referred to as Project 2), Project 46231 (hereafter referred to as Project 3), and Project P120709 (hereafter referred to as Project 4). However, we do not have ground truth data, i.e., International Rough-

ness Index (hereafter referred to as IRI) measurements. As such, we train a set of machine learning models with data from different economies within the Global South (the Republic of Kenya and the Republic of Liberia) that share significant commonalities in their road infrastructure and then use these trained models to perform temporal-spatial inference based on satellite imagery.

To accomplish this objective, we matched patches of satellite imagery with spatial locations of IRI measurements across the three countries. These measurements were recorded across different years and climatic conditions. These IRI measurements represent the diversity and heterogeneity in road quality common to the Global South. By doing this, we can compare the before and after images to better understand how road quality changes over time and how it is affected by different environmental factors. This approach allows us to more accurately measure the impact of road quality on economic outcomes in these countries. For satellite imagery, we use publicly available data scraped from Google Earth [72] (hereafter referred to as GEP) at 0.6m/pixel. Satellite imagery, particularly GEP, is derived from diverse data sources (Maxar, Airbus) with varying spatiotemporal resolutions, sizes, and visibility. These inherent characteristics reflect the realistic conditions under which most remotely sensed data, especially in the Global South, is acquired and made available.

To validate the model predictions in DRC, we use a combination of monthly reports by Humanitarian Organizations such as the World Food Program (WFP) and International Peace Information Services (IPIS). These organizations provide a wealth of reliable data encompassing areas like impacted communities, new roadblocks, locations of new artisanal mining sites, and the intensity of conflict, empowering us to verify the precision of our predictive models. Furthermore, implementing a human-in-the-loop strategy permits identifying and rectifying any inaccuracies or inherent biases present within the predictive outcomes of our model.

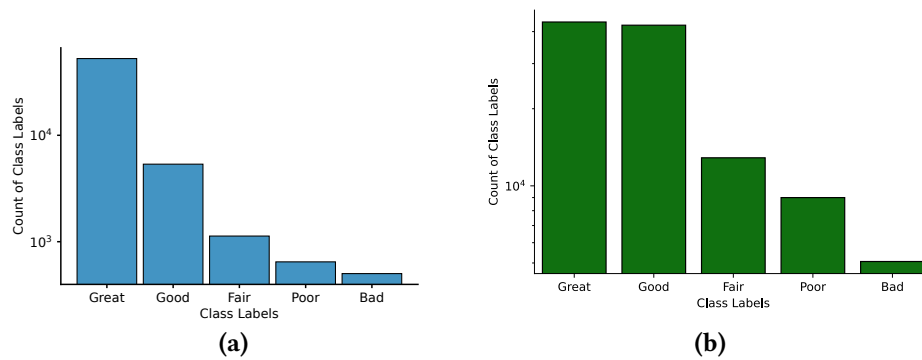




**Figure 5.11:** Map of Democratic Republic of Congo showing the location of the four projects used in the study

### 5.3.3.2 Transfer Learning

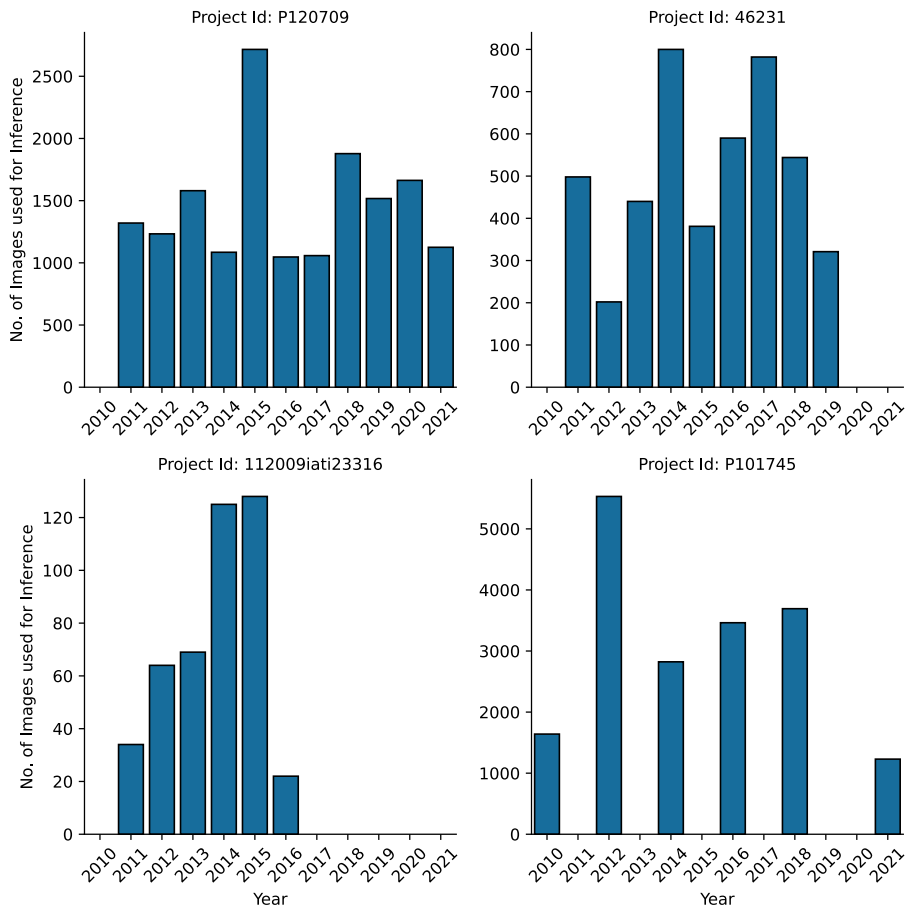
Our ultimate goal is to predict the quality of roads across four major road construction and rehabilitation projects in the eastern part of the democratic republic of Congo from satellite imagery. However, we do not have ground data (IRI measurements) to validate our predictions. We train and validate machine learning models on data from regions with satellite data and ground observations, i.e., Kenya and Liberia. Then, using the learned characteristics, we perform inference via transfer learning. Transfer learning [211, 309] is a machine learning technique that allows us to use the knowledge learned from a task to improve the performance of a model on a related task. In road quality prediction, we can use transfer learning to train a model on a large dataset of road imagery of roads in one region and then fine-tune the model on a smaller dataset of satellite imagery of roads in a different region [33, 295]. This can help improve the model's performance on the new dataset, even if the two datasets are not drawn from the same data distribution (non-I.I.D). In the context of road quality prediction, where we do not have ground observation, transfer learning is useful because it can help overcome the challenges of collecting ground observations. Road quality-related datasets (i.e., IRI) are often expensive and time-consuming to collect across time and space. As such, transfer learning can help to address these challenges by allowing us to train models on smaller, more easily obtained datasets. However, the performance of transfer learning models often depends on the similarity between the source and target tasks. If the two tasks differ, the transfer learning model may not perform as well as a model trained from scratch. Secondly, transfer learning can be computationally expensive, requiring training two models (the source and the target models). Despite these limitations, transfer learning is especially applicable for road quality prediction from satellite imagery without ground data and can improve the performance of machine learning models even when there is limited data available.



**Figure 5.12:** Graphs showing the distribution of labels in the Kenyan(left) and Liberia(right) dataset. Both datasets follow the same distribution. We observed that Labels associated with “great” road quality contributed the largest percentage of the distribution.

### 5.3.3.3 Handling Imbalanced Data

Both of our reference datasets (Kenyan and Liberia) are highly imbalanced in light of the availability of IRI measurements and the limitation of the measurements to measurements collected in 2015 and 2016, respectively. We observed that the most frequent labels are 0 to 7 (Figure 5.12). However, there are various ways to handle imbalanced data in machine learning, including re-weighting class labels, oversampling the smaller distribution, under-sampling the more significant distribution, and data enhancement, among others. We will use a combination of different strategies to increase the label size of minor classes without over-fitting. Specifically, we will explore two approaches: sampling distribution of the majority classes in relation to the size of the minority classes (over/under sampling) and assigning significant weight to minority classes and small weights to majority classes (class weighting). Additionally, we will perform data augmentation: artificially increasing the amount of data by generating new data points from existing data or making minor modifications to data (such as flipping images or reducing their contrast) to increase the size of minor classes in the distribution. It’s important to note that our data extraction from GEP didn’t yield a consistent number of images for each year we performed inference. Consequently, we represent the predictions as percentages across the defined classes to ensure a consistent interpretation.



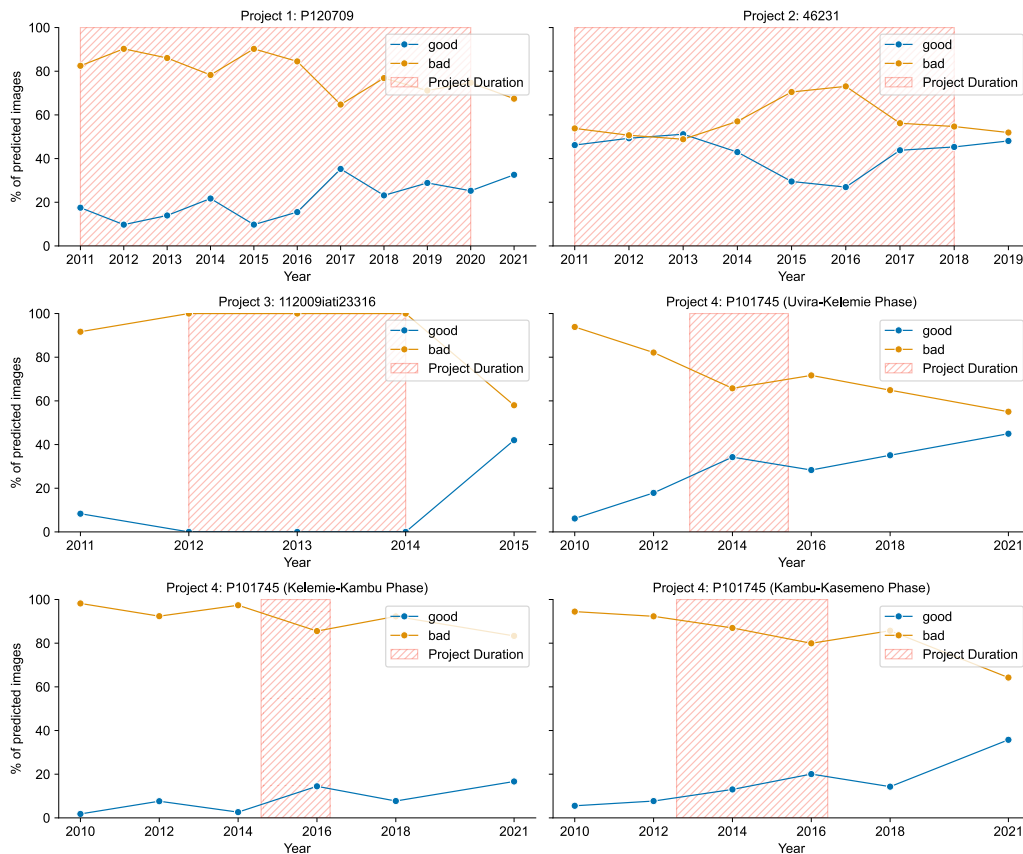
**Figure 5.13:** Bar plots showing the count of images used for inference in each of the four projects over several years. The number of images fluctuates annually, stemming from the temporal-spatial variability in image availability on Google Earth Pro (our data source)

#### 5.3.4 Results and Discussion

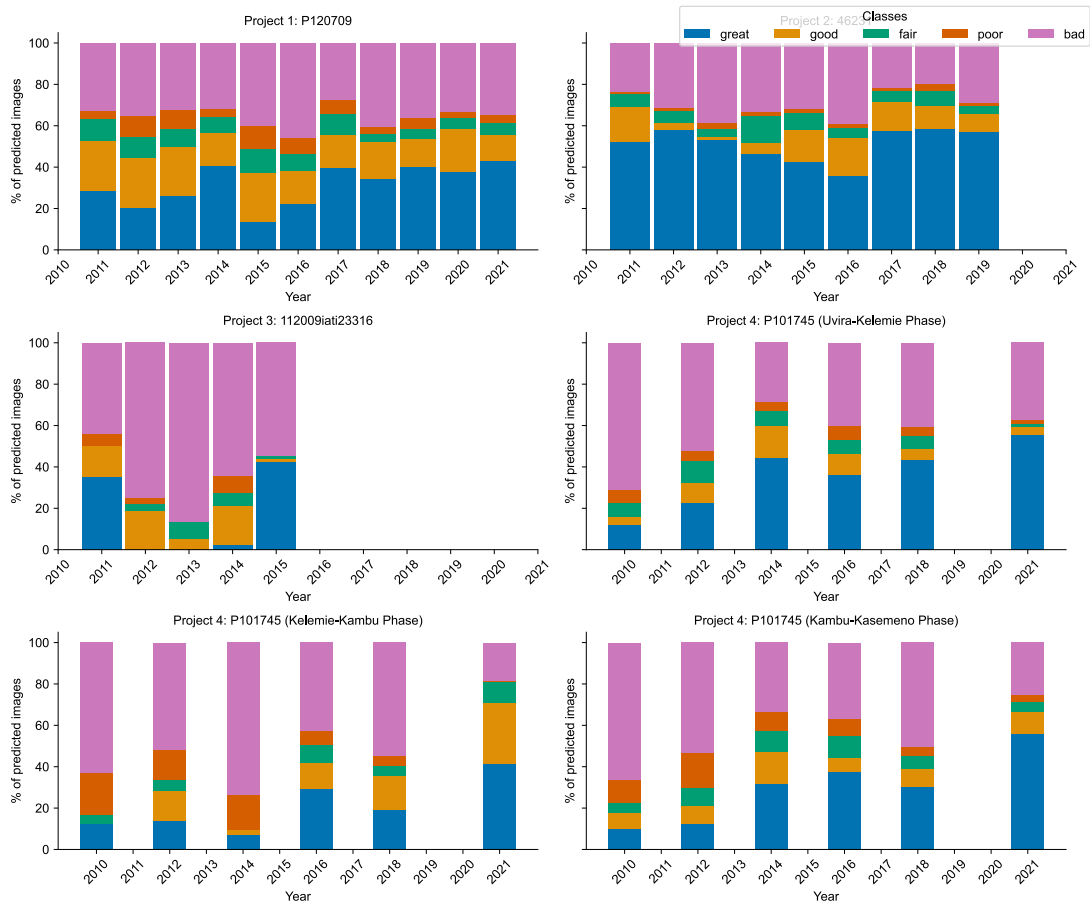
The primary objective of this study is to predict the change in road quality across six major road construction and rehabilitation projects in the eastern part of the Democratic Republic of Congo (DRC) from satellite imagery. Given each project's start and end date, we analyze changes in road quality across time from satellite imagery. The predictions were made using a Data Efficient Transformer (DEiT) machine learning model that was fine-tuned on the road quality-related data (Section 5.3.3.1) from two countries: The Republic of Kenya and the democratic republic of Liberia. This section offers a broader context of road infrastructure development in the eastern DRC, offering insights, discussions, implications for future development, and quantifiable metrics to analyze the impact of good quality infrastructure on peace, rehabilitation, and the socio-economic well-being of people in the eastern DRC.

We performed road quality assessments using two classification methods: binary (Figure 5.14) and five-class (Figure 5.15). Our primary assessment, using binary classification – identifying good segments and bad segments – shows that road quality consistently improved during the construction period for all six projects, with most roads classified as “good” by the end of each project duration. Our five-class classification assessment method provides more granular insights but shows the same trends as the binary approach. While the five-class method offers additional nuances, the binary method will be central to our further interpretation, as it is easier to interpret and has shown better accuracy in classification relative to ground truth in past experiments in Kenya and Liberia. Across the six projects (Figure 5.14), we observe a significant transition from a larger proportion of “bad quality” road predictions to an increase in “good quality” predictions across time (pre-project inception, project duration, and project completion).

The predictions for Project 1 reveal that the road construction and rehabilitation project started with the road in poor condition, as evidenced by a high percentage of “bad” classifications at the project's outset. Throughout the project, there is a notable shift in road



**Figure 5.14:** Graphs showing binary classification predictions for selected projects. The y-axis indicates the percentage of image patches classified as “good” or “bad”. The shaded area marks the project’s timeline from inception to completion. A noticeable trend reveals improvement in road quality throughout individual project durations, with higher percentage of patches predicted as “good” towards and beyond project completion



**Figure 5.15:** Bar plots showing the five-class classification predictions across the six projects. The distribution of predictions among the five classes is depicted on the vertical axis for one year. A consistent uptick is observed in the percentage of patches categorized as “great” over time. This trend is especially evident in projects 2 and 4. In contrast, project 3 exhibits a relatively consistent road quality throughout the project duration.

improvements. The initial years highlight a pressing need for road quality enhancement, which appears to have been progressively addressed throughout the project's span.

The predictions for Project 2 are slightly oscillatory but indicate a general improvement in road quality over time. The oscillations could be indicators of periodic phases of development, other internal/external factors, or noise resulting from image variability. Such variations could correspond to funding cycles, political shifts (e.g., changes in the ruling parties), political instability (wars), changes in climate variables (e.g., long wet seasons), or unforeseen challenges encountered after project inception. In regions such as the eastern Democratic Republic of Congo, these variations are attributed to political transitions, the discovery of new mining sites, or even environmental variations (e.g., long rainy seasons). On the other hand, such variations could indicate that this was a pilot project. Lessons from this project could be invaluable for understanding the dynamics of short-term interventions and their impacts on the socioeconomic and political dynamics of countries with similar patterns.

The predictions for Project 3 show a relatively stable temporal change in road quality with minor fluctuations in the percentage of road patches predicted as "good" and "bad." The temporal consistency suggests a stable phase in the project duration, likely pointing to the fact that the project was either building upon an already established infrastructure (e.g., road upgrades/asphalt pavement rehabilitation) or that the project interventions did not significantly alter the existing road quality. The relative stability observed across Project 3 could also indicate that the road had reached a plateau regarding road quality. This could indicate that while maintaining existing standards is necessary, future investments in this region (or regions within the same socio-economic and political context) could focus on more innovative solutions (e.g., stronger bridges and multi-lane roads) to improve road quality.

Taking a deeper look, we separated the length of Project 4 into three physical stretches for which we had specific duration data for construction phases: Uvira to Kelemie (385Km),



Kelemie to Kambu (146Km), and Kambu to Kasemero (619Km). By overlapping the construction periods of the phases, predictions for Project 4 indicate a significant increase in road patches predicted as “good quality,” starting with the later years of the project duration phase. This change could have resulted from possible interventions or developments during this period. A combination of a more substantial change in the road quality at the project inception and the extended period of the entire project suggests the possibility that the project was potentially either building up a relatively higher quality unpaved road (e.g., grading a Murram road) or the project potentially aimed to upgrade the pavement material (e.g., to an asphalt road). The consistent positive trajectory indicates sustained efforts and successful intervention.

Projects 1 and 4 demonstrate a clear improvement in road quality across the respective project time spans. Such trends have been observed in other infrastructure projects in the Global South [21, 68], where initial investments yield significant improvements in road quality over time. This significant improvement is attributed to the initial low-quality infrastructure getting major upgrades. The upward trend could be a strong indicator that the region is on the cusp of socio-economic and political stability, most likely driven by improvement in infrastructure. As such, the World Bank and other stakeholders could study the strategies for these projects to replicate their successes in similar contexts across other regions.

High-quality roads are pivotal in any region’s socioeconomic and political development. This is especially important for regions in the Global South, such as the eastern part of the Democratic Republic of Congo, which has faced economic hardships and political instability since gaining independence in the 1960s. As such, the upward trends in Projects 1 and 4 might indicate successful infrastructure investment, potentially leading to positive economic outcomes, peace, and political stability in the region. Additionally, the pronounced improvement in some projects indicates the importance of targeted interventions, policy changes, or substantial investments during these periods. As such, understanding the



**Figure 5.16:** The images from Project ID: P101745 depict the condition of various roads before and after construction and rehabilitation. These examples highlight situations where the model saw the most substantial changes in road quality. The model effectively highlights the transformations, including the transition from murrum to asphalt and the grading and widening of the murrum roads.

specifics of these interventions could provide insights into successful strategies for road infrastructure development.

Further, the relative stability in predictions of Project 2 could be indicative of either consistent road quality or consistent methodologies. The former suggests that road maintenance and rehabilitation maintained a consistent quality across time. In contrast, the latter suggests upgrading the prediction methods, such as increasing the data quantity (as shown in Figure 5.13) to capture the evolving road conditions the model has failed to capture. Finally, validating these predictions, such as visual inspection (Figure 5.16), on-ground assessments, and liaising with local communities and non-governmental organizations that work in the regions where these projects have been implemented, is paramount to provide a more holistic picture of the change in road quality across time as well as the resulting impact of these infrastructure changes on the socioeconomic development and political stability of the local communities. Delving deeper into each project's specific details: budget, interventions, technologies, funding policies, and other contextual factors

can provide a richer understanding of the observed trends. This is especially important if these predictions and recommendations will influence future policies and funding opportunities that directly impact the socioeconomic well-being of the local communities.

Our method provides unprecedented detail on road quality in both spatial and temporal dimensions. This can be especially valuable to policymakers, investors, business operators, and researchers for many purposes, including but not limited to monitoring construction quality and contractor performance, assessing the effects of seasonal weather on degradation, recomputing travel times to assess market access, and quantifying the impact of road quality on economic or social outcomes. However, it is important to recognize the limitations of our approach – while our previous work has shown the performance of our binary and five-class techniques to be strong enough for a wide array of applications, we note that those results were obtained in the same country as the training data (though in different regions that were “held out” from the training sample). In this work, we use observations from other countries to predict in an entirely unseen country – this “transfer learning” approach cannot be quantitatively evaluated in the absence of local ground truth data (i.e., from DRC); instead, we resort to validation using imagery (such as in Figure 5.16. Further, the availability of imagery is another notable limitation – in this study, we attempted to obtain imagery every 1 to 2 years for each project. While there may be additional images that can produce further data points for analysis, we do not expect to have substantially more than one image per year for each of the road segments. Ultimately, this limits the ability of our method to quantify seasonality effects on a spatially granular basis. Despite these limitations, we believe that we have demonstrated a truly novel capability that can have valuable impacts on the monitoring and evaluating of transportation projects in remote regions.

