Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI-24)
Special Track on AI for Good

Time-Evolving Data Science and Artificial Intelligence for Advanced Open Environmental Science (TAIAO) Programme

Yun Sing Koh[1], **Albert Bifet**[2,3], **Karin Bryan**[1], **Guilherme Cassales**[2], **Olivier Graffeuille**[1], **Nick Lim**[2], **Phil Mourot**[2,5], **Ding Ning**[4], **Bernhard Pfahringer**[2], **Varvara Vetrova**[4] and **Heitor Gomes**[6]

[1]University of Auckland, New Zealand
[2]AI Institute, University of Waikato, New Zealand
[3]LTCI, Télécom Paris, IP Paris, France
[4]University of Canterbury, New Zealand
[5]Waikato Regional Council, New Zealand
[6]Victoria University of Wellington, New Zealand
{y.koh, karin.bryan, olivier.graffeuille}@auckland.ac.nz,
{abifet, guilherme.cassales, nick.lim, phil.mourot, bernhard}@waikato.ac.nz,
ding.ning@pg.canterbury.ac.nz, varvara.vetrova@canterbury.ac.nz, heitor.gomes@vuw.ac.nz

Abstract

New Zealand's unique ecosystems face increasing threats from climate change, impacting biodiversity and posing challenges to safety, livelihoods, and well-being. To tackle these complex issues, advanced data science and artificial intelligence techniques can provide unique solutions. Currently in its fourth year of a seven-year program, TAIAO focuses on methods for analyzing environmental datasets. Recognizing this urgency, the open-source TAIAO platform was developed. This platform enables new artificial intelligence research for environmental data and offers an open-access repository to enhance reproducibility in the field. This paper will showcase four environmental case studies, artificial intelligence research, platform implementation details, and future development plans.

1 Introduction

The evolving climate landscape presents a pressing motivation for New Zealand to prioritize environmental data science, as it becomes a strategic imperative in understanding and mitigating the impacts of climate change. Against a predominantly stable climate over the past 10,000 years, New Zealand has fostered unique species and ecosystems that define its natural environment [Ogden, 1995]. However, the stability enjoyed for millennia is now under threat due to human activities, both historical and current, which exert pressure on the environment and challenge its capacity to adapt to rapid changes. The repercussions are becoming increasingly evident, with conditions changing faster than ecosystems can naturally adjust, challenging restoration efforts. One of the prominent consequences of human-induced factors is the rapid increase in atmospheric greenhouse gas emissions, resulting in a warming Earth. The global mean surface temperature has escalated by 1.1 degrees Celsius above pre-industrial levels, and projections indicate a likely exceedance of 1.5 degrees Celsius by early 2030 [Pörtner *et al.*, 2023]. Despite the seemingly incremental nature of these temperature changes, they have already triggered substantial and far-reaching impacts.

Environmental data science is vital for research, adaptation, and conservation in this climate-induced context. The New Zealand government's ambitious objectives to enhance freshwater quality and achieve carbon neutrality by 2050 underscore the pivotal role of effective data science in addressing the complex challenges posed by climate change. Environmental time series or data streams, integral components of this scientific approach, play a ubiquitous role across diverse applications in New Zealand. These encompass monitoring observations or modeling outputs related to various environmental parameters, including flow (such as wind, current, water level, ice flow, and ice height), concentration (covering suspended sediment, nutrients, and contaminants), physical properties (spanning temperature, density), and external forcing factors (encompassing gravity and solar radiation).

The unique characteristics of environmental time series data necessitate specific processing techniques, considering evolving information of dynamic processes, the decision-making process over time-based on incomplete information, the historical or spatial context that can enhance predictive power, the significance of rare extreme events, the scarcity of data in some applications, and the multi-scale attributes of the information. Environmental data relevant to a particular problem is often diverse and multi-modal, presenting further challenges in modelling and data management. For example, a flood forecasting tool may require integration from many datasets, including meteorological data, hydrological data, topographic data, and soil moisture data, while a water quality monitoring system may benefit from remote sensing measure-

ment and manual sample collection datasets.

Problem Statement. There is a need for a centralized platform tailored to deal with environmental data problems in the New Zealand environment. A standardized platform facilitates collaboration between machine learning researchers, data scientists, and environmental scientists by managing data, notebooks, and software related to these problems.

The following will introduce the TAIAO program, its team, current case studies, implementation, and future milestones.

2 TAIAO Programme

TAIAO is a seven-year data science programme (2020 - 2027) funded by the Ministry of Business, Innovation, and Employment (MBIE) New Zealand. It will advance the state-of-the-art in environmental data science by developing new machine learning methods for time series and data streams that are able to deal with large quantities of data in real-time, and are tailored to deal with data collected in the New Zealand environment. The programme aims to build a new open-source framework to implement machine learning on time series data, provide an openly available repository with datasets to improve reproducibility in environmental data science, and build capability in fundamental and applied data science accessible to all New Zealanders.

