Mapping Woody Invasive Alien Plant Species (MAPWAPS)

Presented by:
Natasha Soldin

Prepared for:
Dr. Amit Mishra
Department of Electrical Engineering
University of Cape Town

Submitted to the Department of Electrical Engineering at the University of Cape Town in partial fulfilment of the academic requirements for a Bachelor of Science degree in Electrical and Computer Engineering

October 29, 2023

Declaration

1. I know that plagiarism is wrong. Plagiarism is to use another's work and pretend that it is one's own.
2. I have used the IEEE convention for citation and referencing. Each contribution to, and quotation in, this report from the work(s) of other people has been attributed, and has been cited and referenced.
3. This report is my own work.
4. I have not allowed, and will not allow, anyone to copy my work with the intention of passing it off as their own work or part thereof.

Signature:  .....

N. Soldin

Date: October 29, 2023

Word Count: 16866

Foreword and Acknowledgement

First and foremost I would like to thank my project supervisors: Professor Amit Mishra of the University of Cape Town, for his invaluable guidance and shared love of LaTex; and Dr. Alanna Rebelo of the Agricultural Research Council of South Africa, for her assistance and patience in helping me navigate an unfamiliar domain and for introducing me to and facilitating my contribution towards an important and worthwhile initiative. I would also like to acknowledge Liam Cogill, a PhD student working with Dr. Rebelo, for granting me access to South African data that he is in the process of collecting.

This research project is the culmination of my undergraduate degree at the University of Cape Town (UCT), a degree that would not have been possible without my family and friends. Firstly to my mom, Gayle Sherman, for her unwavering support and unconditional love. She epitomises kindness, resilience and excellence and is my greatest role model to whom I credit all of my accomplishments.

I am also grateful to my brother, Alex Soldin. Over the past four years, he has happily fielded every phone call and question. His assistance has not only helped me understand and endure but has been instrumental in my perseverance and completion of this degree. Alongside him, I would also like to thank Dalia Solomon.

Finally, to my friends and chosen family. To those that preceded my university career, despite our dispersion after high school, I have continued to feel your love and support over the past four years. You remain as near and dear to me as when we shared the same continent.

And to the friends that I gained at UCT, although temporally robbed by the COVID pandemic, we have and will continue to make up for time lost through ongoing friendship. Although my university career was greatly impacted by many people, I would like to make special mention of: Ryan Jones, Kamryn Norton, Trisyn Ferreiro, Mishay Naidoo and Daniel Jones, without whom I could not have imagined my time at UCT. You have left a permanent imprint of our countless hours spent in white lab and I am grateful to each of you ⊣⊢.

Abstract

Invasive Alien Plants (IAPs) threaten South Africa's water security. The Mapping of Woody Invasive Plant Species (MAPWAPS) initiative aims to leverage freely available satellite imagery to map the distribution of IAPs and estimate their water usage using Evapotranspiration (ET) measurements. This project constitutes a component of MAPWAPS and focuses on deriving ET estimates from Remote Sensing (RS) data using Machine Learning (ML) techniques.

The ML dataset was acquired by combining Landsat 8 satellite data and American flux tower ground truth data, because of the lack of South African ET data. This dataset was used to train five ML models, namely: Neural Network (NN), Recurrent Neural Network (RNN), Long Short-Term Memory (LSTM), Random Forest (RF) and Support Vector Machine (SVM), to determine the best performing model.

During the training phase, multiple variations of the dataset were input in an attempt to improve ML model results. This resulted in numerous combinations of the following: changes in dataset size, inclusion or exclusion of the date variable and feature engineering techniques such as application of normalisation and the modification of variable datatype.

The best performing model was the RF with an R^2 value of 0.8566 when trained on a dataset variation comprising a smaller dataset size, date inclusion, no application of normalisation and numeric manipulation of the date variable. As anticipated, applying this model to the limited South African ground truth data yielded a poor performance considering ET's high spatial variability and the fact that the model was trained on American data. This project did however lay the groundwork and establish future recommendations for achieving smart monitoring of ET once enough local ground truth data can be collected.

Contents

List of Figures

List of Tables

Abbreviations

AI Artificial Intelligence

ATP Acceptance Test Protocol

DoA Department of Agriculture

ESA European Space Agency

ET Evapotranspiration

GEE Google Earth Engine

IAP Invasive Alien Plant

LE Latent Heat Flux

LSTM Long Short-Term Memory

MAE Mean Absolute Error

MAPWAPS Mapping of Woody Invasive Plant Species

ML Machine Learning

MODIS Moderate Resolution Imaging Spectroradiometer

MSE Mean Squared Error

MTEF Medium Term Expenditure Framework

NASA National Aeronautics and Space Administration

NIR Near Infrared

NN Neural Network

NOAA National Oceanic and Atmospheric Administration

PC Personal Computer

RE Red Edge

RF Random Forest

RNN Recurrent Neural Network

RS Remote Sensing

SVM Support Vector Machine

SVR Support Vector Regression

SWIR Short Wave Infrared

TOA Top of Atmosphere

VIs Vegetation Indices

WFW Working for Water

WRC Water Research Commission

WRS Worldwide Reference System

Chapter 1

Introduction

1.1 Motivation and Background

The presence of an Invasive Alien Plant (IAP) can cause detrimental damage to its inhabited ecosystem. They reduce native vegetation, alter soil properties as well as fire regimes, homogenize biodiversity, decrease natural water resources and even go so far as to impact a country's economy [20]. South Africa is among the many countries recognising the potential and significant damage that IAPs' can enact and is placing increasing importance on IAP mitigation and elimination initiatives. Unlike other countries, South Africa is home to a particularly biodiverse environment as it is the 'third most species-rich country in the world' [21] making it that much more susceptible to IAP induced damage. South Africa is also facing an ongoing water crisis and cannot afford the ecosystem disruption and resulting water resource depletion caused by IAP presence.

The Western Cape's annual overview of provincial revenue and expenditure report, stresses the importance of exacting 'water resilience interventions' [22]. It further states that the 'Province's most cost-effective way to increase water supply is to invest in eliminating invasive alien plants'. The Department of Agriculture (DoA) has been allocated ZAR 121.161 million over the 2023 Medium Term Expenditure Framework (MTEF) to deal with this very problem.

However, past efforts have been seemingly futile as they have not made a difference to IAP spread and impact. The presence and distribution of IAPs is incredibly difficult to keep track of due to their fluid, ever-changing and invasive nature. The country needs improved and more targeted monitoring and tracking methods that will grant them the upper hand in the fight to eradicate IAPs and restore habitats.

1.2 Objective

The objective of this study is to grant our country this upper hand (or at least contribute). This thesis forms a small part of a larger government initiative: the Mapping of Woody Invasive Plant Species (MAPWAPS). This national initiative aims to provide detailed and reliable maps of IAP spread and distribution as well as an estimate of their water usage, to better inform IAP mitigation and eradication programs. MAPWAPS plans to leverage freely available and high-resolution satellite imagery to identify and distinguish IAPs and estimate their water usage [23]. This scope is graphically depicted in Figure 1.1 below.

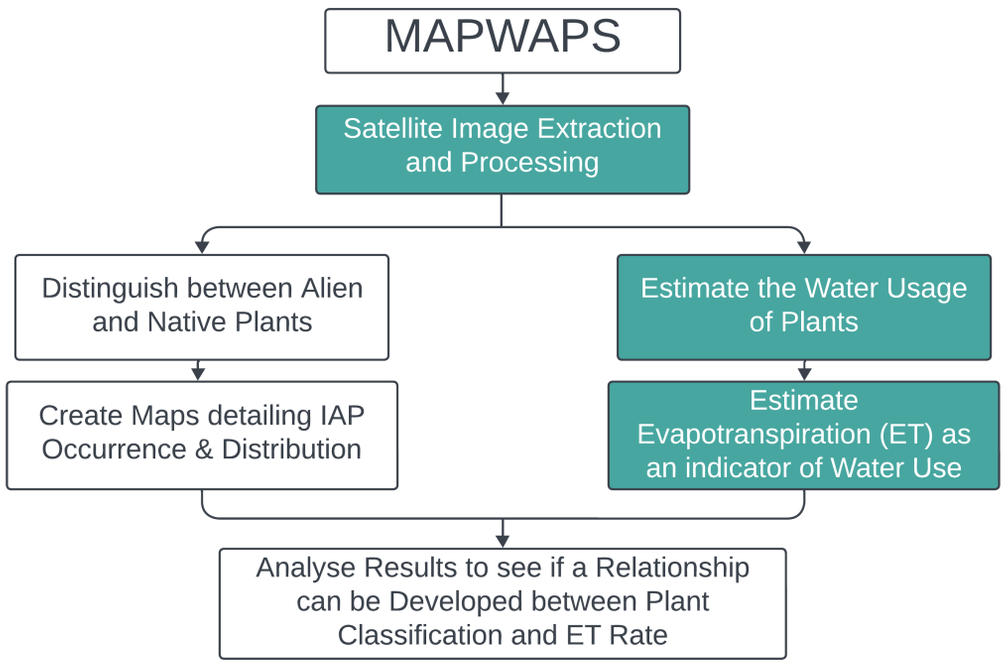


Figure 1.1: Mapping of Woody Invasive Plant Species (MAPWAPS) Initiative

Unfortunately, this initiative's scope is rather extensive and for practicality purposes was reduced to suit that of an honours thesis: the estimation of plant water usage from satellite imagery - as shown by the green blocks in Figure 1.1 above. This project aims to obtain a daily water usage or Evapotranspiration (ET) rate from a satellite image. The significance of this scope is two-fold: (1) Evapotranspiration (ET) estimation is rapidly gaining popularity as a key indicator of water use and further investigation into its significance is globally relevant and (2) using satellite Remote Sensing (RS) would allow for this project's application in resource-constrained countries as well as regions that are inaccessible to humans.

1.3 Scope and Limitations

As mentioned above and shown in Figure 1.1, the scope of this project is the estimation of ET from extracted satellite imagery. The limitations of this project, apparent at the project's start, which have the potential for complicating or hindering progress are as follows:

- The time constraint of a single university semester restricted the work and progress that could be completed and greatly narrowed the scope.
- The limited knowledge and lack of familiarity of an engineering student concerning hydrological and environmental sciences, adds to project complexity.
- The remote nature of such surveillance introduces geographical limitations because reaching certain locations in the pursuit of ground-truth data for verification is difficult and potentially dangerous to human life.
- The generalizability of this project is limited to the satellite chosen, the specific algorithms used and the South African context in which the project will be executed (i.e. native vegetation).

As the project progresses, it is to be expected that further limitations will be discovered. These will be detailed in the Conclusion and methods for addressing them will be in the Recommendations for Future Work section.

1.4 Plan of Development

The report outline is as follows: it begins with Chapter 2 which outlines the relevant literature pertaining to invasive vegetation and their impact as well as the application of RS, and ET estimation to the same or similar project contexts. Chapter 3 delves deeper into some theoretical background to further illuminate relevant concepts.

Chapter 4 outlines the system's modular breakdown and its requirements as well as its derived specifications and acceptance test protocols. The design of the project is laid out in Chapter 5 and implemented in Chapter 6. The results are presented and discussed in Chapter 7. Chapter 8 provides a concise summary of the project's work and offers recommendations for future work within the field.

Chapter 2

Literature Review

2.1 Invasive Alien Plants

An Invasive Alien Plant (IAP) is defined as vegetation that is not indigenous to the region in which it is found but has been relocated either intentionally or accidentally [21] and would otherwise not occur there naturally [24]. To illustrate the introduction and invasion of IAPs one can consider exotic plants introduced for beautification or ornamental value, crops imported for the sake of broadening the available produce variety [25] as well as trees planted for the procurement of lumber to compensate for a country's lack of natural forestry [26].

It is noteworthy that not all alien species are considered to be invasive and to fulfil the title of 'invasive', a plant must adapt quickly to its new environment, spread easily without human intervention and cause significant harm to its surrounding ecosystem [27]. Invasion of a species occurs in three stages, namely: arrival, establishment and integration [28]. There are different categories for invasive plants depending on the severity of their effects and the extent of the threat that they pose [24]. Category 1 describes IAPs that require immediate actions towards eradication and once identified these plants need to be removed and destroyed. Category 2 classifies IAPs that are regulated by an area and a permit of demarcation is required to possess, plant, breed, sell or buy these types of plants and Category 3 denotes IAPs that are regulated by an activity whereby an individual plant permit is required in order to import, possess, plant, breed, sell or buy these types of plants.

The main methods for controlling IAPs are four-fold namely biological, chemical, manual and mechanical [29]. Biological control involves the introduction of an IAPs' natural enemies like insects or pathogens, chemical control utilises herbicides or bioherbicides to

target IAPs, manual control refers to the physical removal of small IAPs by hand and mechanical refers to the automated removal of larger IAPs through cutting them down or ring-barking which is the process of removing or damaging a strip of bark of the tree thus disrupting its vascular system (flow of nutrients and water) and resulting in its death.

2.1.1 Consequences of Invasive Alien Plants

There are several negative effects that the presence of IAPs can have on a given ecosystem. These include: water resource depletion, fire intensity and risk increase, soil content contamination, carbon loss, biodiversity loss as well as threatening the agricultural and tourism industries [21, 23].

The invasion by IAPs into an ecosystem increases the overall Evapotranspiration (ET) and thus the water requirements of that region. If that requirement is too great, exceeding that which is coming into the ecosystem (i.e. rainfall), there is the potential to deplete the available water resources. The presence of IAPs also increases the fire risk and intensity for the area that they inhabit. This is due to their large size, rapid growth and softening boundaries between wildlands and urban occupied areas.

IAPs attempt to out-compete the indigenous vegetation around them, whether that competition be for available water, land or nutrients. This causes a significant decline in the native plantation taxa and consequently a loss in biodiversity [30]. IAPs are also known to release chemicals into the soil, changing its composition and making it inhospitable to native plants. The soil corruption is also contributed to by the increased fire intensity which scorches the soil, thereby consuming its organic matter. This results in a net carbon loss in the environment due to the reduction in the soil's organic carbon and native plant's biomass storage.

IAPs' invasion into and competition with crop fields has posed a great threat towards the agricultural industry, both in terms of activity and productivity. The loss in biodiversity also impacts the tourism industry because many ecosystems are losing their native vegetation which attracts tourists to view them.

2.1.2 Invasive Alien Plants in South Africa

South Africa is well known for its vegetation and significant biodiversity, holding the title for the 'third most species-rich country in the world' [21] and the first in Africa. This biodiversity is a result of the varied topography and climate of the country which

allows it to play host to nine different biomes, namely: Albany thicket, desert, forest, fynbos, grassland, Indian Ocean coastal belt, nama-karoo, savanna and succulent karoo [31] and around 20000 different plant species [32]. South Africa's climate can be described as semi-arid with annual precipitation of 464mm [30] and temperatures ranging between 15°C - 36°C during summer months (DJFM) and -2°C - 26°C during winter months (JJA) emphasising its climate variability [16]. This fluctuation in temperature and precipitation is summarised in table 2.1 below.

Observed Seasonal Mean Temperature				
Units	°C			
Months	DJF	MAM	JJA	SON
South Africa	23.59	18.41	12.37	18.98
Highest Point - Limpopo	25.10	20.97	15.77	22.41
Lowest Point - Eastern Cape	21.55	16.95	11.74	16.59

(a) Observed Seasonal Mean Temperature

Observed Seasonal Precipitation				
Units	mm			
Months	DJF	MAM	JJA	SON
South Africa	135.63	108.26	39.26	112.54
Highest Point - Limpopo	259.79	171.45	55.67	249.34
Lowest Point - Eastern Cape	69.80	73.20	24.59	39.32

(b) Observed Seasonal Precipitation

Table 2.1: South African Climate (1991-2020) [16]

South Africa has a long and complex history with IAPs [33]. It has been reported that roughly 10 million hectares or 8.28% of South Africa has been invaded by IAPs [34]. This figure is only increasing with time and is not reliably updated due to IAPs' unpredictable and distributive nature coupled with inadequate tracking systems.

IAPs pose a significant threat to South Africa's vegetation for several reasons as explored in Section 2.1.1 above and South Africa as a country recognises this and has started and continues to take measures to mitigate the harm of IAPs. The country spends up to ZAR 6.5 billion annually addressing the issue of IAP presence [35]. The cost of IAPs is not just limited to the economy but to another resource of great worth - water.

Water Scarcity in South Africa

South Africa's water resource scarcity is a result of the lack of physical rainfall as well as a sub-optimal infrastructure [36]. Therefore, the lack of consistent and significant rainfall in past years coupled with the lack of maintenance performed on South Africa's water distribution systems, resulting in faulty or leaking pipes has left the country in a state of water crisis. Water scarcity has even so far as been identified as the key constraint of South Africa's economic development [26].

Although there are many factors contributing to this water crisis including but not limited to urbanisation, pollution, agriculture and mining - for this project the cause that will be focused on in this section is water depletion as a result of the presence of IAPs. IAPs

have been described as 'the single biggest long-term threat' [21] to South Africa's water resource security. In 2018 it was reported that up to 30% of South Africa's major cities' water supply was at risk due to IAPs [35]. More recently, in a 2021 paper, it was estimated that the regional stream-flow of a catchment situated in the southwest of South Africa had been reduced by 304 million m^3 or 4,14% annually, due to the effects of IAPs [37]. Overall, it has been estimated that South Africa loses 1.44 million m^3 of water every year due to IAPs. This amount of water is equivalent to a year's worth of water provided to 3.38 million households or a year's worth of irrigation to a cropland of 120,000 hectares [38].

South Africa's already limited water resources are being depleted by significant IAP infestations, which are known for affecting both the quantity as well as the quality of water within their reach [39]. This resource depletion has happened to such a notable extent that it has catalysed the creation of programs and initiatives within the country.

The initiative most commonly referred to in the reviewed literature seems to be the Working for Water (WFW) campaign. It was launched nationally in 1995 and is one of the largest IAP targeted initiatives worldwide. In the first 12 years of its operation (1995 - 2007), it is credited with clearing up to 1.6 million hectares of IAPs, at a total cost of ZAR 3.2 billion [40]. The WFW program makes use of chemical, biological and physical IAP removal methods. In pursuing its objective of protecting South Africa's unique biodiversity and limited water resources, it has also provided the added socio-economic benefit of job creation [34].

2.2 MAPWAPS

It is the conclusion of many previously executed studies that efforts to mitigate the spread and damage of IAPs in South Africa are failing to keep apace [33]. A key tool in improving this would be to accurately detect and map the occurrence and spatial distribution of IAPs [30]. Such a map could inform management and rehabilitation efforts as well as provide an overview allowing for appropriate prioritisation. Traditionally, IAP data collection had been carried out by means of field surveys and aerial photographs. However, due to the unpredictable, ever-changing and widespread behaviour and distribution of IAPs as well as excessive resources, capital and time required for such methods, they have proved to be insufficient [20]. A new approach for the necessary data collection is using satellite Remote Sensing (RS) which would allow for the construction of reliable and extensive IAP maps.

This is the basic motivation behind the Mapping Woody Invasive Alien Plant Species initiative, MAPWAPS for short. The project was initiated by Dr. Alanna Rebelo, a senior researcher at the Agricultural Research Council of South Africa, and funded by the Water Research Commission (WRC) of South Africa [23]. This project aims to use freely available satellite imagery to map the occurrence and spread of IAPs at strategic water sources. These are defined as water sources that provide a disproportionate amount of water relative to their area within South Africa. The project extends further to adopt the same RS strategy in determining the water usage of IAPs relative to native plants - a key indicator of the damage caused by IAP presence.

The four site locations chosen for this project's application are the Luvuvhu River catchment in Limpopo (-22.468961824786202, 30.935705064360043), the Sabie and Crocodile River catchments in Mpumalanga [(-25.044497669134017, 30.919627926830245) and (-25.270651919979613, 31.801455450894075) respectively], the Tukhela River catchment in KwaZulu-Natal (-28.67297871113665, 30.160667656458283) and the uMzimvubu River catchment in the Eastern Cape (-31.60962545188555, 29.54052324551523). The selection process was based on four components: (1) water security, in that the catchment needs to be a strategic water source and hold importance for water provision to its surrounding region, (2) alien trees, the catchment must be home to IAPs, (3) spatial variation, meaning that they needed to be in different provinces of South Africa and be of differing biomes and (4) data scarcity, the catchments lack adequate data and need increased surveying [23, 35].

2.3 Remote Sensing

The popularity of spectroscopy, more commonly named Remote Sensing (RS), has been fueled by the recent technological developments in satellite imagery capture, with emphasis on the improved quality and resolution [33]. It has great environmental applications including: climate change analysis, land cover classification and mapping water and air quality management as well as biodiversity conservation [41]. The use of satellite imagery and derived variables for environmental analytics have their advantages and disadvantages [42]. The advantages are the remote large-scale cover, access and data collection of any region under satellite jurisdiction - including areas that are otherwise inaccessible to humans. Satellites can provide real-time data that is unhindered by bias caused by human intervention into an otherwise undisrupted ecosystem. The disadvantages however are factors such as noise or errors, susceptibility to obscurities as a result of atmospheric conditions as well as the challenge of storing and accessing such a large-scale data collection while retaining resolution quality.

As further expanded upon in Section 2.4.1 below, the use of RS has gained significant popularity in the monitoring of IAPs [20] and has largely been integrated into ET estimation. Even so far as entire empirical models being altered such that they are satellite-based. In other words, although RS cannot provide direct ET measurements, it can provide factors that influence and relate to ET, allowing for its proxy determination [2]. Multispectral satellites have proved to be the most useful for the application as their bands store Electromagnetic Spectrum data that is relevant and informative to ET estimation. Hyperspectral satellites have recently gained more traction within the field as unlike multispectral, which offer around 10-13 bands, hyperspectral can offer around 50-200 more narrow or fine-scale bands. These satellites are more expensive and have an added layer of complexity as a result of their higher spectral resolution (i.e. number of bands). It is for this reason and due to their infancy in the field with very little available literature, that they will not be explored in the upcoming sections.