### 5.3.5 Conclusion

Road quality prediction is more than just a measure of infrastructure; it is a cornerstone of socio-economic development and a potential harbinger of peace, particularly in the Democratic Republic of Congo, which has been marred by political instability for over 60 years. A well-maintained road network fosters connectivity, facilitates trade, and can deter regional conflicts by enhancing accessibility and mutual dependencies. However, roads are susceptible to degradation due to the vagaries of climate, especially in tropical regions like the Eastern part of the Democratic Republic of Congo. Recognizing this, institutions like the World Bank allocate millions of dollars annually towards constructing and rehabilitating infrastructure in these regions. Given the profound impact of road quality on socio-economic metrics and peacekeeping, it becomes imperative to monitor road quality across time accurately. This foresight enables timely maintenance, repairs, and rehabilitation, ensuring the continued benefits of a robust infrastructure. The selection of projects highlighted in this study was driven by their political and socio-economic implications, geographical diversity, and varied project durations. The insights and lessons learned from these projects can guide the World Bank and other stakeholders in crafting strategies for infrastructure planning, execution, and adaptability across diverse economies in the Global South. In order to maximize the efficacy of future infrastructure projects, combining these insights with other political and socio-economic indicators is crucial. This integrated approach can hopefully be instrumental in shaping the infrastructure endeavors of the World Bank and emerging economies.

## CHAPTER 6

### CONVERTING FISHING BOATS FOR ELECTRIC MOBILITY TO SERVE AS MINI-GRID ANCHOR LOADS

#### 6.1 Motivation

<sup>1</sup>Electricity is an increasingly indispensable ingredient to modern lifestyles. While governments, donors, non-profits, and private companies have made enormous strides in recent decades in improving electricity access around the world, over 1 billion people still remain without access to electricity at their homes [132]. Historically, the only pathway to electrification was extension of existing grids, but the emergence of decentralized generation and storage, coupled with sensing and communication technologies for better system management, have enabled a variety of electrification pathways beyond grid extension. Chief among these are solar home systems – typically comprised of an individual solar panel paired with a battery and a captive set of appliances – and minigrids – typically a microcosm of a centralized grid but with a smaller footprint, obviating the need for high voltage transmission. While solar home systems have recently gained substantial traction, with tens of millions of systems deployed globally already and an accelerating market [98], this electrification pathway faces a ceiling in providing for high-power “productive uses” of electricity. On the other hand, minigrids aim to find the “sweet spot” between the scalability of centralized grids and the decentralization of solar home systems. In fact,

---

<sup>1</sup>The thesis contributions in this chapter are a result of long term project on which Aggrey Muhebwa was the second author. This information has been included with permission and full cooperation from the first author (June Lukuyu [jlukuyu@uw.edu]. Aggrey Muhebwa’s role on the project included; designing and implementing the software used for monitoring and tracking the location of fishing boats, assembling and deploying the tracking devices on the fishing boats, collecting data and resetting the devices daily, and analyzing the collected data. Aggrey Muhebwa also assisted June Lukuyu in conducting surveys and liaising with stakeholders during the two weeks deployment on Lolwe Island in Lake Victoria)



**Figure 6.1:** Example fishing boat on Lolwe Island. Inset: boat tracking device attached to the boat before deployment.

the International Energy Agency (IEA) projects minigrids to be the linchpin for meeting universal electrification goals by 2030, as laid out in United Nations Sustainable Development Goal #7. In their World Energy Outlook 2017, the IEA’s “Energy for All” scenario for 2030 calls for 450 million people to receive access via nearly 200,000 minigrids, outpacing both grid extension and solar home systems as the most common source of new access [132]. Despite this rosy outlook, minigrids continue to face critical headwinds in gaining widespread traction. At present, the World Bank estimates that 47 million people are connected to 19,000 minigrids, most of which are hydro- or diesel-powered [75]. Few of these are privately owned, which is the primary ownership structure likely needed to enable such expansive scale [290]. The foremost impediment to achieving this scale is a lack of investment capital, driven by concerns about high costs for construction and operation as well as low demand for electricity, limiting revenue potential of these systems. In this paper, we focus on techniques for one strategy towards making private minigrid business models viable: demand stimulation. Encouraging growth in electricity consumption has crucial benefits to each stakeholder: electricity service companies increase revenues that allow for lower per-unit costs of electricity (ideally further increasing demand

for electricity) and electricity consumers can apply electricity towards improving livelihoods, both for household uses as well as for productive uses that grow income. While the research community has not yet been able to prove a causal link between electricity and economic growth [209], the correlation is undeniably strong; these measures track in lock-step worldwide: there are no low-consumption, high-income economies [36].

While demand stimulation has long historical precedent (as described in Section 6.2.2), advances in technology and changes in behavior are pushing demand stimulation techniques to evolve. In this paper, we study demand stimulation on a minigrid via electrifying fishing boats for a hybrid 600 *kWp* solar-battery-diesel minigrid on an island in Lake Victoria.

While our systems study deeply examines a particular setting and its attendant design and deployment challenges, we believe that our work has generalized utility. Minigrids have long sought anchor loads (e.g., telecom towers or irrigation pumps [226, 234]) to provide predictable demand and increased revenue. We extend this line of inquiry by studying tradeoffs among the multiple load classes that a financially-sustainable minigrid may encounter. Using this lens, our study maps to a variety of demand stimulation strategies that can be applied to the great range of settings where minigrids are found. Additionally, we characterize electric boats, a previously unstudied electric mobility load class that offers substantial promise for strengthening the electric systems of coastal communities worldwide.

Our study proceeds as follows: 1) We conduct surveys among fishing boat operators and outfit a set of fishing boats with custom tracking devices to understand the potential for adoption of this relatively new electricity technology and better understand boat usage patterns. 2) We use the insights from this in-the-field activity to size and identify an electric outboard motor and battery pack candidate, which we then incorporate into a model of electric mobility that embodies the range of usage patterns derived from our dataset. 3) We then evaluate electric mobility both technically – understanding the ability of this

system to meet user needs – as well as financially – characterizing the payback of such a system within the context of a privately-operated minigrid in our target environment, as both of these perspectives are crucial for adoption in such a setting. 4) We examine tradeoffs in incorporating this demand stimulation technique with the rest of the minigrid, which includes a range of domestic and small commercial customers as well as an ice manufacturing operation. 5) Having modeled the operation of the entire minigrid, we also consider the benefits of demand response via scheduled charging of boat batteries and the implications of an alternative target depth of discharge after charging. 6) We then discuss design considerations for a boat monitoring system given our observations from the target environment and conclude the study.

## **6.2 Background and Related Work**

In this section, we present background on the challenges of minigrid business models and demand stimulation and discuss some of the prior work on electric mobility as flexible loads and demand response strategies for optimizing grid operations.

### **6.2.1 Minigrids – Financial and Operational Challenges**

While there is substantial research literature on minigrids, the grand majority considers only grid-connected minigrids, which have few overlaps with the characteristics of grid-disconnected minigrids used in energy access scenarios like those we consider.

Despite the growing popularity of minigrids as an attractive alternative to grid expansion for providing power to rural and underserved communities, the long-term operation and management of minigrids to provide electricity to the poor faces considerable financial and operational challenges. A sustainable minigrid business model would require that the capital expenditure (CapEx) and the operating expenses (OpEx) be recovered from either initial connection costs, cost-reflective tariffs, or subsidy schemes. However, to date, there are few examples of established minigrids that are operating sustainably in



Africa [216]. Since most non-electrified households are poor and located in rural areas, minigrad business models typically must involve low connection costs. With the challenge of high capital costs and limited to non-existent financing and subsidy schemes for minigrads, this leaves only one pathway for minigrad developers – charging significantly higher “cost-reflective” tariffs to recover their investments, at levels that only around 10 to 15% of rural customers can afford [222]. Given the already diminished consumption levels of newly-connected customers, the high tariffs exacerbate the problem of low capacity utilization.

### **6.2.2 Why Demand Stimulation?**

The main goal of electrification is to enable activities that use power. However, to achieve this goal, electricity service must be reliable and affordable. In addition, customers should be able to access and afford domestic and commercial appliances that make use of the electricity provided. Electricity access programs in sub-Saharan Africa have made great strides in increasing electricity generation and customer connections. However, many newly-connected customers consume limited amounts of electricity, with limited growth over time [85], resulting in cumulative demand that is far less than supply. Some of the key reasons why new customers are unable to grow their consumption is that they often have limited access to and/or cannot afford appliances that consume electricity and have limited income to support electricity purchases. To alleviate this chicken-and-egg problem, it is imperative to implement demand stimulation programs that either facilitate new customers to organically grow their consumption over time or develop ancillary businesses that consume electricity directly from the grid/minigrad, such as the strategies considered in this study. Adding consumption to the system creates a virtuous cycle whereby developers are able to recoup system costs from increased revenue, and can therefore afford to lower the unit cost of power, enabling customers to afford to further grow their consumption, and ultimately realizing the intended benefits of electricity access.

Demand stimulation has a rich history. With the rapid increase in the rural electrification rate in the US in the 1930s, the government paired grid extension and supply expansion with demand stimulation programs through the Electric Home and Farm Authority, providing financing for farmers to purchase home appliances and equipment [42]. More recently, demand stimulation efforts in sub-Saharan Africa have gained traction. As an example, in 2016, JUMEME Rural Power Supply Ltd., a minigrid developer in Tanzania, provided financing for 12 of their business customers to purchase appliances. In 2018, they identified a new business opportunity on Ukara island in Lake Victoria that involves using their own electricity to run a fish freezing and delivery system to serve local markets, which improved capacity while providing an additional revenue stream [7].

### **6.2.3 Electric Mobility as Flexible Demand**

Use of electric mobility loads as flexible demand is a well-studied topic in the literature. Previous work has used a variety of techniques for improving use of this flexible resource, including better predictions of arrivals using fluid dynamic models [13], optimal charging schedules when faced with unknown future demand [265], a Markov Decision Process (MDP) framework for making charging decisions [305], heuristic algorithms for an NP-hard construction of the EV charging problem [308], genetic algorithms that consider grid parameters [8], model-free reinforcement learning techniques to coordinate multiple charging stations [156], and a technique based on Distributed Resource Allocation for smoothing grid operations using EVs [186]. While some or all of these techniques may be applicable for our scenario, they have been largely evaluated on centralized electricity grids, which exhibit substantially different constraints than decentralized minigrids, and for electric cars, which have different usage patterns and requirements as compared to boats.

In minigrids, demand response (DR) may be used for optimizing grid operations through load scheduling and load control to achieve higher efficiencies, saving fuel for backup

needs, decreasing operation expenses, providing grid resilience, and delaying the need for further investments [194]. Various DR strategies using EV charging have been explored, such as influencing the behavior of the users by reducing the EVs' trip distance and/or trip time shifting [256], day-ahead planning of EV schedules due to fluctuations in daily renewable generation [41], time-varying pricing intended to shift load from peak to off-peak periods [196], incentive-based DR programs that support vehicle-to-grid for load shifting and congestion management [229] and controlled EV charging DR programs [255].

### **6.3 Data and Methodology**

In this section, we describe our modeling approach, shown in Fig. 6.2. We begin by describing our data collection process and describe the datasets collected in Section 6.3.1. Next, in Section 6.3.2, we discuss our methodology in sizing an electric outboard motor and battery based on fishing boat movement patterns. We then describe the components of the electric load on the island in Sections 6.3.3, and 6.3.4, from residential and small commercial connections and the ice factory respectively. We then present an iterative minigrid operation model with a stochastic electric boat charging load algorithm developed to determine the maximum electric boat charging load each day over a year, while minimizing the charging infrastructure based on the capacity constraints of the minigrid, as well as an economic analysis of the system in Section 6.3.5.

#### **6.3.1 Data collection and description**

Our study takes place on Lolwe Island, which is situated in Lake Victoria in Eastern Uganda and has an estimated population of 14,841 people. Fishing is the major economic activity on the island, home to a vibrant fishing hub of over 1,000 boats. Figure 6.1 shows an example fishing boat. Currently, the minigrid developer for Lolwe Island is planning for a minigrid with an installed capacity of 600 *kWp* of solar PV, 650 *kWh* of lithium-ion

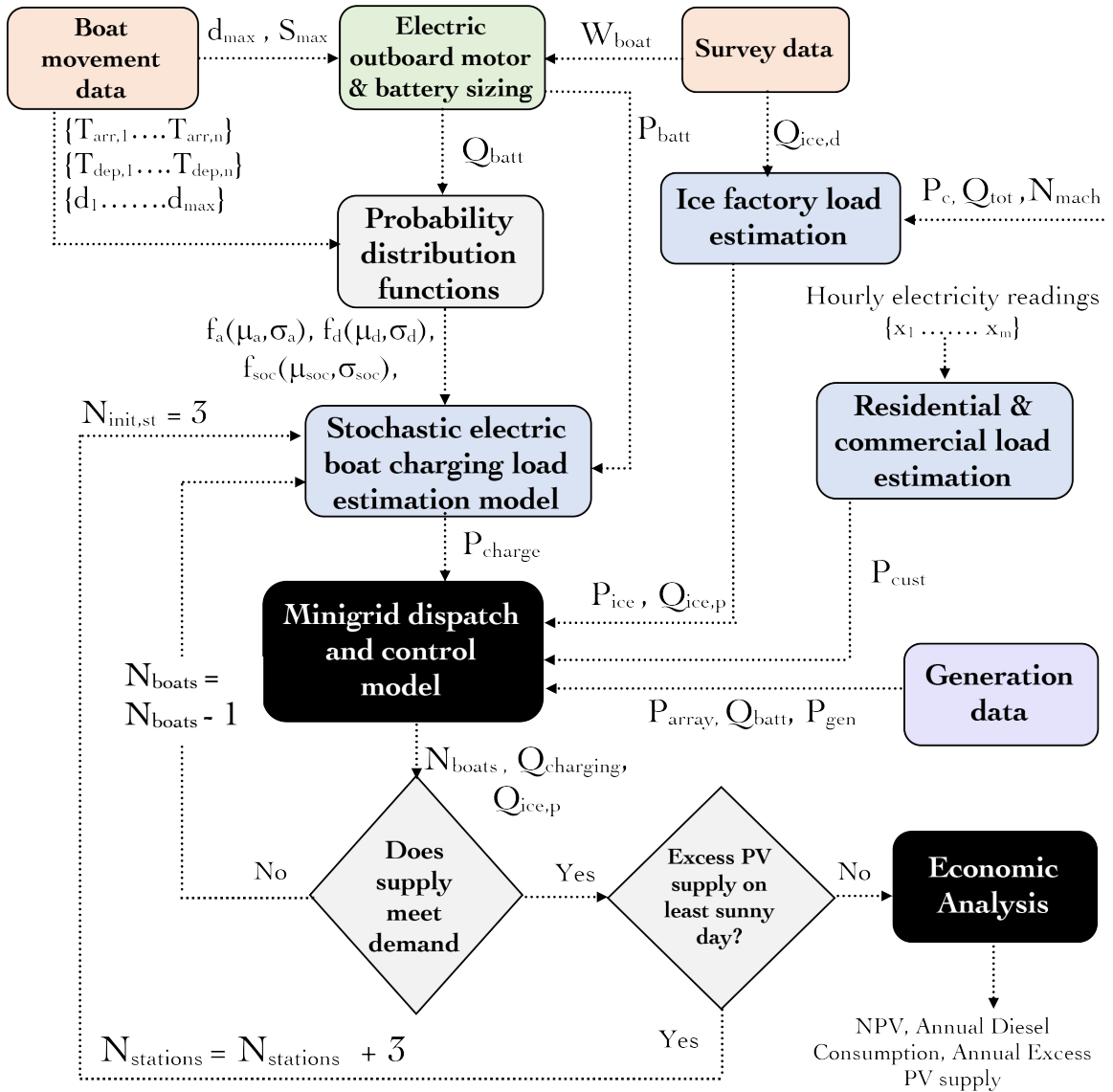


Figure 6.2: Flowchart of the study methodology.

battery bank capacity, and a 120 kW backup diesel generator [221]<sup>2</sup>. The developer is also setting up an industrial park on the island to mitigate the challenges of fish storage. It will include an ice factory to provide affordable ice for preserving Nile perch and a fish drying factory to enable efficient drying of Silver fish, which is currently done under the sun or using firewood [221]. For this study, we limited our power demand model of the industrial park to the ice factory only.

**Survey data.** We conducted a survey with 69 respondents in three villages on the island as part of our study to learn more about the socioeconomic and demographic characteristics of the fishing community, as well as their fishing habits<sup>3</sup>. Fishing boat owners on the island on average have a fleet of 3 - 4 boats and a majority of them handle the management of their fleets and the marketing of their catch, as opposed to fishing themselves. Instead, they hire young men to fish and operate their boats. The island depends on diesel to power fishing boats. On average, each boat consumes about 20 liters of fuel per trip, which translates to approximately 20,000 liters of fuel for the entire island for every boat to make a fishing trip. On average, fishing takes place during 6 days each week. This is expensive, unreliable, and has a negative impact on the environment.

The fishing boats in use on the island are v-shaped bottom boats (see Figure 6.1), locally constructed with wood, which has a density of 440 kg/m<sup>3</sup>. We measured three random boats, whose lengths ranged between 9m and 13m, with a width of 1.85m. Based on these dimensions, the shape of the boat and density of the wood, we estimated that the boats weigh between 1,209kg and 2,156kg unloaded. At the beginning of each fishing trip, we observed that each boat was loaded with fishing gear and two male fishing boat operators and on return, there was additional weight from the fish caught. Fishing boat operators in two of three villages we surveyed fish Nile perch, whose peak season is from July to December. They catch an average of 16 kg/day on a day with sparse catch and an

---

<sup>2</sup>The planned minigrid is slated for commissioning in 2020.

<sup>3</sup>We obtained approval for our study from our university's Institutional Review Board.

average of 98 *kg/day* when there is bountiful catch. Fishing boat operators in the third village catch Silver fish. Peak Silver fish season is between March and June, during which operators reported catching on average about 35,000 basins of Silver fish and as few as 900 basins during low season. We therefore estimated a maximum weight,  $W_{boat}$ , of about 3000kg for a loaded fishing boat.

Boat owners in the Nile perch villages reported to purchase between 30 and 1500 *kg* of ice a day. We therefore estimated an average of 10,000 - 15,000 *kg/day* per day of ice demand,  $Q_{ice,d}$ , on the island. Transporting ice from the mainland to the island and storing it before use results in substantial ice loss and embodied energy consumption. Typically, 20% of any ice that is purchased is lost before use for fish preservation.

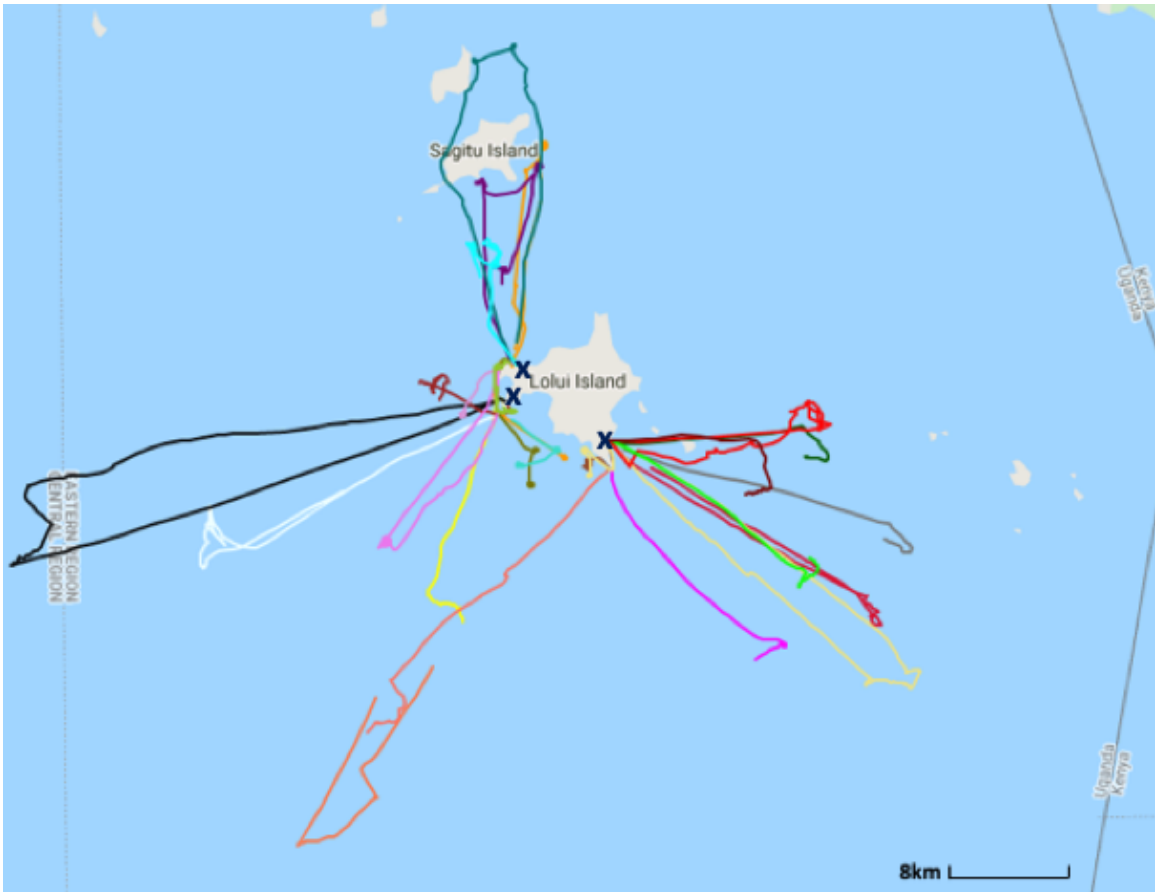
**Boat movement data.** To collect information about boat movements and the communications environment, we constructed 20 custom boat tracking devices. Each device consisted of a custom mobile application on a low-cost mobile phone (Motorola G5) to sense and log different metrics including GPS location coordinates, accelerometer and gyroscope readings, and cellular signal strength. The mobile phone was sealed in a water-tight enclosure for deployment. We favored a low-cost mobile phone for its configurability over an off-the-shelf micro-controller device, the Particle Electron IoT suite, because the Particle system is designed to store data in the cloud as it is collected. However, the cellular network on the island and in the lake is slow and unreliable, and thus a constant uplink to upload data was unavailable. While we could have stored data locally, the expansion slots on the Electron do not provide a direct way of expanding the memory storage. For our deployment, we also planned to extract logged data on a daily basis, which was easier to do from the mobile device than the Electron. A third reason is that the Electron typically uses an external 1,800mAh battery, which we projected would not have been able to last the duration of a 12-16 hour fishing trip. On the other hand, the Motorola G5 uses a 2,800mAh battery, which can last up to 24 hours when the phone's functions are limited to the core requirements. Finally, the mobile phone provided all the required sen-

sors already assembled and a readily available application programming interface (API) to interact with them, compared to the Particle Electron. Note that we made these decisions given our limited deployment purpose, size, and timeline – we discuss design considerations for a longer-term deployment in Section 6.5.