Data are essential for research, understanding, and setting policies to manage New Zealand's environment. Still, environmental data presents many challenges that require new data science methods to overcome. This programme is currently developing new methods and building the required capability. In particular, the programme will focus on developing methods to deal with environmental datasets collected in large volumes over time. It must, therefore, be dealt with as streams that are analyzed incrementally, as they are measured, rather than as collections of data that can be analyzed all at once. These methods will address underlying characteristics of the data that evolve over time (*e.g.* due to climatic or ecological changes) and data that are collected at a range of time intervals and spatial scales ranging from broadscale satellite images to single-point measurements on the ground in the water or air. The methods we develop will be interpretable and explainable (to help users understand why an algorithm produces some particular output), identify and understand anomalies (to distinguish 'normal' from 'unusual' measurements) and quantify uncertainty in algorithm output (to help decision-makers understand how confident they can be in conclusions drawn from the data science methods). We will further discuss the research carried out in the programme in Section 4.

To deliver the methods we develop in a form that environmental scientists and managers can use, we build a new open-source framework, the TAIAO platform, to carry out machine learning on time series data and provide an open-access repository of environmental datasets to improve reproducibility in environmental data science.

2.1 TAIAO Platform

TAIAO is an online platform[1] designed to facilitate collaboration, resource-sharing, and community engagement for environmental scientists, data scientists, academics, and the broader scientific community. TAIAO is a research hub, providing access to datasets, Jupyter Notebooks, software, and tutorials, fostering a culture of collaboration and knowledge exchange [Lim *et al.*, 2023].

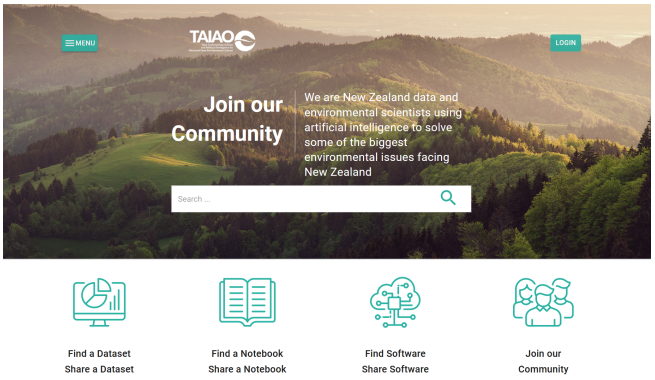

Figure 1: TAIAO Platform

Users can benefit from a user-friendly interface, allowing efficient downloading and exploring resources curated to meet their research needs, as shown in Figure 1. To navigate TAIAO effectively, users can leverage the search function and metadata filtering options for datasets and notebooks. This ensures a targeted approach to accessing relevant information. The commitment to open licenses for all shared data and notebooks ensures accessibility and promotes an ethos of collaboration.

For contributors, TAIAO offers a structured process for accessing datasets or notebooks. Beyond that, TAIAO extends its support with the tutorials page, offering insights into using Jupyter Notebooks, Python, and other tools, especially in applications for machine learning. The dataset page showcases the links to available open datasets, creating a unified location for data access. Table 1 provides information on the current datasets available. The software page provides free access to various data science software and packages, enriching the scientific toolkit.

2.2 United Nations Sustainable Development Goals (UNSDGs)

The platform's role addresses several UNSDGs [2015]. These include climate change impacts and supporting conservation research positions within Climate Action (UNSDG 13) and Sustainable Cities and Communities (UNSDG 11). Additionally, TAIAO directly contributes to Life Below Water and Life on Land (UNSDGs 14 and 15) by supporting research that can influence terrestrial ecosystems and promoting open licenses for responsible sharing of scientific resources. Beyond that, we contribute to the sustainable management of water (UNSDG 6). In promoting innovation in data science

[1] https://taiao.ai/

Domain	Dataset	Data Format
Ocean	GPATS Oceania Lightning Feed	JSON
	LINZ tidal gauge data	CSV
Freshwater	Bay of Plenty Water Quality and Discharge Data	CSV, JSON, Other, XLS
	Coromandel River and Rain Gauge Time Series	CSV
Biodiversity	Mt Karioi Predator Project	CSV, Other
	Kahikatea aerial imagery	Other
	Crown-of-Thorns Starfish	JPG
Atmosphere	Moana hydrodynamic hindcast	netCDF
	Sentinel-3	netCDF
Climate	McMurdo Dry Valley Foehn Wind Dataset	CSV
	MetService weather stations	CSV
Land	LCDB and Waikato Regional Aerial Photography	Other
	Sentinel Satellite snapshot of Waikato region	Other

Table 1: Sample datasets linked on the TAIAO platform

and environmental research, TAIAO supports Industry, Innovation, and Infrastructure (UNSDG 9).

2.3 Project Team

The program fosters a robust collaborative team, uniting data science researchers with key partners such as Beca, MetService, the University of Waikato, the University of Auckland, the University of Canterbury, and regional council environmental scientists throughout New Zealand. This collective effort is characterized by bidirectional collaborations, where the expertise of data scientists is enriched through a deepened understanding of the datasets, opportunities, and challenges specific to environmental research. Simultaneously, environmental scientists actively engage with data science methodologies, gaining insights into how these tools can illuminate data and test hypotheses related to underlying environmental processes. This collaborative synergy creates a strong and harmonious team where both groups' collective knowledge and skills contribute to the program's success.

3 Case Studies

Here, we describe four examples of case studies that the programme has mounted. Table 2 summarizes the current case studies active in the programme.

3.1 River Flood Forecasting in the Coromandel Peninsula, Waikato Region

Floods are one of the most common natural disasters and probably the most affected by climate change. The acceleration of warming temperatures increases atmospheric moisture levels, leading to more severe heavy rainfall and associated extreme weather events [Douville *et al.*, 2022].