2.3.1 Satellite Options

Table A.1 in Appendix A below provides a non-exhaustive summary of some available satellite options and their characteristics, with specific emphasis placed on those that have been used in vegetation RS applications, specifically ET estimation.

Note: the following header definitions in Table A.1 are clarified: spatial resolution refers to the smallest level of detail that the satellite can discern or the pixel size of an image; spectral resolution refers to the number, width and range of spectral bands that the satellite collects and temporal resolution refers to the frequency at which the satellite collects data of a given region [43]. Temporal resolution is not the same as revisit time although they are related concepts. The revisit time refers to how often the satellite observes a given region but data collection of that region is not guaranteed.

Sentinel vs. Landsat

The two most well-known multispectral satellites, frequently referred to in reviewed literature, are Sentinel-2 and Landsat 8/9. The European Space Agency (ESA) launched the Copernicus Sentinel-2 satellite, or more specifically a constellation of two satellites Sentinel-2A and B, in 2015 and 2017 respectively [19, 44]. NASA and the U.S. Geological Survey are responsible for managing and maintaining the Landsat program, which launched their first satellite (Landsat 1) in 1975 and relevant to this project, launched Landsat 8 in 2013 [45]. When comparing their specifications, Sentinel-2 has a spatial resolution of 10-20m and a temporal resolution of 5 days while Landsat 8 has a spatial resolution of

30m and a temporal resolution of 16 days.

In the literature, many studies have found that Sentinel-2 satellite-produced images are better for vegetation discrimination and classification. Mtengwana et al. found that when compared to the classification performance obtained from using Landsat 8 and Sentinel-2 images, the accuracy was 63% and 71% respectively [30] and similarly Holden et al. concluded that Sentinel-2's available bands provide invaluable information for geological or hydrological examination of plants [33]. Both of the aforementioned studies placed emphasis on the Near Infrared (NIR), Red Edge (RE) and Short Wave Infrared (SWIR) bands which aid in plant trait distinction and identification of water absorption features. Sentinel's spectral resolution (i.e. bands) have even been described as providing 'novel capabilities' [19] for vegetation detection applications.

2.3.2 Image Bands and Vegetation Indices (VIs)

The strategically positioned [20] multi-spectral bands of both the satellites discussed above make them particularly applicable to vegetation detection, discrimination and characteristic derivation. Although the bands do not provide a direct measure they provide insights into light absorption, leaf biochemistry, reflectance, biomass and water absorption features [33, 30]. Through mathematical manipulation of these bands and inclusion of other data, various VIs can be calculated as shown in Table A.2 in Appendix A below. These indices provide ecological detail regarding the surveyed area and are often used as inputs to varying Evapotranspiration Measurement methods including ML model training [2].

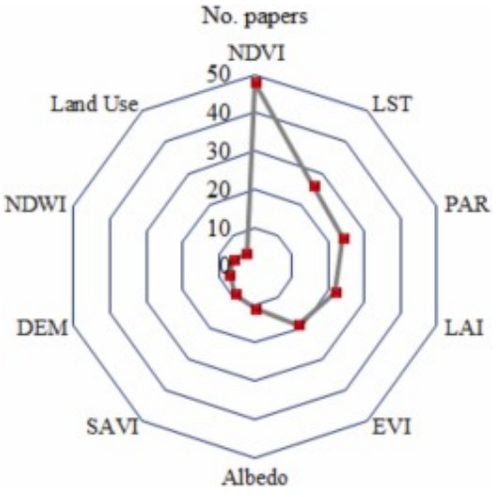
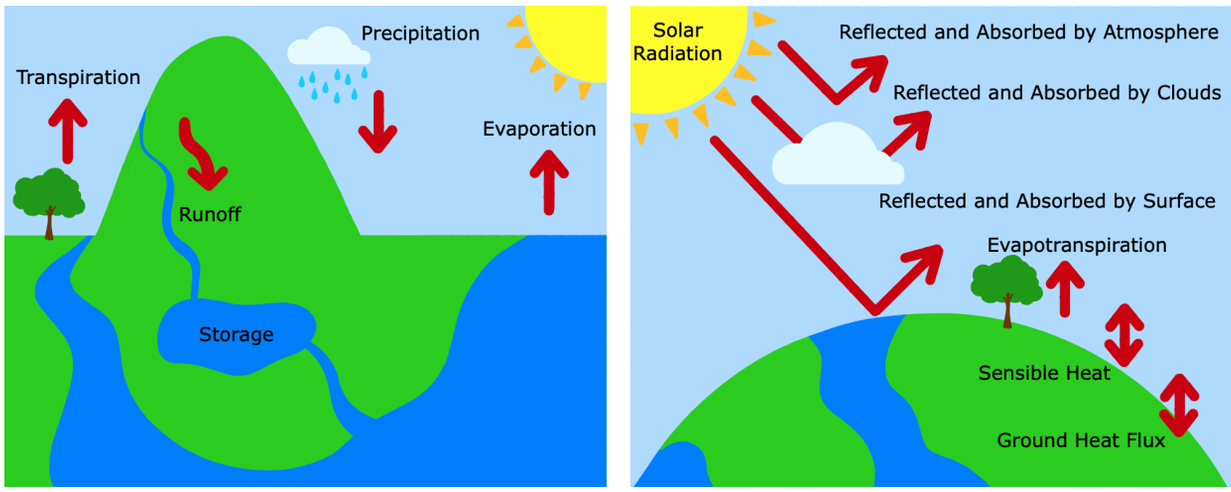


Figure 2.1: Number of Papers Mentioning Specific RS derived VIs that are used in ET Estimation

Figure 2.1 alongside shows a graph emphasising which VIs have been deemed important for ET estimation in a RS context, based upon their frequency of appearance in the literature [2]. Many of these VIs are detailed in Table A.2 in Appendix A below.

2.4 Evapotranspiration Measurement

ET is a term describing the combination of evaporation and transpiration and is the sum total of water lost from a given surface. This can occur from available water or soil moisture evaporation or plant tissue transpiration [34]. It is a variable of great significance within the hydrological cycle [46], seen in Figure 2.2a below, as it is the second largest component after its inverse operation, precipitation [47]. ET is responsible for returning up to 70% of precipitation back to the atmosphere annually [46]. Accurate estimation and ongoing monitoring of ET is of vital importance as it can inform water resource management decisions and thus improve water use efficiency [3] as well as aid in the understanding of present ecological and hydrological processes and how they contribute to the relationship between the atmosphere, biosphere and hydrosphere [2]. This importance is magnified in arid or semi-arid regions such as Africa, South Asia, Australia and the Middle East due to the evident climatic and water resource problems.



(a) Water Balance Model (b) Energy Balance Model [48]

Figure 2.2: Environmental Models that relate to Evapotranspiration (ET) (not drawn to scale)

Various factors can influence ET values. Both processes of evaporation and transpiration are driven by ecological as well as meteorological conditions such as air temperature, humidity, solar radiation and windspeed [3]. Plant factors such as crop characteristics and methods of cultivation also impact plant transpiration. Dzikiti et al. noted that some factors that impact the measurement of ET are climate, soil composition as well as landscape topography and heterogeneity [46]. The paper further emphasises that given that South Africa is home to a diverse and heterogeneous environment and a semi-arid climate, these factors will contribute to the difficulty in ET measurement.

ET variability is also subject to spatial and temporal factors. In different locations, which

are home to different ecosystem characteristics, plant species and weather conditions, ET will differ spatially. At different times during a day (morning or night) or year (seasonally) where ecological variables change with time, ET will differ temporally. The spatial variability can also be attributed to varying ET limiting factors. Zhang et al. propose that a region's ET rate is controlled by one of three limiting factors: (1) Demand, which relates to air temperature, humidity and wind speed, (2) Supply, which refers to precipitation and (3) Energy, which refers to radiation and cloud cover [1]. A graphical worldwide representation of this is shown in Figure 2.3 below. Regions with different limiting factors will have ET that 'behaves' in very different manners and vice versa.

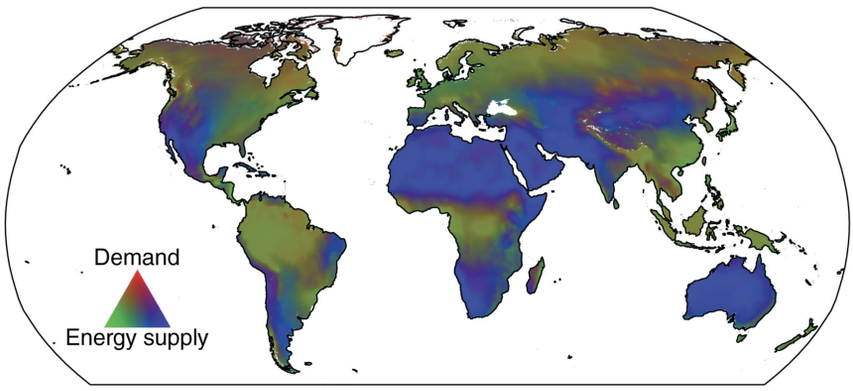


Figure 2.3: Geographic Representation of Evapotranspiration (ET) Limiting Factors [1]

Unlike the hydrological cycle's other water balance components (i.e. precipitation and run-off), ET suffers the greatest measurement difficulty. This can largely be attributed to its 'invisibility'. It is incredibly difficult to measure what one cannot see, in this case, the vaporisation of water [5]. ET values are also unpredictable and mobile in that they often change regularly and sometimes even drastically depending on the aforementioned influencing factors - environmental, spatial, temporal and otherwise. Bai et al. reported that ET measurement inaccuracies can be anywhere between 10% - 30% [5].

On a more technical note, instrumentation selection and data integration contribute enormously to ET's measurement difficulties. As further explored in Section 2.4.1 below, many measurement methods require specialised hardware whose deployment, use and maintenance can be difficult and expensive. Some methods also require data variables from external datasets and sources such as weather stations or satellite images which can be complex to integrate into a singular ET measurement method. These external datasets also have limitations and can contribute to overall measurement inaccuracies (e.g. improper weather sensor calibration or limited satellite imagery access due to orbital revisit time).

Many existing measurement methods can only provide plant or local scale measurements.

Therefore large observational errors [5] and measurement inaccuracies can occur when these products are scaled for larger or regional applications [49]. Although a method may prove to have reasonable measurement accuracy when applied to the environment in which it was obtained (i.e. region whose data was used to train it or whose conditions it was modelled after), this does not mean it will prove to have the same standard of accuracy when applied to a different environment. This is especially true if the two regions differ too drastically in terms of inherent plant and climate characteristics.

All these measurement hurdles result in limited ET data being available and reliable. This places significant constraints on ET estimation, commonly resulting in over or under-estimation and hindering model or method improvements [49]. The measurement difficulties have also ensured that ET is the least understood process of the hydrological cycle [50].

2.4.1 Measurement Methods

Due to the difficulties in measuring ET, indirect or proxy measurement methods as well as empirical estimation models are used to quantify ET of a point or region. Different literature categorises ET measurement methods differently but Figure 2.4 below is a simplified and non-exhaustive summary of some of the most common ET measurement methods.

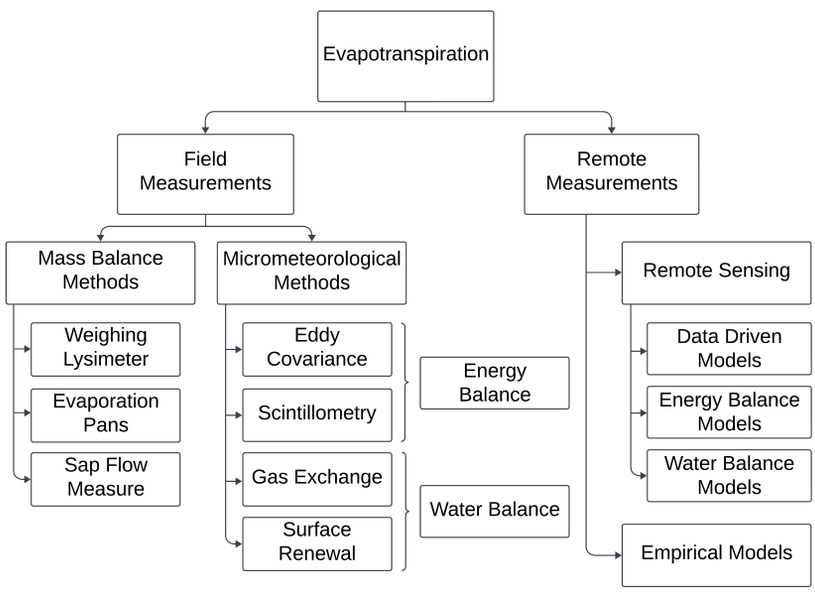


Figure 2.4: Classification of Evapotranspiration (ET) Measurement Methods [2, 3, 4, 5]

This classification is the combination of two published categorisations of ET measurement methods. Yang separates the methods into 'field' and 'remote', where 'field' refers to the need for physical access to the environment to obtain ET measurements and 'remote' refers to methods where ET can be calculated without environmental intervention [4]. Yang delves further into field measurement which he separates into mass balance and micrometeorological methods. Bai deals with remote measurements and separates RS measurements into data-driven, water balance-based or energy balance-based models [5].

Water Balance Methods

Some ET measurement methods are based directly on the hydrological cycle, as depicted by Figure 2.2a above and governed by equation 2.1 below [51]:

$$P = Q + E \pm S \tag{2.1}$$

where P is precipitation or rainfall, Q is run-off, E is Evapotranspiration (ET) and S is change in storage. This method aims to determine E by first determining or calculating the values of P, Q and S. Related methods or those used in conjunction, would be gas exchange and surface renewal as described in Figure 2.4.

Energy Balance Methods

Some ET measurement methods are based directly on the energy balance model, as depicted by Figure 2.2b above and governed by equation 2.2 below [48]:

$$R_n = ET + H + G = \frac{LE}{\lambda} + H + G \tag{2.2}$$

where R_n is the net radiation which is the difference between incoming solar radiation and outgoing reflected radiation, ET is Evapotranspiration (ET), H is the temperature felt on Earth's surface as a result of heat exchange between Earth and the atmosphere, G is the ground heat flux which is heat transferred in or out of the ground, LE is the latent heat flux which is the energy released or absorbed during phase changes (i.e. liquid $\leftrightarrow$ gas in condensation and evaporation) and λ is the conversion factor between ET and LE. This method aims to determine ET by first determining or calculating the values of R_n, H and G and in some cases LE and λ. Related methods or those used in conjunction, would be eddy covariance and scintillometry as described in Figure 2.4.

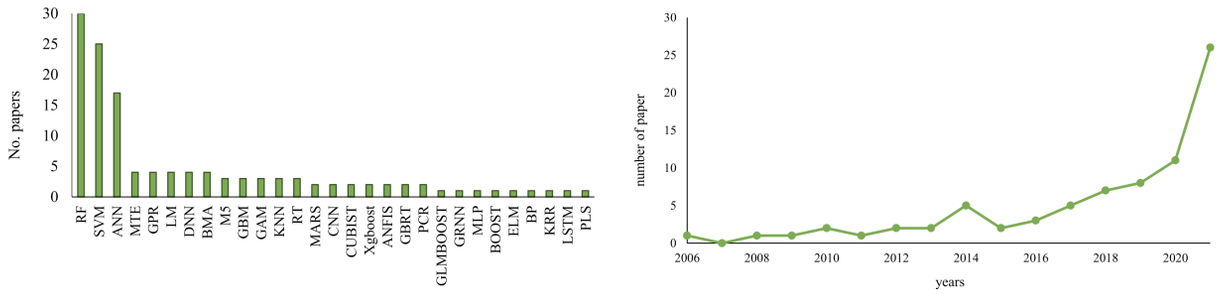
LE and ET are closely related because they both deal with water's phase change. It is common practise to measure LE through eddy covariance systems or flux towers and convert it to attain ET values. The two values are related by the latent heat of vaporization constant λ. The reported calculation of this constant varies between sources but a common one is that it is dependent on the environment's air temperature T_a and expressed by equation 2.3 [52]:

$$\lambda = 2.501 - (2.36 \cdot 10^{-3}) \cdot T_a \tag{2.3}$$

Many of the existing empirical models are based on the Water Balance Methods or the Energy Balance Methods or some combination of the two. Examples of the most commonly used are: Penman-Monteith, Priestly–Taylor, Stanghellini or Hargreaves Models - these models will not be further expanded upon but various sources are available for greater detail and explanation [3]. An existing global product that incorporates ET estimation with RS is MOD16. This product uses the Moderate Resolution Imaging Spectroradiometer (MODIS) satellite, as detailed in Table A.1 in Appendix A and the aforementioned Penman-Monteith ET estimation method. Ramoelo et al. inferred that MOD16 does not perform accurately in a South African context and recommended a 'locally parameterised and improved product' for performance improvement [53].

2.4.2 ET Estimation through Machine Learning

Unlike other ET estimation methods, a Machine Learning (ML) algorithm requires far fewer input variables and parameters than traditional methods and empirical models [2]. It also relies on far fewer assumptions regarding the homogeneity of an ecosystem, crop or environmental variable coefficients or initial condition and reference values. Therefore, it may prove to be a powerful tool in ET measurement, especially in resource-constrained regions. In recent years, as a result of ML's development, ML has been used more and more frequently in ET estimation. This is evident in Figure 2.5b below where there has been a significant increase in its use in literature. It is also noteworthy that some ML model types are better suited, and therefore more frequently tested, to ET measurement as conveyed by the bar graph in Figure 2.5a below. The three most commonly used have been Random Forest (RF), Support Vector Machine (SVM) and Neural Network (NN) ML models.



(a) Number of Papers Mentioning Specific ML ET Estimation Models using RS Data

(b) Number of Papers Mentioning ET ML Estimation Models (2006 - 2021)

Figure 2.5: Evapotranspiration Estimation Machine Learning Model Publications [2]

The performance accuracy and overall generalizability of a ML algorithm is greatly dependent on the ET ground truth data or input variables that are supplied to the model. The identification of such model parameters, specifically relating to their relevance and availability, is imperative in ensuring adequate model performance [2]. In RS applications, the common inputs include the multispectral band values and the VIs that can be derived from them.

2.4.3 OpenET

A relevant organisation worthwhile highlighting is OpenET. They are an American-based company developing models to estimate ET using RS satellite imagery. Their slogan reads: "Filling the Biggest Data Gap in Water Management", emphasising the worldwide recognition that ET is gaining as a significant indicator of water usage [47].

They offer an ensemble of ET estimation models that are based on the Surface Energy Balance (SEB), as described in Section 2.4.1 above. The six models offered include: (1) Atmosphere-Land Exchange Inverse / Disaggregation of the Atmosphere-Land Exchange Inverse (ALEXI/DisALEXI), (2) Google Earth Engine implementation of the Mapping Evapotranspiration at high Resolution with Internalized Calibration model (eeMETRIC), (3) Google Earth Engine implementation of the Surface Energy Balance Algorithm for Land (geeSEBAL), (4) Priestley-Taylor Jet Propulsion Laboratory (PT-JPL), (5) Satellite Irrigation Management Support (SIMS) and (6) Operational Simplified Surface Energy Balance (SSEBop). As input to these models they use Landsat 8 satellite, weather station or crop type and land use data [47].

Their ethos is a shared approach in that they provide public access to their code, via a GitHub Repository, and methodology to further the global effort towards accurate ET measurement.

Chapter 3

Theory

The following chapter outlines some of the relevant theoretical information pertaining to the project. This theory section is non-exhaustive and further explanations can be gained by accessing the sources used as references.

3.1 Satellite

The term satellite refers to any object that orbits a celestial body (i.e. planet or star). For instance, Earth is a natural satellite as it orbits the sun. The term however has by the majority been co-opted by artificial or man-made satellites which are instruments that are purposefully placed within the orbit of a celestial body. Satellites have varying purposes including weather forecasting, earth observation, navigation and communication [54]. Satellites house sensors that are sensitive to different parts of the Electromagnetic Spectrum and these sensors 'convert energy to imagery' [55]. Depending on the application and purpose of a given satellite, its sensors will only focus on and capture images from specific portions or channels of the Electromagnetic Spectrum. Various types of satellites provide different imaging for different applications, including Synthetic Aperture Radar (SAR), Light Detection and Ranging (LiDAR) and Panchromatic Imaging. However, this brief theory section will focus on multispectral imaging. As per Section 2.3.1 above, the majority of the relevant literature about this project's context refers to the use of multispectral satellites and their compatibility with environmental or vegetation-based applications. Multispectral imaging captures channels on the Electromagnetic spectrum into several labelled bands, which differ in quantity and channel type between satellites as explored in Section 3.1.1 below.

3.1.1 Electromagnetic Spectrum

The Electromagnetic Spectrum, as depicted by Figure 3.1 below, is the range of all Electromagnetic Radiation [6] or in other words, all light that exists. This ranges from radio waves to gamma rays and is arranged according to wavelength which is also the metric used to distinguish between the regions of a spectrum.