The data collection device was attached to a fishing boat at the start of a fishing trip with the mobile application (app) running in the foreground, as seen in Figure 6.1. Putting into consideration the hardware constraints and software delays, the app continuously logged sensor data by sampling at 30 second intervals. We limited data logging to only the required sensors to maintain the privacy of boat operators. Over the course of six days, boat owners and operators were randomly approached each day to host the devices on their boats for the duration of their upcoming fishing trip. On each day, we deployed the devices at the start of the fishing trips in the afternoon. We then returned the next morning to detach and collect devices at the end of each fishing trip to recharge and deploy them in a different village in the afternoon. Deployment took place twice in each village. While we initially aimed to deploy all 20 devices each day, the number of devices deployed each day varied depending on a number of logistical factors such as the turnaround time of charging the phones, which was limited to about three hours each day when the island’s diesel generator was running. We logged the departure time and level of fuel in the boat’s fuel containers at the beginning of each fishing trip and the arrival time as well as the corresponding level of fuel on arrival at the end of the trip.

We tracked 77 fishing trips in total. Figure 6.4 shows the probability distribution functions of the departure times,  $f_d(\mu_d, \sigma_d)$ , and arrival times  $f_a(\mu_a, \sigma_a)$  of the 77 fishing trips. We can see the distinct differences of times based on the type of fish pursued: Silver fish boats tend to operate for shorter journeys in the pitch dark of the night. Additionally, we can see that the window for departure times is narrow, while arrival times are more spread out. This has positive implications by enabling fewer charge stations to charge all of the boats. Due to challenges with environmental conditions, logistics, and a software

bug, we only captured near-complete GPS traces of 27 fishing trips. Figure 6.3 shows a select number of traces that we captured. The three distinct origin points on the island are the shores of the three participating villages.



**Figure 6.3:** Select daily traces of fishing boats. Three departure villages are each denoted with an X.

We utilize the GPS data to calculate the distance travelled by each boat between subsequent coordinates using Vincenty’s solution for the distance between points on an ellipsoidal earth model [268]. We calculated that the longest fishing trip,  $d_{max}$  covered 62km round trip and the shortest trip covered 12km. Despite the Nile perch fishing trips lasting longer than the Silver fish fishing trips, we do not observe much difference in the distances covered. The mean distances of the Nile perch and Silver fish fishing trips are 24.8km and 24.5km respectively, with 80% of the fishing trips covering under 35km. We also see a

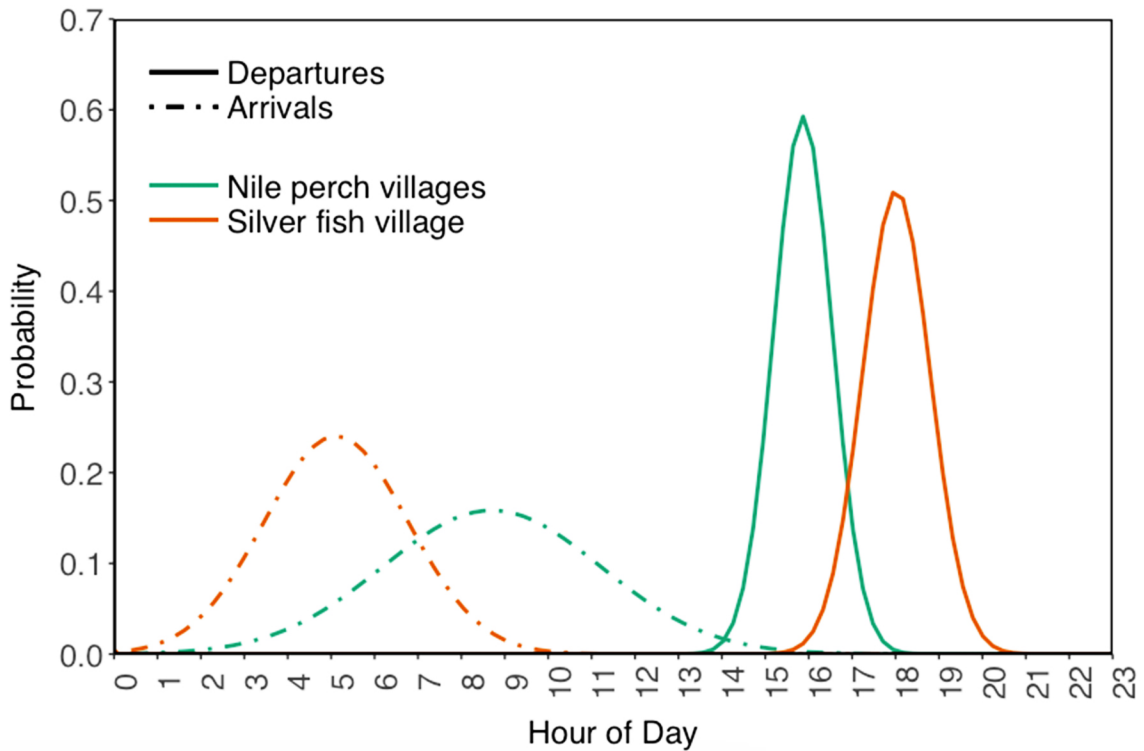


distinct portion of the trip before the return trip begins which we surmise is when fishing is taking place. We unfortunately were not be able to determine whether the engine was on or off during this period due to noisy inertial measurement unit data.

We calculated the speeds of the boat over the course of the fishing trips based on the GPS coordinates data at a one-minute resolution using Eq. 6.1.

$$s(t) = \sum_{i=1}^{n_t} (d_{\lambda_1\phi_1, \lambda_2\phi_2})_i * 60 \quad (6.1)$$

During the times when the boats were travelling to and from the fishing grounds we calculated a range of maximum speeds,  $S_{max}$  between 4.5 m/s and 5.2 m/s. In between these two portions of the trips, we calculated speeds lower than 0.5 m/s, which could be indicative of a stationary or drifting boat.



**Figure 6.4:** Probability distributions of the arrival and departure times of all 77 recorded fishing trips.

### **6.3.2 Sizing of an electric outboard motor and battery system**

The first step in the sizing an electric outboard motor is determining the required power output of the electric motor. A number of resources [184, 231] explain in detail how to calculate the force/thrust required to propel a boat based on the ship resistance model. However, for the purpose of this study we used a simplified estimation of the power requirement of the propulsion system. We utilize a simplified thrust-to-weight relationship [30] to estimate the thrust requirement based on the calculated maximum weight of the boats.

Though the thrust of the electric motor is important, the range of the battery is equally important and often overlooked when choosing an electric motor. We therefore also consider the range estimates of batteries compatible with the list of potential electric motors. The final selection of an electric outboard motor has sufficient thrust to propel the boat, and is compatible with a battery that has sufficient capacity and range to cover the distance of the longest fishing trip measured in our dataset.

### **6.3.3 Residential and commercial demand estimation**

A crucial step in the deployment of any minigrid is the assessment of electricity demand prior to implementation. Although electricity demand is hard to predict, especially in a village that has never had access to electricity, different methodologies are used to carry out electricity demand assessment of minigrids to provide a baseline for a well-founded project design. Two fairly common practices have been used in previous studies; using primary data collected through pre-electrification surveys [115] and second, using existing demand data from other minigrid projects in a similar socioeconomic and cultural context [146, 289].

We utilize the second approach to estimate the load profiles of prospective minigrid customers on the island using electricity consumption data from customers of existing minigrids in East Africa because we did not have the primary data collected for the island.

These customers are categorized based on three main connection types: residential customers, small commercial customers, and residential customers who run businesses in their homes. We begin by determining the daily load profile patterns of the existing minigrid customers, by applying a  $k$ -means clustering approach to the normalized hourly consumption readings of customers in each of the three mentioned categories. The consensus among a number of studies [227, 84, 297] is that it is the best known and most frequently applied partitioning clustering technique to analyze daily load profile patterns of electricity consumers. The objective of clustering is to improve the accuracy of the predicted load profile of the prospective minigrid customers. The clustering technique divides the customers,  $(x_1, x_2, \dots, x_n)$  in each connection type,  $j$ , into  $k$  clusters such that similar load profile patterns are placed in the same cluster  $x_i; x_j \in C_k$  and dissimilar load profile patterns are grouped into different clusters. We determine the optimal number of clusters,  $k$ , in each dataset using the NbClust approach [43]. Next, we get the average load profile of each cluster,  $P_{k,avg}$  using Eq. 6.2.

$$P_{k,avg} = \frac{\sum_{i,j \in C_k} (x_i; x_j)}{N_{C_k}} \quad (6.2)$$

We then use a weighted allocation method to allocate the number of prospective customers in each connection,  $N_j$  to each cluster and then calculate the total hourly load of each cluster,  $C_k$ , in each connection,  $j$ , which we sum up to generate an aggregate load profile,  $P_{cust}$  for all the prospective minigrid customers using Eq. 6.3.

$$P_{cust,1} \dots P_{cust,24} = \sum_{j=1}^3 \sum_{C_k \in j} \frac{P_{k,avg} * N_{C_k}}{N_j * N_{pc,j}} \quad (6.3)$$

#### 6.3.4 Modeling ice factory power demand

We modeled the ice factory as a series of similar small ice machines, which operate during hours of PV supply – that is, between the hours of 9 am and 6 pm. Assuming perfect knowledge of the next day's demand for ice  $Q_{ice,d}$ , we calculate the number of machines

started up for the day to meet this demand. We also assume that there is enough storage for all the ice produced for at least 24 hours. We use the technical data of one ice machine, including its capacity  $Q_{tot}$ , in kg/day, and power drawn by the compressor,  $P_c$ , in addition to the number of hours in a day the machines run,  $h_m$  to determine the number of machines required,  $N_{mach}$ , to meet demand,  $Q_{ice,d}$ . Lastly, we generate the demand profile of the ice machine for various levels of ice demand as summarized in Eq. 6.4

$$(P_9, P_2, \dots, P_{18}) = P_c * [N_{mach} = \frac{Q_{ice,d}}{Q_{tot}} * \frac{24}{h_m}] \quad (6.4)$$

### 6.3.5 Minigrid operation model considering stochastic electric boat charging load

We present an iterative dispatch and control model for the proposed minigrid over a 24-hour period for each day over a year considering the stochastic nature of electric boat charging and PV supply in Algorithm 1. While Markov Chain models have been widely used in the stochastic generation of EV charging load [103, 257], it is a decision-based time and state model that models vehicle traffic flows and as such, not applicable to the boat movement patterns in this study. We therefore propose a stochastic method based on Monte Carlo simulations that considers the boat movement patterns we observe in our data. This method has many advantages such as possibility of simultaneous consideration of many probabilistic factors and ease of implementation. The boat movement data collected is processed to identify the probability distribution functions of battery State of Charge (SoC),  $f_{soc}(\mu_{soc}, \sigma_{soc})$ , hours of boat arrival,  $f_a(\mu_a, \sigma_a)$  and hours of boat departures,  $f_d(\mu_d, \sigma_d)$ .

We initialize the algorithm with three charging stations,  $N_{ch}$  one at each shore in each village. We assume that the charging stations are fast charging and the power output from the chargers,  $P_{ch}$  is constant. We run the simulation for an entire year. During each day,  $k$ , we estimate the charging demand of the electric boats during the charging window determined by the boat arrival and departure times. We use 1 hour as the sampling

interval time. At each time step of the day, we consider the total load from the ice factory,  $P_{ice}$ , customer connections,  $P_{cust}$  and the charging stations,  $P_{charge}$  as well as the total generation from the PV array,  $P_{array}$ , minigrid battery bank storage,  $Q_{batt}$  and backup diesel generator,  $P_{gen}$ . We used HOMER software's algorithm [74] to generate synthetic hourly solar data for Lolwe Island for an entire year by combining monthly averaged solar insolation data and the clearness index for the coordinates corresponding to Lolwe averaged over a 35-year period, from 1983 - 2018, available from NASA's Prediction of Worldwide Energy Resource (POWER) project [2]. The PV supply is first used to meet the total load at each hour, and any excess PV supply is used to recharge the battery bank if needed. Any excess PV supply after this is curtailed. During hours of insufficient PV supply, the battery bank is discharged to meet the remaining load. The Depth-of-Discharge (DoD) is limited to 80%. When the battery bank reaches its DoD, the backup generator then ramps up to meet the remaining load.

If there is still additional minigrid capacity during the least sunny day of the year, the number of charging stations at each shore is incremented by one during each iteration to allow for additional boats to charge until the maximum capacity of the minigrid is reached. The model therefore minimizes the charging infrastructure required to maximize boats charging within the constraints of the charging window and capacity of the minigrid. In addition to the technical operation of the minigrid, we estimate the profitability of the system characterized by Net Present Value (NPV), a parameter that expresses the initial capital investment and all future cash flows arising from operating the system over its lifetime as an equivalent amount at present time, summarized by Eq. 6.5.

$$NPV = \sum_{t=0}^n \frac{A_t}{(1+d)^t} \quad (6.5)$$

where  $A_t$  is the project's revenues minus costs in time  $t$ , from year 0 to year  $n$  and  $d$  is the discount rate. We calculate the NPV over a period of 20 years, which is the average lifetime of a PV system and discounted at a rate of 14%, which is the reported discount

---

**Algorithm 1:** Minigrid dispatch and control algorithm with stochastic boat charging load

---

**Result:** Maximum number of boats charged in a day,  $N_{b,max}$ , Annual diesel consumption from backup generator, Annual excess PV generation,  $Q_{excess}$

**foreach** *day of year, k* **do**

Initialize  $N_b = N_{ch}$ ;

**foreach** *charging station,  $N_{ch,j}$*  **do**

1. Arrival time,  $T_{a,i} = \text{np.random}(\mu_a, \sigma_a, n)$ . = Connection hour of first boat,  $T_{conn,i}$ ;
2. Battery state of charge on arrival,  $B_{soc,i} = \text{np.random}(\mu_{soc}, \sigma_{soc}, n)$ ;
3. Departure time,  $T_{d,i} = \text{np.random}(\mu_d, \sigma_d, n)$ ;
4. Charging duration,  $T_{ch,i} = [0.8 - B_{soc,i}] * Q_{boat} / P_{ch}$  ;
5. Disconnection hour  $T_{disc,i} = T_{conn,i} + T_{ch,i}$  ;
6. Available charging time,  $T_{avail} = T_{d,i} - T_{disc,i}$ ;

**while**  $T_{avail} \neq 0$  **do**

- a.  $T_{conn,i+1} = T_{disc,i}$ ;
- b.  $T_{d,i+1} = \text{np.random}(\mu_d, \sigma_d, n)$ ;
- c.  $B_{soc,i+1} = \text{np.random}(\mu_{soc}, \sigma_{soc}, n)$ ;
- d.  $T_{disc,i+1} = T_{conn,i+1} + T_{ch,i+1}$  ;
- e.  $T_{avail} = T_{d,i+1} - T_{disc,i+1}$ ;
- f. Populate connection and disconnection matrices to keep track of the number of electric boats in charging status at every given hour,  $N_{b,hr}$ .

**end**

7. Total charging load,  $P_{charge,hr} = P_{ch} * N_{b,hr}$  ;

**end**

Sum all boats charged:  $N_b = N_b + N_{b,conn}$ ;

**foreach** *hour, i* **do**

$P_{tot,hr} = P_{ice,hr} + P_{cust,hr} + P_{charge,hr}$  ;

**if**  $P_{array,hr} > P_{tot,hr}$  **then**

$P_{excess} = P_{array,hr} - P_{tot,hr} - [650 - Q_{batt}]$  ;

**else**

**if**  $[P_{tot,hr} - P_{array,hr}] < (\eta_{batt} Q_{batt})$  **then**

$Q_{batt} = Q_{batt} - [P_{tot,hr} - P_{array,hr}]$

**else**

$[P_{tot,hr} - P_{array,hr} - Q_{batt}] = P_{gen}$

**end**

**end**

**end**

If  $P_{gen,required} > P_{gen,max}$ , decrease number of boats charged per day,  $N_b$  and rerun algorithm.

**end**

---

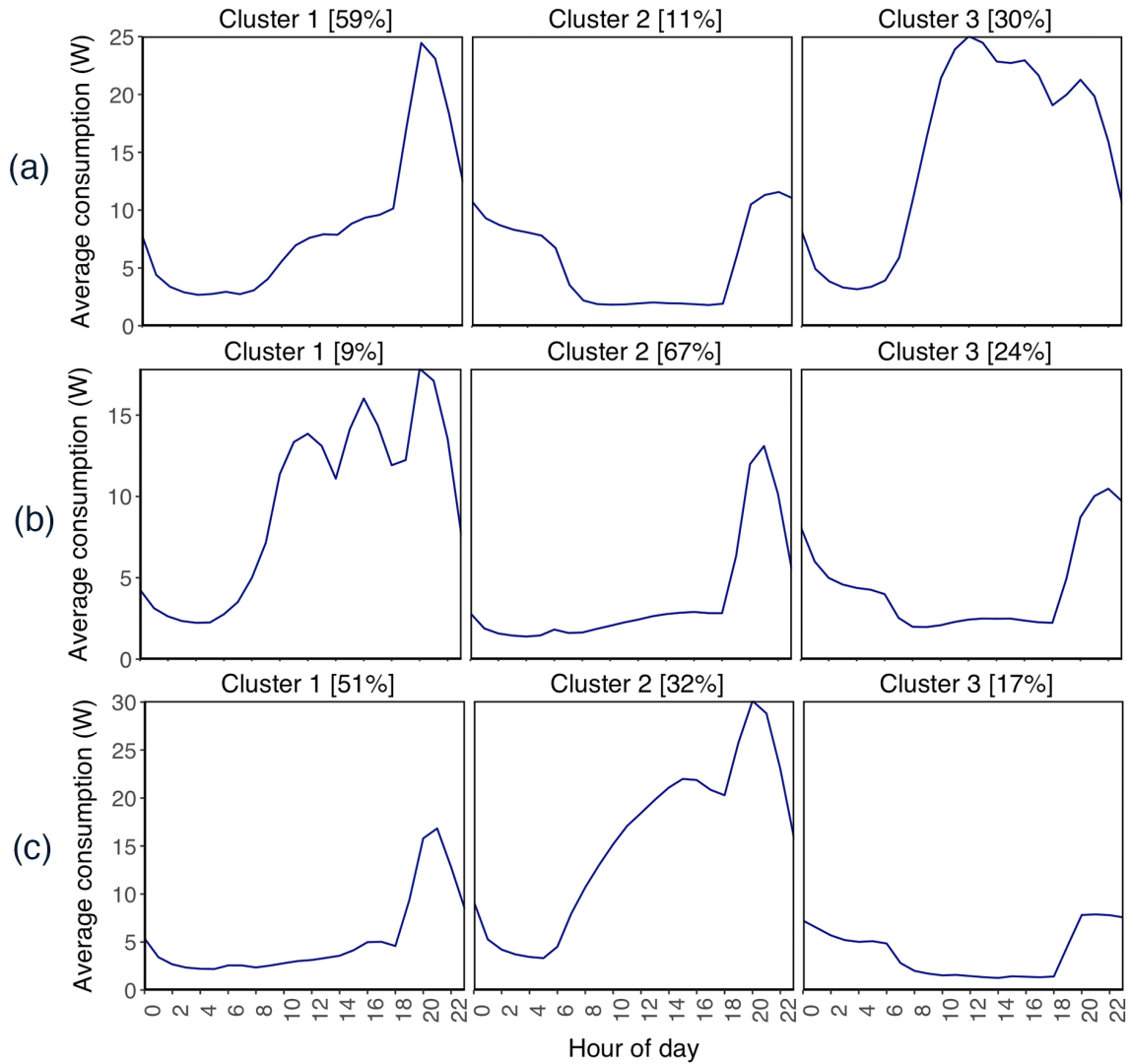
rate for Uganda [195]. For this study, we assumed that the system without infrastructure for ice production and electric boat charging has a zero NPV, which means that the project breaks even.

## **6.4 Analysis**

In this section we present a techno-economic feasibility analysis of adding electric boat charging and ice factory load to the proposed minigrid, the impact of infrastructure planning on the maximum electric boat charging load, the financial benefit to the boat owners and finally the impact of demand response on the operation of the minigrid.