In New Zealand, the cost of flooding has steadily increased over the last five years, peaking in 2023 with the Auckland floods in January and cyclone Gabrielle in February, with over thirty times the average cost, according to the Insurance Council of New Zealand [ICNZ, 2023]. Regional councils oversee rainfall and water levels, issue flood alerts, and maintain flood protection infrastructure. Civil defence planning is the responsibility of district councils. The initial signs of impending floods typically come from severe weather warnings provided by meteorologists. Moreover, regional councils maintain independent networks of rain gauges and river levels, which can prompt automatic alerts when a rapid increase in rainfall or river levels occurs [Potter *et al.*, 2021]. In 2021, the National Institute of Water and Atmospheric Research (NIWA) in New Zealand developed the first national-scale streamflow forecasting system. Still a proof-of-concept, the forecasting tool is a physics-based model that provides hourly forecasts with 48 hour lead time [Cattoën *et al.*, 2022]. However, the forecast uses a weather model (NZCSM), which needs heavy computational resources and runs only every 6 hours.

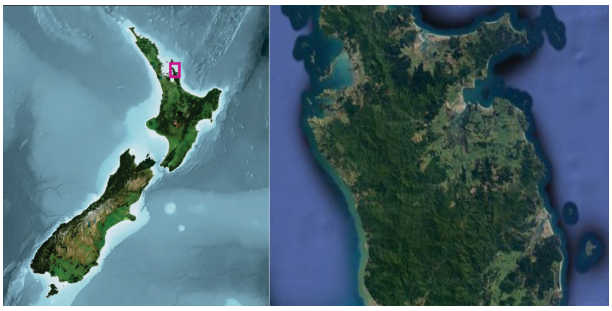

Figure 2: The region of study, The Coromandel Peninsula, Waikato, New Zealand

Accurate and timely warnings are critical for mitigating flood risks. Forecasting the magnitude and timing of floods in real-time is a challenging problem for planning how to respond quickly in emergencies. In recent years, using machine learning for flood forecasting has gained traction as a growing area of research [Mosavi *et al.*, 2018]. The main advantage of machine learning over traditional techniques is its capability to handle high-dimensional and complex non-linear datasets.

Our case study examined the Coromandel Peninsula on New Zealand's North Island (Figure 2), an area prone to severe storms and flooding due to its diverse hydrometeorological and topographic features. Our initial focus was on predicting the water levels of the three main rivers in the region with a lead time of 12 hours, using only the river stage data from the past decade. As we delved into the project, we aimed to minimise forecast errors and lag time. After evaluating various neural network architectures, we found significant improvement using Long Short-Term Memory (LSTM) networks, as suggested by [Kratzert *et al.*, 2019], along with incorporating additional datasets. Given the limited coverage of just two rain gauge stations across the expansive area, we used rain radar images provided by MetService. The rain radar data is updated every 7.5 minutes and covers a 50x70 km2 region over three elevation intervals (sea level, 500 m, 1000 m). Access to these data through several APIs allowed

Project Description	Stakeholders	UNSDG
Develop accurate water quality remote sensing machine learning models to improve water quality monitoring capabilities.	Waikato Regional Council	6, 14
Develop air quality prediction for wood smoke pollution in towns in the South Island, NZ	National Institute of Water and Atmospheric Research (NIWA)	13
Develop reliable and accurate flood forecasting in the Coromandel Peninsula, New Zealand	Waikato Regional Council	6, 11
Develop accurate forecasts of water usage and tree growth in plantation forests in New Zealand.	New Zealand Forest Research Institute Limited (SCION research) and Forest Flows project	6, 13, 15
Develop a video-based inventorying and classification system for benthic habitats in New Zealand	Department of Conservation	14
Develop a satellite and aerial imagery-based classification and segmentation system for land use inventorying in New Zealand	Waikato Regional Council	15

Table 2: Case-studies from the TAIAO programme

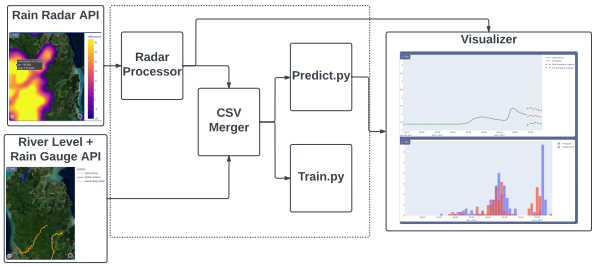

Figure 3: Data flow and design of the river-stage predictions

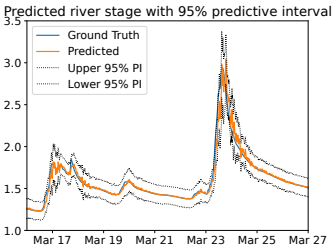

Figure 4: Visualization of the 95% predictive interval

for an automated real-time forecasting process. We developed a dashboard to visualise the neural network's output, issue warning messages, and evaluate performance (Figure 3) [Mourot *et al.*, 2022].

To estimate the uncertainty of river stage predictions, we employed Bayesian LSTM techniques, finding that the predictive intervals aligned well with theoretical 95% intervals (Figure 4). Ongoing efforts are directed towards refining these predictive intervals while preserving accuracy in mean predictions. Additionally, we explore model explainability techniques to gain deeper insights into how rivers respond to meteorological activity.