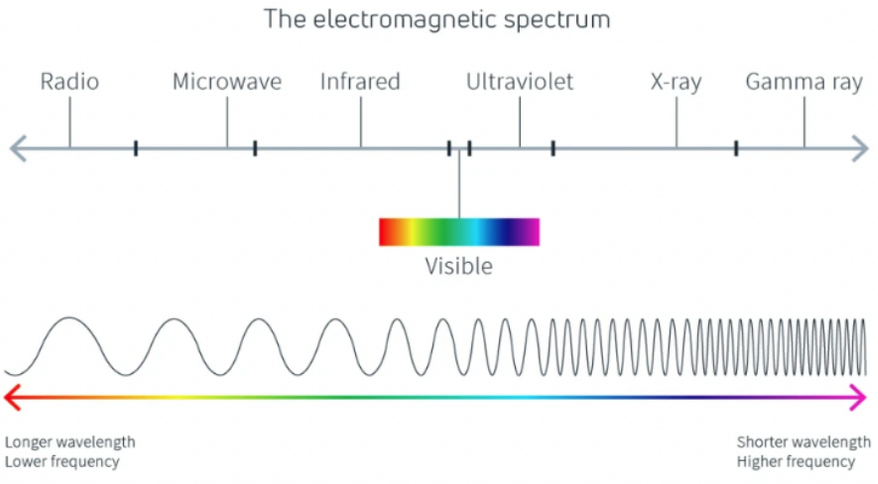


Figure 3.1: Electromagnetic Spectrum [6]

It is from this range that a given multispectral imaging satellite will 'choose' which portions to capture in their bands. The bands will capture a portion (or subrange) of the available wavelengths. The central wavelength refers to the value at the centre of that sub-range. To illustrate this point Tables 3.1, 3.2 and 3.3 below summarise the multispectral bands and central wavelengths of Landsat 7, Landsat 8 and Sentinel-2 multispectral satellites. Figure 3.2 compares the spectrum 'coverage' between these three satellites.

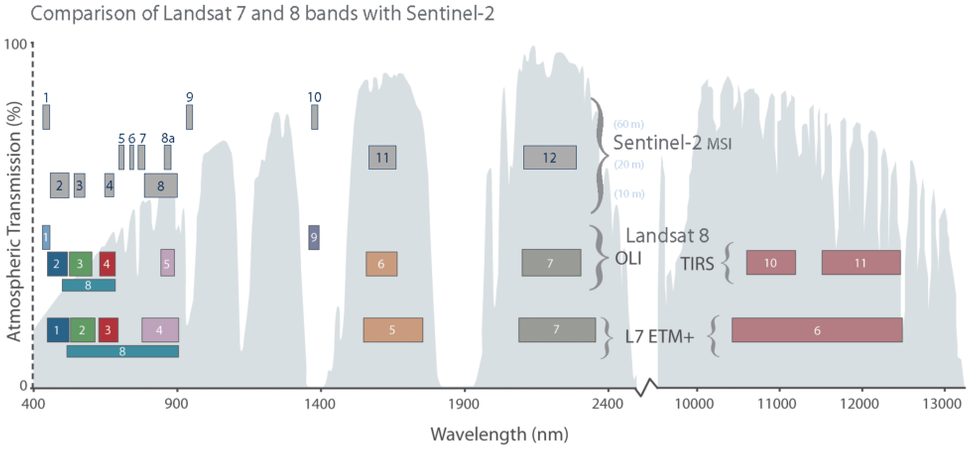


Figure 3.2: Electromagnetic Spectrum Coverage Comparison between Landsat 7, Landsat 8 and Sentinel-2 Multispectral Satellites [7]

Band	Label	Central Wavelength (nm)
B1	Blue	485
B2	Green	560
B3	Red	660
B4	Near Infrared (NIR)	825
B5	Shortwave Infrared (SWIR) 1	1650
B6	Thermal	11080
B7	Shortwave Infrared (SWIR) 2	2110
B8	Panchromatic	770

Table 3.1: Landsat 7 Bands [17]

Band	Label	Central Wavelength (nm)
B1	Ultra Blue (Coastal and Aerosol)	443
B2	Blue	482
B3	Green	561
B4	Red	655
B5	Near Infrared (NIR)	865
B6	Shortwave Infrared (SWIR) 1	1610
B7	Shortwave Infrared (SWIR) 2	2110
B8	Panchromatic	655
B9	Cirrus	1365
B10	Thermal Infrared (TIRS) 1	10980
B11	Thermal Infrared (TIRS) 2	12410

Table 3.2: Landsat 8 Bands [17]

Band	Label	Central Wavelength (nm)
B1	Ultra Blue (Coastal and Aerosol)	443
B2	Blue	490
B3	Green	560
B4	Red	665
B5	Visible and Near Infrared (VNIR)	705
B6	Visible and Near Infrared (VNIR)	740
B7	Visible and Near Infrared (VNIR)	783
B8	Visible and Near Infrared (VNIR)	842
B8a	Visible and Near Infrared (VNIR)	865
B9	Short Wave Infrared (SWIR)	940
B10	Short Wave Infrared (SWIR)	1375
B11	Short Wave Infrared (SWIR)	1610
B12	Short Wave Infrared (SWIR)	2190

Table 3.3: Sentinel-2 Bands [7]

3.2 Machine Learning

Machine Learning (ML), a component of Artificial Intelligence (AI), is the exploration and development of algorithms that can make predictions or decisions based on the data they have been provided to 'learn' from. A trained model should act without explicit programming or human-driven code [8].

ML has broadly been separated into three categories: supervised learning, unsupervised learning and re-enforcement learning [8, 56]. A brief explanation of each is as follows [8, 9]:

- Supervised Learning: a model is trained off of labelled ground truth data whereby the input has a known output. These models aim to develop a relationship between input and output such that they can predict an output given any input. The two main branches of this category are classification and regression, the difference being the nature of their output that is either discrete or continuous, respectively.

- Unsupervised Learning: a model is trained off of unlabelled data whose nature or over-arching purpose is often unknown. It is the model's job to intuitively observe and determine patterns or relationships present in the data. The two main branches of this category are clustering, which groups similar data points together, and dimensionality reduction which, as the name alludes to, decreases the dimension or complexity of a dataset.

- Reinforcement Learning: a model, or agent, attempts to learn how to behave in a given environment by receiving feedback. Ultimately this is best described as the model 'making decisions' and receiving either a positive or negative response from which it learns to improve its decision-making strategy.

These ML categories, sub-categories and their respective most commonly used algorithms can be seen in Figure 3.3. Note: this is a very brief overview of ML which is sufficient for this project and more detailed explanations are available in various resources.

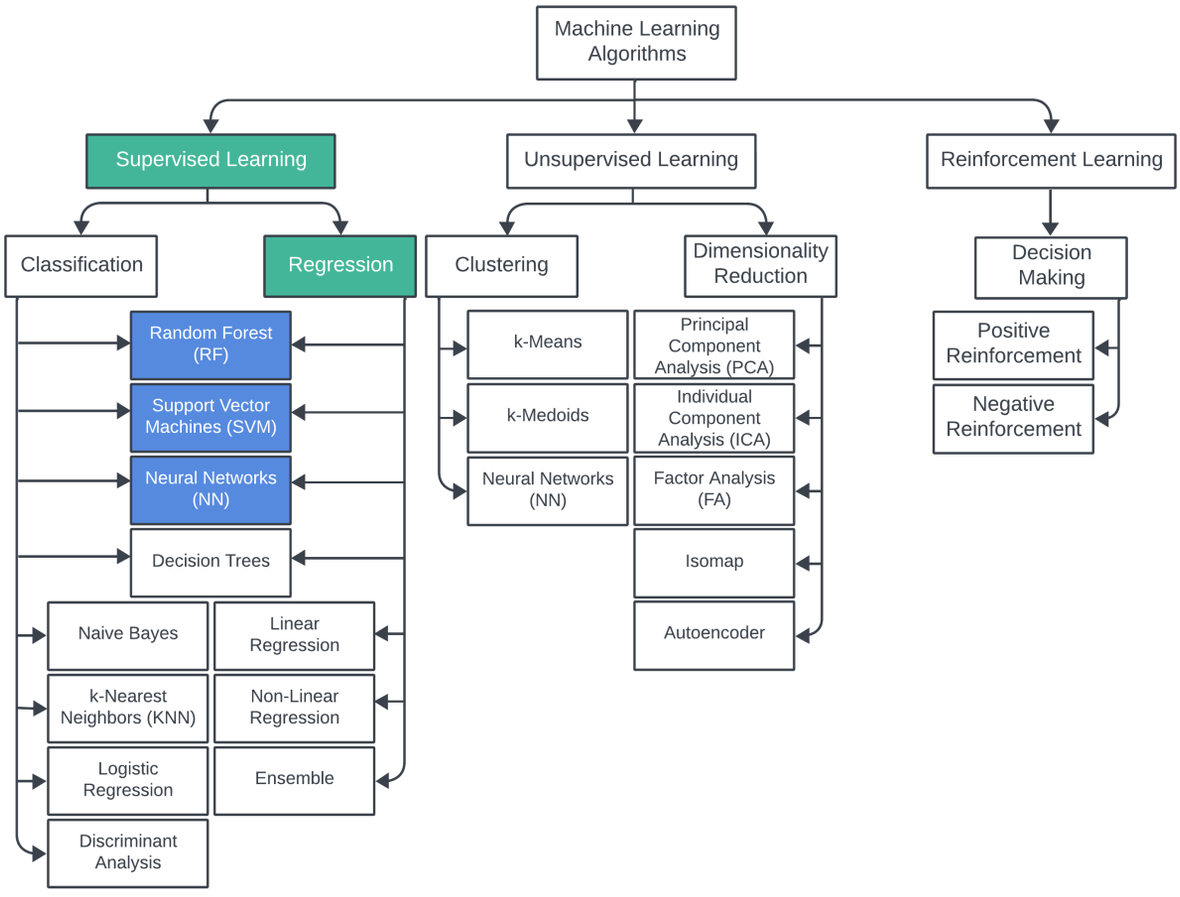


Figure 3.3: Classification of Machine Learning Algorithms [8, 9, 10, 11]

3.2.1 Regression Machine Learning

As demonstrated by the green blocks in Figure 3.3 above, this project's scope falls within the supervised learning category and more specifically, a regression task. With regression being the context of this project, further explanation of this facet of ML is needed. As explored above, regression ML aims to develop a relationship between two variables: a dependent (or predicted) and an independent (or known). Although there are various types of regression and the nature of the regression (i.e. nature in which the variables relate to each other) remains unknown, it is common practice to begin with or assume linear regression [57].

This means the model makes an initial baseline assumption that the two variables relate to each other in a linear manner and determines the 'line of best fit'. The 'line of best fit' is that which best describes the data's overall trend (which is extracted from a scatter plot produced when the dependent and independent variables are plotted against each other). This assumed linear line will be of the equation:

$$\hat{y} = b_0 + b_1 \cdot x \tag{3.1}$$

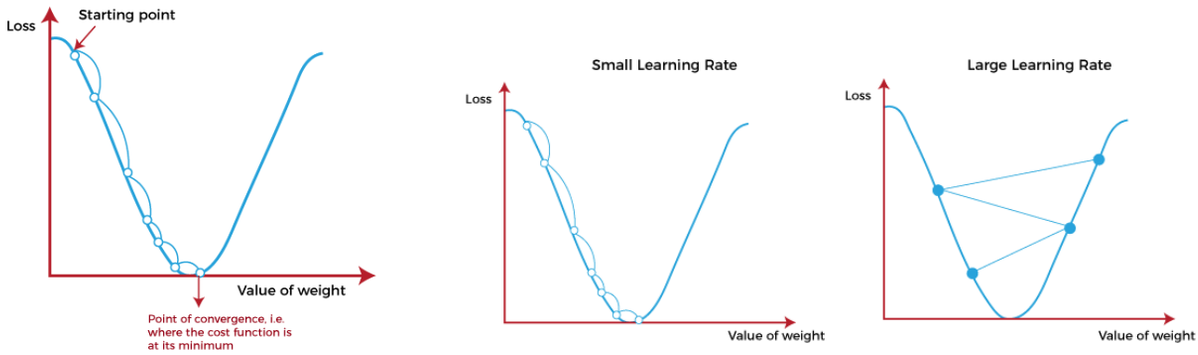
Where $\hat{y}$ is the predicted value (dependent variable), b_0 is the y-axis intercept or vertical shift, b_1 is the line's gradient (signalling the change in $\hat{y}$ for a single unit change in x) and x is the known value (independent).

The 'line of best fit' is categorised by coefficients b_0 and b_1, whose values are determined through minimization of a cost function; meaning values that ensure the predicted value $\hat{y}$ is as close to the known value y as possible (i.e. the model's predictions are as accurate or close to ground truth as possible). The cost function for linear regression tasks is generally the Mean Squared Error (MSE), shown by Equation 3.2 below, which determines the average of the squared difference between the known value and the predicted value (i.e. the model's error).

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{3.2}$$

This cost function can be minimised through optimisation algorithms such as gradient descent. This method computes the gradient of the cost function with respect the the coefficients b_0 and b_1. It then intuitively and iteratively adjusts these coefficients, using the calculated gradient as a guide, until the gradient descent algorithm converges signalling the optimal point or lowest achievable MSE as shown by Figure 3.4a below. In this

case 'converges' refers to the point where further changes made to the coefficients do not decrease MSE. It is also relevant to note the learning rate and its importance as it determines the speed at which the algorithm converges to a minimum value and if chosen incorrectly can result in un-optimised results as shown in Figure 3.4b below.



(a) Operation of Gradient Descent Cost Function Optimisation

(b) Impact of Learning Rate on the Efficacy of the Gradient Descent Algorithm

Figure 3.4: Gradient Descent Algorithm used to Minimise Cost Function [12]

Regression Metrics

The performance of ML models is often bench-marked using metrics. The relevance of a given metric varies between different ML models and for a regression task, the most popular are [58]: Mean Squared Error (MSE), Mean Absolute Error (MAE) and Regression Co-efficient, also known as the Coefficient of Determination (R^2). Although there are various other metrics with merit to such an application, these will be the ones focused on in this project.

MSE, as described above and expressed by Equation 3.2 can be used for both model training (i.e. cost function optimisation) as well as performance benchmarking. MAE, expressed by Equation 3.3 is similar to MSE except that is does not square the error. The key difference between these two metrics is MAE grants 'equal weight' to each error regardless of their magnitudes while MSE provides a greater 'weight' to the larger errors and lesser 'weight' to smaller ones placing more importance on the larger errors [59]. R^2, expressed by Equation 3.4, is an indication of how well the regression model 'fits' the data, in other words, how much variation exists around the model's 'line of best fit' [58]. It ranges between 0 and 1 which means it ranges from no correlation to perfect correlation, respectively. There are certain cases, when the model is poorly fit to the data or inadequate for the application when R^2 will exceed 1 or even display a negative correlation value.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \qquad (3.3)$$

$$R^2 = 1 - \frac{RSS}{TSS} \qquad (3.4)$$

Where $\hat{y}$ is the predicted value (dependent variable), y is the known value (independent variable), n is the number of data points, RSS is the residual sum of squares (i.e sum of squared differences between the predicted and actual value) and TSS is the total sum of squares (i.e. the sum of squared differences between each actual value and average of all actual values).

3.2.2 Machine Learning Models

The regression specific Machine Learning (ML) models used in this project, as alluded to by the blue blocks in Figure 3.3 above, are Neural Network (NN) (with two variations being Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM)), Random Forest (RF) and Support Vector Machine (SVM). A brief explanation of these models and their manner of operation is provided below.

Neural Network (NN)

A NN is a ML algorithm that mimics the structure and functionality of a human brain [60]. It comprises a series of neurons (or nodes) linked together by weighted connections and separated into different layers. During a NNs training phase, these weights are adjusted to minimize the difference between the model's predicted value and the actual value [61]. A RNN is a type of NN that implements sequential processing and places emphasis on the order of data elements.

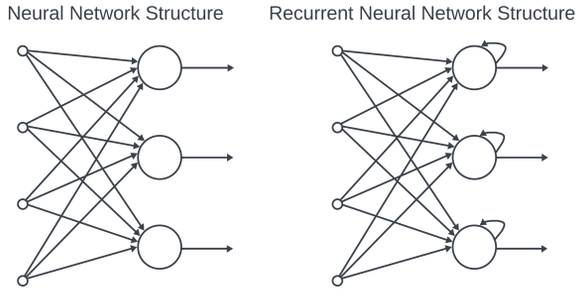


Figure 3.5: Neural Network (NN) vs. Recurrent Neural Network (RNN) Structure [13]

The difference between their structures, as highlighted in Figure 3.5, is that RNNs have a connection that loops back to themselves allowing them to retain the memory of previous inputs and outputs to better inform current predictions. This is conducive for sequence-relevant applications [13].

The RNN does suffer from a significant performance-hindering problem known as the vanishing gradient problem. This problem is apparent in other types of NNs where during the backpropagation stages when the gradient of the loss function is computed with respect to the model's parameters in pursuit of loss function minimization, the gradient becomes exponentially small as a result of being backpropagated through the model's layers. This is particularly prominent when a structure repeatedly applies the same parameter weightings as in the case of the RNN [62]. An exceptionally small gradient, or seemingly vanishing one, updates the weightings or the earlier layers to very small values resulting in slow or less learning taking place within those layers. This hinders the model's ability to fully comprehend and capture the data and in the case of the RNN prohibits it from retaining long-term information and therefore detracts from its unique property. To address this issue a variation of a RNN model was created called the LSTM, which has a more robust and sophisticated memory cell.

The difference between the structure of a RNN versus a LSTM is the presence of a forget gate. This allows the LSTM to discard irrelevant information being stored in its 'memory' and thereby obtaining clearer views of data patterns and dependencies [63]

When applied to regression tasks, NNs and its variations (i.e. RNNs and LSTMs) approach the data with a 'traditional regression methodology' as described in Section 3.2.1 above. In broad terms, they aim to minimize a cost function (through gradient descent, backpropagation or related methods) to best parameterise the model and improve it's fit' to the provided data.

Random Forest (RF)

RF is a type of ensemble ML algorithm meaning that it is a combination of multiple ML models - in this case decision trees [64]. There are two categories of ensemble methods: bagging where the models work in parallel and their results are combined or aggregated and boosting when the models work sequentially. RF is the former where each decision tree is constructed and trained individually and receives as input a random subset of the training data. This is a result of a process called bootstrapping which creates these subsets by randomly sampling the training data with replacement. Each decision tree produces an output, all of which are averaged to obtain the RF's final output. The structure of a RF model is shown in Figure 3.6 below.

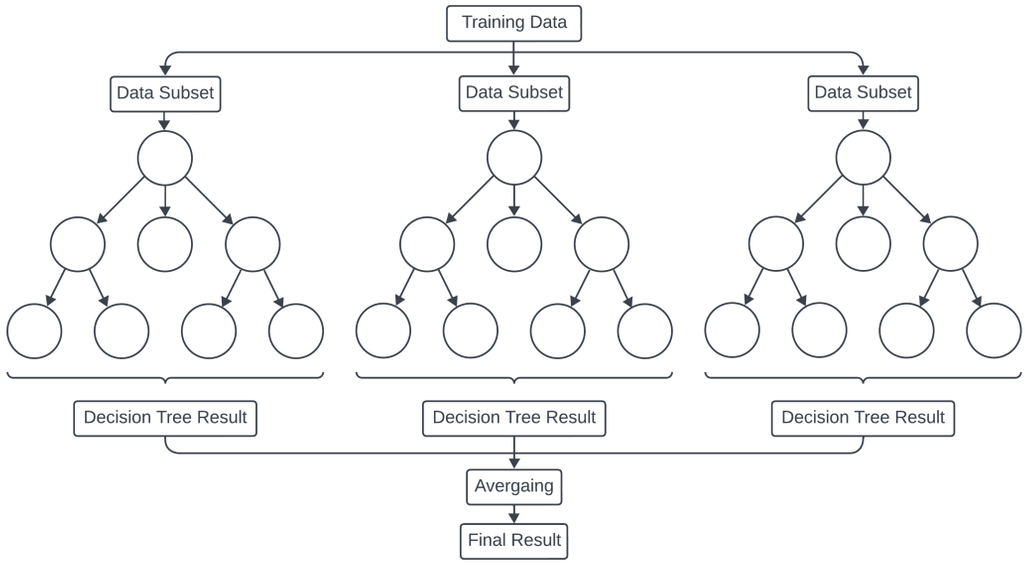


Figure 3.6: Random Forest (RF) Structure [14]

The advantages of this model are that it is incredibly flexible and therefore adaptable to many applications and it is stable and avoids the problem of over-fitting by training off of unique and randomly produced data subsets and averaging to obtain a result.

When applied to regression tasks, RFs do not opt for 'traditional regression methodology' as their operation is not through cost function minimisation or parameter optimisation. Each decision tree will construct its regression model and these will be averaged to obtain an overall regression model. However, a RF model can still be optimised through hyperparameter tuning including variables: number of decision trees, maximum number of 'leaf nodes' that a tree can have, size of data subsets, etc.

Support Vector Machine (SVM)

When an SVM is used for a regression task it is often referred to as a Support Vector Regression (SVR). A SVM works by finding a hyper-plane, usually in high dimensionality data that best separates data into respective classes while minimizing error margin but an SVR uses the same technique to find a hyper-plane that represents the underlying trend or 'regression' of the data [65]. Much like RF, SVR does not follow 'traditional regression methodology', instead it maps the input data to a higher dimensional space where its behaviour can be best understood.

Chapter 4

Requirements Analysis

4.1 Overall Project

4.1.1 System Description

The overall project can be graphically understood by Figure 4.1 below - a simplified version of Figure 1.1 above. The project is broken into four submodules each representing one of the user requirements, as further detailed in Section 4.1.3, which are based on the needs and expectations of the client.