### **6.4.1 Residential and small commercial demand profile**

We began by analyzing the hourly electricity consumption data of customers of 18 minigrids in East Africa, including those described in previous work [289]. 56% of these customers have a residential connection, 36% have a small business connection, and the remaining 8% run a business from their residential premises. From the results of the clustering algorithm, we found three distinct load patterns for each category as shown in Figure 6.5.

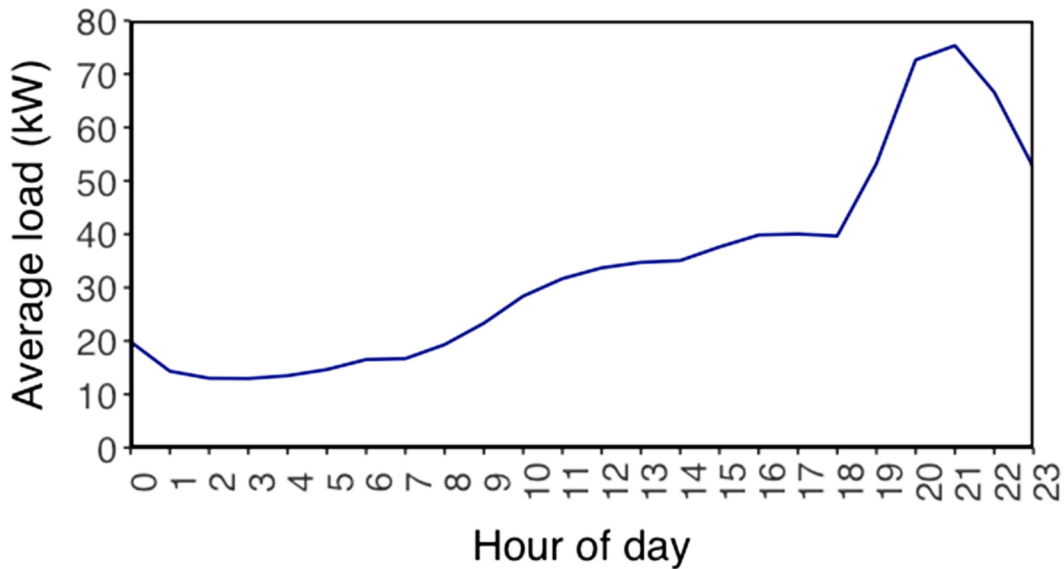


**Figure 6.5:** Average daily load profile of clusters among (a) small business, (b) residential, and (c) home business customers.

The minigrid developer estimates a total of 3000 prospective residential connections and 700 small commercial connections on Lolwe island [221]. Of the residential connections, using data from the Kenya Integrated Household Budget Survey [200], which indicates a proportion of rural households that run business out of their homes, we estimate 68% prospective residential connections and the remaining 12% as residential connections with businesses on their premises. Figure 6.6 shows the resulting demand profile of all prospec-



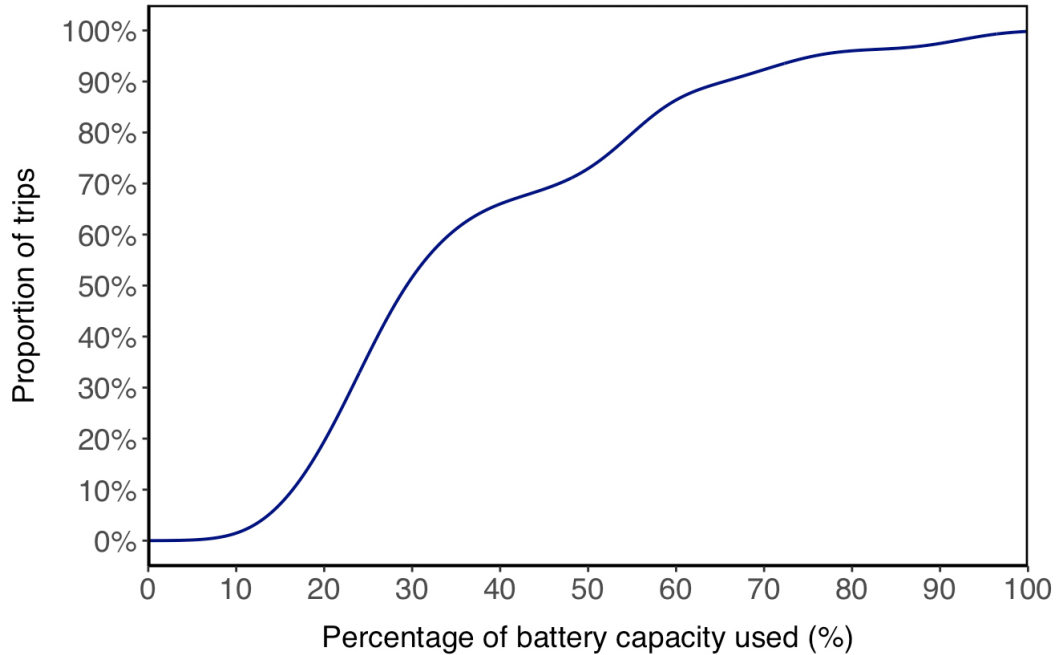
tive customers. We observe high demand during the evening hours of 8pm - 10pm, with a peak of  $75kW$  at 9pm.



**Figure 6.6:** Daily load profile of residential and small business customers.

#### 6.4.2 Electric outboard motor and battery sizing

Based on the estimated weight of a loaded fishing boat we calculated the required thrust to be  $458lbs$ , which results in about  $11kW$  ( $15HP$ ) of propulsive power at the maximum speed of  $5.2m/s$ . The 15 - 30 HP electric outboard motors were not compatible with batteries with sufficient capacity to last the duration of the longest trip,  $62km$ . The  $40HP$  Torqeedo Deep Blue 25 RL electric outboard motor, with a propulsive power of  $16kW$ , met our requirements [272]. It is compatible with two  $9.1kWh$  BMW i8 lithium ion battery packs connected in parallel. This battery setup has a range of between  $32km$  and  $86km$  at a speed of  $12m/s$  and  $2m/s$  respectively. It draws  $3.7kW$  at a  $240V$  fast charging station. Based on these characteristics, we calculated that the total battery capacity is sufficient to last the duration of all monitored trips, with 90% of the trips using less than 65% of the battery capacity as shown in Figure 6.7. It is necessary to slightly oversize the battery to prevent stranded boats.



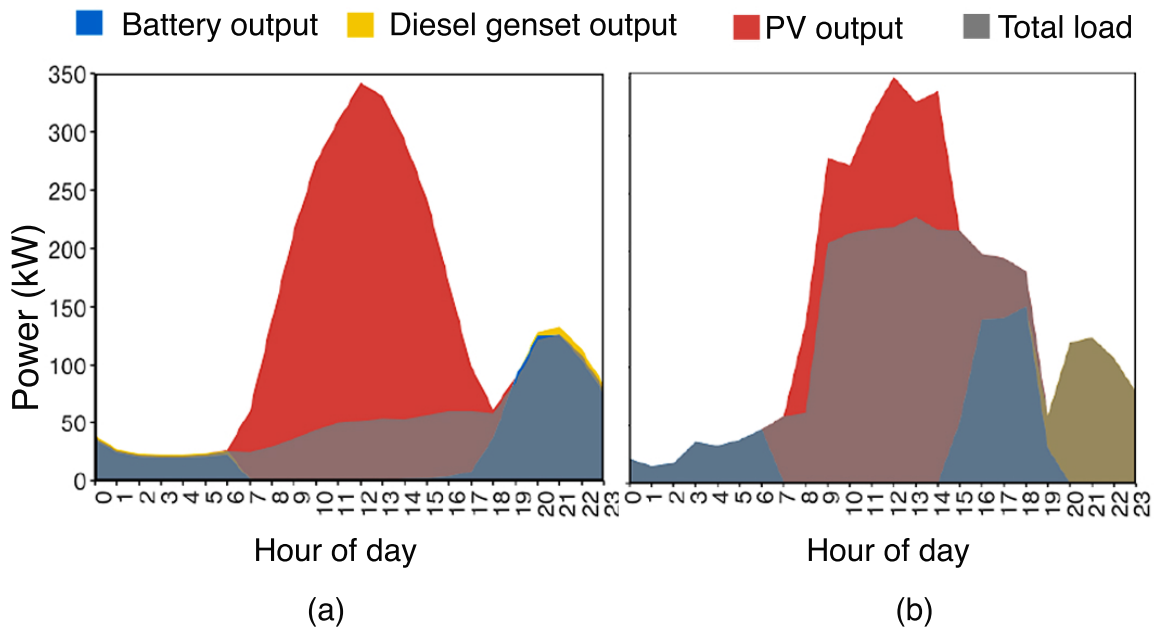
**Figure 6.7:** Cumulative distribution of boat battery usage from two  $9.1kWh$  BMW i8 lithium ion battery packs connected in parallel.

### 6.4.3 Impact on minigrid operation

The ice factory comprises of multiple  $5000\text{ kg/day}$  flake ice machines with compressor power of  $17.5kW$ , each costing \$15,000 [163]. We assume that the machines operate everyday between the hours of 9am and 6pm. In our simulation of minigrid operations, we assume that ice production takes precedence over recharging the electric fishing fleet batteries. It is becoming common practice to limit EV battery depth-of-discharge to about 20% (i.e., SoC to 80%), which reduces battery degradation and increases longevity. Our simulations therefore limit the electric boat battery charging to 80%, which we observed in Figure 6.7 to be sufficient for over 95% percent of the trips.

The average daily ice demand on the island was estimated as  $13,000\text{ kg/day}$ . At this level of ice production, the minimum number of charging stations required to maximize boat charging on the day of the year with minimum PV supply was determined to be 15. This is the most risk averse charging infrastructure plan, where the minigrid opera-

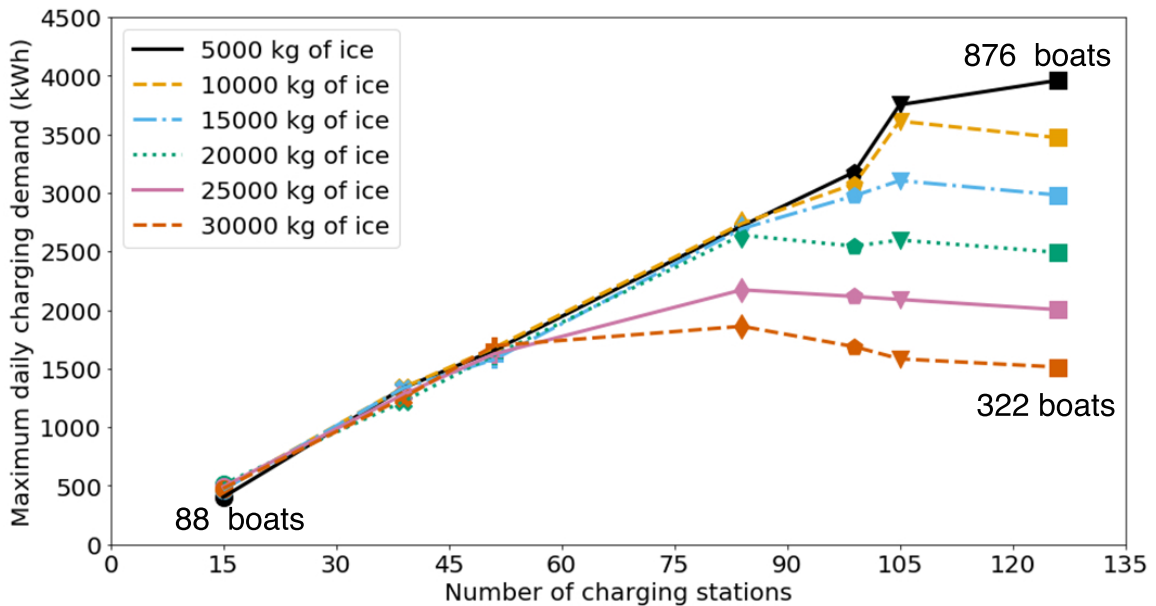
tor would install just enough charging stations such that on any given day of the year, all the charging stations are in use during the entire available charging window without resulting in inadequate supply to meet the charging demand. During the charging window, a maximum of 102 boats are able to charge in a day, adding 466 kWh of load to the system (17% of total load), in addition to 1,350 kWh from ice production (51% of total load). As shown in Figure 6.8, the capacity utilization of the minigrid increases, reducing the amount of curtailed PV supply by about 20%. However, generation from the backup diesel generator increases.



**Figure 6.8:** Minigrid supply and demand curve on an average day of the year (a) without ice factory and electric boat charging load and (b) with ice production of 13,000 kg/day of and 102 boats charged at 15 charging stations.

We also carried out a sensitivity analysis, to observe how the maximum daily charging load that the minigrid can serve at different quantities of ice production between 5,000kg and 30,000kg is impacted by charging infrastructure planning. As shown in Figure 6.9, we find that a more audacious charging infrastructure plan that increases the number of charging stations to 51, increases the maximum daily boat charging demand by about

240%. We also observe that below 60 charging stations, the number of charging stations limit the maximum daily charging demand on days with high PV supply, therefore we see very little variation with the quantity of ice produced. Above this, the quantity of ice production is pivotal to the number of boats that can recharge in a day. For example, we observe that when 126 charging stations are installed, almost 3 times as many boats can be charged when ice production is decreased from 30,000kg to 5,000kg. These results help to elucidate some of the load tradeoffs between ice manufacturing and electric mobility, as well as charging infrastructure planning.



**Figure 6.9:** Change in maximum daily electric boat charging demand a function of number of charging stations and ice produced

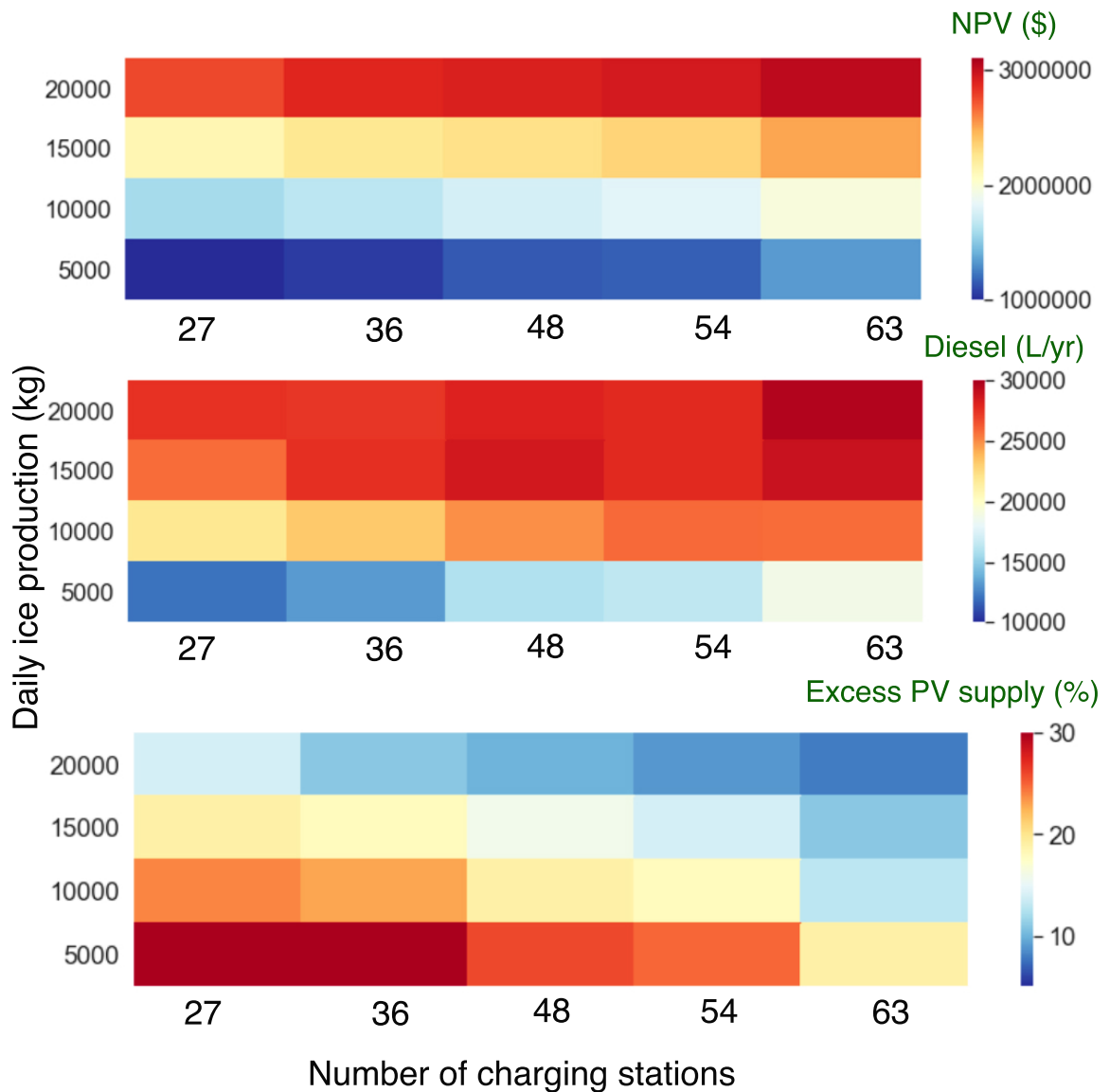
#### 6.4.4 Impact on economics of minigrid project

Technical performance by itself is insufficient for the viability of this system; financial sustainability is also crucial for system viability. We quantify the impact of ice factory and charging demand on the economics of the system, as well as capacity utilization. For the NPV calculation, we made a number of assumptions:

- The base case system, *i.e.*, without ice production and boat charging, has an NPV of zero
- A standard tariff of \$0.40 per kWh is charged for electric boat charging
- Each charging stations costs \$200 to install
- We assumed that all the ice produced is sold at \$0.068 per *kg*, which is the average cost of ice quoted by the boat owners.
- Lastly, we did not account for the operating costs of maintaining the ice machines and the charging stations.

The aforementioned cost assumptions are competitive numbers based on discussions with mini-grid developers in the area. The price of diesel on the island is reported to fluctuate between \$1.08 per liter and \$1.32 per liter. We used the highest price (\$1.32 per liter) to calculate the costs associated with the fuel consumption of the backup generator.

As shown in Figure 6.10, there is a tradeoff between maximizing the NPV of the system, and minimizing both the diesel fuel consumption from running the backup generator and the amount of PV supply that is curtailed. We observe that any level of ice production, planning the charging infrastructure to accommodate the charging of more boats per day would increase the NPV and improve capacity utilization but would also require higher usage of the backup generator. The planning of charging infrastructure ultimately depends on the goal of the minigrid developer. Preference for a higher NPV, while ensuring adequate charging infrastructure to service a large fleet would mean compromising by adding diesel consumption, which is vulnerable to price fluctuations, not to mention the environmental costs of fossil fuel use.



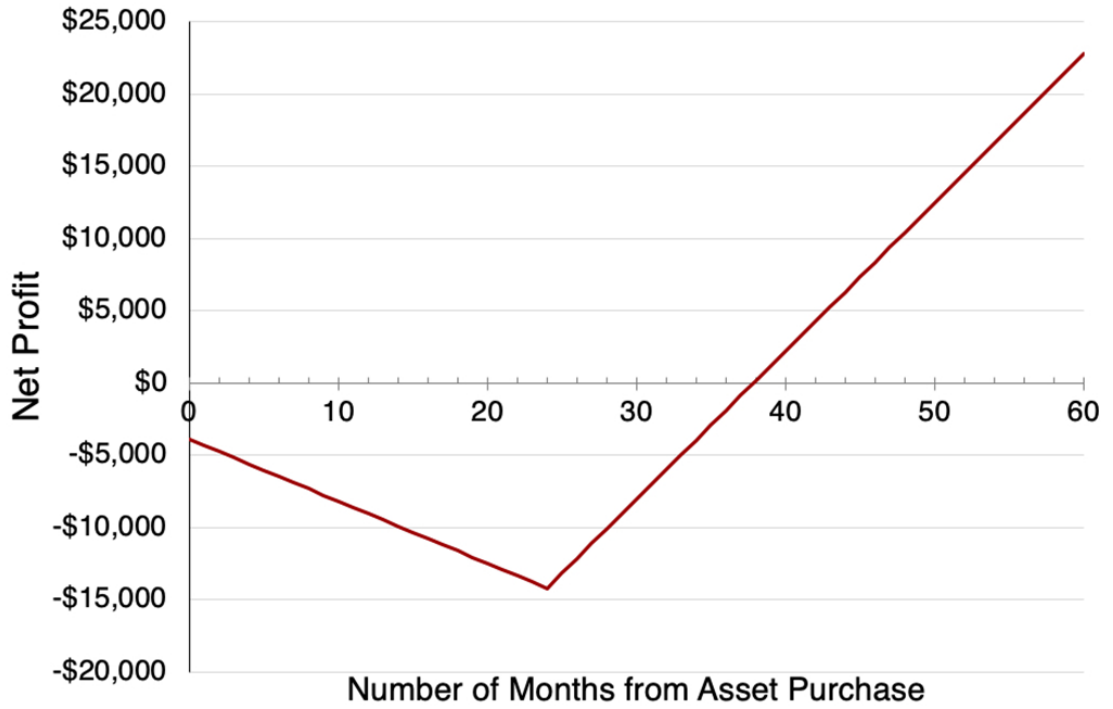
**Figure 6.10:** Net Present Value, annual diesel consumption and annual excess PV supply as a function of number of charging stations and daily ice production. Note that the number of boats are maximized in each scenario.