3.2 Forest Flows

Forests are important for several tasks, such as biodiversity conservation and regulating water, carbon, energy and nutrient cycling [Bar-On *et al.*, 2018; Keenan and Williams, 2018]. However, extreme events and changes in societal demand pose a challenge to the maintenance of such tasks [Bonan, 2016]. Therefore, understanding how the complex ecological processes inside the forests interact with the surrounding environment due to unforeseen climate conditions has received increased research attention [Bennett *et al.*, 2009].

The main goal in such a situation is to acquire valuable forecast information regarding the behaviour of the forest and its connected systems that enable stakeholders to respond accordingly. However, the amount of data generated by the many sensors and the sheer size of the forests create a challenging data-intensive situation. Thus, efficiently using this data is important to provide the knowledge required to achieve the full potential that forests have to act as a bio-based solution for global climate change [Seddon, 2022].

The creation and deployment of the infrastructure to collect and store data was performed by Forest Flows, NIWA, and SCION. We have collaborated in this process by guiding the tools that would facilitate the data analysis. Since then, we have been closely collaborating with the Forest Flows project to provide insights into how such systems work through the exploration of model explainability and the deployment of several ML techniques to accurately forecast forest growth and future water usage.

The forest growth forecast uses dendrometer readings as a proxy (Figure 5), which, coupled with soil and weather data, provides a powerful method to predict the growth of the forest. We measure the error in the forecast by comparing the forecast value against the reading whenever it becomes available. The model explainability is being used to understand what drives the growth of the trees and compare the findings with the forestry literature to make sure it is biologically acceptable.

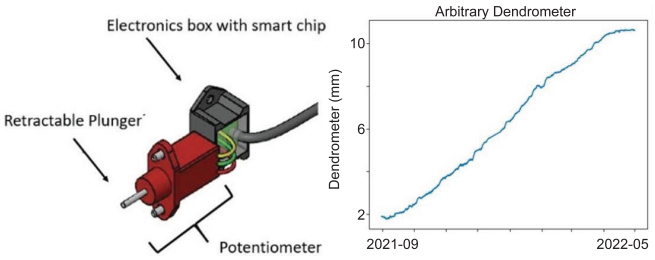

Figure 5: Diagram and data sample of a dendrometer sensor.

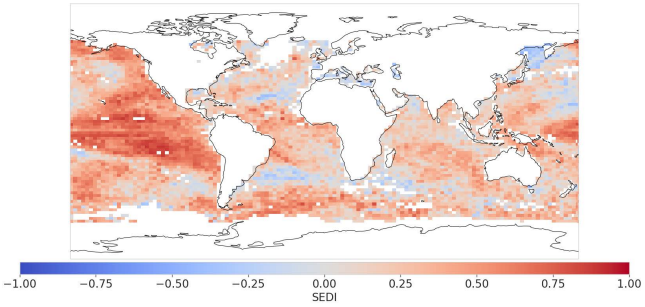

Figure 6: A four-month-ahead global MHW forecast produced by a suite of GraphSAGE models trained via a temporal diffusion process with a custom weighted MSE loss function, applied to SSTA graphs derived from the ERA5 SST reanalysis. Forecast performance at each node is evaluated for the test years (2011-2022) using the SEDI, where red indicates a positive SEDI, blue a negative SEDI, and white an undefined SEDI.

3.3 Marine Heatwaves

Marine heatwaves (MHWs) are observed around the world and have strong impacts on marine ecosystems. Such impacts include shifts in species ranges, local extinctions, and can have a follow-on economic impact on seafood industries [Hobday *et al.*, 2016]. The devastating impact on marine ecosystems caused by MHWs could bring irreversible loss of species or foundation habitats [Oliver *et al.*, 2019], for example, mass coral bleaching and substantial declines in kelp forests and seagrass meadows [Holbrook *et al.*, 2020]. MHWs also affect aquaculture businesses and area-restricted fisheries because of the change in the distribution of sea life with follow on effects on production [Hobday *et al.*, 2018], such as mussel, oyster and salmon farms.

In this case study, we aimed to advance global monthly to seasonal MHW forecasts by proposing a unified deep learning approach encompassing anomaly, spatial, and temporal aspects. First, we viewed MHW prediction, characterized by sea surface temperature anomalies (SSTAs), as an imbalanced regression task, where instances above the 90$^{\text{th}}$ percentile are underrepresented. We evaluated regression loss functions with fully-connected networks, including the MSE, MAE, Huber, focal-R [Yang *et al.*, 2021], balanced MSE [Ren *et al.*, 2022], and a custom weighted MSE, to identify the most effective one in improving performance metrics like the critical success index [Schaefer, 1990]. We integrated the selected loss functions with advanced neural network architectures and used the symmetric extremal dependence index (SEDI) [Ferro and Stephenson, 2011] for a comprehensive evaluation of MHW prediction [Jacox *et al.*, 2022].

Second, we examined graph re-sampling and graph neural networks (GNNs) to address spatial pattern learning. Graphs, representing a more generalized data structure than the commonly used grids, offer advantages in modeling global climate teleconnections, handling missing values, and avoiding issues related to receptive fields [Luo *et al.*, 2016] and the Earth's spherical properties [Defferrard *et al.*, 2019]. We developed tools to convert the ERA5 SST reanalysis [Hersbach *et al.*, 2020] into graph structures. Then we evaluated several GNN classes [Ning *et al.*, 2023] and found that the GraphSAGE model [Hamilton *et al.*, 2017] provided robust SST and SSTA forecasts. Furthermore, we specifically concentrated on forecasting at historical MHW hotspots [Oliver *et al.*, 2021].