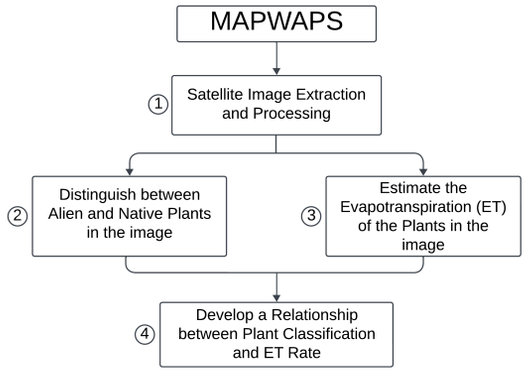


Figure 4.1: System Flow Chart

A simplified project explanation derived from this diagram is that an image will be extracted from a suitable satellite and have two algorithms applied to it. The first will differentiate between the alien and the native vegetation and the second will estimate the Evapotranspiration (ET) level experienced by the vegetation in that image. Finally, a relationship between the type of plant and the amount of water usage will be developed.

4.1.2 High Level Description

The high-level description of such a system is one that can remotely monitor a given region, and determine the extent of alien plant species invasion as well as its overall water usage. These factors are monitored to determine if intervention is warranted i.e. if a region is invaded by IAPs and is experiencing a great demand for water resources then intervention is required to remove the IAPs and re-stabilize the ecosystem. This type of 'smart' monitoring will allow for IAP management programs to better understand the effect of IAPs on a given invaded environment and make better use of their time and resources in areas with greater priority. These would be the worst affected areas where water security is most threatened.

4.1.3 Requirements, Specifications and Acceptance Test Protocols

Table 4.1 below details the overall project's user requirements derived from the project's description and shown in the system flow chart in Figure 4.1 above and their corresponding specifications and ATPs

R#	Requirement	S#	Specification	A#	Acceptance Test Protocol
UR1	A satellite image must be obtained at specified co-ordinates.	US1	A Python script will accept co-ordinates and extract the corresponding Landsat 8 satellite image using GEE API to access the Landsat database.	UATP1	Using co-ordinates of a known region, extract the Landsat 8 image and display it to visually confirm correct image extraction.
UR2	The native and alien plants in the image must be distinguished.	US2	A plant classification ML algorithm will categorize the plantation appearing in the image with an overall accuracy of 85% or a species specific accuracy of 70%.	UATP2	The plant classification ML algorithm will be assessed by applying the algorithm to an image with associated plant classification ground truths and measuring model accuracy.
UR3	The ET of the plants in the image must be estimated.	US3	An ET estimation regression ML algorithm will estimate the water use of the plantation appearing in the image with a co-efficient of regression >0.8.	UATP3	The ET estimation regression ML algorithm will be assessed by applying the algorithm to an image with associated ET ground truths and measuring model co-efficient of regression.
UR4	A relationship between plant types and water use must be developed	US4	A function or 'rule of thumb' will be ascertained that can relate plant classification to estimated water use or ET	UATP4	Combining UATP2 and UATP3 above, the derived function will be tested to see if it 'holds true' on different images with different plant classifications and ET rates.

Table 4.1: Overall System User Requirement Analysis

4.2 Project Scope

As stated in the Introduction, the scope of this report is only a portion of the Overall Project. Therefore, this project only addresses the user requirements associated with satellite image extraction and processing and ET estimation, denoted as UR1 and UR3 in Table 4.1 above. In addition to these above user requirements, a more in-depth analysis can be performed regarding the application requirements of the project. These are known as the Functional Requirements denoted as FR. Table 4.2 below describes the project scope-specific functional requirements and their corresponding specifications and ATPs.

R#	Requirement	S#	Specification	A#	Acceptance Test Protocol
FR1	The chosen satellite's data must be easily accessed.	FS1	Landsat 8 can be accessed via GEE's Python API.	FATP1&2	Write a Python script, using GEE, to extract an image of a known region and display it to visually confirm correct image extraction and adequate quality.
FR2	The choice of satellite must provide data of sufficient quality.	FS2	Landsat 8 has a spatial resolution of 15 - 30m.		
FR3	The chosen satellite must provide data that is relevant to environmental or vegetation applications.	FS3	Landsat 8 has multispectral bands which hold information regarding vegetation and land cover.	FATP3	Use the bands to calculate vegetation indices to determine if they contribute to the performance of the ML model (i.e. useful for plant application).
FR4	The chosen satellite must regularly visit a region under observance.	FS4	Landsat 8 has a temporal resolution of 16 days and a revisit time of 8 days.	FATP4	Extract satellite images over a known range (e.g. a year) and see how many images outputted to determine frequency of region surveyance.
FR5	A database on which to train a ML algorithm off of must be constructed.	FS5	AmeriFlux flux tower data containing LE values will be used and merged with corresponding satellite data to produce a combined dataset.	FATP5	Process the data using Python functions and check (step by step) that each function operates as intended and that the final dataset is correct.
FR6	The project's coding language and environment must be conducive to large data storage and processing.	FS6.1	Python will be used as the programming language of choice as it is highly applicable to data analytic and processing applications.	FATP6	Code the project and see if any issues arise with the chosen language or environment.
		FS6.2	Jupyter notebooks will be run from Google Colab to leverage this cloud based platform's computational resources.		
		FS6.3	Google Drive will be used to store data as it too is cloud based, can accommodate large storage capacities and is compatible with working in conjunction with Google Colab.		

Table 4.2: Project Scope Functional Requirement Analysis

Chapter 5

Design

The project's methodology and design can be graphically described by the flow chart diagram in Figure 5.1 below which is followed by an in-depth description.

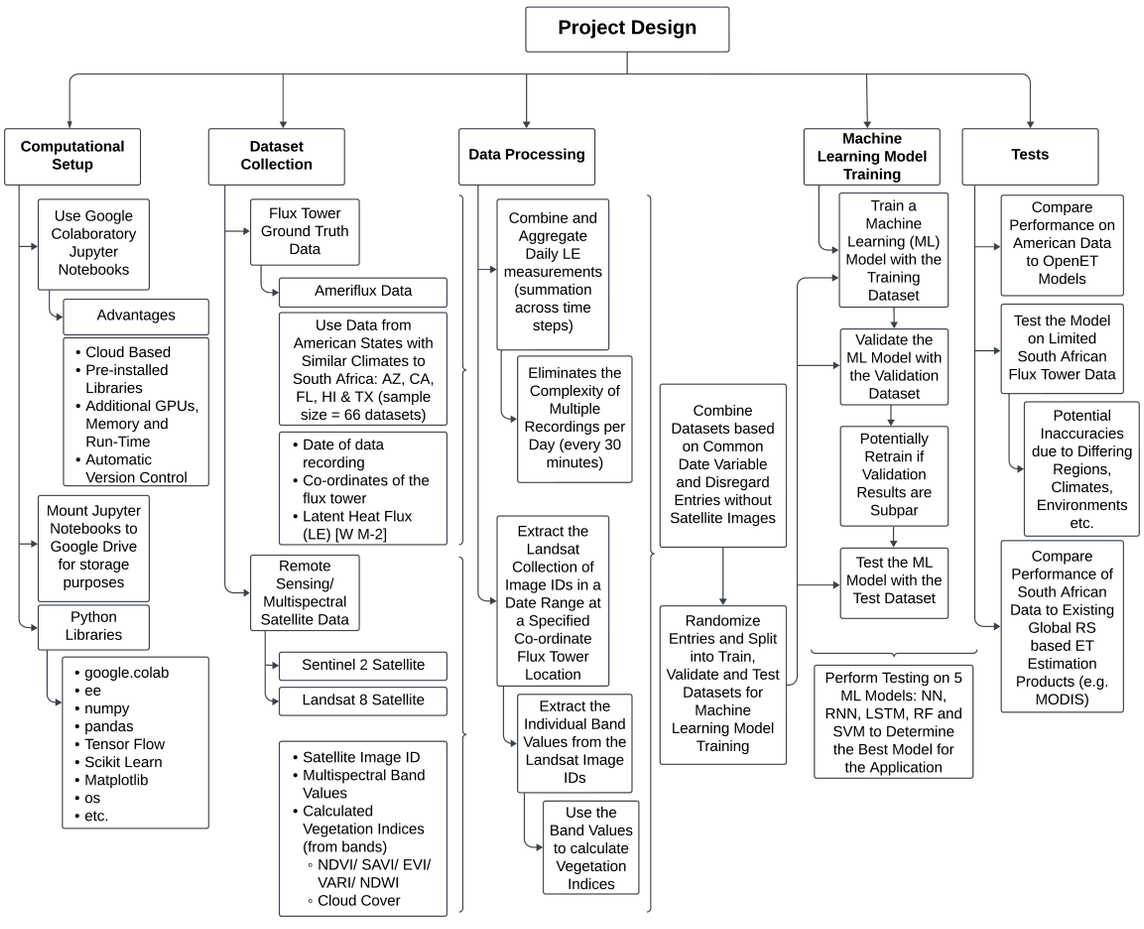


Figure 5.1: Project Methodology and Design Flow Chart

5.1 Computational Environment Setup

This project is purely software-based and will reside within Google Colaboratory-based Jupyter Notebooks. The advantage of using Google Colab is that it is a cloud-based service that allows a local machine to utilize its processing capabilities, run time and additional memory that would otherwise pose limitations to a Personal Computer (PC) based coding project. These additional computational resources are well suited to Artificial Intelligence (AI) and Machine Learning (ML) applications [66] which inherently have high processing demands. Python is the programming language of choice due to its high compatibility with ML applications, numerous libraries and pre-defined functions that aid in the training, validation, testing and performance improvement of ML models. The following Table 5.1 details the Python libraries that will be used through the entirety of this project:

Python Library	**Purpose**
ee	This library is the Python API for Google Earth Engine (GEE), it grants access to a wide variety of earth observational data and in this case Landsat 8 satellite imagery. It requires authentication with user interaction and an externally generated authentication token as well as initialization before use.
geemap	This library is a derivative of ee and allows for the visualization of GEE imagery in an interactive capacity.
google.colab	This library, exclusive to Google Colab, grants Jupyter notebook the ability to connect to Google Drive for data retrieval and storage.
IPython.display	This library allows for the display of 'rich output' including images, audio or HTML.
Matplotlib	Python library used for visualizations and plots.
numpy	This library allows for the use of arrays (or multi-dimensional data structures) for storage, traversal or manipulation.
pandas	This library provides a DataFrame in which csv file data can be uploaded, analysed and manipulated. All data processing and collation will be performed within this library's framework.
rasterio	This library allows for the writing and reading of rasterio (or geospatial) data.
re	The regular expression library allows for pattern recognition and string manipulation.
Scikit-learn	This library is used for machine learning tasks, most commonly splitting a dataset into train and test segments but also machine learning ensemble models, pre-processing and metrics.
Seaborn	This library, often used in conjunction with Matplotlib, is used specifically for the visualization and plotting of statistical data.
Tensor Flow	This library grants access to open-source machine learning framework, including: models, metrics, optimizers, add-ons, losses and layers.

Table 5.1: Python Libraries to be used in this Project [18]

A noticeable disadvantage is that Google Colab is intended for interactive use and is therefore prone to 'timing out' when functions take too long to run. For example, a function that accesses Google Earth Engine (GEE) to extract Landsat imagery for multiple datasets. The code will therefore be written to accommodate occasional runtime stops such that it 'picks up where it left off' as opposed to restarting entirely.

5.2 Dataset Selection and Acquisition

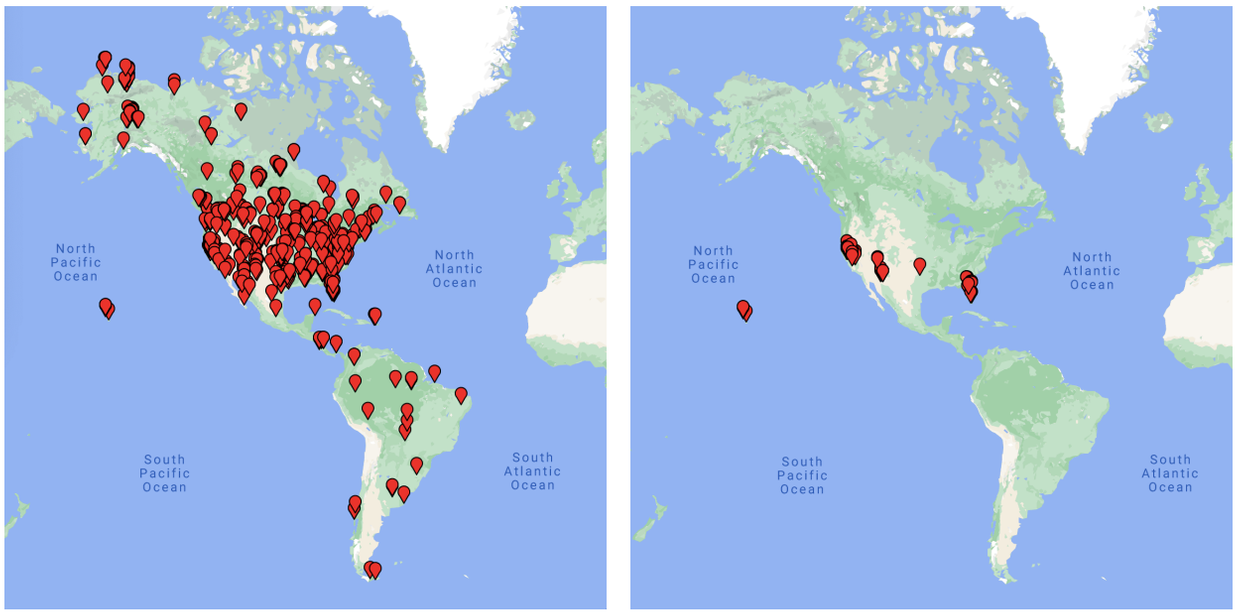
Two separate dataset selections and extractions are required: flux tower ground truth datasets which store known ET, or proxy measurement LE, values for a specified location and date as well as satellite image datasets which store multi-spectral satellite image data with corresponding location and date. These two datasets will later be merged, as detailed in Section 5.3 to provide a single ML dataset. This will link satellite image properties to a known ET estimate. The decisions made for which flux tower and satellite databases to obtain data from, as well as how the data will be acquired, are detailed below.

5.2.1 Flux Tower Ground Truth Data

Due to the lack of flux tower stations and resulting data resources available in South Africa, there is not enough ground truth data available for accurate and reliable training of an ET estimation ML model. The solution to such an obstacle is to train the model on readily available and large-scale datasets that provide enough information to obtain an acceptable level of model performance. The data will be selected to closely resemble the data that would be expected to come from South African flux towers. If, in the worst-case scenario, the data available results in a model that does not test well in a South African application, at the very least this project lays out a process that, when enough South African flux tower data has accumulated, can be implemented in South Africa (i.e. the model created in this project can be trained on different data but following the same method and hopefully attain similar results).

AmeriFlux is an organisation situated in the United States of America that observes and records weather data using a network of eddy flux towers situated in and around the United States of America [15]. This network was first launched in 1996 and has grown from 15 operational sites in 1997 to over 110 active sites today. The distribution of the towers is illustrated in Figure 5.2a below. This company has been endorsed by many organisations including the National Aeronautics and Space Administration (NASA), the United States Forest Services and the National Oceanic and Atmospheric Administration (NOAA). Its contribution is significant in that it supplies long-term and open-source data that can be used to track weather, climate and ecosystem characteristics at a fine-scaled vantage.

Each flux tower differs in the variables that it senses as well as the years it has been operational. This is due to the flux towers differing in the sensors that they house as well as the reason for their presence in a given region. An exhaustive list of all possible variables contained within AmeriFlux's individual flux tower datasets can be seen in Table B.1 in Appendix B.



(a) All AmeriFlux Flux Towers (b) Selected AmeriFlux Flux Towers

Figure 5.2: AmeriFlux Flux Tower Distribution Map

The selection of which datasets to utilise in this project's context will be based upon the ultimate application to a South African climate. It has been reported that the American provincial states that most closely resemble South Africa with respect to weather and climate are: Arizona (AZ), California (CA), Florida (FL), Hawaii (HI) and Texas (TX) [67]. The AmeriFlux flux tower distribution for these selected states is shown in Figure 5.2b above. This similarity in climate also ensures that these regions' ET behave in similar ways as they are governed by the same ET limiting factors [1] as mentioned in Section 2.4 above.

Along with regional states or site characteristics, the other factors that can be used to filter the data are data variables (i.e. the sensed variables) and data characteristics (i.e. data use policy). These filters are easily applied in the AmeriFlux Portal as described below and shown in Figure B.1 in Appendix B.

AmeriFlux Portal

AmeriFlux offers a user-friendly and intuitive interface that allows for data information, filtering and download. Figures B.1 and B.2 in Appendix B show the Site Search and Data Download platforms of the AmeriFlux website respectively.

As per advice given by Dr. Alanna Rebelo, henceforth referred to as the client, the variables that will be used to filter the data are (1) that the dataset must contain the sensed variable Latent Heat Flux (LE), (2) a Creative Commons by Attribution 4.0 International (CC-BY-4.0) [68] data use policy, (3) the AmeriFlux BASE product and (4) a regional limitation to the above mentioned five provincial states as shown by Figure 5.2b above. This results in 66 individual AmeriFlux flux tower datasets, as seen in Figure B.2 in Appendix B - all of which will be used in the upcoming Data Processing and Preparation section and subsequent Model Training.

The dataset acquisition process described above is graphically summarised in Figure 5.3 below. The aforementioned four filtering criteria will be applied to the AmeriFlux database which will result in several, or in this case 66, individual AmeriFlux flux tower datasets as well as a single AmeriFlux site overview dataset which describes the site characteristics of the resulting individual flux tower sites. A list of all variables contained within the AmeriFlux site overview dataset can be seen in Table B.2 in Appendix B.

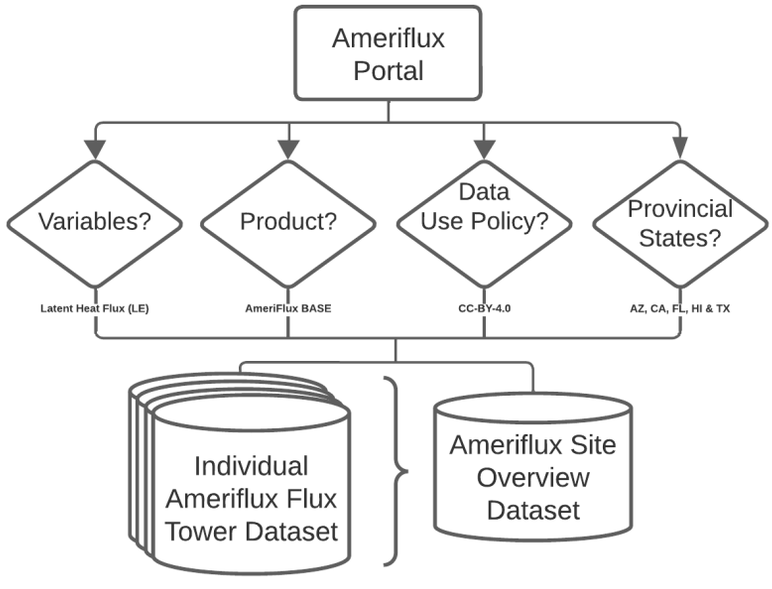


Figure 5.3: Flux Tower Dataset Acquisition Flow Chart

5.2.2 Satellite Remote Sensing Data

The design choice was made to utilize Landsat 8 as the project's satellite. Landsat 8 was chosen due to several factors, some of which being: (1) ease of extraction as GEE has a Python API that allows for easy image and image data extraction, (2) project compatibility since this satellite has had historical success when used in vegetation or ET related applications, (3) the possession of thermal band information lends to its pertinence in this context and (4) because it is the satellite used by OpenET, making performance comparison to their models easier.

The Landsat image naming convention can be broken down into two parts: Landsat collection and specific Landsat image ID, the format of which is best understood with an example:

LANDSAT/LC08/C01/T1_TOA/LC08_044033_20210116

The collection specified by **LANDSAT/LC08/C01/T1_TOA/**, states the Landsat satellite number, image collection number, tier number and processing method applied. More specifically this collection is from Landsat 8, collection 1, tier 1 and has had Top of Atmosphere (TOA) processing applied. This is the collection that will be used in this project's application because tier 1 holds images of high quality and TOA has had atmospheric and radiometric correction applied such that it indicates true reflectance values. The specific image ID specified by **LC08_044033_20210116**, states the Landsat satellite number, Worldwide Reference System (WRS) path and row and the date of image capture. More specifically this image was taken by Landsat 8 satellite on the 16th of January 2021 and its position in the WRS is in path 044, row 033.

A satellite image of the region in which a flux tower is situated, in other words, a region where LE ground truth is known, is required. This is in the ultimate pursuit of using satellite imagery to predict LE values and by extension ET values. The variables and parameters extracted from the satellite will make up the majority of the ML dataset and includes: the date on which the satellite image was taken; co-ordinates of the flux tower; unique Landsat Image ID; the 11 multi-spectral band values; percentage cloud cover and several Vegetation Indices (VIs) calculated from the band values.

A decision was made not to include the actual visual image as a ML model input. This, as per client recommendations, is because each flux tower will be dealt with as a single co-ordinate point or image pixel, which can hopefully be extrapolated further if a regional ET value is needing to be obtained. Therefore the satellite image which covers the region in which the flux tower is situated is of no use to a point-scale ML approach. However,

the multispectral band values can be obtained at a fine co-ordinate scale and therefore are necessary input parameters. These band values can be obtained from the Landsat image ID instead of the Landsat image itself.

Excluding images from the analysis removes a certain level of computational complexity as images are not only larger in memory size but also require specific ML methods such as computational vision and in this case some form of geo-spatial referencing (to link satellite images to co-ordinate specified ground truth data).