#### 6.4.5 Economic impact for boat owners

To consider the potential benefit of converting diesel-powered fishing boats to electric for the boat owners, we consider the payback period, which is the amount of time it takes to recover the cost of an investment. We consider the cost of purchasing the engine

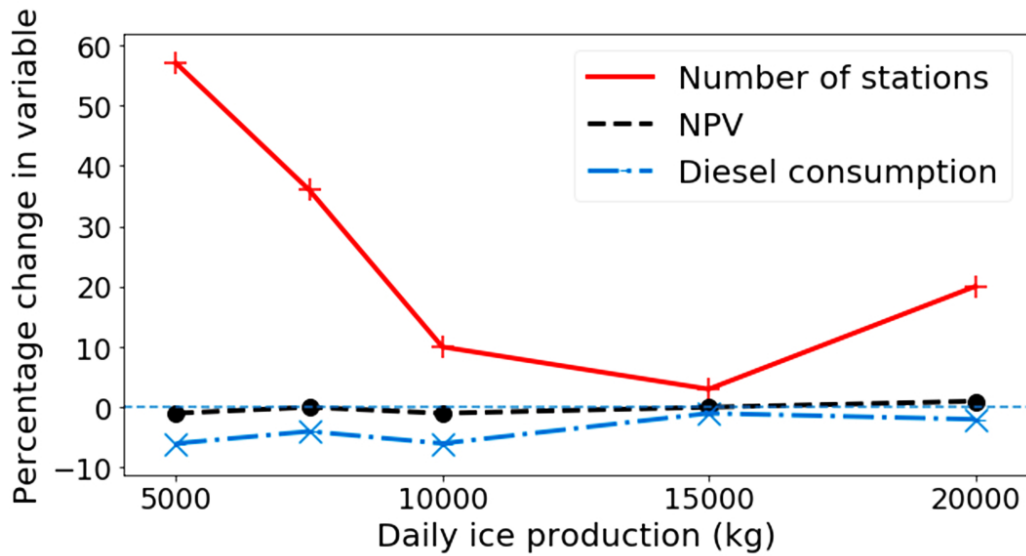
and battery system through an asset financing scheme with a loan term of 24 months, a 10% down payment and a monthly interest rate of 2.55% per month. These terms are comparable with other asset financing programs in East Africa. As part of the survey, we also examined fishing boat owners' willingness to participate in asset financing. 90% of owners have access to either a mobile money platform or a bank account. 25% of owners reported to have requested for a loan within the past year, mainly for expenses related to their fishing boat, with a majority of those reported to have borrowed from family/friends or from a savings group. However, only 45% reported to have completed loan repayments. They all reported a willingness to take a loan in the near future for further investment in their fishing business.

We also consider the cost of recharging the battery at \$ 0.40 per kWh. We compare these costs to the savings from not purchasing diesel at \$1.32 per litre and the cost of maintaining a diesel engine which we determined to be \$50 per month from the survey. The repayment amount on a 40 *HP* Torqeedo Deep Blue engine with the corresponding 18.2 kWh lithium ion battery system and charger, which costs \$ 26,200 [176] would therefore be \$38,766. As shown in Figure 6.11 , a boat owner with an average round-trip of 25 km per fishing trip, using an average of 200 litres of fuel per week would recover the cost of their investments after about 3 years and by 5 years they could potentially see about \$ 20,000 in savings. Shortening this payback period would entail exploring cheaper or pre-owned options for an electric engine/battery system that still meet the requirements of the boat owners.



**Figure 6.11:** Payback period on purchase of 40 HP Torqeedo Deep Blue electric outboard motor and 18.2 kWh lithium-ion battery system.

#### 6.4.6 Demand response



**Figure 6.12:** Impact of shifting electric boat charging load.



Considering that the charging window of the electric fishing fleet does not coincide with peak demand hours of residential and commercial load, the approach we took with regard to DR is to restrict electric boat charging to the hours of PV production (7am - 6pm). This means that early boat arrivals delay connection of their batteries to charging stations until 7am. This DR strategy would require up to 56% increase in the charging infrastructure to serve the same number of boats as shown in Figure 6.12, which reduces the NPV by at most 1%. However, about 5% of fuel can be saved. This DR strategy would be useful to a minigrid operator whose main goal is decreasing fuel consumption from running the backup generator.

Controlling the level to which EV batteries charge is also a common DR strategy. While we have established an 80% charging baseline for the electric boats in this study, we evaluated how the economic and operational performance of the grid would change if the charging limit was relaxed to allow boats to charge to 100%. We expect the number of boats that charge in a day to decrease, given all the other factors that influence the number of boats charged remain unchanged. This hypothesis is validated as shown in Figure 6.13. We observe that there is minimal difference in NPV between the two cases. However, we find that between 5000kg and 15000kg ice production, there is between 5 - 10% in fuel savings. Based on these results, it would seem favorable to the minigrid developer to limit charging to 80% as this allows up to 60% more boats to be charged at a minimal cost to their financial and operation goals and with little (but nonzero) risk for stranded boats.

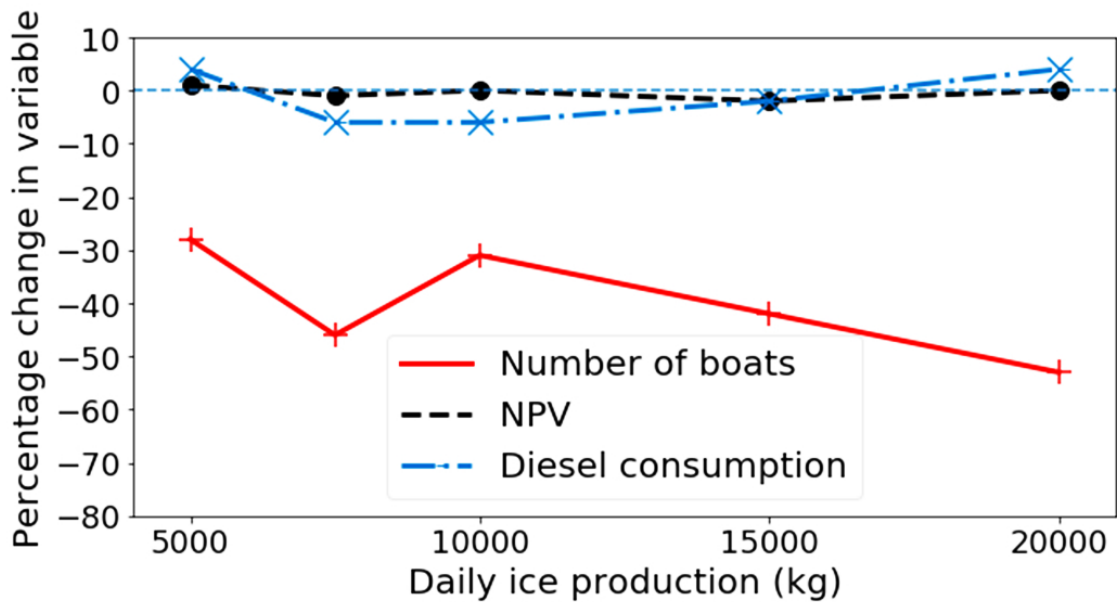


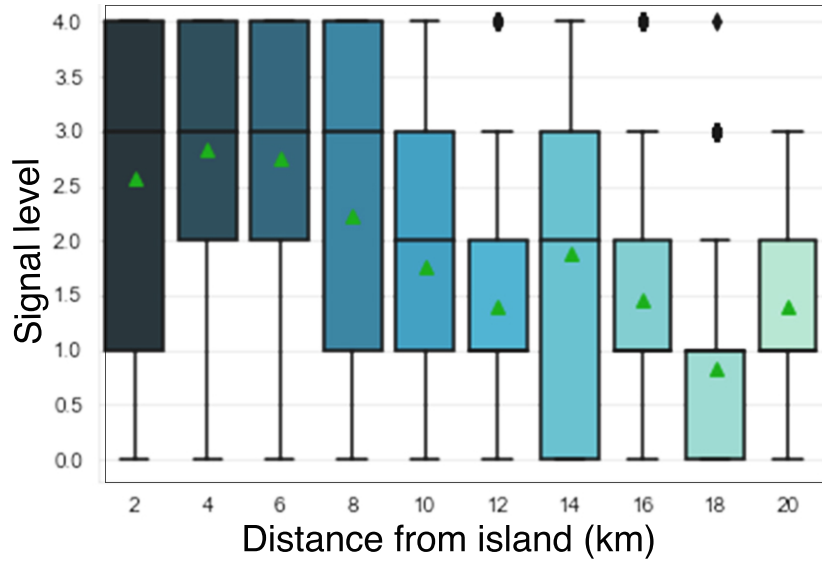
Figure 6.13: Impact of charging electric boats to 100% vs. 80%.

## 6.5 Discussion and Future Work

To test this project at scale there is need to explore the potential for charging station networks on other surrounding islands to address range anxiety. We are also yet to explore the potential for a boats-to-grid and boat-to-boat demand response strategy. The risk factors affecting adoption of this project, such as policy and political factors, have to be detailed and strategies to mitigate these risks explored. Real-time boat monitoring could be explored, as it allows for predictive management of boat charging, improving the operational efficiency of the minigrid. To that end, we leverage our deployment experience to discuss design considerations for a long-term electric boat monitoring system:

While a low-cost mobile phone is a better solution for collecting data in the short-term, it may not be feasible in the long run. A fishing expedition can last up to 72 hours. As such, even a 2800mAh battery that can hold up to 24 hours of battery charge would not suffice for this application. Therefore, an additional power bank that can potentially store up to threefold the phone battery power would be an ideal back-up. Second, a mobile device is susceptible to being tampered with and used otherwise, thus, building an embedded

device would be a better solution in the long run. Third, while storing data on the device is a better solution because of network unreliability, data can easily be lost when the device comes into contact with water or gets stolen, thus, by routinely monitoring the cellular signal, the data collected can be backed up to a secure cloud service each time the device is within a better signal range. Fourth, while the rest of the world has embraced 4G and is gearing towards 5G, most places in rural sub-Saharan Africa oscillate between EDGE and 3G. Thus, the device to be deployed should cater for typically unreliable networks, and adapt accordingly. For example, from our deployment, Figure 6.14 shows that the signal strength quickly deteriorates with increasing distance from the island. There is an option of using LAN beacons atop buoys, which would then upload data to the cloud via a satellite link. These are more reliable, albeit expensive. Therefore, this approach would only work if the project is to be scaled beyond a single island. Finally, some of the mobile sensors (e.g., gyroscope and accelerometer) are very sensitive to small changes in the movement of fishing boats. The data collected by these sensors is noisy as the sensor readings are affected by weather changes, boat drifting, and other uncoordinated boat movements. Going forward, we could explore using commercial-grade, albeit expensive sensors, if there is a need to collect certain metrics whose sensors are prone to noise.



**Figure 6.14:** Distribution of signal strength as a function of distance from the island. High signal strength = 4; No signal strength = 0.

## 6.6 Conclusion

To summarize our contribution, we studied the potential for electric fishing boats to provide valuable and flexible load to a decentralized minigrid system on an island in Lake Victoria, with the potential to improve outcomes for fishing boat owners and operators as well as minigrid developers alike. We applied a survey, a low-cost boat tracking system, and substantial system modeling to create a large-scale model for understanding technical and financial tradeoffs in the minigrid. Our work shows the significant scale of load possible from a modest deployment of electric boats and the crucial value from adding relatively trivial control to the boat charging system. We also outline the considerations for a future boat tracking system, laying the groundwork for a much larger future operational deployment. We intend for this effort to serve as unique guidance to minigrid developers for incorporating electric mobility to ensure the technical and financial viability of their systems, paving the way for sustainable progress towards universal electrification and its associated economic empowerment.

## CHAPTER 7

### CONCLUSION

In this thesis, we developed data-driven methods and techniques to support critical infrastructure measurement and sustainable development. We focused on three key infrastructures: rivers, roads, and electricity access.

In Chapter 3, we proposed a new approach for global river discharge prediction based on Long Short-Term Memory (LSTM) Recurrent Neural Networks (RNNs) and spatiotemporal hydrologic knowledge. Our approach outperforms the latest state-of-the-art data assimilation and lumped machine learning models in ungauged basins. Our work is especially important in hydrology, a field that has traditionally relied on process-based models. This is because our technique incorporates the topological physics of river flows into the data modeling phase, which improves the generalization of models in ungauged basins and improves the overall performance of the models. This work can potentially improve methods for predicting river discharge on a global scale and advance our understanding of the cascading impacts of anthropogenic climate change on global water resources. Finally, this work sets the stage to examine the constraints of process-based modeling approaches and better characterize how machine learning-based models can be used to model physical processes in hydrology and other physical sciences.

In Chapter 4, we proposed a new approach for explaining machine learning models for river discharge prediction, which can be extended beyond hydrology to other areas of physical science. In order to understand how machine learning models work, we use statistical methods, such as cooperative game theory tools, to analyze the internal workings and how they make their predictions. This is especially important in hydrology, a field

that uses physics-based models and has been reluctant to adopt machine-learning techniques due to their opaque nature (black box models). In high-stakes applications such as water resource allocation and flood forecasting, the inability to explain ML models makes it difficult to trust their predictions. Using explainable machine learning, hydrologists can better understand how ML models make predictions and use this knowledge to improve river discharge predictions. Additionally, explainable machine learning can help hydrologists identify and address the limitations of physics-based models: physics-based models are not able to accurately capture all the complex processes that occur in a hydrometeorological cycle. Finally, explainable machine learning can help hydrologists to identify the processes that physics-based models are not accurately capturing and to develop new physics-based models that are more accurate. Thus, explainable AI has the potential to improve the accuracy and reliability of machine learning models, which could be beneficial for a variety of applications, such as flood forecasting, water resource management, and climate change research.

In Chapter 5, we proposed a new temporal-spatial road quality prediction approach using satellite imagery. Our approach uses convolutional neural networks (CNNs) and vision transformers to predict road quality from high-resolution and medium-resolution satellite imagery. Combining these techniques, we achieve substantial predictions even in regions with limited and low-resolution data. Road quality measurement is critical for socio-economic and political development. Good roads can lead to increased economic activity, improved access to education and healthcare, and reduced poverty. They can also help to promote political stability and democracy. Predictions and recommendations by our models are currently used by policymakers, such as the World Bank, to influence policies geared towards funding for the construction and rehabilitation of infrastructure in developing countries, which could lead to several benefits for socio-economic and political development.

In chapter 6, we proposed tools to support sustainable development initiatives. Here, we detail our contributions to creating software tools to stimulate electricity demand in off-grid communities, with a case study on Lolwe Island, Lake Victoria (East Africa). These tools can help optimize electric boats' charging schedules and reduce diesel fuel use. Converting diesel-based fishing boats to electric motor and battery-based systems can be a valuable way to stimulate demand for sustainable mini-grids and reduce environmental pollution. Electric boat charging can contribute at least 17% more daily consumption, resulting in substantial technical and financial value to the mini-grid system while reducing greenhouse gas emissions and other forms of environmental pollution.

Machine learning (ML) is rapidly transforming society, becoming an essential tool for addressing global challenges, such as climate change on water. ML can decipher complex patterns and predict future scenarios, providing actionable insights that empower us to make informed decisions and policies to support socio-economic and political development, especially for countries in the global south. However, ML is still a relatively new field with much room for growth and innovation. Thus, this thesis has focused on areas where such innovations will play a critical role in benefiting society and humanity, particularly in the global south. These methods have the potential to help us make more informed decisions and policies, which can lead to a more sustainable and equitable future for all.

## BIBLIOGRAPHY

- [1] Adshead, Daniel, Thacker, Scott, Fuldauer, Lena I, and Hall, Jim W. Delivering on the sustainable development goals through long-term infrastructure planning. *Global Environmental Change* 59 (2019), 101975.
- [2] Aeronautics, National, and Administration, Space. Nasa prediction of worldwide energy resources data access viewer, Feb. 2020.
- [3] Agostinelli, Forest, Hoffman, Matthew, Sadowski, Peter, and Baldi, Pierre. Learning activation functions to improve deep neural networks. *arXiv preprint arXiv:1412.6830* (2014).
- [4] Akinyemi, Edward O, and Zuidgeest, Mark HP. Managing transportation infrastructure for sustainable development. *Computer-Aided Civil and Infrastructure Engineering* 17, 3 (2002), 148–161.
- [5] Albert, Adrian, Kaur, Jasleen, and Gonzalez, Marta C. Using convolutional networks and satellite imagery to identify patterns in urban environments at a large scale. In *Knowledge Discovery and Data Mining* (2017).
- [6] Ali, Rubaba, Barra, Alvaro Federico, Berg, Claudia N, Damania, Richard, Nash, John D, and Russ, Jason. Infrastructure in conflict-prone and fragile environments: evidence from the democratic republic of congo. *World Bank Policy Research Working Paper*, 7273 (2015).
- [7] All, Sustainable Energy For. Jumeme’s unique mini-grid model gains traction in tanzania, Sept. 2018.
- [8] Alonso, Monica, Amaris, Hortensia, Germain, Jean Gardy, and Galan, Juan Manuel. Optimal charging scheduling of electric vehicles in smart grids by heuristic algorithms. *Energies* 7, 4 (2014), 2449–2475.
- [9] Andreadis, KM, Brinkerhoff, CB, and Gleason, CJ. Constraining the assimilation of swot observations with hydraulic geometry relations. *Water Resources Research* 56, 5 (2020), e2019WR026611.
- [10] Arsenault, Richard, and Brissette, François P. Continuous streamflow prediction in ungauged basins: The effects of equifinality and parameter set selection on uncertainty in regionalization approaches. *Water Resources Research* 50, 7 (2014), 6135–6153.



- [11] Aziz, Omar I Abdul, and Burn, Donald H. Trends and variability in the hydrological regime of the mackenzie river basin. *Journal of hydrology* 319, 1-4 (2006), 282–294.
- [12] Bachmann, Jan, and Schouten, Peer. Concrete approaches to peace: infrastructure as peacebuilding. *International Affairs* 94, 2 (2018), 381–398.
- [13] Bae, Sungwoo, and Kwasinski, Alexis. Spatial and temporal model of electric vehicle charging demand. *IEEE Transactions on Smart Grid* 3, 1 (2011), 394–403.
- [14] Bahdanau, Dzmitry, Cho, Kyunghyun, and Bengio, Yoshua. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473* (2014).
- [15] Bai, Yutong, Mei, Jieru, Yuille, Alan L, and Xie, Cihang. Are transformers more robust than cnns? *Advances in Neural Information Processing Systems* 34 (2021), 26831–26843.
- [16] Bao, Jining, Zhang, Yunzhou, Su, Xiaolin, and Zheng, Rui. Unpaved road detection based on spatial fuzzy clustering algorithm. *Journal on Image and Video Processing* 26 (2018).
- [17] Baroni, Gabriele, Schalge, Bernd, Rakovec, Oldrich, Kumar, Rohini, Schüler, Lennart, Samaniego, Luis, Simmer, Clemens, and Attinger, Sabine. A comprehensive distributed hydrological modeling intercomparison to support process representation and data collection strategies. *Water Resources Research* 55, 2 (2019), 990–1010.
- [18] Basijokaite, R, and Kelleher, C. Time-varying sensitivity analysis reveals relationships between watershed climate and variations in annual parameter importance in regions with strong interannual variability. *Water Resources Research* 57, 1 (2021), e2020WR028544.
- [19] Beaudoin, H, and Rodell, M. Gldas noah land surface model l4 3 hourly 0.25× 0.25 degree v2. 0, greenbelt, maryland, usa, goddard earth sciences data and information services center (ges disc)[data set], 2019.
- [20] Belvederesi, Chiara, Zaghoul, Mohamed S, Achari, Gopal, Gupta, Anil, and Hassan, Quazi K. Modelling river flow in cold and ungauged regions: a review of the purposes, methods, and challenges. *Environmental Reviews* 30, 1 (2022), 159–173.
- [21] Berg, Claudia N, Blankespoor, Brian, and Selod, Harris. Roads and rural development in sub-saharan africa. In *The Transformation of Rural Africa*. Routledge, 2020, pp. 80–100.
- [22] Bergstra, James, and Bengio, Yoshua. Random search for hyper-parameter optimization. *Journal of machine learning research* 13, 2 (2012).
- [23] Berrar, Daniel. Cross-validation., 2019.

- [24] Beuran, Monica, Gachassin, Marie Castaing, and Raballand, Gael. Are there myths on road impact and transport in sub-saharan africa?
- [25] Beven, Keith J. Uniqueness of place and process representations in hydrological modelling. *Hydrology and earth system sciences* 4, 2 (2000), 203–213.
- [26] Bhatt, Umang, Xiang, Alice, Sharma, Shubham, Weller, Adrian, Taly, Ankur, Jia, Yunhan, Ghosh, Joydeep, Puri, Ruchir, Moura, José MF, and Eckersley, Peter. Explainable machine learning in deployment. In *Proceedings of the 2020 conference on fairness, accountability, and transparency* (2020), pp. 648–657.
- [27] Biancamaria, Sylvain, Lettenmaier, Dennis P, and Pavelsky, Tamlin M. The swot mission and its capabilities for land hydrology. In *Remote sensing and water resources*. Springer, 2016, pp. 117–147.
- [28] Bickel, Peter J, Li, Bo, Tsybakov, Alexandre B, van de Geer, Sara A, Yu, Bin, Valdés, Teófilo, Rivero, Carlos, Fan, Jianqing, and van der Vaart, Aad. Regularization in statistics. *Test* 15, 2 (2006), 271–344.
- [29] Birhanu, Dereje, Kim, Hyeonjun, and Jang, Cheolhee. Effectiveness of introducing crop coefficient and leaf area index to enhance evapotranspiration simulations in hydrologic models. *Hydrological Processes* 33, 16 (2019), 2206–2226.
- [30] Boats, Ruban Bleu Electric. How to choose an electric propulsion system for your boat?, Apr. 2017.
- [31] Boyle, Douglas P, Gupta, Hoshin V, Sorooshian, Soroosh, et al. Multicriteria calibration of hydrologic models. *Calibration of Watershed Models, edited by: Duan, Q., Gupta, H., Sorooshian, S., Rousseau, A., Turcotte, R., AGU* (2003), 185–196.
- [32] Brants, Thorsten, Papat, Ashok C, Xu, Peng, Och, Franz J, and Dean, Jeffrey. Large language models in machine translation.
- [33] Brewer, Ethan, Lin, Jason, Kemper, Peter, Hennin, John, and Runfola, Dan. Predicting road quality using high resolution satellite imagery: A transfer learning approach. *Plos one* 16, 7 (2021), e0253370.
- [34] Brinkerhoff, CB, Gleason, CJ, Feng, D, and Lin, P. Constraining remote river discharge estimation using reach-scale geomorphology. *Water Resources Research* 56, 11 (2020), e2020WR027949.
- [35] Brinkerhoff, CB, Gleason, CJ, and Ostendorf, DW. Reconciling at-a-station and at-many-stations hydraulic geometry through river-wide geomorphology. *Geophysical Research Letters* 46, 16 (2019), 9637–9647.
- [36] Brown, James, Burnside, William R, Davidson, Ana D, and Delong, John P. Energetic limits to economic growth. *Bioscience* 61, 19 (Jan. 2011), 19–26.