Third, while short-term SSTA and MHW forecasts could be obtained using the sliding window method, long-term forecasts are challenging, and models usually have underfitting issues, especially when predicting anomalies. We studied recurrent methods and a temporal diffusion method to improve longer-lead forecasts. The recurrent approaches included using the LSTM aggregator within the GraphSAGE and adding LSTM layers. The diffusion method was analogue to conventional diffusion networks for image learning, where one or several interpolators with stochasticity and one forecaster are trained iteratively [Cachay *et al.*, 2023]. In summary, by integrating anomaly, spatial, and temporal aspects into a unified deep learning framework, our models have been able to provide MHW forecasts with reasonable goodness-of-fit at most locations worldwide, up to five months ahead. Figure 6 shows an example forecast.

3.4 Algal Bloom Monitoring in Lakes

Harmful algal blooms are the rapid increase in algae in water bodies, often caused by excess nutrients from fertilizer or wastewater runoff. Certain algae are toxic and can be harmful to local ecosystems, aquaculture and human health (UNSDG 6). Manual data collection is expensive, with low spatial and temporal resolution. Remote sensing algorithms instead estimate the concentration of chlorophyll-a pigment in lakes from the water colour to be used as an indicator of water quality and better monitor these events (Figure 7).

We have collaborated with the Waikato Regional Council and other local governments to combine and process data records into high-quality datasets. In partnership with freshwater scientists from Xerra Earth Observation Institute, we have defined practical outcomes in water quality monitoring and developed algorithms that achieve these outcomes.

To tackle the limited ground-truth data scarcity associated with this task due to data collection costs, we developed a semi-supervised learning algorithm that leverages unlabelled remote sensing data to improve algorithmic performance. Specifically, we use abundant satellite data without co-situated in-situ sample measurements to improve the performance of Mixture Density Networks (MDNs), a neural network architecture which captures uncertainty in predictions and is useful for water quality remote sensing due to the ill-posed nature of this task [Graffeuille *et al.*, 2022]. Our

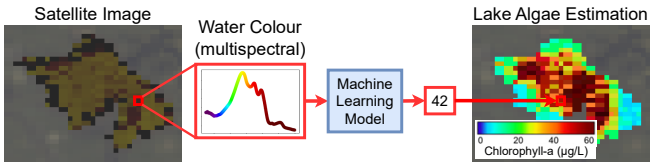

Figure 7: Water Quality Remote Sensing in Lake Waikare.

work is the first to leverage unlabelled data with MDNs, and this improves model performance equivalently to a 50% increase in the quantity of gathered labelled data. In a case study on Lake Waikare, Waikato, our model detected a sequence of algal blooms over the summer of 2021. Performance improvement and cost reduction in remote sensing technologies will allow for more universal and accurate monitoring of these environmental risks.

4 Contribution to Current Research

Our research initiative is driven by the recognition that tackling contemporary environmental challenges requires machine learning advancements.

Environmental Solutions through Novel ML Research. In the area of research focused on Machine Learning for Data Streams and Time Series, we aim to develop novel techniques specifically tailored for dynamic data streams. The goal is to address current challenges and lay the groundwork for new ML methodologies that can effectively handle real-world environmental data, incorporating advancements such as anomaly detection, clustering, and event detection. For example, consider the work on online continual learning within our wider team, led by Zhang et al. [2022]. This method aims to train neural networks incrementally from a non-stationary data stream with a single pass-through data, revisiting rehearsal dynamics in online settings. The team provides theoretical insights into the inherent memory overfitting risk from the biased and dynamic empirical risk minimization viewpoint. This example showcases our commitment to exploring innovative approaches and pushing the boundaries of ML methodologies.

Contribution to Predictive Capabilities for Environmental Extremes. In the research on Weak Signals and Extreme Events, we acknowledge the urgency of embedding predictive capabilities into existing ML algorithms. We aim to solve current environmental challenges and pave the way for more effective, adaptive ML models capable of addressing the escalating frequency of extreme environmental events.

As an example, our wider team [Milz *et al.*, 2023] has delved into understanding and predicting Foehn winds, such as those impacting the McMurdo Dry Valleys in Antarctica. Foehn winds are accelerated, warm, and dry winds with significant environmental impacts, including ice and glacial melt and the destabilization of ice shelves, potentially leading to rising sea levels. Conventional automatic detection methods rely on rule-based methodologies with static thresholds, which struggle to define the complex and varied patterns of Foehn winds in different alpine valleys worldwide. The research introduces and compares the first unsupervised machine-learning approaches for detecting Foehn wind events. Most existing machine learning approaches to this problem follow a supervised learning paradigm, relying on labels generated by imprecise static rule-based algorithms. The proposed unsupervised approaches overcome this limitation, providing a more dynamic definition of Foehn wind events independent of the location. The first approach is based on multivariate time-series clustering, while the second utilizes a deep autoencoder-based anomaly detection method. Remarkably, our best model achieves an f1-score of 88%, matching or surpassing previous machine-learning methods. This approach enhances accuracy and provides a more flexible and inclusive definition of Foehn events, showcasing the potential of unsupervised machine learning in advancing our understanding of complex meteorological phenomena.