Landsat Image Data Extraction

Google Earth Engine (GEE), a cloud-based platform used for planetary and geospatial analysis and observation, will be used to extract Landsat Satellite data. Using GEE's Python API as included in Table 5.1 above, will allow access to Landsat's database.

In order to obtain satellite data relevant for this application, the AmeriFlux site overview dataset will be used. This dataset, detailing the flux tower site characteristics, provides the four necessary variables for satellite image extraction. These variables are start year and end year (i.e. operational date range of a flux tower) as well as latitude and longitude (i.e. flux tower site location). The start year and end year will be converted to start date and end date as this is the form required for Landsat 8 satellite image extraction. For simplification purposes, the start date will be set to the 1st of January [start year] and the end date to the 31st of December [end year] to ensure complete inclusivity. This step is taken as the AmeriFlux site overview dataset does not store date-specific information and to obtain it would mean delving into the individual AmeriFlux flux tower datasets which would lend unnecessary complexity to the code. Although this may lead to redundant image extraction (i.e. satellite images would be extracted for April but the flux tower only started operating in May), it is considered negligible and a worthwhile trade-off in the pursuit of reducing code complexity and ensuring code generalisability. This will not impact the code's accuracy as the redundant satellite datasets collected will be disregarded when merged with the flux tower datasets.

These four variables - start date, end date, latitude and longitude - now in the correct format, will be used in conjunction with GEE to allow for Landsat image ID extraction, image cloud cover percentage determination and plant index calculation. These are the variables that constitute the satellite image datasets.

The dataset acquisition process described above can be graphically summarised in Figure 5.4 below. The aforementioned important four variables will be extracted from the AmeriFlux site overview dataset and used to extract a list of Landsat image IDs that correspond to images taken of a specific flux tower (specified by co-ordinates in latitude and longitude) during its period of operation (specified by a start and end date). A separate dataset is then created for each of the 66 flux towers which contains Landsat image IDs and their derived band values, cloud cover percentage and VIss. Thus, each individual flux tower will have an individual and corresponding satellite dataset (i.e. the result will be 66 individual satellite datasets).

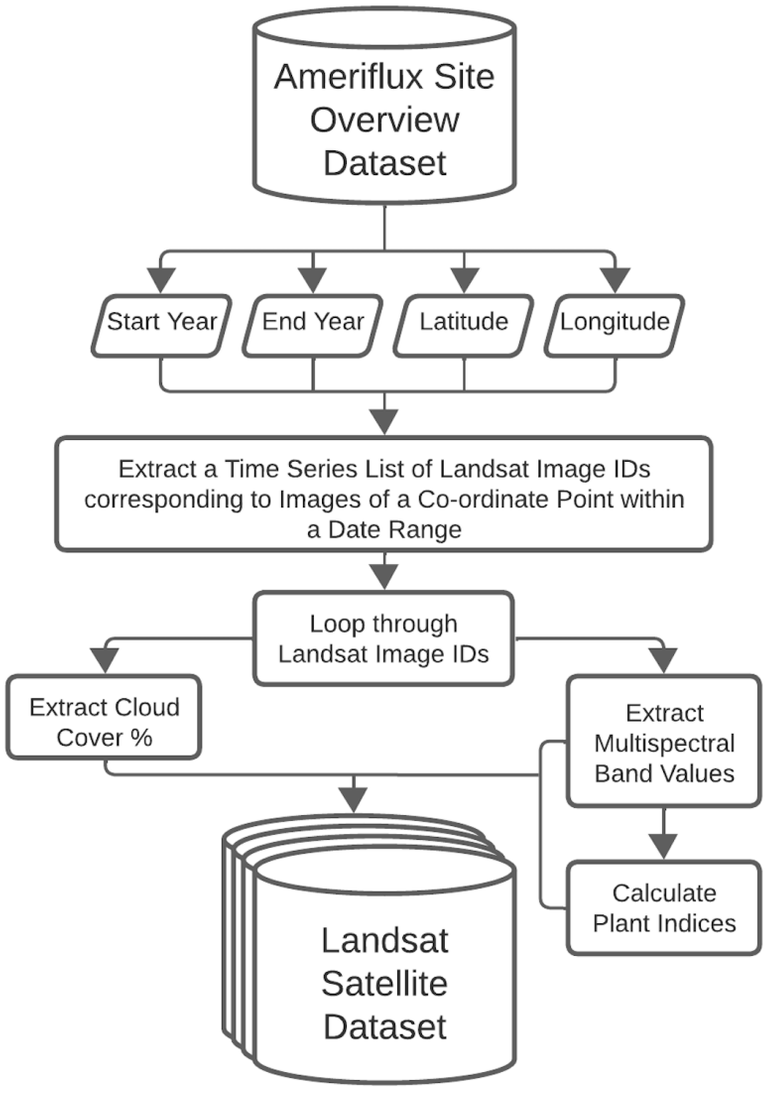


Figure 5.4: Satellite Dataset Acquisition Flow Chart

5.3 Data Processing and Preparation

The dataset construction will form a significant element of this project's scope. The proposed method of data acquisition, processing and collation can be graphically shown, as a continuation and combination of Figures 5.3 and 5.4 above, by Figure 5.5 below:

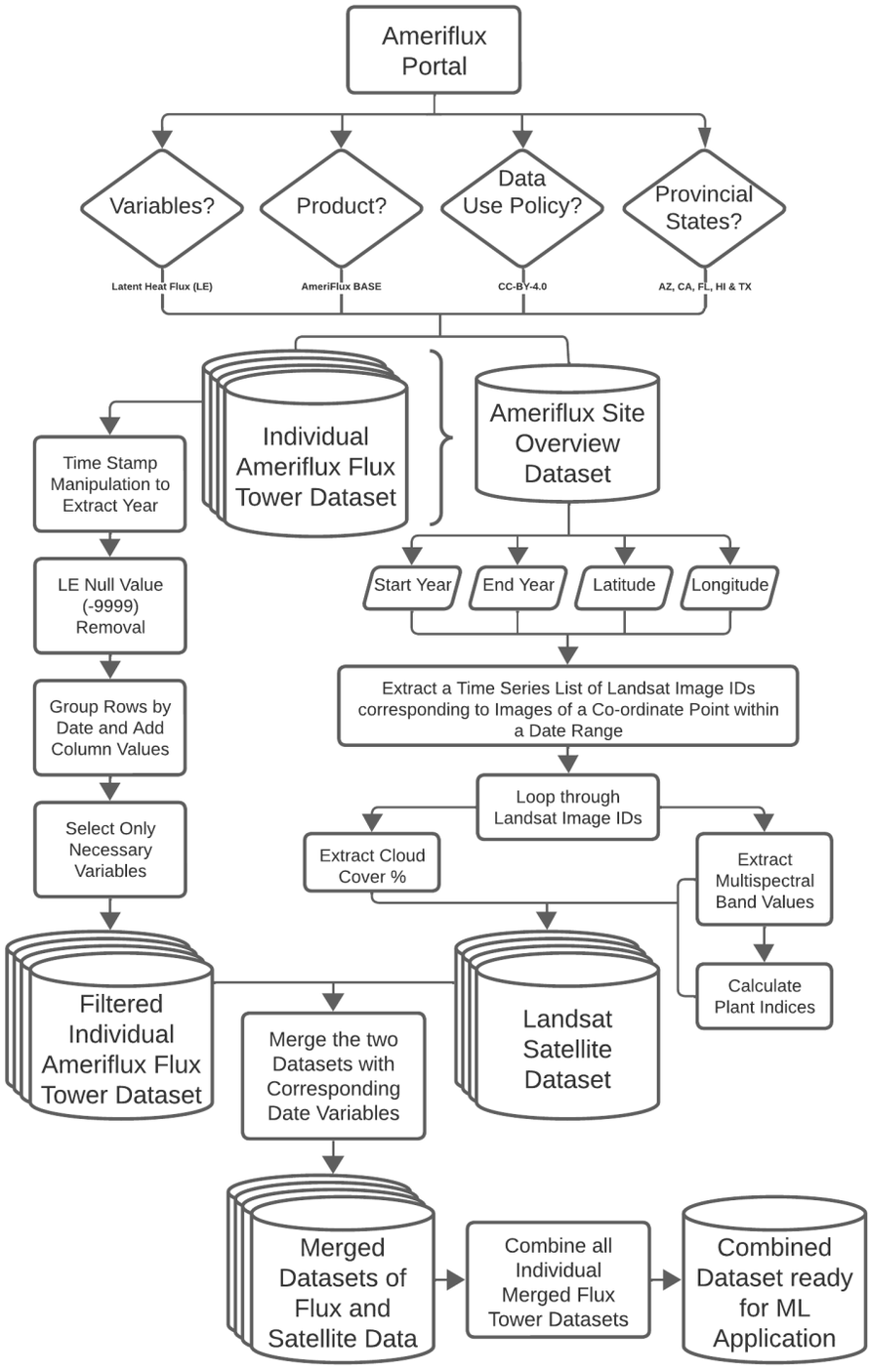


Figure 5.5: Dataset Acquirement, Processing and Collation

The written explanation of the above flow chart is as follows: once the flux tower data has been extracted from the AmeriFlux database and the corresponding Landsat satellite data has been extracted with the aid of GEE as described in Section 5.2 above, the following processing steps will take place:

1. The flux tower datasets will be manipulated, filtered and simplified as follows:
   (a) All null and/or missing LE data entries are removed. It is common for certain ecological sensors to return null value results (which in AmeriFlux's database is represented by a -9999 value).
   (b) The data will be manipulated such that it contains a complete set of daily LE measurements as the ultimate goal is the estimation of daily ET. However, accommodations must be made due to the LE null value removal as this will leave dates without a 'full day's worth of data' - these entries will be removed as they are not a true daily representation. In other words, all daily LE values will be the summation of the same number of measurements taken throughout a day's duration.
   (c) Only the important variables of the flux tower datasets will be retained and the rest will be excluded for simplification.
2. The individual flux tower datasets will be merged with their corresponding satellite datasets. This is done on the date variable. Due to a satellite's orbit and revisit time, it is likely that many flux tower entries are without a satellite image rendering them useless for this application - these entries are therefore removed.
3. All individual merged datasets will be combined to form a single dataset that will be used for ML model use. The dataset will be split for training, validating and testing of ML models.

These processing steps are hypothesised to decrease the size of the final dataset prepared for ML application. If this reduction is too drastic then alternative methods will be investigated, including: spatially increasing the dataset by removing the provincial state filter restriction; temporally increasing the dataset by making use of Landsat 7 whose operation, unlike Landsat 8, precedes 2013; data patching to minimize the number of null or missing data points or just a generally less 'aggressive' data processing strategy.

5.4 Machine Learning Model

In pursuit of the best achievable performance, several ML models will be implemented and tested and their results will be compared. The models will be trained and tested similarly to allow for informative comparison. Further work, optimisation or attempts at model improvement will occur on the best-performing model to achieve as high a model accuracy as possible.

5.4.1 Model Choice

The models were chosen based on their compatibility for regression tasks as well as their historical use in ET estimation algorithms. The following five ML models fulfilled these criteria: Random Forest (RF), Support Vector Machine (SVM), Neural Network (NN), Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM). Each of these models, as explained in Section 3.2.2 above, are applicable to regression tasks whether that be in the traditional sense or not and, as explained in Section 2.4.2 above and Figure 2.5a, some have been used before in ET estimation applications. These are NN, RF and SVM. RNN and LSTM have had far less application in this area but it was decided that given the nature of the data being sequential (i.e. time series data) and the hypothesised importance of the date variable given ET's temporal variation, there may be a possibility of improved performance.

5.4.2 Model Training

The model training will be done by splitting the data into training, validating and testing datasets such that the model can be trained on the training set, its performance fine-tuned using the validating set, and ultimately evaluated for generalization on the testing set. The validation step is not always common practice but will be done to ensure as robust a training approach as possible.

Of the models that use optimiser functions in their training processes, namely: NN, RNN and LSTM, the optimiser Adaptive Moment Estimation (ADAM) was chosen. ADAM is a variation on gradient descent optimisation, which as discussed in Section 3.2.1 above is ideal for regression task application. Unlike gradient descent optimisation, ADAM can dynamically adapt its learning rate during the training process to improve model training.

All ML models will be run three times to ensure accurate and robust performance analysis.

This is intended to account for variability and randomness that may occur especially from models that used optimiser functions due to the random initialisation of base parameters. This will indicate a model's true performance capabilities and ensure reproducibility.

Feature Engineering

Feature engineering is the processing of transforming raw data to extract the most relevant features to ensure as accurate and efficient a ML model as possible [69]. Effectively it is providing a ML model a 'leg up' so that when it begins its learning it does not start from scratch nor does it have to deal with 'unhelpful' data.

Many methods are encompassed within the practice of feature engineering but the one that will be used in this project is normalisation. This is the process of adjusting all values in a dataset to a common scale, usually, such that they fall into the range 0 - 1, to help the ML model 'put things into perspective' [70]. The type of normalisation that will be employed is Min-Max scaling which is achieved by Equation 5.1 below:

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}} \tag{5.1}$$

Where X_{min} and X_{max} are the minimum and maximum values of the variable, respectively.

Normalisation is not always guaranteed to improve a ML model's performance and so in the implementation and testing of this project, it will be included and excluded to determine any benefit or harm.

5.5 Proposed Testing

The ML model's performance can be compared to that of OpenET models. Given that the ML models will be trained off of American data, the two methods (this project's application and OpenET's models) can be directly compared.

Once an ML model has trained and produces adequate results, it will be applied to the very limited South African data. It is highly likely that because the ML models were trained off of American data, their results will not generalize well to the application on South African data. The model performance achieved on South African data can then be compared to global ET products like MOD16 which are well established but have historically performed poorly in a South African application to determine if any improvement was achieved.

Chapter 6

Implementation

Ethics clearance was obtained from the university prior to project implementation as can be seen in Figure E.1 in Appendix E.

The overall project implementation was separated into two parts: the first being data handling and the second being machine learning application. The data handling portion, as detailed in Section 6.1 below, includes the Dataset Selection and Acquisition as well as the Data Processing and Preparation as described above. The machine learning portion, as detailed in Section 6.2 below, trains and tests the various ML models chosen in Section 5.4.1 above.

6.1 Data Acquisition and Dataset Curation

The implementation of the data handling was done in three modules: (1) AmeriFlux flux tower data filtering, (2) Landsat 8 satellite image data extraction and (3) data combination and ML input preparation. To ensure proper operation, each module was designated its own Jupyter Notebook, the complete code of which can be accessed through the project's GitHub Repository as detailed in Appendix C, where the module and its tailored functions were created and tested:

1. `Flux Tower Data Processing`: this notebook was created to process a single AmeriFlux flux tower dataset - to later be extrapolated. The description of its operation is in Section 6.1.1 below and its functions are shown in Table 6.2.

2. `Landsat Satellite Data Processing`: this notebook was created to extract the satellite data corresponding to a single AmeriFlux flux tower dataset - to later

be extrapolated. The description of its operation is in Section 6.1.2 below and its functions are shown in Table 6.3.

3. `Data Acquisition, Processing and Collation`: this notebook was created to extrapolate the previous modules' operations and apply them to a collection of AmeriFlux flux tower datasets. It also has the additional functionality of merging and combining datasets as described in Section 6.1.3 below, with its functions shown in Table 6.3.

The dataset filtering, manipulating, creation and combination can be seen in Figure 6.1 below. In the forthcoming sections and processing explanations, this figure will be referenced and utilized as a visual aid to the data processing process.

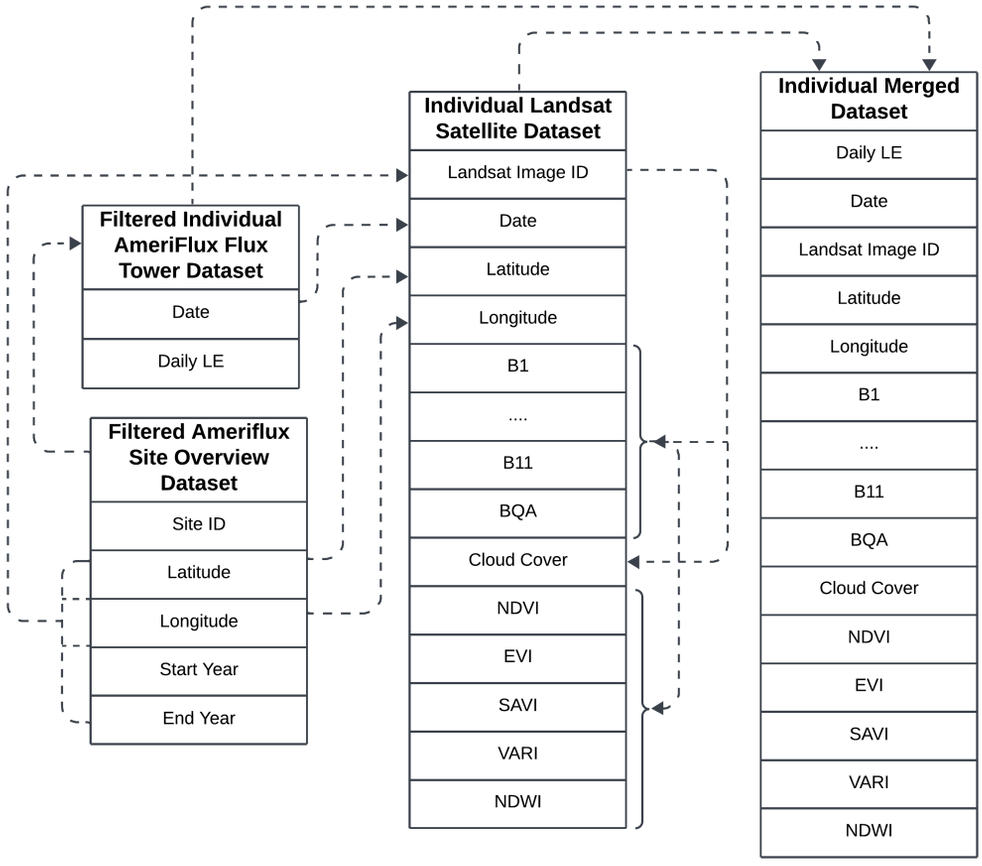


Figure 6.1: Data Acquisition and Dataset Curation Diagram

For 'house-keeping' purposes the miscellaneous function detailed in Table 6.1 below is used through all three Jupyter Notebooks. It allows for a dataset to be saved from a Python environment to a specified Google Drive file path.

Miscellaneous Functions	
Function	**Description**
`save_df_to_drive`	Function that saves a Pandas DataFrame as a csv file to a Google Drive folder specified by a Google Drive file path.

Table 6.1: Miscellaneous Functions

6.1.1 Flux Tower Data Processing

The following functions were created to process a single AmeriFlux flux tower dataset.

Flux Tower Data Functions	
Function	**Description**
`timestamp_separate`	Function that uses string handling techniques to separate the TIME_STAMP (YYYYMMDDHHMM) into TIME_STAMP_DATE (YYYYMMDD) and TIME_STAMP_TIME (HHMM) in order to isolate the date variable for later use.
`check_LE`	Function that checks if a dataset has an LE column and if not finds a LE column derivation (i.e. "LE+*") and renames it LE for simplified use.
`remove_null`	Function that removes any entry with an LE null value (represented by -9999).
`group_df`	Function that groups a Pandas DataFrame by the date variable and adds up the other columns to obtain daily entries.

Table 6.2: Flux Tower Data Functions

The only variables in a single individual AmeriFlux flux tower dataset relevant to this project were the date and LE value. The issues with the original state of the dataset were that the date was in the incorrect format, the dataset's time interval was taken every 30 minutes, and the LE columns were inconsistently named and had many null value entries.

The processing addresses these issues in the following manner: the `timestamp_separate` function was used to separate the time stamp variable into its date and time components, to isolate the desired variable - date. Before any LE variable processing could take place, there first needed to be an assurance that a LE column existed. As described in Section 5.2.1 above, when extracting AmeriFlux datasets a data variable filter was used to only extract datasets that included LE values (i.e. from flux towers that recorded LE). However, the column name is not guaranteed to be LE and may be a variation of such in the case where a positional qualifier is used. A positional qualifier indicates the position of the LE sensor on the flux tower concerning its horizontal position (H), vertical position (V) and replicates (R) - in this case, the LE column name becomes LE_H_V_R

(e.g. LE_1_1_2). For consistency and simplification, the `check_LE` function ensures that the column name is just LE, so if a positional qualifier is used, it is removed and the column is renamed. The `remove_null` function removes any entry that contains a null LE value as, for this application, these entries are irrelevant.

Finally, the `group_df` function was applied to group all entries taken in a day and add these up. This is to ensure the desired format of daily LE values. Throughout this process, there was a count variable that stored the number of entries per day. Taken at 30-minute intervals, there were originally 48 ($24 \cdot 2$) entries and when grouped/ added together the count variable becomes 2304 ($48 \cdot 48$). However, due to the null value removal, certain dates did not have a 'full day's worth of data' and their daily LE values were incomplete. These data entries were removed by the `group_df` function. This function also excludes all unnecessary columns resulting in a **Filtered Individual AmeriFlux Flux Tower Dataset** as shown in Figure 6.1 above.

6.1.2 Satellite Data Processing

The following functions were created to generate a satellite dataset for a single AmeriFlux flux tower dataset.