- [37] Buck, Samuel F. A method of estimation of missing values in multivariate data suitable for use with an electronic computer. *Journal of the Royal Statistical Society: Series B (Methodological)* 22, 2 (1960), 302–306.
- [38] Burnash, Robert JC, Ferral, R Larry, and McGuire, Robert A. *A generalized stream-flow simulation system: Conceptual modeling for digital computers*. US Department of Commerce, National Weather Service, and State of California . . . , 1973.
- [39] Cadamuro, Gabriel, Muhebwa, Aggrey, and Taneja, Jay. Assigning a grade: Accurate measurement of road quality using satellite imagery. *arXiv preprint arXiv:1812.01699* (2018).
- [40] Cadamuro, Gabriel, Muhebwa, Aggrey, and Taneja, Jay. Street smarts: measuring intercity road quality using deep learning on satellite imagery. In *Proceedings of the 2nd ACM SIGCAS Conference on Computing and Sustainable Societies* (2019), pp. 145–154.
- [41] Cai, Hui, Du, W, Yu, XP, Gao, S, Littler, T, and Wang, HF. Day-ahead optimal charging/discharging scheduling for electric vehicles in micro-grids. *Prot Control Mod Power Syst* 3, 9 (Oct. 2018).
- [42] Carmody, John M. Rural electrification in the united states. *The ANNALS of the American Academy of Political and Social Science* 201, 1 (Jan. 1939), 82–88.
- [43] Charrad, Malika, Ghazzali, Nadia, Boiteau, Véronique, and Niknafs, Azam. Nbclust: An r package for determining the relevant number of clusters in a data set. *Journal of Statistical Software* 61, 6 (Oct. 2014), 236–244.
- [44] Chaudhari, Sneha, Mithal, Varun, Polatkan, Gungor, and Ramanath, Rohan. An attentive survey of attention models. *ACM Transactions on Intelligent Systems and Technology (TIST)* 12, 5 (2021), 1–32.
- [45] Che, Zhengping, Purushotham, Sanjay, Cho, Kyunghyun, Sontag, David, and Liu, Yan. Recurrent neural networks for multivariate time series with missing values. *Scientific reports* 8, 1 (2018), 1–12.
- [46] Chung, Junyoung, Gulcehre, Caglar, Cho, KyungHyun, and Bengio, Yoshua. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555* (2014).
- [47] Claesen, Marc, and De Moor, Bart. Hyperparameter search in machine learning. *arXiv preprint arXiv:1502.02127* (2015).
- [48] Clair, Thomas A, and Ehrman, James M. Using neural networks to assess the influence of changing seasonal climates in modifying discharge, dissolved organic carbon, and nitrogen export in eastern canadian rivers. *Water Resources Research* 34, 3 (1998), 447–455.

- [49] Clark, Martyn P, Nijssen, Bart, Lundquist, Jessica D, Kavetski, Dmitri, Rupp, David E, Woods, Ross A, Freer, Jim E, Gutmann, Ethan D, Wood, Andrew W, Gochis, David J, et al. A unified approach for process-based hydrologic modeling: 2. model implementation and case studies. *Water Resources Research* 51, 4 (2015), 2515–2542.
- [50] Clark, Martyn P, Rupp, David E, Woods, Ross A, Zheng, Xiaogu, Ibbitt, Richard P, Slater, Andrew G, Schmidt, Jochen, and Uddstrom, Michael J. Hydrological data assimilation with the ensemble kalman filter: Use of streamflow observations to update states in a distributed hydrological model. *Advances in water resources* 31, 10 (2008), 1309–1324.
- [51] Clark, Martyn P, Schaeffli, Bettina, Schymanski, Stanislaus J, Samaniego, Luis, Luce, Charles H, Jackson, Bethanna M, Freer, Jim E, Arnold, Jeffrey R, Moore, R Dan, Istanbuluoglu, Erkan, et al. Improving the theoretical underpinnings of process-based hydrologic models. *Water Resources Research* 52, 3 (2016), 2350–2365.
- [52] Coghlan, Benjamin, Ngoy, Pascal, Mulumba, Flavien, Hardy, Colleen, Bemo, Valerie Nkamgang, Stewart, Tony, Lewis, Jennifer, and Brennan, Richard J. Update on mortality in the democratic republic of congo: results from a third nationwide survey. *Disaster medicine and public health preparedness* 3, 2 (2009), 88–96.
- [53] Cramer, Wolfgang, Guiot, Joël, Fader, Marianela, Garrabou, Joaquim, Gattuso, Jean-Pierre, Iglesias, Ana, Lange, Manfred A, Lionello, Piero, Llasat, Maria Carmen, Paz, Shlomit, et al. Climate change and interconnected risks to sustainable development in the mediterranean. *Nature Climate Change* 8, 11 (2018), 972–980.
- [54] Cui, Xintong, Guo, Xiaoyu, Wang, Yidi, Wang, Xuelei, Zhu, Weihong, Shi, Jianghong, Lin, Chunye, and Gao, Xiang. Application of remote sensing to water environmental processes under a changing climate. *Journal of Hydrology* 574 (2019), 892–902.
- [55] Cuiyun, Cheng, and Chazhong, Ge. Green development assessment for countries along the belt and road. *Journal of environmental management* 263 (2020), 110344.
- [56] Damania, Richard, Barra, Alvaro Federico, Burnouf, Mathilde, and Russ, Jason Daniel. Transport, economic growth, and deforestation in the democratic republic of congo.
- [57] D’Amour, Alexander, Heller, Katherine, Moldovan, Dan, Adlam, Ben, Alipanahi, Babak, Beutel, Alex, Chen, Christina, Deaton, Jonathan, Eisenstein, Jacob, Hoffman, Matthew D, et al. Underspecification presents challenges for credibility in modern machine learning. *arXiv preprint arXiv:2011.03395* (2020).
- [58] Dartus, Denis, Courivaud, JM, and Dedecker, L. Use of a neural net for the study of a flood wave propagation in an open channel. *Journal of Hydraulic Research/Journal de Recherches Hydraulique* 31, 2 (1993), 161–170.

- [59] Dembélé, Moctar, Hrachowitz, Markus, Savenije, Hubert HG, Mariéthoz, Grégoire, and Schaepli, Bettina. Improving the predictive skill of a distributed hydrological model by calibration on spatial patterns with multiple satellite data sets. *Water resources research* 56, 1 (2020), e2019WR026085.
- [60] Demir, Ilke, Koperski, Krzysztof, Lindenbaum, David, Pang, Guan, Huang, Jing, Basu, Saikat, Hughes, Forest, Tuia, Devis, and Raskar, Ramesh. Deepglobe 2018: A challenge to parse the earth through satellite images. In *DeepGlobe Workshop at CVPR (DeepGlobe 2018)* (2018).
- [61] Deng, Jia, Dong, Wei, Socher, Richard, Li, Li-Jia, Li, Kai, and Fei-Fei, Li. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition* (2009), Ieee, pp. 248–255.
- [62] Dey, Sumi, and Fuentes, Olac. Predicting solar x-ray flux using deep learning techniques. In *2020 International Joint Conference on Neural Networks (IJCNN)* (2020), IEEE, pp. 1–7.
- [63] Dickey, David A, and Pantula, Sastry G. Determining the order of differencing in autoregressive processes. *Journal of Business & Economic Statistics* 5, 4 (1987), 455–461.
- [64] DigitalGlobe. Basemap +Vivid Product. [dg - cms - uploads - production.s3.amazonaws.com / uploads / document / file / 2 / DG.Basemap\\_Vivid\\_DS.1.pdf](https://s3.amazonaws.com/dg-cms-uploads-production/uploads/document/file/2/DG.Basemap_Vivid_DS.1.pdf).
- [65] Dimitriadis, Panayiotis, Koutsoyiannis, Demetris, Iliopoulou, Theano, and Papanicolaou, Panos. A global-scale investigation of stochastic similarities in marginal distribution and dependence structure of key hydrological-cycle processes. *Hydrology* 8, 2 (2021), 59.
- [66] Dirscherl, Mariel, Dietz, Andreas J, Kneisel, Christof, and Kuenzer, Claudia. Automated mapping of antarctic supraglacial lakes using a machine learning approach. *Remote Sensing* 12, 7 (2020), 1203.
- [67] Döll, Petra, Kaspar, Frank, and Lehner, Bernhard. A global hydrological model for deriving water availability indicators: model tuning and validation. *Journal of Hydrology* 270, 1-2 (2003), 105–134.
- [68] Dorosh, Paul, Wang, Hyoung Gun, You, Liangzhi, and Schmidt, Emily. Road connectivity, population, and crop production in sub-saharan africa. *Agricultural Economics* 43, 1 (2012), 89–103.
- [69] Dosovitskiy, Alexey, Beyer, Lucas, Kolesnikov, Alexander, Weissenborn, Dirk, Zhai, Xiaohua, Unterthiner, Thomas, Dehghani, Mostafa, Minderer, Matthias, Heigold, Georg, Gelly, Sylvain, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).

- [70] Durand, Michael, Chen, Curtis, de Moraes Frasson, Renato Prata, Pavelsky, Tamlin M, Williams, Brent, Yang, Xiao, and Fore, Alex. How will radar layover impact swot measurements of water surface elevation and slope, and estimates of river discharge? *Remote Sensing of Environment* 247 (2020), 111883.
- [71] Durand, Michael, Gleason, CJ, Garambois, Pierre-André, Bjerklie, David, Smith, LC, Roux, H el ene, Rodriguez, Elizandro, Bates, Paul D, Pavelsky, Tamlin M, Monnier, Jerome, et al. An intercomparison of remote sensing river discharge estimation algorithms from measurements of river height, width, and slope. *Water Resources Research* 52, 6 (2016), 4527–4549.
- [72] Earth, Google. Google earth pro, 2017–.
- [73] Eck, Douglas, and Schmidhuber, Juergen. A first look at music composition using lstm recurrent neural networks. *Istituto Dalle Molle Di Studi Sull Intelligenza Artificiale* 103 (2002), 48.
- [74] Energy, HOMER. Generating synthetic solar data, Feb. 2020.
- [75] (ESMAP), Energy Sector Management Assistance Program. Mini grids for half a billion people: Market outlook and handbook for decision makers. Technical Report 014/19, World Bank, Washington, DC, 2019. Executive Summary.
- [76] Fan, Jerome, Upadhye, Suneel, and Worster, Andrew. Understanding receiver operating characteristic (roc) curves. *Canadian Journal of Emergency Medicine* 8, 1 (2006), 19–20.
- [77] Fantini, C. *Infrastructure for Peacebuilding: The Role of Infrastructure in Tackling the Underlying Drivers of Fragility*. UNOPS, 2020.
- [78] Fatichi, Simone, Vivoni, Enrique R, Ogden, Fred L, Ivanov, Valeriy Y, Mirus, Benjamin, Gochis, David, Downer, Charles W, Camporese, Matteo, Davison, Jason H, Ebel, Brian, et al. An overview of current applications, challenges, and future trends in distributed process-based models in hydrology. *Journal of Hydrology* 537 (2016), 45–60.
- [79] Feng, Dapeng, Fang, Kuai, and Shen, Chaopeng. Enhancing streamflow forecast and extracting insights using long-short term memory networks with data integration at continental scales. *Water Resources Research* 56, 9 (2020), e2019WR026793.
- [80] Feng, Dapeng, Lawson, Kathryn, and Shen, Chaopeng. Mitigating prediction error of deep learning streamflow models in large data-sparse regions with ensemble modeling and soft data. *Geophysical Research Letters* 48, 14 (2021), e2021GL092999.
- [81] Feng, Dongmei, Gleason, Colin J, Lin, Peirong, Yang, Xiao, Pan, Ming, and Ishitsuka, Yuta. Recent changes to arctic river discharge. *Nature Communications* 12, 1 (2021), 1–9.

- [82] Feng, Dongmei, Gleason, Colin J, Yang, Xiao, and Pavelsky, Tamlin M. Comparing discharge estimates made via the bam algorithm in high-order arctic rivers derived solely from optical cubesat, landsat, and sentinel-2 data. *Water Resources Research* 55, 9 (2019), 7753–7771.
- [83] Feng, Dongmei, Gleason, Colin J, Yang, Xiao, and Pavelsky, Tamlin M. Comparing discharge estimates made via the bam algorithm in high-order arctic rivers derived solely from optical cubesat, landsat, and sentinel-2 data. *Water Resources Research* 55, 9 (2019), 7753–7771.
- [84] Flath, Christoph, Nicolay, David, Conte, Tobias, van Dinther, Clemens, and Filipova-Neumann, Lilia. Cluster analysis of smart metering data - an implementation in practice. *Business and Information Systems Engineering* 1 (Mar. 2011), 31–39.
- [85] Fobi, Simone, Deshpande, Varun, Ondiek, Samson, Modi, Vijay, and Taneja, Jay. A longitudinal study of electricity consumption growth in kenya. *Energy Policy* 123 (Dec. 2018), 569–578.
- [86] Forslöf, Lars, and Jones, Hans. Roadroid: Continuous road condition monitoring with smart phones. *Journal of Civil Engineering and Architecture* 9, 4 (2015), 485–496.
- [87] Fraiwan, Luay, and Alkhodari, Mohanad. Investigating the use of uni-directional and bi-directional long short-term memory models for automatic sleep stage scoring. *Informatics in medicine unlocked* 20 (2020), 100370.
- [88] Frasson, Renato Prata de Moraes, Pavelsky, Tamlin M, Fonstad, Mark A, Durand, Michael T, Allen, George H, Schumann, Guy, Lion, Christine, Beighley, R Edward, and Yang, Xiao. Global relationships between river width, slope, catchment area, meander wavelength, sinuosity, and discharge. *Geophysical Research Letters* 46, 6 (2019), 3252–3262.
- [89] Fry, Timothy J, and Maxwell, Reed M. Using a distributed hydrologic model to improve the green infrastructure parameterization used in a lumped model. *Water* 10, 12 (2018), 1756.
- [90] Fujita, Koji. Effect of precipitation seasonality on climatic sensitivity of glacier mass balance. *Earth and Planetary Science Letters* 276, 1-2 (2008), 14–19.
- [91] Furey, Peter R, and Gupta, Vijay K. Effects of excess rainfall on the temporal variability of observed peak-discharge power laws. *Advances in Water Resources* 28, 11 (2005), 1240–1253.
- [92] Ghojogh, Benyamin, and Crowley, Mark. The theory behind overfitting, cross validation, regularization, bagging, and boosting: tutorial. *arXiv preprint arXiv:1905.12787* (2019).

- [93] Ghosh, Shalini, Vinyals, Oriol, Strophe, Brian, Roy, Scott, Dean, Tom, and Heck, Larry. Contextual lstm (clstm) models for large scale nlp tasks. *arXiv preprint arXiv:1602.06291* (2016).
- [94] Gleason, Colin J, and Durand, Michael T. Remote sensing of river discharge: A review and a framing for the discipline. *Remote Sensing* 12, 7 (2020), 1107.
- [95] Gleason, Colin J, and Hamdan, Ali N. Crossing the (watershed) divide: Satellite data and the changing politics of international river basins. *The Geographical Journal* 183, 1 (2017), 2–15.
- [96] Gleason, Colin J, and Smith, Laurence C. Toward global mapping of river discharge using satellite images and at-many-stations hydraulic geometry. *Proceedings of the National Academy of Sciences* 111, 13 (2014), 4788–4791.
- [97] Gleason, Colin J, Smith, Laurence C, and Lee, Jinny. Retrieval of river discharge solely from satellite imagery and at-many-stations hydraulic geometry: Sensitivity to river form and optimization parameters. *Water Resources Research* 50, 12 (2014), 9604–9619.
- [98] GOGLA, Global, Lighting, for Access Coalition, Efficiency, and Berenschot. Global off-grid solar market report semi-annual sales and impact data. Tech. rep., GOGLA, The Netherlands, 2019.
- [99] Goldstein, Richard. Atmospheric limitations to repeat-track radar interferometry. *Geophysical research letters* 22, 18 (1995), 2517–2520.
- [100] Goodfellow, Ian, Bengio, Yoshua, and Courville, Aaron. *Deep learning*. MIT press, 2016.
- [101] Gorelick, Noel, Hancher, Matt, Dixon, Mike, Ilyushchenko, Simon, Thau, David, and Moore, Rebecca. Google earth engine: Planetary-scale geospatial analysis for everyone. *Remote sensing of Environment* 202 (2017), 18–27.
- [102] governmentOfCanada. Canadian Water Office. [https : / / wateroffice.ec.gc.ca/](https://wateroffice.ec.gc.ca/).
- [103] Grahn, P., Alvehag, K., and Söder, L. Plug-in-vehicle mobility and charging flexibility markov model based on driving behavior. In *2012 9th International Conference on the European Energy Market* (2012), IEEE, pp. 1–8.
- [104] Graves, Alex, and Schmidhuber, Jürgen. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural networks* 18, 5-6 (2005), 602–610.
- [105] Gu, Jiuxiang, Wang, Zhenhua, Kuen, Jason, Ma, Lianyang, Shahroudy, Amir, Shuai, Bing, Liu, Ting, Wang, Xingxing, Wang, Gang, Cai, Jianfei, et al. Recent advances in convolutional neural networks. *Pattern recognition* 77 (2018), 354–377.



- [106] Gueguen, Lionel, and Hamid, Raffay. Large-scale damage detection using satellite imagery. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2015).
- [107] Guidotti, Riccardo, Monreale, Anna, Ruggieri, Salvatore, Turini, Franco, Giannotti, Fosca, and Pedreschi, Dino. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)* 51, 5 (2018), 1–42.
- [108] Gupta, Hoshin V, Kling, Harald, Yilmaz, Koray K, and Martinez, Guillermo F. Decomposition of the mean squared error and nse performance criteria: Implications for improving hydrological modelling. *Journal of hydrology* 377, 1-2 (2009), 80–91.
- [109] Gwilliam, Ken, and Shalizi, Zmarak. Road Funds, User Charges, and Taxes. *The World Bank Research Observer* 14, 2 (1999), 159–186.
- [110] Haddeland, Ingjerd, Heinke, Jens, Biemans, Hester, Eisner, Stephanie, Flörke, Martina, Hanasaki, Naota, Konzmann, Markus, Ludwig, Fulco, Masaki, Yoshimitsu, Schewe, Jacob, et al. Global water resources affected by human interventions and climate change. *Proceedings of the National Academy of Sciences* 111, 9 (2014), 3251–3256.
- [111] Hagemann, MW, Gleason, CJ, and Durand, MT. Bam: Bayesian amhg-manning inference of discharge using remotely sensed stream width, slope, and height. *Water Resources Research* 53, 11 (2017), 9692–9707.
- [112] Halff, Albert H, Halff, Henry M, and Azmoodeh, Masoud. Predicting runoff from rainfall using neural networks. In *Engineering hydrology* (1993), ASCE, pp. 760–765.
- [113] Hallegatte, Stéphane, and Przulski, Valentin. The economics of natural disasters: concepts and methods. *World Bank Policy Research Working Paper*, 5507 (2010).
- [114] Hardt, Michaela, Chen, Xiaoguang, Cheng, Xiaoyi, Donini, Michele, Gelman, Jason, Gollaprolu, Satish, He, John, Larroy, Pedro, Liu, Xinyu, McCarthy, Nick, et al. Amazon sagemaker clarify: Machine learning bias detection and explainability in the cloud. *arXiv preprint arXiv:2109.03285* (2021).
- [115] Hartvigsson, Elias, and Ahlgren, Erik O. Comparison of load profiles in a mini-grid: Assessment of performance metrics using measured and interview-based data. *Energy for Sustainable Development* 43 (Feb. 2018), 186–195.
- [116] He, Kaiming, Zhang, Xiangyu, Ren, Shaoqing, and Sun, Jian. Deep residual learning for image recognition. In *IEEE conference on Computer Vision and Pattern Recognition (CVPR)* (2016).
- [117] He, Kaiming, Zhang, Xiangyu, Ren, Shaoqing, and Sun, Jian. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2016), pp. 770–778.