Environmental Data Science with Deep Learning. In the area of research focused on Deep Learning, recognizing its increasing role in environmental data science, we delve into the challenges of explainability, modeling evolving interaction networks, and quantifying predictive uncertainty. This research is not just about solving immediate issues; it is about advancing ML methodologies to provide more accurate, interpretable, and adaptable tools for addressing the complex and evolving landscape of environmental problems. As an example, ongoing investigations by Jia et al. [2021], we evaluate explanations using a metric based on the area under the ROC curve (AUC). This evaluation treats expert-provided image annotations as ground-truth explanations and quantifies the correlation between model accuracy and explanation quality during image classifications with deep neural networks. The experiments span two diverse image datasets: the CUB-200-2011 dataset and a newly introduced Kahikatea dataset [Jia *et al.*, 2021]. They compare and evaluate seven neural networks paired with four explainers for each dataset, considering accuracy and explanation quality. Furthermore, they delve into how explanation quality evolves with changes in loss metrics throughout the training iterations of each model. The compelling findings from these experiments highlight a robust correlation between model accuracy and explanation quality. This example underscores our commitment to advancing the explainability of deep learning models. It showcases the exploration of diverse datasets and model-explainer combinations, contributing to a deeper understanding of the interplay between accuracy and interpretability in environmental data science applications.

5 Implementation

In this section, we describe the data policies and procedures for the platform. We also discuss the annual workshops, including partnership and outreach.

5.1 Data Policies and Procedures

The TAIAO programme is dedicated to establishing robust data policies and procedures, outlining a comprehensive framework for collecting, storing, using, re-using, accessing, and retaining data with potential impact on various partners, stakeholders, and communities. These policies aim to advocate and adopt sound data management and governance prac-

tices, thereby fortifying engagement, trust, and collaboration with all entities involved.

The core values guiding the TAIAO programme encompass a commitment to developing new and innovative data science methodologies and empowering environmental researchers and practitioners to utilize data science effectively. Additionally, the project focuses on practical accessibility, ensuring that data science is made available and useful to communities for the betterment of the environment. Openness is a key value, emphasizing the inclusivity of a wider community and the appropriate sharing of data, tools, and methods. Collaborative co-design of data science with input from the broader community is also highly valued.

These values, along with the ongoing refinement of associated policies and procedures, are underpinned by the following principles, specifically the Principles of Māori Data Sovereignty (MDSov) and adherence to the FAIR principles (Findable, Accessible, Interoperable, Reusable) and CARE principles (Collective Benefit, Authority to Control, Responsibility, Ethical).

5.2 Annual Workshops

The programme runs an annual workshop with an open invitation to disseminate and encourage collaboration. The workshop objectives:

- Ignite engaging conversations and collaborations as we delve into the latest and future trends in data science.
- Discover the TAIAO's innovative environmentally driven use cases, sparking connections with potential organizations and stakeholders.
- Unleash knowledge and insights, establish network and valuable relationships.

In 2023, the annual workshop was held in Tauranga, New Zealand. The workshop has participants from researchers from academia, AI and environmental science, local councils, and Crown Research Institutes. The 2023 event schedule is available at[2]. Some of the key highlights of the 2023 workshop include 11 case studies, an open discussion and a panel. We provide several examples as follows. First, Professor Karin Bryan addressed the pressing challenge of planning for climate changes in coastal environments. Coastal managers grappled with complex and detailed modeling systems, often unsuitable for projections at relevant timescales. Prof Bryan proposed using classification and neural networks to augment numerical modeling results, specifically focusing on Ōhiwa Harbour as a case study. Second, Dr Varvara Vetrova shed light on the significance of anomaly detection in environmental contexts. From biosecurity to climate monitoring, timely anomaly detection was crucial. Dr Vetrova discussed applying deep learning methods, emphasizing their role in preventing invasive species in urban areas and understanding climate system extremes, such as those observed in Antarctica. Her talk underscored the diverse scales at which anomaly detection could play a vital role in environmental management.

[2]https://taiao.ai/pages/taiao-workshop-2023.en/

The open panel discussion within the community entitled "Growing Capabilities in Environmental Data Science", is divided into three categories; *opportunities and people*, *data*, and *emerging trends and future*. A sample of questions discussed included:

- How has the evolution of technology transformed the landscape of environmental data science, and what new opportunities does it offer for addressing pressing environmental challenges?
- What strategies can be employed to bridge the gap between academia, industry, and policy circles to ensure that environmental data science research leads to practical and impactful outcomes?
- What innovative strategies are employed to ensure the quality, reliability, and interoperability of environmental datasets collected from various sources?

6 Future Timeline

The programme aims to build more collaboration and mount new collaborative projects every year. The plans include:

- Expanding Collaborative Projects: Building on the new discussions with future partners such as wildlife.ai and Maungatautari, the program intends to initiate new collaborative projects each year. The aim is to mount two new collaborations every year.
- Capacity Building and Training: Recognizing the importance of skill development in AI and environmental science, the program aims to provide training and capacity-building opportunities for stakeholders involved in collaborative projects similar to those attempted in the first four years.
- Scaling Up and Replication: As successful collaborative models emerge from the program, there will be a focus on scaling up successful interventions and replicating them in different geographical regions or ecosystems. This could involve partnerships with government agencies, NGOs, and local communities to leverage resources and expand the reach of conservation efforts.