Satellite Data Functions	
Function	**Description**
`get_landsat_bands`	Function that extracts the multispectral bands from a Landsat 8 satellite image of a co-ordinate specified location.
`get_cloud_cover`	Function that extracts the cloud cover percentage from a satellite image specified by a Landsat 8 image ID.
`calculate_ndvi`	Function that calculates the Normalised Difference Vegetation Index (NDVI) from the multispectral bands of a satellite image.
`calculate_vari`	Function that calculates the Visible Atmospherically Resistant Index (VARI) from the multispectral bands of a satellite image.
`calculate_savi`	Function that calculates the Soil Adjusted Vegetation Index (SAVI) from the multispectral bands of a satellite image.
`calculate_ndwi`	Function that calculates the Normalised Difference Water Index (NDWI) from the multispectral bands of a satellite image.
`calculate_evi`	Function that calculates the Enhanced Vegetation Index (EVI) from the multispectral bands of a satellite image.
`get_collection_landsat_image_ids`	Function that extracts a list of Landsat 8 image IDs of a co-ordinate specified location within a date range.
`create_df_from_image_ids`	Function that creates a Pandas DataFrame from the Landsat 8 image collection: Image ID, Date, Co-ordinates, Band Values, Cloud Cover Percentage and Vegetation Indices.

Table 6.3: Satellite Data Functions

A flux tower's site characteristics of years operational (i.e. start year and end year) and co-ordinate specified location (i.e. latitude and longitude) are required to extract the necessary satellite image data. These variables are stored in a **Filtered Ameriflux Site Overview Dataset**, as seen in Figure 6.1. This is just a simplification of the original Ameriflux Site Overview Dataset by isolating relevant variables (Site ID, Latitude, Longitude, Start Year, End Year). The `get_collection_landsat_image_ids` function uses the GEE API to obtain a list of Landsat 8 image IDs corresponding to images taken at the flux tower's location, within the date range that the flux tower was/ is operational.

For each Landsat 8 image ID the following was done: the multispectral band values were extracted using the `get_landsat_bands` function, the cloud cover percentage was obtained using the `get_cloud_cover` function and the VIs were calculated from the multispectral band values, using functions `calculate_ndvi`, `calculate_vari`, `calculate_savi`, `calculate_ndwi` and `calculate_evi`. It is noteworthy that the multispectral band values are point scale and related to the co-ordinates of the flux tower but the cloud cover percentage relates to the satellite image as a whole and may not be a true indication of the cloud cover at the flux tower's exact location.

To put the satellite dataset together the `create_df_from_image_ids` function executes the above sequence for each Landsat 8 image ID in a list as well as appending the date that the satellite image was taken and the co-ordinate specified location (i.e latitude and longitude). The result of this is an **Individual Landsat Satellite Dataset**, that corresponds to a single flux tower with an associated **Individual AmeriFlux Flux Tower Dataset**, as shown in Figure 6.1 above.

6.1.3 Combined Data Processing

The following functions were created to apply the Flux Tower Data Processing and the Satellite Data Processing as detailed above to all 66 AmeriFlux datasets.

Combined Application Functions	
Function	**Description**
`get_date_range_and_coordinates`	Function that obtains the co-ordinates and operational date range of a flux tower from the AmeriFlux Site Overview dataset to be used as input parameters in extracting the Landsat satellite image dataset.
`data_aquisition`	Function that creates and stores a filtered flux tower dataset and a generated Landsat 8 satellite dataset for each Ameriflux flux tower and saves them to specified Google Drive file paths.
`merge_df`	Function that merges the individual filtered AmeriFlux flux tower dataset and the Landsat 8 satellite dataset for each Ameriflux flux tower and saves them to Google Drive file paths.

Table 6.4: Combined Data Processing Application Functions

The `get_date_range_and_coordinates` function extracts the date range and co-ordinate specified location variables from the **Ameriflux Site Overview Dataset** for a given **Individual AmeriFlux Flux Tower Dataset**. This is done via the individual AmeriFlux flux tower's dataset name containing the flux tower's site ID which is a variable stored in the AmeriFlux site overview dataset. This function therefore 'knows' which flux tower's data to extract.

The overall application loops through a Google Drive folder, specified by a file path, that contains all 66 individual AmeriFlux flux tower datasets and does the following: obtains the date range and co-ordinate specified location using the `get_date_range_and_coordinates` function and applies the `data_aquisition` function. The `data_aquisition` function filters each individual AmeriFlux flux tower dataset as described in section 6.1.1 above and generates the corresponding individual Landsat satellite dataset as described in section 6.1.2 above. These two datasets for each of the 66 flux towers are saved to a Google Drive folder, specified by a file path such that they can be accessed by the `merge_df` function which merges the two datasets based on the date variable. The function uses an inner merge which means that any flux tower data that does not have corresponding satellite data (i.e. a common date variable) is excluded and an **Individual Merged Dataset** is produced as shown in Figure 6.1 above.

This Jupyter Notebook's additional code goes on to combine all 66 individual merged datasets into a single combined dataset and excludes the Landsat Image ID, latitude and longitude variables to produce the machine learning dataset. Two versions of this dataset were created: one including the date variable and one excluding it. This allowed for experimentation to determine what impact the date variable had on the MLs' performances.

6.2 Machine Learning

Each ML model was implemented in a separate Jupyter Notebook in a Google Colab environment. The complete code can be accessed through the project's GitHub Repository, as detailed in Appendix C. The notebooks accessed Google Drive to acquire the prepared ML dataset as described in Section 6.1 above. The ML models were implemented using Python's pre-defined libraries: `Tensor Flow` and `Scikit-learn`, both of which are detailed in Table 5.1 above. The NN, RNN and LSTM models were implemented using `Tensor Flow` and the RF and SVM models with `Scikit-learn`.

Two parameters characterise the ML training datasets: the inclusion of the date variable and the application of ML technique normalisation. This results in four combinations. Each model was tested on all four combinations and as detailed in Section 5.4.2 above, the testing process was repeated three times to ensure a model's correct performance was obtained. This means that there will be $4 \cdot 5 \cdot 3 = 60$ tests performed in total.

As described in Section 3.2.1 above, regression-specific metrics were used to benchmark the performance of the models. The NN, RNN and LSTM models were benchmarked using loss, MAE and R^2 while the RF and SVM models were benchmarked with MAE, MSE and R^2. These were the metrics deemed most applicable to the respective models.

Plots generated using `MatplotLib` show the training and validation loss of the NN, RNN and LSTM models as well as regression plots for all models, generated by plotting the predicted value against the actual value and demonstrating the R^2 value.

6.3 Testing

During the implementation phase of this project, the proposed testing as outlined in Section 5.5 above evolved. Upon further investigation into OpenET models, it was found that some were only at their own implementation phase of their design process and would not be available for comparison or would not provide informative comparison results. OpenET also makes use of empirical models applied at regional scale whereas this implementation was ML model-based and at point scale.

Instead, priority was placed on model implementation with South African data. However, due to the infancy of the MAPWAPS initiative, this data was still being collected. The decision made was, that if the SA data was not ready by the end of this project's lifespan, then it would fall in the Recommendations for Future Work section.

Chapter 7

Results and Discussion

The following section details the results of the above implementation. The results section, in keeping with the structure of the implementation section, will be separated into two: data handling and machine learning.

7.1 Data Results

7.1.1 Data Processing

To demonstrate the efficacy of the data processing Python functions outlined in Section 6.1 above, one of the 66 AmeriFlux flux tower datasets was used to illustrate the result of applying each function to it. The chosen dataset has a file name: 'AMF_US-DS3_BASE_-HH_1-5'. The Python Pandas DataFrame functions `df.head()` and `df.info()` will be used to show the first few rows of the DataFrame and the DataFrame's characteristics respectively, to confirm proper code operation and output.

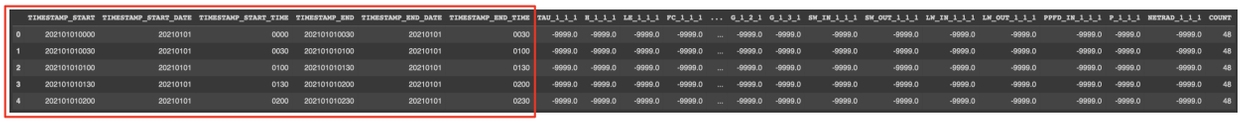
Flux Tower Data Processing

First, the flux tower data is filtered and processed. Figure 7.1 shows the initial `df.head()` output as a result of loading the CSV file into a DataFrame.

	TIMESTAMP_START	TIMESTAMP_END	TAU_1_1_1	H_1_1_1	LE_1_1_1	FC_1_1_1	WS_1_1_1	WS_MAX_1_1_1	WD_1_1_1	USTAR_1_1_1	...	G_1_1_1	G_1_2_1	G_1_3_1	SW_IN_1_1_1	SW_OUT_1_1_1	LW_IN_1_1_1	LW_OUT_1_1_1	PPFD_IN_1_1_1	P_1_1_1	NETRAD_1_1_1
0	202101010000	202101010030	-9999.0	-9999.0	-9999.0	-9999.0	-9999.0	-9999.0	-9999.0	-9999.0	...	-9999.0	-9999.0	-9999.0	-9999.0	-9999.0	-9999.0	-9999.0	-9999.0	-9999.0	-9999.0
1	202101010030	202101010100	-9999.0	-9999.0	-9999.0	-9999.0	-9999.0	-9999.0	-9999.0	-9999.0	...	-9999.0	-9999.0	-9999.0	-9999.0	-9999.0	-9999.0	-9999.0	-9999.0	-9999.0	-9999.0
2	202101010100	202101010130	-9999.0	-9999.0	-9999.0	-9999.0	-9999.0	-9999.0	-9999.0	-9999.0	...	-9999.0	-9999.0	-9999.0	-9999.0	-9999.0	-9999.0	-9999.0	-9999.0	-9999.0	-9999.0
3	202101010130	202101010200	-9999.0	-9999.0	-9999.0	-9999.0	-9999.0	-9999.0	-9999.0	-9999.0	...	-9999.0	-9999.0	-9999.0	-9999.0	-9999.0	-9999.0	-9999.0	-9999.0	-9999.0	-9999.0
4	202101010200	202101010230	-9999.0	-9999.0	-9999.0	-9999.0	-9999.0	-9999.0	-9999.0	-9999.0	...	-9999.0	-9999.0	-9999.0	-9999.0	-9999.0	-9999.0	-9999.0	-9999.0	-9999.0	-9999.0

Figure 7.1: Original DataFrame `df`

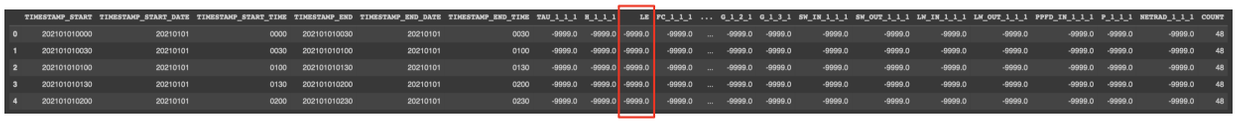
The successful timestamp manipulation resulting from the `timestamp_separate` function, as detailed in Table 6.2 above, can be seen in Figure 7.2 below:



	TIMESTAMP_START	TIMESTAMP_START_DATE	TIMESTAMP_START_TIME	TIMESTAMP_END	TIMESTAMP_END_DATE	TIMESTAMP_END_TIME	TAU_1_1_1	H_1_1_1	LE_1_1_1	FC_1_1_1	...	G_1_2_1	G_1_3_1	SW_IN_1_1_1	SW_OUT_1_1_1	LW_IN_1_1_1	LW_OUT_1_1_1	PPFD_IN_1_1_1	P_1_1_1	NETRAD_1_1_1	COUNT
0	202101010000	20210101	0000	202101010030	20210101	0030	-9999.0	-9999.0	-9999.0	-9999.0	...	-9999.0	-9999.0	-9999.0	-9999.0	-9999.0	-9999.0	-9999.0	-9999.0	-9999.0	48
1	202101010030	20210101	0030	202101010100	20210101	0100	-9999.0	-9999.0	-9999.0	-9999.0	...	-9999.0	-9999.0	-9999.0	-9999.0	-9999.0	-9999.0	-9999.0	-9999.0	-9999.0	48
2	202101010100	20210101	0100	202101010130	20210101	0130	-9999.0	-9999.0	-9999.0	-9999.0	...	-9999.0	-9999.0	-9999.0	-9999.0	-9999.0	-9999.0	-9999.0	-9999.0	-9999.0	48
3	202101010130	20210101	0130	202101010200	20210101	0200	-9999.0	-9999.0	-9999.0	-9999.0	...	-9999.0	-9999.0	-9999.0	-9999.0	-9999.0	-9999.0	-9999.0	-9999.0	-9999.0	48
4	202101010200	20210101	0200	202101010230	20210101	0230	-9999.0	-9999.0	-9999.0	-9999.0	...	-9999.0	-9999.0	-9999.0	-9999.0	-9999.0	-9999.0	-9999.0	-9999.0	-9999.0	48

Figure 7.2: Timestamp Separated DataFrame `df_time_separated`

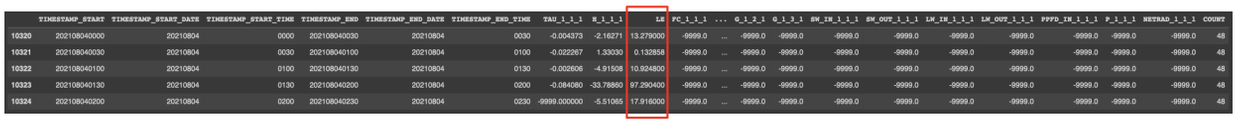
The assurance of a LE column presence through column renaming using the `check_LE` function, as detailed in Table 6.2 above, can be seen in Figure 7.3 below:



	TIMESTAMP_START	TIMESTAMP_START_DATE	TIMESTAMP_START_TIME	TIMESTAMP_END	TIMESTAMP_END_DATE	TIMESTAMP_END_TIME	TAU_1_1_1	H_1_1_1	LE	FC_1_1_1	...	G_1_2_1	G_1_3_1	SW_IN_1_1_1	SW_OUT_1_1_1	LW_IN_1_1_1	LW_OUT_1_1_1	PPFD_IN_1_1_1	P_1_1_1	NETRAD_1_1_1	COUNT
0	202101010000	20210101	0000	202101010030	20210101	0030	-9999.0	-9999.0	-9999.0	-9999.0	...	-9999.0	-9999.0	-9999.0	-9999.0	-9999.0	-9999.0	-9999.0	-9999.0	-9999.0	48
1	202101010030	20210101	0030	202101010100	20210101	0100	-9999.0	-9999.0	-9999.0	-9999.0	...	-9999.0	-9999.0	-9999.0	-9999.0	-9999.0	-9999.0	-9999.0	-9999.0	-9999.0	48
2	202101010100	20210101	0100	202101010130	20210101	0130	-9999.0	-9999.0	-9999.0	-9999.0	...	-9999.0	-9999.0	-9999.0	-9999.0	-9999.0	-9999.0	-9999.0	-9999.0	-9999.0	48
3	202101010130	20210101	0130	202101010200	20210101	0200	-9999.0	-9999.0	-9999.0	-9999.0	...	-9999.0	-9999.0	-9999.0	-9999.0	-9999.0	-9999.0	-9999.0	-9999.0	-9999.0	48
4	202101010200	20210101	0200	202101010230	20210101	0230	-9999.0	-9999.0	-9999.0	-9999.0	...	-9999.0	-9999.0	-9999.0	-9999.0	-9999.0	-9999.0	-9999.0	-9999.0	-9999.0	48

Figure 7.3: LE Column ensured DataFrame `df_with_LE`

All null LE values were removed by the `remove_null` function, as detailed in Table 6.2 above, and can be seen in Figure 7.4 below:



	TIMESTAMP_START	TIMESTAMP_START_DATE	TIMESTAMP_START_TIME	TIMESTAMP_END	TIMESTAMP_END_DATE	TIMESTAMP_END_TIME	TAU_1_1_1	H_1_1_1	LE	FC_1_1_1	...	G_1_2_1	G_1_3_1	SW_IN_1_1_1	SW_OUT_1_1_1	LW_IN_1_1_1	LW_OUT_1_1_1	PPFD_IN_1_1_1	P_1_1_1	NETRAD_1_1_1	COUNT
10320	202108040000	20210804	0000	202108040030	20210804	0030	-0.004373	-2.16271	13.279000	-9999.0	...	-9999.0	-9999.0	-9999.0	-9999.0	-9999.0	-9999.0	-9999.0	-9999.0	-9999.0	48
10321	202108040030	20210804	0030	202108040100	20210804	0100	-0.022267	1.33030	0.132858	-9999.0	...	-9999.0	-9999.0	-9999.0	-9999.0	-9999.0	-9999.0	-9999.0	-9999.0	-9999.0	48
10322	202108040100	20210804	0100	202108040130	20210804	0130	-0.002606	-4.91508	10.924800	-9999.0	...	-9999.0	-9999.0	-9999.0	-9999.0	-9999.0	-9999.0	-9999.0	-9999.0	-9999.0	48
10323	202108040130	20210804	0130	202108040200	20210804	0200	-0.084080	-33.78860	97.290400	-9999.0	...	-9999.0	-9999.0	-9999.0	-9999.0	-9999.0	-9999.0	-9999.0	-9999.0	-9999.0	48
10324	202108040200	20210804	0200	202108040230	20210804	0230	-9999.000000	-5.51065	17.916000	-9999.0	...	-9999.0	-9999.0	-9999.0	-9999.0	-9999.0	-9999.0	-9999.0	-9999.0	-9999.0	48

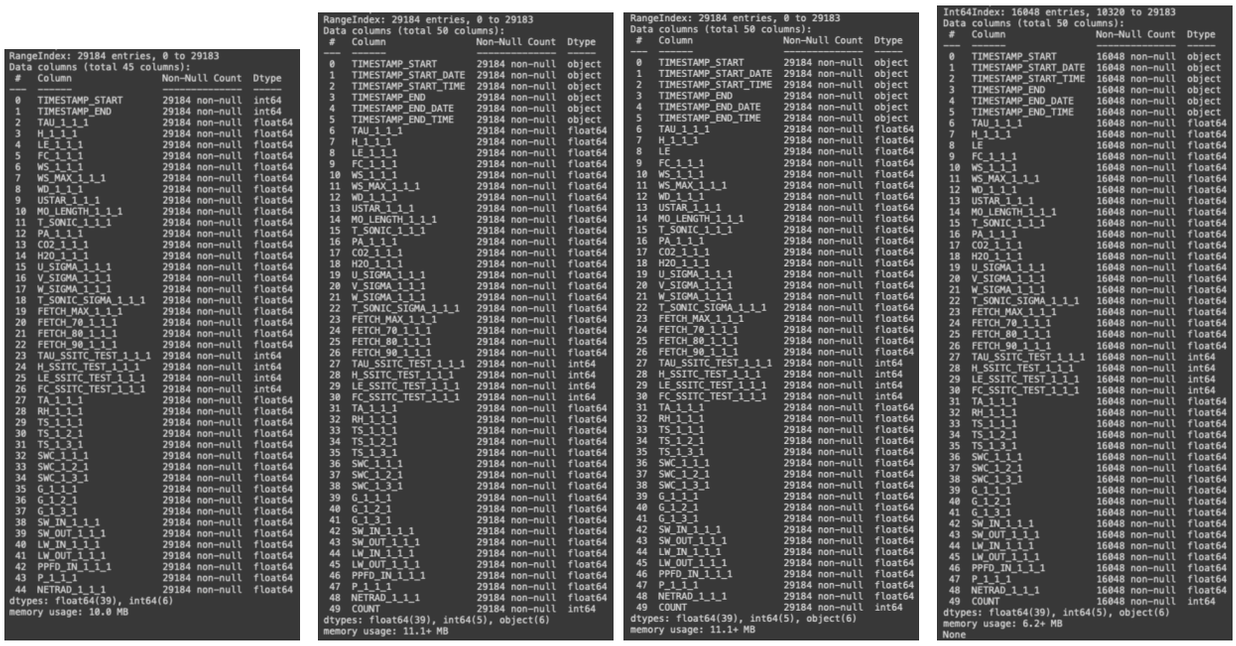
Figure 7.4: Null LE value Removed DataFrame `df_without_LE_null_values`

Figure 7.5 below shows the expected result of the `group_df` function, as detailed in Table 6.2 above, since the DataFrame has been grouped by date variable, has daily LE values, has had entries with 'incomplete days worth of data' removed and has been simplified to only the necessary columns.

	DATE	DAILY LE	DAILY COUNT
0	20210804	10642.718258	2304
1	20210805	8691.474500	2304
8	20210812	11809.477100	2304
9	20210813	10825.376100	2304
13	20210817	7467.851920	2304

Figure 7.5: Grouped DataFrame `df_grouped`

The follow Figures 7.6a, 7.6b, 7.6c and 7.6d show the DataFrame characteristics for each of the above altered DataFrames obtained using the `df.info()` function.



(a) `df` (b) `df_time_separated` (c) `df_with_LE` (d) `df_without_LE_null_values`

Figure 7.6: Characteristics of DataFrames undergoing Processing

The final processed dataset's characteristics are shown in Figure 7.7 below.

```
Int64Index: 142 entries, 0 to 383
Data columns (total 3 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   DATE         142 non-null    object
 1   DAILY LE     142 non-null    float64
 2   DAILY COUNT  142 non-null    int64
dtypes: float64(1), int64(1), object(1)
memory usage: 4.4+ KB
None
```

Figure 7.7: `df_grouped` DataFrame Characteristics

The above processing of the individual 'AMF_US-DS3_BASE_HH_1-5' dataset can be graphically summarised by the diagram in Figure 7.8 below. Where `df` refers to the original 'AMF_US-DS3_BASE_HH_1-5' dataset and `df_grouped` refers to the final **Filtered Individual Flux Tower Dataset** as shown in Figure 6.1 above. The diagram also emphasises the reduction in dataset size by specifying the number of rows and columns (rows x columns) as obtained from Figures 7.6a, 7.6b, 7.6c, 7.6d and 7.7.