- [118] Herderschee, Johannes, Kaiser, Kai-Alexander, and Samba, Daniel Mukoko. *Resilience of an African giant: boosting growth and development in the Democratic Republic of Congo*. World Bank Publications, 2011.
- [119] Hinton, Geoffrey E, Srivastava, Nitish, Krizhevsky, Alex, Sutskever, Ilya, and Salakhutdinov, Ruslan R. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580* (2012).
- [120] Hirpa, Feyera A, Gebremichael, Mekonnen, Hopson, Thomas M, Wojick, Rafal, and Lee, Haksu. Assimilation of satellite soil moisture retrievals into a hydrologic model for improving river discharge. *Remote sensing of the terrestrial water cycle 206* (2014), 319.
- [121] Hirpa, Feyera A, Salamon, Peter, Beck, Hylke E, Lorini, Valerio, Alfieri, Lorenzo, Zsoter, Ervin, and Dadson, Simon J. Calibration of the global flood awareness system (glofas) using daily streamflow data. *Journal of Hydrology* 566 (2018), 595–606.
- [122] Hirsch, Robert M, and Costa, John E. Us stream flow measurement and data dissemination improve. *Eos, Transactions American Geophysical Union* 85, 20 (2004), 197–203.
- [123] Hirschberg, Julia, and Manning, Christopher D. Advances in natural language processing. *Science* 349, 6245 (2015), 261–266.
- [124] Hochreiter, Sepp, and Schmidhuber, Jürgen. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.
- [125] Hosking, Jonathan RM. Modeling persistence in hydrological time series using fractional differencing. *Water resources research* 20, 12 (1984), 1898–1908.
- [126] Hrachowitz, Markus, Savenije, HHG, Blöschl, G, McDonnell, JJ, Sivapalan, M, Pomeroy, JW, Arheimer, Berit, Blume, Theresa, Clark, MP, Ehret, U, et al. A decade of predictions in ungauged basins (pub)—a review. *Hydrological sciences journal* 58, 6 (2013), 1198–1255.
- [127] Hsu, Kuo-lin, Gupta, Hoshin Vijai, and Sorooshian, Soroosh. Artificial neural network modeling of the rainfall-runoff process. *Water resources research* 31, 10 (1995), 2517–2530.
- [128] Hu, Caihong, Wu, Qiang, Li, Hui, Jian, Shengqi, Li, Nan, and Lou, Zhengzheng. Deep learning with a long short-term memory networks approach for rainfall-runoff simulation. *Water* 10, 11 (2018), 1543.
- [129] Huang, Qi, Long, Di, Du, Mingda, Han, Zhongying, and Han, Pengfei. Daily continuous river discharge estimation for ungauged basins using a hydrologic model calibrated by satellite altimetry: Implications for the swot mission. *Water Resources Research* 56, 7 (2020), e2020WR027309.

- [130] Hunink, Johannes E, Eekhout, Joris PC, de Vente, Joris, Contreras, Sergio, Droogers, Peter, and Baille, Alain. Hydrological modelling using satellite-based crop coefficients: A comparison of methods at the basin scale. *Remote Sensing* 9, 2 (2017), 174.
- [131] Iandola, Forrest N, Han, Song, Moskewicz, Matthew W, Ashraf, Khalid, Dally, William J, and Keutzer, Kurt. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and; 0.5 mb model size. *arXiv preprint arXiv:1602.07360* (2016).
- [132] IEA. Energy access outlook 2017, 2017. Special report.
- [133] Immerzeel, Walter W, Van Beek, Ludovicus PH, and Bierkens, Marc FP. Climate change will affect the asian water towers. *science* 328, 5984 (2010), 1382–1385.
- [134] Ishitsuka, Yuta, Gleason, Colin J, Hagemann, Mark W, Beighley, Edward, Allen, George H, Feng, Dongmei, Lin, Peirong, Pan, Ming, Andreadis, Konstantinos, and Pavelsky, Tamlin M. Combining optical remote sensing, mcfl discharge estimation, global hydrologic modeling, and data assimilation to improve daily discharge estimates across an entire large watershed. *Water Resources Research* 57, 3 (2021), e2020WR027794.
- [135] Jamshidian, Mortaza, and Mata, Matthew. Advances in analysis of mean and covariance structure when data are incomplete. In *Handbook of latent variable and related models*. Elsevier, 2007, pp. 21–44.
- [136] Jean, Neal, Burke, Marshall, Xie, Michael, Davis, W Matthew, Lobell, David B, and Ermon, Stefano. Combining satellite imagery and machine learning to predict poverty. *Science* 353, 6301 (2016), 790–794.
- [137] Jiang, Dejuan, and Wang, Kun. The role of satellite-based remote sensing in improving simulated streamflow: A review. *Water* 11, 8 (2019), 1615.
- [138] Jones, JA. Hydrologic processes and peak discharge response to forest removal, regrowth, and roads in 10 small experimental basins, western cascades, oregon. *Water Resources Research* 36, 9 (2000), 2621–2642.
- [139] Kabeja, Crispin, Li, Rui, Guo, Jianping, Rwatangabo, Digne Edmond Rwabuhungu, Manyifika, Marc, Gao, Zongting, Wang, Yipu, and Zhang, Yuxiang. The impact of reforestation induced land cover change (1990–2017) on flood peak discharge using hec-hms hydrological model and satellite observations: A study in two mountain basins, china. *Water* 12, 5 (2020), 1347.
- [140] Kahraman, Aysegul, Hou, Peng, Yang, Guangya, and Yang, Zhile. Comparison of the effect of regularization techniques and lookback window length on deep learning models in short term load forecasting. In *Proceedings of 2021 International Top-Level Forum on Engineering Science and Technology Development Strategy* (2022), Springer, pp. 655–669.

- [141] Karunanithi, Nachimuthu, Grenney, William J, Whitley, Darrell, and Bovee, Ken. Neural networks for river flow prediction. *Journal of computing in civil engineering* 8, 2 (1994), 201–220.
- [142] Khan, Salman, Naseer, Muzammal, Hayat, Munawar, Zamir, Syed Waqas, Khan, Fahad Shahbaz, and Shah, Mubarak. Transformers in vision: A survey. *ACM computing surveys (CSUR)* 54, 10s (2022), 1–41.
- [143] Kirchner, James W. Getting the right answers for the right reasons: Linking measurements, analyses, and models to advance the science of hydrology. *Water Resources Research* 42, 3 (2006).
- [144] Kittel, Cecile MM, Arildsen, Anne L, Dybkjær, Stine, Hansen, Emilie R, Linde, Ida, Slott, Emma, Tøttrup, Christian, and Bauer-Gottwein, Peter. Informing hydrological models of poorly gauged river catchments—a parameter regionalization and calibration approach. *Journal of Hydrology* 587 (2020), 124999.
- [145] Knoben, Wouter JM, Freer, Jim E, and Woods, Ross A. Inherent benchmark or not? comparing nash–sutcliffe and kling–gupta efficiency scores. *Hydrology and Earth System Sciences* 23, 10 (2019), 4323–4331.
- [146] Kolhe, Mohan L., Ranaweera, K.M. Iromi Udumbara, and Gunawardana, A.G.B. Sisara. Techno-economic sizing of off-grid hybrid renewable energy system for rural electrification in sri lanka. *Sustainable Energy Technologies and Assessments* 11 (Mar. 2015), 53–64.
- [147] Kompas, Tom, Pham, Van Ha, and Che, Tuong Nhu. The effects of climate change on gdp by country and the global economic gains from complying with the paris climate accord. *Earth’s Future* 6, 8 (2018), 1153–1173.
- [148] Konapala, Goutam, Mishra, Ashok K, Wada, Yoshihide, and Mann, Michael E. Climate change will affect global water availability through compounding changes in seasonal precipitation and evaporation. *Nature communications* 11, 1 (2020), 1–10.
- [149] Kratzert, Frederik, Herrnegger, Mathew, Klotz, Daniel, Hochreiter, Sepp, and Klambauer, Günter. Neuralhydrology—interpreting lstms in hydrology. In *Explainable AI: Interpreting, explaining and visualizing deep learning*. Springer, 2019, pp. 347–362.
- [150] Kratzert, Frederik, Klotz, Daniel, Herrnegger, Mathew, Sampson, Alden K, Hochreiter, Sepp, and Nearing, Grey S. Toward improved predictions in ungauged basins: Exploiting the power of machine learning. *Water Resources Research* 55, 12 (2019), 11344–11354.
- [151] Kratzert, Frederik, Klotz, Daniel, Herrnegger, Mathew, Sampson, Alden K, Hochreiter, Sepp, and Nearing, Grey S. Toward improved predictions in ungauged basins: Exploiting the power of machine learning. *Water Resources Research* 55, 12 (2019), 11344–11354.

- [152] Kratzert, Frederik, Klotz, Daniel, Shalev, Guy, Klambauer, Günter, Hochreiter, Sepp, and Nearing, Grey. Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets. *Hydrology and Earth System Sciences* 23, 12 (2019), 5089–5110.
- [153] Kratzert, Frederik, Klotz, Daniel, Shalev, Guy, Klambauer, Günter, Hochreiter, Sepp, and Nearing, Grey. Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets. *Hydrology and Earth System Sciences* 23, 12 (2019), 5089–5110.
- [154] Krizhevsky, Alex, Sutskever, Ilya, and Hinton, Geoffrey E. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems (NIPS)* (2012).
- [155] Krizhevsky, Alex, Sutskever, Ilya, and Hinton, Geoffrey E. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems (NIPS)* (2012).
- [156] Lahariya, Manu, Sadeghianpourhamami, Nasrin, and Devellder, Chris. Reduced state space and cost function in reinforcement learning for demand response control of multiple ev charging stations. In *Proceedings of the 6th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation* (2019), ACM Inc, pp. 344–345.
- [157] Lakkaraju, Himabindu, Bach, Stephen H, and Leskovec, Jure. Interpretable decision sets: A joint framework for description and prediction. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (2016), pp. 1675–1684.
- [158] Lakkaraju, Himabindu, Kamar, Ece, Caruana, Rich, and Leskovec, Jure. Faithful and customizable explanations of black box models. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society* (2019), pp. 131–138.
- [159] Larnier, Kevin, and Monnier, Jerome. Hybrid neural network–variational data assimilation algorithm to infer river discharges from swot-like data. *Nonlinear Processes in Geophysics Discussions* (2020), 1–30.
- [160] LeCun, Yann, Bengio, Yoshua, and Hinton, Geoffrey. Deep learning. *nature* 521, 7553 (2015), 436–444.
- [161] Liang, Xu, Lettenmaier, Dennis P, Wood, Eric F, and Burges, Stephen J. A simple hydrologically based model of land surface water and energy fluxes for general circulation models. *Journal of Geophysical Research: Atmospheres* 99, D7 (1994), 14415–14428.
- [162] Lim, Sunghoon, Kim, Sun Jun, Park, YoungJae, and Kwon, Nahyun. A deep learning-based time series model with missing value handling techniques to predict various types of liquid cargo traffic. *Expert Systems with Applications* 184 (2021), 115532.

- [163] Limited, Alibaba Group Holding. Flake ice machine fish storage, 2020.
- [164] Lin, Peirong, Pan, Ming, Beck, Hylke E, Yang, Yuan, Yamazaki, Dai, Frasson, Renato, David, Cédric H, Durand, Michael, Pavelsky, Tamlin M, Allen, George H, et al. Global reconstruction of naturalized river flows at 2.94 million reaches. *Water resources research* 55, 8 (2019), 6499–6516.
- [165] Liu, Zhuang, Mao, Hanzi, Wu, Chao-Yuan, Feichtenhofer, Christoph, Darrell, Trevor, and Xie, Saining. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022), pp. 11976–11986.
- [166] Long, Mingsheng, Zhu, Han, Wang, Jianmin, and Jordan, Michael I. Deep transfer learning with joint adaptation networks. In *International conference on machine learning* (2017), PMLR, pp. 2208–2217.
- [167] Lu, Zhiying, Xie, Hongtao, Liu, Chuanbin, and Zhang, Yongdong. Bridging the gap between vision transformers and convolutional neural networks on small datasets. *arXiv preprint arXiv:2210.05958* (2022).
- [168] Lundberg, Scott, and Lee, Su-In. A unified approach to interpreting model predictions. *arXiv preprint arXiv:1705.07874* (2017).
- [169] Lundberg, Scott M, Erion, Gabriel, Chen, Hugh, DeGrave, Alex, Prutkin, Jordan M, Nair, Bala, Katz, Ronit, Himmelfarb, Jonathan, Bansal, Nisha, and Lee, Su-In. From local explanations to global understanding with explainable ai for trees. *Nature machine intelligence* 2, 1 (2020), 56–67.
- [170] Lundberg, Scott M, Nair, Bala, Vavilala, Monica S, Horibe, Mayumi, Eisses, Michael J, Adams, Trevor, Liston, David E, Low, Daniel King-Wai, Newman, Shu-Fang, Kim, Jerry, et al. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nature biomedical engineering* 2, 10 (2018), 749–760.
- [171] Ma, Kai, Feng, Dapeng, Lawson, Kathryn, Tsai, Wen-Ping, Liang, Chuan, Huang, Xiaorong, Sharma, Ashutosh, and Shen, Chaopeng. Transferring hydrologic data across continents—leveraging data-rich regions to improve hydrologic prediction in data-sparse regions. *Water Resources Research* 57, 5 (2021), e2020WR028600.
- [172] Mai, Juliane, Craig, James R, Tolson, Bryan A, and Arsenault, Richard. The sensitivity of simulated streamflow to individual hydrologic processes across north america. *Nature communications* 13, 1 (2022), 1–11.
- [173] Malemo Kalisya, Luc, Lussy Justin, Paluku, Kimona, Christophe, Nyavandu, Kavira, Mukekulu Eugenie, Kamabu, Jonathan, Kasereka Muhindo Lusi, Claude, Kasereka Masumbuko, and Hawkes, Michael. Sexual violence toward children and youth in war-torn eastern democratic republic of congo. *PloS one* 6, 1 (2011), e15911.

- [174] Mamirkulova, Gulnara, Mi, Jianing, Abbas, Jaffar, Mahmood, Shahid, Mubeen, Ri-aqa, and Ziapour, Arash. New silk road infrastructure opportunities in developing tourism environment for residents better quality of life. *Global Ecology and Conservation* 24 (2020), e01194.
- [175] Marcinkevičs, Ričards, and Vogt, Julia E. Interpretability and explainability: A machine learning zoo mini-tour. *arXiv preprint arXiv:2012.01805* (2020).
- [176] Marine, Pennine. Torqeedo - deep blue - electric outboard engines, 2012.
- [177] Marshall, Lucy, Nott, David, and Sharma, Ashish. Hydrological model selection: A bayesian alternative. *Water resources research* 41, 10 (2005).
- [178] Mastorakis, Georgios. Human-like machine learning: limitations and suggestions. *arXiv preprint arXiv:1811.06052* (2018).
- [179] McLaughlin, Dennis. Recent developments in hydrologic data assimilation. *Reviews of Geophysics* 33, S2 (1995), 977–984.
- [180] McMillan, HK, and Westerberg, IK. Rating curve estimation under epistemic uncertainty. *Hydrological Processes* 29, 7 (2015), 1873–1882.
- [181] Mhaskar, Hrushikesh Narhar, and Micchelli, Charles A. How to choose an activation function. *Advances in neural information processing systems* 6 (1993).
- [182] Mitchell, Tom M, and Mitchell, Tom M. *Machine learning*, vol. 1. McGraw-hill New York, 1997.
- [183] Mnih, Volodymyr, and Hinton, Geoffrey E. Learning to detect roads in high-resolution aerial images. In *European Conference on Computer Vision (ECCV 2010)* (2010).
- [184] Molland, Anthony F., Turnock, Stephen R., and Hudson, Dominic A. *Ship resistance and propulsion [electronic resource] : practical estimation of ship propulsive power*. Cambridge University Press, 32 Avenues of the Americas, New York, NY 10013-2473 USA, 2011.
- [185] Montanari, Alberto, and Koutsoyiannis, Demetris. A blueprint for process-based modeling of uncertain hydrological systems. *Water Resources Research* 48, 9 (2012).
- [186] Moyalán, J, Sawant, M, Bhagyashree, U, Sheikh, A, Wagh, S, and Singh, N. Electric vehicle-power grid incorporation using distributed resource allocation approach. In *2019 18th European Control Conference (ECC)* (2019), IEEE, pp. 3034–3039.
- [187] Muhammad, Ameer, Evenson, Grey R, Stadnyk, Tricia A, Boluwade, Alaba, Jha, Sanjeev Kumar, and Coulibaly, Paulin. Impact of model structure on the accuracy of hydrological modeling of a canadian prairie watershed. *Journal of Hydrology: Regional Studies* 21 (2019), 40–56.

- [188] Mullins, Christopher W, and Rothe, Dawn L. Gold, diamonds and blood: International state-corporate crime in the democratic republic of the congo. *Contemporary Justice Review* 11, 2 (2008), 81–99.
- [189] Nadkarni, Prakash M, Ohno-Machado, Lucila, and Chapman, Wendy W. Natural language processing: an introduction. *Journal of the American Medical Informatics Association* 18, 5 (2011), 544–551.
- [190] Naseer, Muhammad Muzammal, Ranasinghe, Kanchana, Khan, Salman H, Hayat, Munawar, Shahbaz Khan, Fahad, and Yang, Ming-Hsuan. Intriguing properties of vision transformers. *Advances in Neural Information Processing Systems* 34 (2021), 23296–23308.
- [191] Nash, J Eamonn, and Sutcliffe, Jonh V. River flow forecasting through conceptual models part i—a discussion of principles. *Journal of hydrology* 10, 3 (1970), 282–290.
- [192] Nearing, Grey S, Kratzert, Frederik, Sampson, Alden Keefe, Pelissier, Craig S, Klotz, Daniel, Frame, Jonathan M, Prieto, Cristina, and Gupta, Hoshin V. What role does hydrological science play in the age of machine learning? *Water Resources Research* 57, 3 (2021), e2020WR028091.
- [193] Nepal, Santosh, Flügel, Wolfgang-Albert, and Shrestha, Arun Bhakta. Upstream-downstream linkages of hydrological processes in the himalayan region. *Ecological Processes* 3, 1 (2014), 1–16.
- [194] Neves, Diana, and A.Silva, Carlos. Optimal electricity dispatch on isolated mini-grids using a demand response strategy for thermal storage backup with genetic algorithms. *Energy* 82 (Mar. 2015), 436–445.
- [195] Newsdesk, Central Banking. Bank of uganda implements 100bp rate cut, Oct. 2019.
- [196] Nezamoddini, Nasim, and Wang, Yong. Risk management and participation planning of electric vehicles in smart grids for demand response. *Energy* 116, 1 (Oct. 2016), 836–850.
- [197] Nichols, Gordon, Lake, Iain, and Heaviside, Clare. Climate change and water-related infectious diseases. *Atmosphere* 9, 10 (2018), 385.
- [198] Niu, Zhaoyang, Zhong, Guoqiang, and Yu, Hui. A review on the attention mechanism of deep learning. *Neurocomputing* 452 (2021), 48–62.
- [199] Ntegeka, Victor, Baguis, Pierre, Roulin, Emmanuel, and Willems, Patrick. Developing tailored climate change scenarios for hydrological impact assessments. *Journal of Hydrology* 508 (2014), 307–321.
- [200] of Statistics, Kenya National Bureau. Kenya integrated household budget survey 2015-2016, Apr. 2018.



- [201] Ortiz-Bobea, Ariel, Ault, Toby R, Carrillo, Carlos M, Chambers, Robert G, and Lobell, David B. Anthropogenic climate change has slowed global agricultural productivity growth. *Nature Climate Change* 11, 4 (2021), 306–312.
- [202] Osborne, Martin J, et al. *An introduction to game theory*, vol. 3. Oxford university press New York, 2004.
- [203] O’Shea, Keiron, and Nash, Ryan. An introduction to convolutional neural networks. *arXiv preprint arXiv:1511.08458* (2015).
- [204] Oubanas, Hind, Gejadze, Igor, Malaterre, P-O, Durand, M, Wei, Rui, Frasson, Renato PM, and Domeneghetti, Alessio. Discharge estimation in ungauged basins through variational data assimilation: The potential of the swot mission. *Water Resources Research* 54, 3 (2018), 2405–2423.
- [205] Oubanas, Hind, Gejadze, Igor, Malaterre, Pierre-Olivier, and Mercier, Franck. River discharge estimation from synthetic swot-type observations using variational data assimilation and the full saint-venant hydraulic model. *Journal of Hydrology* 559 (2018), 638–647.
- [206] Oudin, Ludovic, Andréassian, Vazken, Perrin, Charles, Michel, Claude, and Le Moine, Nicolas. Spatial proximity, physical similarity, regression and ungauged catchments: A comparison of regionalization approaches based on 913 french catchments. *Water Resources Research* 44, 3 (2008).
- [207] Ouyang, Wenyu, Lawson, Kathryn, Feng, Dapeng, Ye, Lei, Zhang, Chi, and Shen, Chaopeng. Continental-scale streamflow modeling of basins with reservoirs: Towards a coherent deep-learning-based strategy. *Journal of Hydrology* 599 (2021), 126455.
- [208] Owen, Guillermo. *Game theory*. Emerald Group Publishing, 2013.
- [209] Ozturk, Ilhan. A literature survey on energy–growth nexus. *Energy Policy* 38 (Oct. 2010), 340–349.
- [210] Pagliero, Liliana, Bouraoui, Fayçal, Diels, Jan, Willems, Patrick, and McIntyre, Neil. Investigating regionalization techniques for large-scale hydrological modelling. *Journal of hydrology* 570 (2019), 220–235.
- [211] Pan, Sinno Jialin, and Yang, Qiang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering* 22, 10 (2009), 1345–1359.
- [212] Pan, Sinno Jialin, and Yang, Qiang. A survey on transfer learning iee transactions on knowledge and data engineering. 22 (10): 1345 1359 (2010).
- [213] Pandey, Shailesh, Agarwal, Tushar, and Krishnan, Narayanan C. Multi-task deep learning for predicting poverty from satellite images. In *Proceedings of the AAAI Conference on Artificial Intelligence* (2018), vol. 32.