7 Conclusion

The TAIAO program represents a strategic response to the urgent need for advanced environmental data science in the face of accelerating climate change. By developing new machine learning methods tailored for real-time analysis of large environmental datasets, TAIAO aims to provide actionable insights for research, adaptation, and conservation efforts. Through the TAIAO platform, researchers can collaborate and access resources, accelerating progress toward achieving the United Nations Sustainable Development Goals related to climate action, partnerships for sustainable development, and environmental conservation. The program's case studies demonstrate its real-world impact, showcasing its ability to provide predictive capabilities for environmental extremes, improve water quality monitoring, and enhance understanding of ecological processes.

Ethical Statement

There are no ethical issues.

Acknowledgments

This programme was supported by MBIE Strategic Science Investment Fund (SSIF) Data Science platform - Time-Evolving Data Science / Artificial Intelligence for Advanced Open Environmental Science (UOWX1910).

References

[Bar-On *et al.*, 2018] Yinon M. Bar-On, Rob Phillips, and Ron Milo. The biomass distribution on earth. *Proceedings of the National Academy of Sciences*, 115(25):6506–6511, 2018.

[Bennett *et al.*, 2009] Elena M. Bennett, Garry D. Peterson, and Line J. Gordon. Understanding relationships among multiple ecosystem services. *Ecology Letters*, 12(12):1394–1404, 2009.

[Bonan, 2016] Gordon B. Bonan. Forests, climate, and public policy: A 500-year interdisciplinary odyssey. *Annual Review of Ecology, Evolution, and Systematics*, 47(1):97–121, 2016.

[Cachay *et al.*, 2023] Salva Rühling Cachay, Bo Zhao, Hailey James, and Rose Yu. Dyffusion: A dynamics-informed diffusion model for spatiotemporal forecasting. *arXiv preprint arXiv:2306.01984*, 2023.

[Cattoën *et al.*, 2022] Céline Cattoën, Jono Conway, Nava Fedaeff, Daniel Lagrava, Paula Blackett, Kelsey Montgomery, Ude Shankar, Trevor Carey-Smith, Stuart Moore, Andrea Mari, et al. A national flood awareness system for ungauged catchments in complex topography: The case of development, communication and evaluation in new zealand. *Journal of Flood Risk Management*, page e12864, 2022.

[Defferrard *et al.*, 2019] Michaël Defferrard, Nathanaël Perraudin, Tomasz Kacprzak, and Raphael Sgier. DeepSphere: towards an equivariant graph-based spherical CNN. *arXiv preprint arXiv:1904.05146*, 2019.

[Douville *et al.*, 2022] Hervé Douville, Saïd Qasmi, Aurélien Ribes, and Olivier Bock. Global warming at near-constant tropospheric relative humidity is supported by observations. *Communications Earth & Environment*, 3(1):237, 2022.

[Ferro and Stephenson, 2011] Christopher AT Ferro and David B Stephenson. Extremal dependence indices: Improved verification measures for deterministic forecasts of rare binary events. *Weather and Forecasting*, 26(5):699–713, 2011.

[Graffeuille *et al.*, 2022] Olivier Graffeuille, Yun Sing Koh, Jörg Wicker, and Moritz K Lehmann. Semi-supervised conditional density estimation with wasserstein laplacian regularisation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 6746–6754, 2022.

[Hamilton *et al.*, 2017] Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *Proceedings of the 30th Advances in Neural Information Processing Systems*, 2017.

[Hersbach *et al.*, 2020] Hans Hersbach, Bill Bell, Paul Berrisford, Shoji Hirahara, András Horányi, Joaquín Muñoz-Sabater, Julien Nicolas, Carole Peubey, Raluca Radu, Dinand Schepers, et al. The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730):1999–2049, 2020.

[Hobday *et al.*, 2016] Alistair J Hobday, Lisa V Alexander, Sarah E Perkins, Dan A Smale, Sandra C Straub, Eric CJ Oliver, Jessica A Benthuysen, Michael T Burrows, Markus G Donat, Ming Feng, et al. A hierarchical approach to defining marine heatwaves. *Progress in Oceanography*, 141:227–238, 2016.

[Hobday *et al.*, 2018] Alistair J Hobday, Claire M Spillman, J Paige Eveson, Jason R Hartog, Xuebin Zhang, and Stephanie Brodie. A framework for combining seasonal forecasts and climate projections to aid risk management for fisheries and aquaculture. *Frontiers in Marine Science*, page 137, 2018.

[Holbrook *et al.*, 2020] Neil J Holbrook, Alex Sen Gupta, Eric CJ Oliver, Alistair J Hobday, Jessica A Benthuysen, Hillary A Scannell, Dan A Smale, and Thomas Wernberg. Keeping pace with marine heatwaves. *Nature Reviews Earth & Environment*, 1(9):482–493, 2020.

[ICNZ, 2023] ICNZ. New Zealand Insurance Council: Cost of disaster events in New Zealand. https://www.icnz.org.nz/industry/cost-of-natural-disasters/, 2023. Accessed: 2024-06-06.

[Jacox *et al.*, 2022] Michael G Jacox, Michael A Alexander, Dillon Amaya, Emily Becker, Steven J Bograd, Stephanie Brodie, Elliott L Hazen, Mercedes Pozo Buil, and Desiree Tommasi. Global seasonal forecasts of marine heatwaves. *Nature*, 604(7906):486–490, 2022.