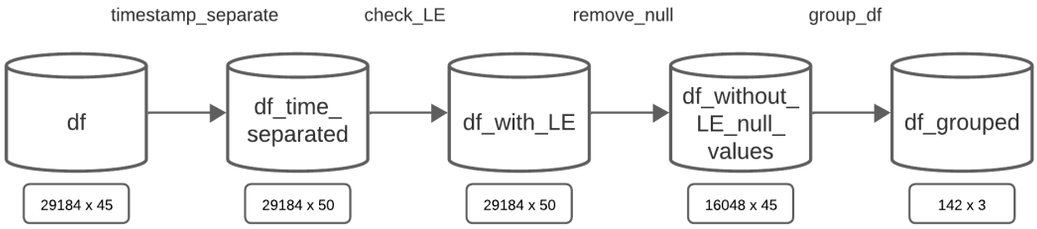


Figure 7.8: Flux Tower Data Processing Diagram of an Individual AmeriFlux Flux Tower Dataset

Satellite Data Processing

Thereafter the satellite data was extracted and processed.

```
Landsat Image IDs:
LANDSAT/LC08/C01/T1_TOA/LC08_044033_20210116
LANDSAT/LC08/C01/T1_TOA/LC08_044033_20210201
LANDSAT/LC08/C01/T1_TOA/LC08_044033_20210217
LANDSAT/LC08/C01/T1_TOA/LC08_044033_20210305
LANDSAT/LC08/C01/T1_TOA/LC08_044033_20210321
LANDSAT/LC08/C01/T1_TOA/LC08_044033_20210406
LANDSAT/LC08/C01/T1_TOA/LC08_044033_20210422
LANDSAT/LC08/C01/T1_TOA/LC08_044033_20210508
LANDSAT/LC08/C01/T1_TOA/LC08_044033_20210524
LANDSAT/LC08/C01/T1_TOA/LC08_044033_20210609
LANDSAT/LC08/C01/T1_TOA/LC08_044033_20210625
LANDSAT/LC08/C01/T1_TOA/LC08_044033_20210711
LANDSAT/LC08/C01/T1_TOA/LC08_044033_20210727
LANDSAT/LC08/C01/T1_TOA/LC08_044033_20210812
LANDSAT/LC08/C01/T1_TOA/LC08_044033_20210828
LANDSAT/LC08/C01/T1_TOA/LC08_044033_20210913
LANDSAT/LC08/C01/T1_TOA/LC08_044033_20210929
LANDSAT/LC08/C01/T1_TOA/LC08_044033_20211015
LANDSAT/LC08/C01/T1_TOA/LC08_044033_20211031
LANDSAT/LC08/C01/T1_TOA/LC08_044033_20211116
LANDSAT/LC08/C01/T1_TOA/LC08_044033_20211202
LANDSAT/LC08/C01/T1_TOA/LC08_044033_20211218
LANDSAT/LC08/C01/T1_TOA/LC08_044034_20210116
LANDSAT/LC08/C01/T1_TOA/LC08_044034_20210201
LANDSAT/LC08/C01/T1_TOA/LC08_044034_20210217
LANDSAT/LC08/C01/T1_TOA/LC08_044034_20210305
LANDSAT/LC08/C01/T1_TOA/LC08_044034_20210321
LANDSAT/LC08/C01/T1_TOA/LC08_044034_20210406
LANDSAT/LC08/C01/T1_TOA/LC08_044034_20210422
LANDSAT/LC08/C01/T1_TOA/LC08_044034_20210508
LANDSAT/LC08/C01/T1_TOA/LC08_044034_20210524
LANDSAT/LC08/C01/T1_TOA/LC08_044034_20210609
LANDSAT/LC08/C01/T1_TOA/LC08_044034_20210625
LANDSAT/LC08/C01/T1_TOA/LC08_044034_20210711
LANDSAT/LC08/C01/T1_TOA/LC08_044034_20210727
LANDSAT/LC08/C01/T1_TOA/LC08_044034_20210812
LANDSAT/LC08/C01/T1_TOA/LC08_044034_20210828
LANDSAT/LC08/C01/T1_TOA/LC08_044034_20210913
LANDSAT/LC08/C01/T1_TOA/LC08_044034_20210929
LANDSAT/LC08/C01/T1_TOA/LC08_044034_20211015
LANDSAT/LC08/C01/T1_TOA/LC08_044034_20211031
LANDSAT/LC08/C01/T1_TOA/LC08_044034_20211116
LANDSAT/LC08/C01/T1_TOA/LC08_044034_20211202
LANDSAT/LC08/C01/T1_TOA/LC08_044034_20211218
```

Figure 7.9: Landsat Image ID List produced for Individual AmeriFlux flux tower

Looking at the **Filtered Ameriflux Site Overview Dataset** as shown in Figure 6.1 above, the flux tower with Site ID: US-DS3 has an operational date range of 2021 - 2022 and a co-ordinate specified location of (38.1235, -121.5490). Figure 7.9 alongside shows that using these variables (start date, end date, latitude and longitude), the `get_collection_landsat_image_ids` function detailed in Table 6.3 above, produces a list of all Landsat image IDs that were taken over that flux towers operational date range and at that flux tower's location. The produced list had 44 Landsat image IDs, implying that the satellite 'revisited' this location 44 times during the year. The revisit time of the Landsat 8 satellite is 8 days. Therefore, on average, during a year the satellite should revisit the same location around 45.63 times. Thus, 44 is justifiable and confirms proper program operation.

The individual satellite dataset corresponding to the individual AmeriFlux flux tower dataset is created using the `create_df_from_image_ids` function, as detailed in Table

6.3 above. This uses the Landsat image IDs shown in Figure 7.9 above and extracts from each Landsat image ID: the multispectral band values, percentage cloud cover and vegetation indices. The resulting DataFrame can be seen in Figure 7.10 below and its DataFrame characteristics in Figure 7.11 below. This corresponds to the **Individual Landsat Satellite Dataset** as shown in Figure 6.1 above.

	Landsat Image ID	Date	Longitude	Latitude	B1	B2	B3	B4	B5	B6	...	B9	B10	B11	BQA	Cloud Cover	NDVI	EVI	SAVI	VARI	NDWI
0	LANDSAT/LC08/C01/T1_TOA/LC08_044033_20210116	20210116	-121.549	38.1235	0.134205	0.103966	0.069266	0.053299	0.052942	0.053209	...	0.000847	286.350586	285.287354	2720	7.74	-0.003359	-0.001504	-0.000883	0.858514	0.133577
1	LANDSAT/LC08/C01/T1_TOA/LC08_044033_20210201	20210201	-121.549	38.1235	0.323471	0.290708	0.315990	0.316670	0.368755	0.358714	...	0.007641	268.571747	269.729767	2800	59.27	0.075989	0.119629	0.065907	-0.001989	-0.077058
2	LANDSAT/LC08/C01/T1_TOA/LC08_044033_20210217	20210217	-121.549	38.1235	0.125583	0.101063	0.075981	0.064054	0.077244	0.075666	...	0.006981	278.961884	280.532562	2976	3.75	0.093347	0.046865	0.030851	0.306031	-0.008242
3	LANDSAT/LC08/C01/T1_TOA/LC08_044033_20210305	20210305	-121.549	38.1235	0.126522	0.101538	0.073504	0.063831	0.077139	0.082931	...	0.001972	291.675354	290.650909	2720	7.41	0.094406	0.047626	0.031144	0.270224	-0.024131
4	LANDSAT/LC08/C01/T1_TOA/LC08_044033_20210321	20210321	-121.549	38.1235	0.125971	0.104435	0.080836	0.075885	0.118353	0.123248	...	0.001705	290.048615	289.068756	2720	0.67	0.218635	0.134322	0.091756	0.094687	-0.188346

Figure 7.10: Extracted Satellite DataFrame

```
RangeIndex: 44 entries, 0 to 43
Data columns (total 22 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   Landsat Image ID  44 non-null     object
 1   Date              44 non-null     object
 2   Longitude         44 non-null     float64
 3   Latitude          44 non-null     float64
 4   B1                44 non-null     float64
 5   B2                44 non-null     float64
 6   B3                44 non-null     float64
 7   B4                44 non-null     float64
 8   B5                44 non-null     float64
 9   B6                44 non-null     float64
 10  B7                44 non-null     float64
 11  B8                44 non-null     float64
 12  B9                44 non-null     float64
 13  B10               44 non-null     float64
 14  B11               44 non-null     float64
 15  BQA               44 non-null     int64
 16  Cloud Cover       44 non-null     float64
 17  NDVI              44 non-null     float64
 18  EVI               44 non-null     float64
 19  SAVI              44 non-null     float64
 20  VARI              44 non-null     float64
 21  NDWI              44 non-null     float64
dtypes: float64(19), int64(1), object(2)
memory usage: 7.7+ KB
```

Figure 7.11: Characteristics of the Satellite DataFrame

Nested within the `create_df_from_image_ids` function, are the `get_landsat_bands`, `get_cloud_cover`, `calculate_ndvi`, `calculate_vari`, `calculate_savi`, `calculate_ndwi` and `calculate_evi` functions, as detailed in Table 6.3 above, that extract values for the DataFrame. These values were extracted and calculated for a single Landsat image ID and confirmed using online resources such as GEE.

Combined Data Processing

The combined data processing intends to apply both the Flux Tower Data Processing and the Satellite Data Processing to a collection of individual AmeriFlux flux tower datasets. These methods have been proven to work for a single Ameriflux dataset in Section 7.1.1 above. Therefore the results for this section are structured differently as, unlike the individual flux tower and satellite processing, they are not intended to be applied to a single AmeriFlux flux tower dataset.

The `data_aquisition` function, as detailed in Table 6.4, applies both the Flux Tower Data Processing and the Satellite Data Processing in a single function call. Although it is generally applied to a list of individual AmeriFlux flux tower datasets, when applied to a single one (i.e. 'AMF_US-DS3_BASE_HH_1-5'), it produces the same **Filtered Individual Flux Tower Dataset** and **Individual Landsat Satellite Dataset** as shown in Figures 7.5 and 7.10 above. This proves that when combined, the separate processing methods work in the same manner (i.e. correctly or as intended).

The `get_date_range_and_coordinates` function, as detailed in Table 6.4, is nested within the `data_aquisition` function to facilitate satellite data processing. It extracts the necessary variables from **Filtered Ameriflux Site Overview Dataset** for Landsat image extraction and ultimately Satellite dataset construction. The output of this function is shown in Figure 7.12 below. These variables were manually verified by accessing the **Filtered Ameriflux Site Overview Dataset** and looking at the row with Site ID 'US-DS3'. This function is generalizable such that it works on any given individual AmeriFlux flux tower dataset specified by file name.

```
[38.1235, -121.549, 2021, 2022]
```

Figure 7.12: Output of the `get_date_range_and_coordinates` Function

The `merge_df` function, as detailed in Table 6.4, accepts the **Filtered Individual Flux Tower Dataset** and **Individual Landsat Satellite Dataset** corresponding to the same individual AmeriFlux flux tower dataset and merges them based on date. The flux tower data that does not have a corresponding satellite image is discarded. This operation can be seen by Figure 7.13, a continuation of Figure 7.8.

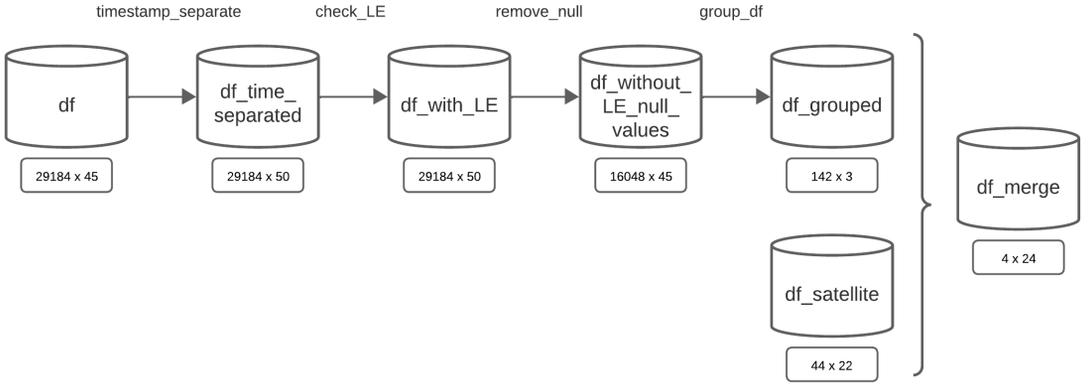


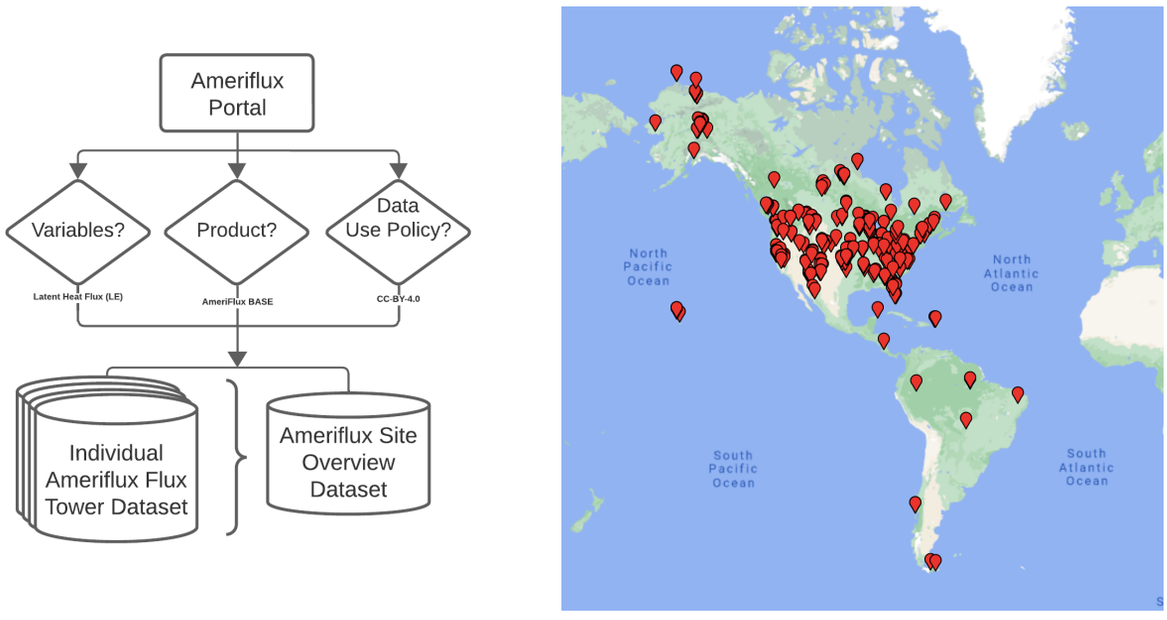
Figure 7.13: Combined Processing Diagram of an Individual AmeriFlux Flux Tower Dataset

This shows that the resulting merged DataFrame only has four entries. Had the flux dataset been 'fully intact', in that no dates had been removed (as per the Flux Tower Data Processing process detailed above), the highest number of rows that the merged DataFrame could have had would have been 44, as there are 44 satellite data entries. However, due to the 'aggressive' flux dataset processing, only four entries exist that have corresponding satellite dataset entries (i.e. with matching dates).

This is a drastic decrease in dataset size and larger than initially anticipated. It was also the case for all 66 individual AmeriFlux flux tower datasets. Once merged, all these individual datasets are combined into a single dataset to be used as ML input. The finalised combined dataset resulted in 1442 datapoints which was concerning as a ML model trained off of insufficient data will have a poor application performance. Therefore the dataset size needed to be increased, and of the suggestions made in the Design Chapter above, spatial increase was chosen as detailed in Section 7.1.2 below.

7.1.2 Spatial Data Increase

In order to achieve a spatial increase in data, the provincial state limitation was removed as can be seen by the updated flux tower acquisition flow chart in Figure 7.14a below in comparison with Figure 5.3 above. This resulted in 375 AmeriFlux datasets as seen in Figure 7.14b below instead of the original 66 as seen in Figure 5.2b above.



(a) Updated Flux Tower Dataset Acquisition Flow Chart

(b) Updated Ameriflux Flux Tower Distribution Map

Figure 7.14: Spatial Increase of the AmeriFlux Data

This introduced flux towers situated in regions whose climate and environmental conditions vary substantially from that of South Africa and from each other. These regions are not governed by the same ET limiting factors [1], as mentioned in Section 2.4 above, and may have drastically different ET behaviours. Therefore, although a larger combined dataset of 7504 datapoints was obtained, the model's performance results may still remain poor because of the increased variability introduced.

This introduced a third variable that parameterises the datasets - dataset size.

Data Analysis

Lifting the spatial limitations and including flux tower data from regions with varying environments and ET limiting factors may introduce data complexity and even inconsistency. Therefore before ML application, it was necessary to compare the distribution of both the smaller dataset (comprised of 66 individual flux tower datasets) and the larger dataset (comprised of 375 individual flux tower datasets).

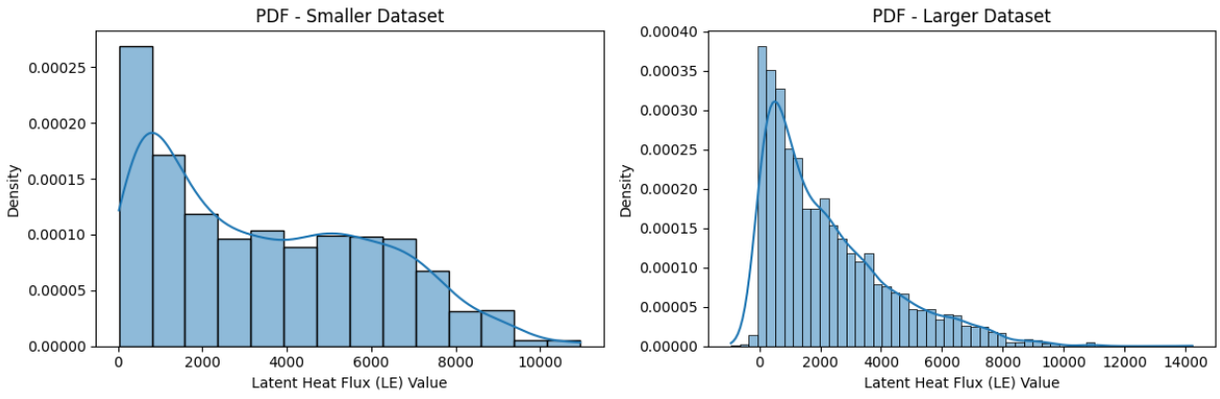


Figure 7.15: Datasets Probability Distribution Functions (PDF)

Figure 7.15 above shows that despite the greater number of data points, both datasets have similar distribution curves and thus should produce similar results when supplied to ML models. This serves to confirm that the potential disparity in ML model performance is not the fault of the data but the ML model configuration. Figure 7.15 demonstrates that the larger dataset included negative values, which after conferring with the client was understood to be a lack of robust data processing, and these values were removed before ML application.

7.2 Machine Learning Results

There are now three parameters that characterise the ML training datasets: the dataset size, the inclusion of the date variable and the application of ML technique normalisation. This results in eight combinations. Each model was tested on all eight combinations and as detailed in Section 5.4.2 above, the testing process was repeated three times to ensure a model's correct performance was obtained. This means that there will be $8 \cdot 5 \cdot 3 = 120$ tests performed in total.

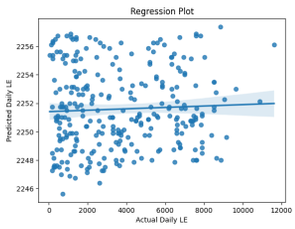
7.2.1 Machine Learning Model Performance

The results for the ML models will be tabulated and displayed separately due to the use of different metrics. Therefore the tabulated results for the NN, RNN and LSTM models are in Table 7.1 below and their loss and regression plots in Figures 7.16, 7.17, 7.18, 7.19, 7.20 and 7.21. The results for the RF and SVM models can be found in Table 7.2 and their regression plots in Figures 7.22 and 7.23 below.

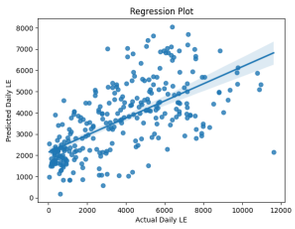
The results Tables 7.1 and 7.2 are the best-performing model result gained from repeating the tests three times. The metric used to determine best performance was R^2. Therefore they only show $8 \cdot 5 = 40$ or $\frac{1}{3}$ of the produced results. The full 120 results obtained from running each model on each data combination three times can be found in Table D.1, D.2, D.3, D.4 and D.5 in Appendix D below.