- [214] Pant, Neeraj, Semwal, Prabhat, Khobragade, Suhas Damodar, Rai, Shive Prakash, Kumar, Sudhir, Dubey, Rajendra Kumar, Noble, Jacob, Joshi, Suneel Kumar, Rawat, Yadhvir Singh, Nainwal, Harish Chandra, et al. Tracing the isotopic signatures of cryospheric water and establishing the altitude effect in central himalayas: A tool for cryospheric water partitioning. *Journal of Hydrology* 595 (2021), 125983.
- [215] Patz, Jonathan A, Campbell-Lendrum, Diarmid, Holloway, Tracey, and Foley, Jonathan A. Impact of regional climate change on human health. *Nature* 438, 7066 (2005), 310–317.
- [216] Peters, Jorg, Sievert, Maximiliane, and Toman, Michael A. Rural electrification through mini-grids: Challenges ahead. *Energy Policy* 132 (May 2019), 27–31.
- [217] Piaggese, Simone, Gauvin, Laetitia, Tizzoni, Michele, Cattuto, Ciro, Adler, Natalia, Verhulst, Stefaan, Young, Andrew, Price, Rhiannan, Ferres, Leo, and Panisson, André. Predicting city poverty using satellite imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (2019), pp. 90–96.
- [218] Pilz, Tobias, Francke, Till, Baroni, Gabriele, and Bronstert, Axel. How to tailor my process-based hydrological model? dynamic identifiability analysis of flexible model structures. *Water resources research* 56, 8 (2020), e2020WR028042.
- [219] Pokhrel, Yadu, Shin, Sanghoon, Lin, Zihan, Yamazaki, Dai, and Qi, Jianguo. potential disruption of flood dynamics in the lower mekong river basin due to upstream flow regulation. *Scientific reports* 8, 1 (2018), 1–13.
- [220] Pool, Sandra, and Seibert, Jan. Gauging ungauged catchments—active learning for the timing of point discharge observations in combination with continuous water level measurements. *Journal of Hydrology* 598 (2021), 126448.
- [221] Power, Equatorial. Off-grid, solar — uganda - lolwe island, Feb. 2019.
- [222] Pueyo, Ana, and DeMartino, Samantha. The impact of solar mini-grids on kenya’s rural enterprises. *Energy for Sustainable Development* 45 (May 2018), 28–37.
- [223] Rácz, Anita, Bajusz, Dávid, and Héberger, Károly. Effect of dataset size and train/test split ratios in qsar/qspr multiclass classification. *Molecules* 26, 4 (2021), 1111.
- [224] Raghu, Maithra, Unterthiner, Thomas, Kornblith, Simon, Zhang, Chiyuan, and Dosovitskiy, Alexey. Do vision transformers see like convolutional neural networks? *Advances in Neural Information Processing Systems* 34 (2021), 12116–12128.
- [225] Ramachandran, Prajit, Zoph, Barret, and Le, Quoc V. Searching for activation functions. *arXiv preprint arXiv:1710.05941* (2017).
- [226] Ramchandran, Neeraj, Pai, Rajesh, and Parihar, Amit Kumar Singh. Feasibility assessment of anchor-business-community model for offgrid rural electrification in india. *Renewable energy* 97 (May 2016), 197–209.

- [227] Ramos, S., Duarte, J. M. M., Soares, J., Vale, Z., and Duarte, F. J. Typical load profiles in the smart grid context. *IEEE Power and Energy Society General Meeting 132* (July 2012), 1–8.
- [228] Rao, R Bharat, Fung, Glenn, and Rosales, Romer. On the dangers of cross-validation. an experimental evaluation. In *Proceedings of the 2008 SIAM international conference on data mining* (2008), SIAM, pp. 588–596.
- [229] Rassaei, Farshad, Soh, Wee-Seng, and Chua, Kee-Chaing. Demand response for residential electric vehicles with random usage patterns in smart grids. *IEEE Transactions on Sustainable Energy* 6, 4 (Oct. 2015), 1367–1376.
- [230] Ray, Patrick A, Yang, Yi-Chen E, Wi, Sungwook, Khalil, Abedalrazq, Chatikavani, Vansa, and Brown, Casey. Room for improvement: Hydroclimatic challenges to poverty-reducing development of the brahmaputra river basin. *Environmental Science & Policy* 54 (2015), 64–80.
- [231] Reabroy, Ratthakrit, Tiaple, Yodchai, Pongduang, Sathit, Nantawong, Tewarat, and Iamraksa, Phansak. The possibility of using electrical motor for boat propulsion system. *Energy Procedia* 79 (Nov. 2015), 1008–1014.
- [232] Rebosio, Michelle, and Wam, Per Egil. Violent conflict and the road sector: Points of interaction.
- [233] Refaeilzadeh, Payam, Tang, Lei, and Liu, Huan. Cross-validation. *Encyclopedia of database systems* 5 (2009), 532–538.
- [234] Robert, F. C., Sisodia, G. S., and Gopalan, S. The critical role of anchor customers in rural microgrids: Impact of load factor on energy cost. In *2017 International Conference on Computation of Power, Energy Information and Commuincation (ICCPEIC)* (2017), IEEE, pp. 398–403.
- [235] Rodell, Matthew, Houser, PR, Jambor, UEA, Gottschalck, J, Mitchell, Kieran, Meng, C-J, Arsenault, Kristi, Cosgrove, B, Radakovich, J, Bosilovich, M, et al. The global land data assimilation system. *Bulletin of the American Meteorological society* 85, 3 (2004), 381–394.
- [236] Roelofs, Rebecca, Shankar, Vaishaal, Recht, Benjamin, Fridovich-Keil, Sara, Hardt, Moritz, Miller, John, and Schmidt, Ludwig. A meta-analysis of overfitting in machine learning. *Advances in Neural Information Processing Systems* 32 (2019).
- [237] Russakovsky, Olga, Deng, Jia, Su, Hao, Krause, Jonathan, Satheesh, Sanjeev, Ma, Sean, Huang, Zhiheng, Karpathy, Andrej, Khosla, Aditya, Bernstein, Michael, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision* 115, 3 (2015), 211–252.

- [238] Rußwurm, Marc, and Korner, Marco. Temporal vegetation modelling using long short-term memory networks for crop identification from medium-resolution multi-spectral satellite images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops* (2017), pp. 11–19.
- [239] Sak, Haşim, Senior, Andrew, Rao, Kanishka, and Beaufays, Françoise. Fast and accurate recurrent neural network acoustic models for speech recognition. *arXiv preprint arXiv:1507.06947* (2015).
- [240] Salehinejad, Hojjat, Sankar, Sharan, Barfett, Joseph, Colak, Errol, and Valaee, Shahrokh. Recent advances in recurrent neural networks. *arXiv preprint arXiv:1801.01078* (2017).
- [241] Sayers, M. W. On the calculation of international roughness index from longitudinal road profile. *Transportation Research Record 1501* (1996), 1–12.
- [242] Schaller, Robert R. Moore’s law: past, present and future. *IEEE spectrum* 34, 6 (1997), 52–59.
- [243] Schouten, Peer, and Bachmann, Jan. Roads to peace? the role of infrastructure in fragile and conflict-affected states. *Danish Institute for International Studies, Copenhagen* (2017).
- [244] Scown, Murray W. The sustainable development goals need geoscience. *Nature Geoscience* 13, 11 (2020), 714–715.
- [245] Segal, Mark R. Machine learning benchmarks and random forest regression.
- [246] Seifert, Axel, and Rasp, Stephan. Potential and limitations of machine learning for modeling warm-rain cloud microphysical processes. *Journal of Advances in Modeling Earth Systems* 12, 12 (2020), e2020MS002301.
- [247] Sheffield, J, Wood, Eric F, Pan, M, Beck, H, Coccia, G, Serrat-Capdevila, A, and Verbist, K. Satellite remote sensing for water resources management: Potential for supporting sustainable development in data-poor regions. *Water Resources Research* 54, 12 (2018), 9724–9758.
- [248] Shen, Chaopeng. A transdisciplinary review of deep learning research and its relevance for water resources scientists. *Water Resources Research* 54, 11 (2018), 8558–8593.
- [249] Shen, Hongren, Tolson, Bryan A, and Mai, Juliane. Time to update the split-sample approach in hydrological model calibration. *Water Resources Research* 58, 3 (2022), e2021WR031523.
- [250] Shi, Zhenwei, and Zou, Zhengxia. Can a machine generate humanlike language descriptions for a remote sensing image? *IEEE Transactions on Geoscience and Remote Sensing* 55, 6 (2017), 3623–3634.

- [251] Siami-Namini, Sima, Tavakoli, Neda, and Namin, Akbar Siami. The performance of lstm and bilstm in forecasting time series. In *2019 IEEE International Conference on Big Data (Big Data)* (2019), IEEE, pp. 3285–3292.
- [252] Sidle, Roy C, Gomi, Takashi, Usuga, Juan Carlos Loaiza, and Jarihani, Ben. Hydrogeomorphic processes and scaling issues in the continuum from soil pedons to catchments. *Earth-Science Reviews* 175 (2017), 75–96.
- [253] Simonyan, Karen, and Zisserman, Andrew. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).
- [254] Simonyan, Karen, and Zisserman, Andrew. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).
- [255] S.M.Mousavi, and Flynn, Damian. Controlled charging of electric vehicles to minimize energy losses in distribution systems. *IIFAC-PapersOnLine* 49, 27 (Oct. 2016), 324–329.
- [256] Soares, João, Morais, Hugo, Sousa, Tiago, Vale, Zita, and Faria, Pedro. Day-ahead resource scheduling including demand response for electric vehicles. *IEEE Transactions on Smart Grid* 4, 1 (Mar. 2013), 596–605.
- [257] Sokorai, Peter, Fleischhacker, Andreas, Lettner, Georg, and Auer, Hans. Stochastic modeling of the charging behavior of electromobility. *World Electric Vehicle Journal* 9, 3 (Oct. 2018), 44–58.
- [258] Song, Mingjun, and Civco, Daniel. Road extraction using svm and image segmentation. *Photogrammetric Engineering & Remote Sensing* 70, 12 (2004), 1365–1371.
- [259] Srivastava, Nitish, Mansimov, Elman, and Salakhudinov, Ruslan. Unsupervised learning of video representations using lstms. In *International conference on machine learning* (2015), PMLR, pp. 843–852.
- [260] Stearns, Jason K. Helping congo help itself: What it will take to end africa’s worst war. *Foreign Aff.* 92 (2013), 99.
- [261] Stone, Mervyn. Cross-validation: A review. *Statistics: A Journal of Theoretical and Applied Statistics* 9, 1 (1978), 127–139.
- [262] Su, Xiaogang, Yan, Xin, and Tsai, Chih-Ling. Linear regression. *Wiley Interdisciplinary Reviews: Computational Statistics* 4, 3 (2012), 275–294.
- [263] Sun, Alexander Y, Jiang, Peishi, Mudunuru, Maruti K, and Chen, Xingyuan. Explore spatio-temporal learning of large sample hydrology using graph neural networks. *Water Resources Research* 57, 12 (2021), e2021WR030394.
- [264] Tan, Chuanqi, Sun, Fuchun, Kong, Tao, Zhang, Wenchang, Yang, Chao, and Liu, Chunfang. A survey on deep transfer learning. In *International conference on artificial neural networks* (2018), Springer, pp. 270–279.

- [265] Tang, Wanrong, and Zhang, Ying Jun Angela. A model predictive control approach for low-complexity electric vehicle charging scheduling: Optimality and scalability. *IEEE transactions on power systems* 32, 2 (2016), 1050–1063.
- [266] Team, Planet. Planet application program interface: In space for life on earth, 2017–.
- [267] The World Bank. Africa’s pulse.
- [268] Thomas, C. M., and Featherstone, W. E. Validation of vincenty’s formulas for the geodesic using a new fourth-order extension of kivioja’s formula. *Journal of surveying engineering* 131, 1 (Feb. 2015), 20–26.
- [269] Thorn, Jessica PR, Bignoli, Diego Juffe, Mwangi, Ben, and Marchant, Robert A. The african development corridors database: a new tool to assess the impacts of infrastructure investments. *Scientific data* 9, 1 (2022), 1–11.
- [270] Thyer, Mark, Renard, Benjamin, Kavetski, Dmitri, Kuczera, George, Franks, Stewart William, and Srikanthan, Sri. Critical evaluation of parameter consistency and predictive uncertainty in hydrological modeling: A case study using bayesian total error analysis. *Water Resources Research* 45, 12 (2009).
- [271] Tian, Siyuan, Tregoning, Paul, Renzullo, Luigi J, van Dijk, Albert IJM, Walker, Jeffrey P, Pauwels, Valentijn RN, and Allgeyer, Sébastien. Improved water balance component estimates through joint assimilation of grace water storage and smos soil moisture retrievals. *Water Resources Research* 53, 3 (2017), 1820–1840.
- [272] Torqeedo. Deep blue 25 rl, 2020.
- [273] Touvron, Hugo, Cord, Matthieu, Douze, Matthijs, Massa, Francisco, Sablayrolles, Alexandre, and Jégou, Hervé. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning* (2021), PMLR, pp. 10347–10357.
- [274] Tran, Quoc Quan, De Niel, J, and Willems, P. Spatially distributed conceptual hydrological model building: A generic top-down approach starting from lumped models. *Water Resources Research* 54, 10 (2018), 8064–8085.
- [275] Tsai, Wen-Ping, Feng, Dapeng, Pan, Ming, Beck, Hylke, Lawson, Kathryn, Yang, Yuan, Liu, Jiangtao, and Shen, Chaopeng. From calibration to parameter learning: Harnessing the scaling effects of big data in geoscientific modeling. *Nature communications* 12, 1 (2021), 1–13.
- [276] Ullah, Amin, Ahmad, Jamil, Muhammad, Khan, Sajjad, Muhammad, and Baik, Sung Wook. Action recognition in video sequences using deep bi-directional lstm with cnn features. *IEEE Access* 6 (2018), 1155–1166.
- [277] Vaswani, Ashish, Shazeer, Noam, Parmar, Niki, Uszkoreit, Jakob, Jones, Llion, Gomez, Aidan N, Kaiser, Łukasz, and Polosukhin, Illia. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

- [278] Volpi, Elena, Di Lazzaro, Michele, Bertola, Miriam, Viglione, Alberto, and Fiori, Aldo. Reservoir effects on flood peak discharge at the catchment scale. *Water Resources Research* 54, 11 (2018), 9623–9636.
- [279] Vorob'Ev, Nicolai N. *Foundations of game theory: noncooperative games*. Birkhäuser, 2012.
- [280] Wagener, Thorsten, and Montanari, Alberto. Convergence of approaches toward reducing uncertainty in predictions in ungauged basins. *Water Resources Research* 47, 6 (2011).
- [281] Wager, Stefan, Wang, Sida, and Liang, Percy S. Dropout training as adaptive regularization. *Advances in neural information processing systems* 26 (2013).
- [282] Walcott-Bryant, Aisha, Bryant, Reginald E, Tatsubori, Michiaki, Emaasit, Daniel, Osebe, Samuel, Wamburu, John, and Fobi, Simone. The living roads project: Giving a voice to roads in developing cities. *Transportation Research Board – 96th Annual Meeting* (2017).
- [283] Wang, Huabin, Cheng, Qian, Wang, Taoyang, Zhang, Guo, Wang, Yunming, Li, Xin, and Jiang, Boyang. Layover compensation method for regional spaceborne sar imagery without gcps. *IEEE Transactions on Geoscience and Remote Sensing* 59, 10 (2021), 8367–8381.
- [284] Wang, Huaijun, Cao, Lei, and Feng, Ru. Hydrological similarity-based parameter regionalization under different climate and underlying surfaces in ungauged basins. *Water* 13, 18 (2021), 2508.
- [285] Wang, Wei, Zheng, Vincent W, Yu, Han, and Miao, Chunyan. A survey of zero-shot learning: Settings, methods, and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)* 10, 2 (2019), 1–37.
- [286] Wanner, Jonas, Herm, Lukas-Valentin, and Janiesch, Christian. How much is the black box? the value of explainability in machine learning models.
- [287] Watts, Nick, Adger, W Neil, Agnolucci, Paolo, Blackstock, Jason, Byass, Peter, Cai, Wenjia, Chaytor, Sarah, Colbourn, Tim, Collins, Mat, Cooper, Adam, et al. Health and climate change: policy responses to protect public health. *The lancet* 386, 10006 (2015), 1861–1914.
- [288] Wilbanks, Thomas, Bhatt, Vatsal, Bilello, Daniel, Bull, Stanley, Ekmann, James, Horak, William, Huang, Y Joe, Levine, Mark D, Sale, Michael J, Schmalzer, David, et al. Effects of climate change on energy production and use in the united states. *US Department of Energy Publications* (2008), 12.
- [289] Williams, N. J., Jaramillo, P., Cornell, B., Lyons-Galante, I., and Wynn, E. Load characteristics of east african microgrids. In *2017 IEEE PES PowerAfrica* (2017), IEEE, pp. 236–241.

- [290] Williams, Nathaniel J., Jaramillo, Paulina, Taneja, Jay, and Ustun, Taha Selim. Enabling private sector investment in microgrid-based rural electrification in developing countries: A review. *Renewable and Sustainable Energy Reviews* 52 (Aug. 2015), 1268–1281.
- [291] Wu, Bichen, Xu, Chenfeng, Dai, Xiaoliang, Wan, Alvin, Zhang, Peizhao, Yan, Zhicheng, Tomizuka, Masayoshi, Gonzalez, Joseph, Keutzer, Kurt, and Vajda, Peter. Visual transformers: Token-based image representation and processing for computer vision, 2020.
- [292] Wu, Hao, and Prasad, Saurabh. Convolutional recurrent neural networks for hyperspectral data classification. *Remote Sensing* 9, 3 (2017), 298.
- [293] Wu, Wenyan, May, Robert J, Maier, Holger R, and Dandy, Graeme C. A benchmarking approach for comparing data splitting methods for modeling water resources parameters using artificial neural networks. *Water Resources Research* 49, 11 (2013), 7598–7614.
- [294] Xie, Kang, Liu, Pan, Zhang, Jianyun, Han, Dongyang, Wang, Guoqing, and Shen, Chaopeng. Physics-guided deep learning for rainfall-runoff modeling by considering extreme events and monotonic relationships. *Journal of Hydrology* 603 (2021), 127043.
- [295] Xie, Michael, Jean, Neal, Burke, Marshall, Lobell, David, and Ermon, Stefano. Transfer learning from deep features for remote sensing and poverty mapping. In *Thirtieth AAAI Conference on Artificial Intelligence* (2016).
- [296] Yamazaki, Dai, Ikeshima, Daiki, Sosa, Jeison, Bates, Paul D, Allen, George H, and Pavelsky, Tamlin M. Merit hydro: A high-resolution global hydrography map based on latest topography dataset. *Water Resources Research* 55, 6 (2019), 5053–5073.
- [297] Yang, Junjing, Ning, Chao, Deb, Chirag, Zhang, Fan, Cheong, David, Lee, Siew Eang, Sekhar, Chandra, and Tham, Kwok Wai. k-shape clustering algorithm for building energy usage patterns analysis and forecasting model accuracy improvement. *Energy and Buildings* 146 (Mar. 2017), 27–37.
- [298] Yang, Xue, Magnusson, Jan, Huang, Shaochun, Beldring, Stein, and Xu, Chong-Yu. Dependence of regionalization methods on the complexity of hydrological models in multiple climatic regions. *Journal of Hydrology* 582 (2020), 124357.
- [299] Yeh, Christopher, Perez, Anthony, Driscoll, Anne, Azzari, George, Tang, Zhongyi, Lobell, David, Ermon, Stefano, and Burke, Marshall. Using publicly available satellite imagery and deep learning to understand economic well-being in africa. *Nature communications* 11, 1 (2020), 1–11.
- [300] Yin, Wenpeng, Kann, Katharina, Yu, Mo, and Schütze, Hinrich. Comparative study of cnn and rnn for natural language processing. *arXiv preprint arXiv:1702.01923* (2017).



- [301] Yoshida, T, Hanasaki, N, Nishina, K, Boulange, J, Okada, M, and Troch, PA. Inference of parameters for a global hydrological model: Identifiability and predictive uncertainties of climate-based parameters. *Water Resources Research* 58, 2 (2022), e2021WR030660.
- [302] Yu, Chen, Li, Zhenhong, Penna, Nigel T, and Crippa, Paola. Generic atmospheric correction model for interferometric synthetic aperture radar observations. *Journal of Geophysical Research: Solid Earth* 123, 10 (2018), 9202–9222.
- [303] Yu, Tong, and Zhu, Hong. Hyper-parameter optimization: A review of algorithms and applications. *arXiv preprint arXiv:2003.05689* (2020).
- [304] Zamir, Amir R, Sax, Alexander, Shen, William, Guibas, Leonidas J, Malik, Jitendra, and Savarese, Silvio. Taskonomy: Disentangling task transfer learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2018), pp. 3712–3722.
- [305] Zhang, Tian, Chen, Wei, Han, Zhu, and Cao, Zhigang. Charging scheduling of electric vehicles with local renewable energy under uncertain electric vehicle arrival and grid power price. *IEEE Transactions on Vehicular Technology* 63, 6 (2013), 2600–2612.
- [306] Zhang, Yue, Liu, Qi, and Song, Linfeng. Sentence-state lstm for text representation. *arXiv preprint arXiv:1805.02474* (2018).
- [307] Zheng, Wendong, Zhao, Putian, Huang, Kai, and Chen, Gang. Understanding the property of long term memory for the lstm with attention mechanism. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management* (2021), pp. 2708–2717.
- [308] Zhu, Ming, Liu, Xiao-Yang, Kong, Linghe, Shen, Ruimin, Shu, Wei, and Wu, Min-You. The charging-scheduling problem for electric vehicle networks. In *2014 IEEE Wireless Communications and Networking Conference (WCNC)* (2014), IEEE, pp. 3178–3183.
- [309] Zhuang, Fuzhen, Qi, Zhiyuan, Duan, Keyu, Xi, Dongbo, Zhu, Yongchun, Zhu, Hengshu, Xiong, Hui, and He, Qing. A comprehensive survey on transfer learning. *Proceedings of the IEEE* 109, 1 (2020), 43–76.
- [310] Zounemat-Kermani, Mohammad, Batelaan, Okke, Fadaee, Marzieh, and Hinkelmann, Reinhard. Ensemble machine learning paradigms in hydrology: A review. *Journal of Hydrology* 598 (2021), 126266.