[Jia *et al.*, 2021] Yunzhe Jia, Eibe Frank, Bernhard Pfahringer, Albert Bifet, and Nick Lim. Studying and exploiting the relationship between model accuracy and explanation quality. In *Machine Learning and Knowledge Discovery in Databases. Research Track: European Conference, ECML PKDD 2021, Bilbao, Spain, September 13–17, 2021, Proceedings, Part II 21*, pages 699–714. Springer, 2021.

[Keenan and Williams, 2018] T.F. Keenan and C.A. Williams. The terrestrial carbon sink. *Annual Review of Environment and Resources*, 43(1):219–243, 2018.

[Kratzert *et al.*, 2019] Frederik Kratzert, Daniel Klotz, Guy Shalev, Günter Klambauer, Sepp Hochreiter, and Grey Nearing. Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets. *Hydrology and Earth System Sciences*, 23(12):5089–5110, 2019.

[Lim *et al.*, 2023] Nick Lim, Albert Bifet, Daniel Bull, Eibe Frank, Yunzhe Jia, Jacob Montiel, and Bernhard

Pfahringer. Showcasing the taiao project: providing resources for machine learning from images of new zealand's natural environment. *Journal of the Royal Society of New Zealand*, 53(1):69–81, 2023.

[Luo *et al.*, 2016] Wenjie Luo, Yujia Li, Raquel Urtasun, and Richard Zemel. Understanding the effective receptive field in deep convolutional neural networks. *Advances in neural information processing systems*, 29, 2016.

[Milz *et al.*, 2023] Tobias Milz, Marte Hofsteenge, Marwan Katurji, and Varvara Vetrova. Foehn wind analysis using unsupervised deep anomaly detection. In *EGU General Assembly Conference Abstracts*, pages EGU–10256, 2023.

[Mosavi *et al.*, 2018] Amir Mosavi, Pinar Ozturk, and Kwok-wing Chau. Flood prediction using machine learning models: Literature review. *Water*, 10(11):1536, 2018.

[Mourot *et al.*, 2022] Phil Mourot, Nick Lim, Bernhard Pfahringer, and Albert Bifet. A regional flood impact prediction tool using machine learning to manage flood risk in real-time. a case study in new zealand. In *EGU General Assembly Conference Abstracts*, pages EGU22–12455, 2022.

[Ning *et al.*, 2023] Ding Ning, Varvara Vetrova, and Karin R Bryan. Graph-based deep learning for sea surface temperature forecasts. *arXiv preprint arXiv:2305.09468*, 2023.

[Ogden, 1995] John Ogden. The long-term conservation of forest diversity in new zealand. *Pacific Conservation Biology*, 2(1):77–90, 1995.

[Oliver *et al.*, 2019] Eric CJ Oliver, Michael T Burrows, Markus G Donat, Alex Sen Gupta, Lisa V Alexander, Sarah E Perkins-Kirkpatrick, Jessica A Benthuysen, Alistair J Hobday, Neil J Holbrook, Pippa J Moore, et al. Projected marine heatwaves in the 21st century and the potential for ecological impact. *Frontiers in Marine Science*, 6:734, 2019.

[Oliver *et al.*, 2021] Eric CJ Oliver, Jessica A Benthuysen, Sofia Darmaraki, Markus G Donat, Alistair J Hobday, Neil J Holbrook, Robert W Schlegel, and Alex Sen Gupta. Marine heatwaves. *Annual Review of Marine Science*, 13:313–342, 2021.

[Pörtner *et al.*, 2023] Hans Pörtner, Debra C Roberts, Camille Parmesan, Helen Adams, Ibidun Adelekan, Carolina Adler, Rita Adrian, Paulina Aldunce, Elham Ali, Rawshan Ara, et al. *IPCC 2022: Technical Summary, Working Group II Impacts, Adaptation and Vulnerability*. PhD thesis, Intergovernmental Panel on Climate Change, 2023.

[Potter *et al.*, 2021] Sally Potter, Sara Harrison, and Peter Kreft. The benefits and challenges of implementing impact-based severe weather warning systems: Perspectives of weather, flood, and emergency management personnel. *Weather, climate, and society*, 13(2):303–314, 2021.

[Ren *et al.*, 2022] Jiawei Ren, Mingyuan Zhang, Cunjun Yu, and Ziwei Liu. Balanced mse for imbalanced visual regression. *arXiv preprint arXiv:2203.16427*, 2022.

[Schaefer, 1990] Joseph T Schaefer. The critical success index as an indicator of warning skill. *Weather and forecasting*, 5(4):570–575, 1990.

[Seddon, 2022] Nathalie Seddon. Harnessing the potential of nature-based solutions for mitigating and adapting to climate change. *Science*, 376(6600):1410–1416, 2022.

[United Nations, 2015] United Nations. Transforming our world: the 2030 agenda for sustainable development, 2015.

[Yang *et al.*, 2021] Yuzhe Yang, Kaiwen Zha, Yingcong Chen, Hao Wang, and Dina Katabi. Delving into deep imbalanced regression. In *Proceedings of the 38th International Conference on Machine Learning*, pages 11842–11851. PMLR, 2021.

[Zhang *et al.*, 2022] Yaqian Zhang, Bernhard Pfahringer, Eibe Frank, Albert Bifet, Nick Jin Sean Lim, and Yunzhe Jia. A simple but strong baseline for online continual learning: Repeated augmented rehearsal. *Advances in Neural Information Processing Systems*, 35:14771–14783, 2022.