ML Model	Dataset Size	Date Included	Normalized	MAE	Loss	Normalised MAE	Normalised Loss	R^2
NN	Smaller	No	Yes	1625.74707	4256189.5	47.6819%	124830.9603%	0.41705
	Smaller	No	No	2192.5822	6818554.0	64.3068%	199983.2582%	0.0660
	Smaller	Yes	Yes	2403.0380	9293380.0	70.0471%	270896.4484%	-0.2914625406265259
	Smaller	Yes	No	2338.5939	7220006.5	68.1686%	210458.8555%	-0.0033344030380249023
	Larger	No	Yes	1198.6232	2501998.75	53.0534%	110743.47036%	0.3777
	Larger	No	No	1305.4387	2845316.25	57.7813%	125939.3897%	0.2923
	Larger	Yes	Yes	2675.5671	8794313.0	118.4259%	389253.8879%	-1.187187671661377
	Larger	Yes	No	1665.5126	5762831.0	73.7189%	255074.42962%	-0.43324363231658936
RNN	Smaller	No	Yes	1935.8814	5512736.0	56.7780%	161684.5605%	0.2449
	Smaller	No	No	2197.22631	6811084.5	64.4430%	199764.1831%	0.0671
	Smaller	Yes	Yes	2344.4726	7209021.0	68.3399%	210138.63477%	-0.0018078088760375977
	Smaller	Yes	No	2328.4892	7082316.5	67.87406%	206445.2746%	0.0157
	Larger	No	Yes	1122.1708	2082334.75	49.6695%	92168.30211%	0.4821
	Larger	No	No	1036.5958	1901509.375	45.8818%	84164.6092%	0.5270
	Larger	Yes	Yes	1653.6815	4075989.75	73.1952%	180411.4610%	-0.013718128204345703
	Larger	Yes	No	1516.8084	3834282.75	67.1369%	169713.0256%	0.0463
LSTM	Smaller	No	Yes	1807.4226	4846433.0	53.0103%	142142.3753%	0.3362
	Smaller	No	No	1913.1228	5269573.5	56.1105%	154552.7802%	0.2782
	Smaller	Yes	Yes	2359.7333	7240822.5	68.7848%	211065.6294%	-0.00622713565826416
	Smaller	Yes	No	2354.7387	7206153.5	68.6392%	210055.0488%	-0.001409292221069336
	Larger	No	Yes	1066.5833	1958038.625	47.2091%	86666.7069%	0.5130
	Larger	No	No	1362.8041	2986077.75	60.3204%	132169.7753%	0.2573
	Larger	Yes	Yes	1739.5117	4265615.5	76.9942%	188804.6761%	-0.06087899208068848
	Larger	Yes	No	1618.7723	4033558.5	71.6501%	178533.3690%	-0.0031652450561523438

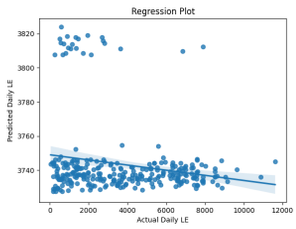
Table 7.1: Neural Network (NN), Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) ML Model Results



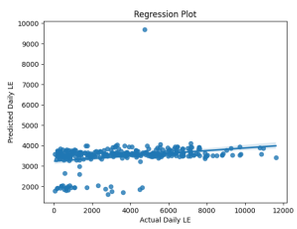
(a) Smaller Dataset, Date Included & Normalisation Applied



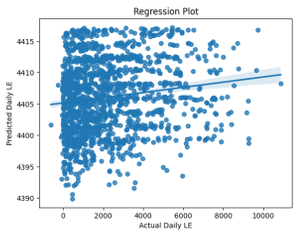
(b) Smaller Dataset, No Date Included & Normalisation Applied



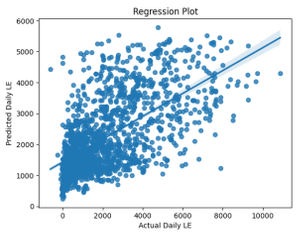
(c) Smaller Dataset, Date Included & No Normalisation Applied



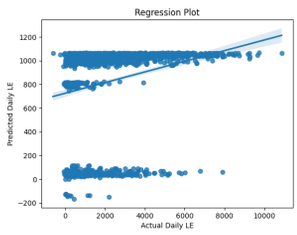
(d) Smaller Dataset, No Date Included & No Normalisation Applied



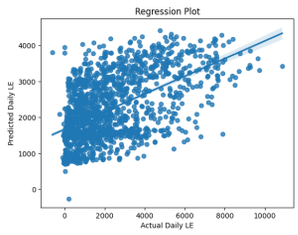
(e) Larger Dataset, Date Included & Normalisation Applied



(f) Larger Dataset, No Date Included & Normalisation Applied

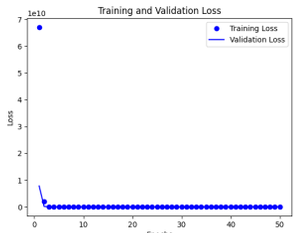


(g) Larger Dataset, Date Included & No Normalisation Applied

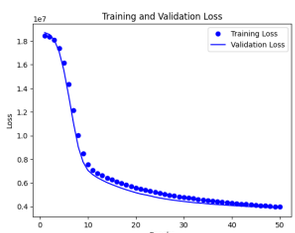


(h) Larger Dataset, No Date Included & No Normalisation Applied

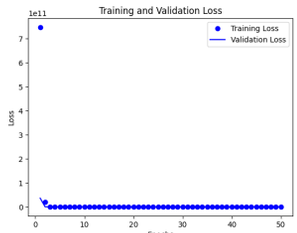
Figure 7.16: Neural Network (NN) Regression Plots



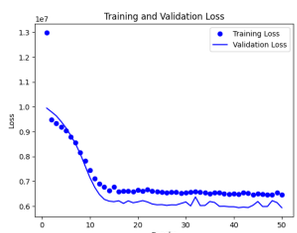
(a) Smaller Dataset, Date Included & Normalisation Applied



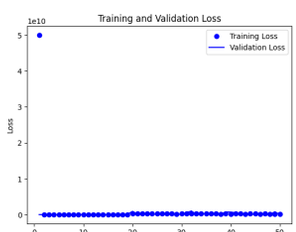
(b) Smaller Dataset, No Date Included & Normalisation Applied



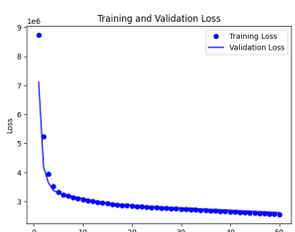
(c) Smaller Dataset, Date Included & No Normalisation Applied



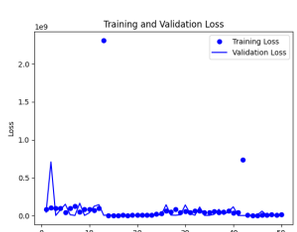
(d) Smaller Dataset, No Date Included & No Normalisation Applied



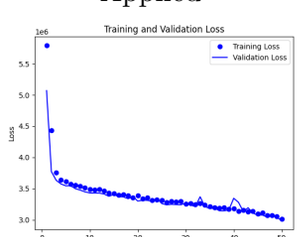
(e) Larger Dataset, Date Included & Normalisation Applied



(f) Larger Dataset, No Date Included & Normalisation Applied

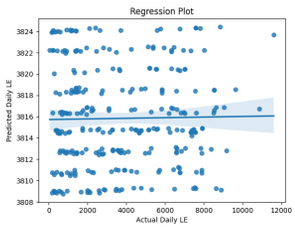


(g) Larger Dataset, Date Included & No Normalisation Applied

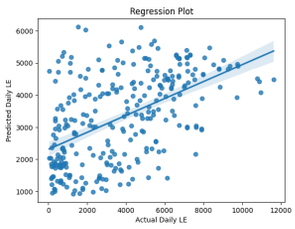


(h) Larger Dataset, No Date Included & No Normalisation Applied

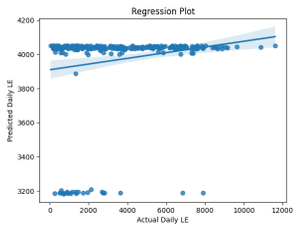
Figure 7.17: Neural Network (NN) Loss Plots



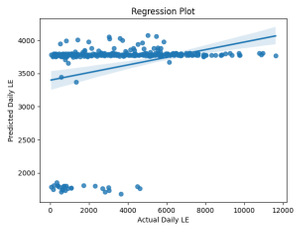
(a) Smaller Dataset, Date Included & Normalisation Applied



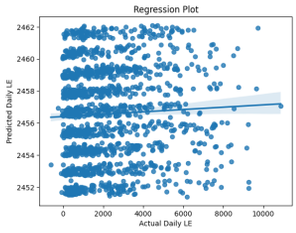
(b) Smaller Dataset, No Date Included & Normalisation Applied



(c) Smaller Dataset, Date Included & No Normalisation Applied

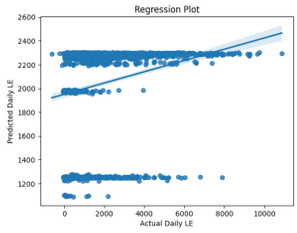


(d) Smaller Dataset, No Date Included & No Normalisation Applied

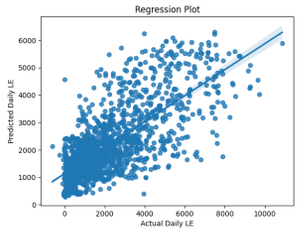


(e) Larger Dataset, Date Included & Normalisation Applied

(f) Larger Dataset, No Date Included & Normalisation Applied



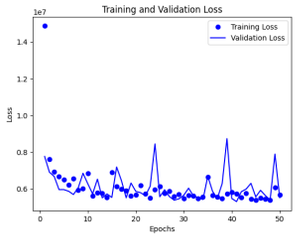
(g) Larger Dataset, Date Included & No Normalisation Applied



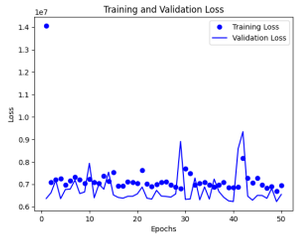
(h) Larger Dataset, No Date Included & No Normalisation Applied

Figure 7.18: Recurrent Neural Network (RNN) Regression Plots

(a) Smaller Dataset, Date Included & Normalisation Applied

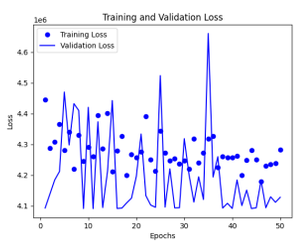


(b) Smaller Dataset, No Date Included & Normalisation Applied

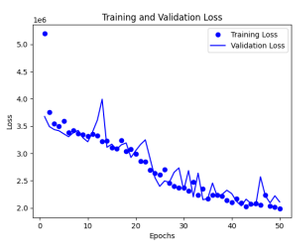


(c) Smaller Dataset, Date Included & No Normalisation Applied

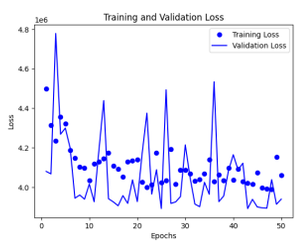
(d) Smaller Dataset, No Date Included & No Normalisation Applied



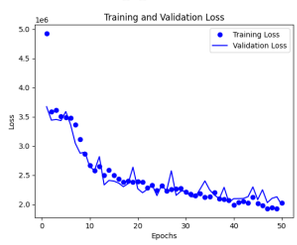
(e) Larger Dataset, Date Included & Normalisation Applied



(f) Larger Dataset, No Date Included & Normalisation Applied



(g) Larger Dataset, Date Included & No Normalisation Applied

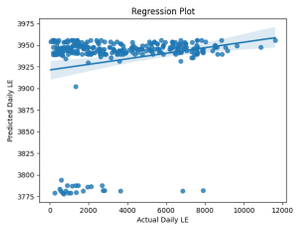


(h) Larger Dataset, No Date Included & No Normalisation Applied

Figure 7.19: Recurrent Neural Network (RNN) Loss Plots

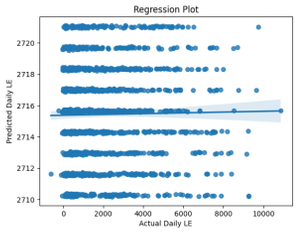
(a) Smaller Dataset, Date Included & Normalisation Applied

(b) Smaller Dataset, No Date Included & Normalisation Applied



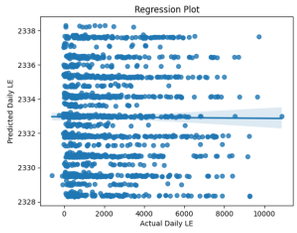
(c) Smaller Dataset, Date Included & No Normalisation Applied

(d) Smaller Dataset, No Date Included & No Normalisation Applied

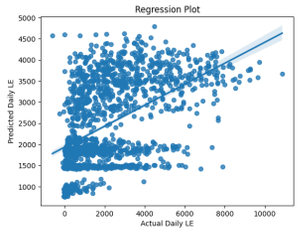


(e) Larger Dataset, Date Included & Normalisation Applied

(f) Larger Dataset, No Date Included & Normalisation Applied

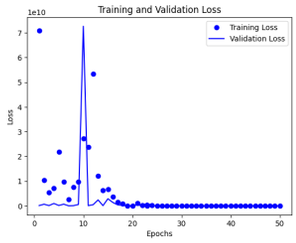


(g) Larger Dataset, Date Included & No Normalisation Applied

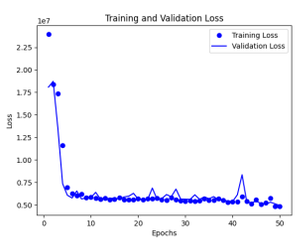


(h) Larger Dataset, No Date Included & No Normalisation Applied

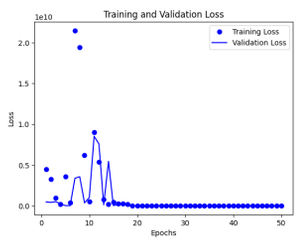
Figure 7.20: Long Short-Term Memory (LSTM) Regression Plots



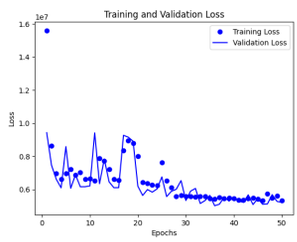
(a) Smaller Dataset, Date Included & Normalisation Applied



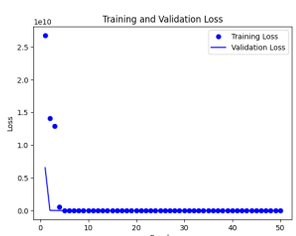
(b) Smaller Dataset, No Date Included & Normalisation Applied



(c) Smaller Dataset, Date Included & No Normalisation Applied

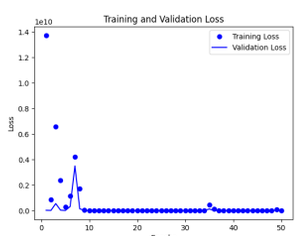


(d) Smaller Dataset, No Date Included & No Normalisation Applied

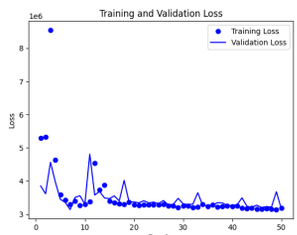


(e) Larger Dataset, Date Included & Normalisation Applied

(f) Larger Dataset, No Date Included & Normalisation Applied



(g) Larger Dataset, Date Included & No Normalisation Applied

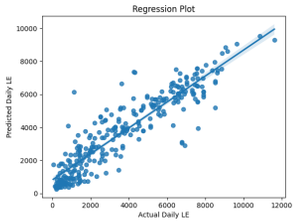


(h) Larger Dataset, No Date Included & No Normalisation Applied

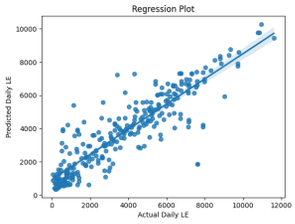
Figure 7.21: Long Short-Term Memory (LSTM) Loss Plots

ML Model	Dataset Size	Date Included	Normalized	MAE	MSE	Normalised MAE	Normalised MSE	R^2
RF	Smaller	No	Yes	806.8974	1479407.0690	23.6657%	43389.9395%	0.7973
	Smaller	No	No	808.8887	1483816.1657	23.7241%	43519.2551%	0.7967
	Smaller	Yes	Yes	789.3202	1224131.9681	23.00823%	35682.7120%	0.8298
	Smaller	Yes	No	787.9831	1220223.9298	22.9692%	35568.7950%	0.8304
	Larger	No	Yes	777.4493	1151251.3127	34.4114%	50956.6863%	0.7136
	Larger	No	No	776.4131	1155595.7674	34.3655%	51148.9806%	0.7125
	Larger	Yes	Yes	766.4599	1133985.4571	33.9250%	50192.4650%	0.7179
	Larger	Yes	No	763.5368	1127794.2726	33.7956%	49918.4308%	0.7195
SVM	Smaller	No	Yes	2279.2645	7812280.7035	66.8491%	229128.5438%	-0.0700
	Smaller	No	No	2196.0951	8518991.4903	64.4098%	249855.8601%	-0.1668
	Smaller	Yes	Yes	29914159.9787	1195008826537548.2	871979.8067%	34833789962149.89%	-166065355.238
	Smaller	Yes	No	77494276.9126	8640780421276447.0	2258911.6542%	251873562453844.7%	-1200772953.120
	Larger	No	Yes	1413.5611	3778283.1359	62.5670%	167234.3707203597%	0.0603
	Larger	No	No	1308.5601	3308838.7838	57.9194%	146455.82448998635%	0.1770
	Larger	Yes	Yes	214464188.7926	5.0061e+16	9492614.0695%	2215804413904137.0%	-12450434806.74
	Larger	Yes	No	N/A	N/A	N/A	N/A	N/A

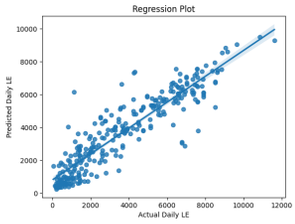
Table 7.2: Random Forest (RF) and Support Vector Machine (SVM) ML Model Results



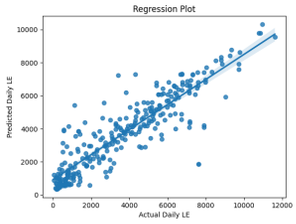
(a) Smaller Dataset, Date Included & Normalisation Applied



(b) Smaller Dataset, No Date Included & Normalisation Applied

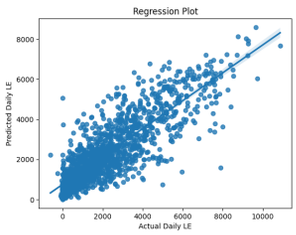


(c) Smaller Dataset, Date Included & No Normalisation Applied



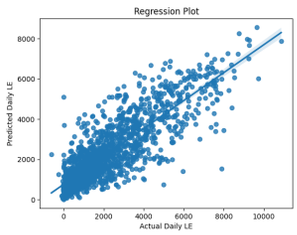
(d) Smaller Dataset, No Date Included & No Normalisation Applied

(e) Larger Dataset, Date Included & Normalisation Applied



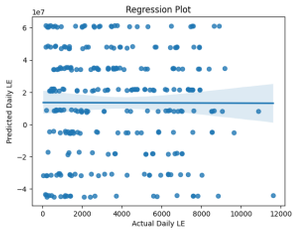
(f) Larger Dataset, No Date Included & Normalisation Applied

(g) Larger Dataset, Date Included & No Normalisation Applied



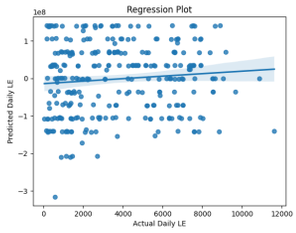
(h) Larger Dataset, No Date Included & No Normalisation Applied

Figure 7.22: Random Forest (RF) Regression Plots

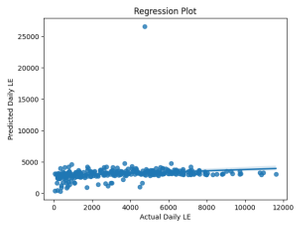


(a) Smaller Dataset, Date Included & Normalisation Applied

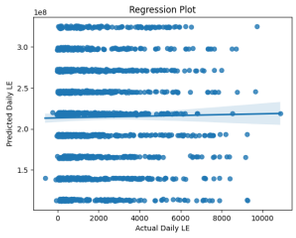
(b) Smaller Dataset, No Date Included & Normalisation Applied



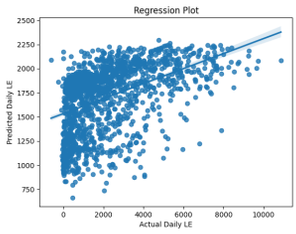
(c) Smaller Dataset, Date Included & No Normalisation Applied



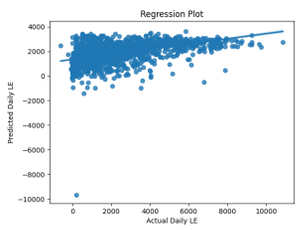
(d) Smaller Dataset, No Date Included & No Normalisation Applied



(e) Larger Dataset, Date Included & Normalisation Applied



(f) Larger Dataset, No Date Included & Normalisation Applied



(g) Larger Dataset, No Date Included & No Normalisation Applied

Figure 7.23: Support Vector Machine (SVM) Regression Plots

It is evident from the above results that the RF model produces the best performance and is, therefore, the most compatible with this project's context. The best achieved R^2 value was 0.8304 which is in the case of smaller dataset size, date inclusion and no normalisation application, as seen by Figure 7.22c. The smaller dataset consisted of ground truth data linked to similar environments and therefore similar ET behaviour which aided in model achievement. Due to RF's 'nontraditional regression methodology', this may mean that the reason it is the most compatible with this regression task is that it is not linear in nature, a hypothesis that should be addressed in Recommendations for Future Work.

All three deep learning-based models NN, RNN and LSTM performed poorly and for the most part that can be attributed to an inherent characteristic of deep learning models requiring large training datasets and this project's dataset size limitation. The RNN and LSTM, although expected to be well suited to this project given their applicability to sequential data, did not perform well with the chosen hyperparameters. The erratic behaviour of the RNN's loss plots shown in Figure 7.19 above can be attributed to the model over-fitting to training data as well as the model's vanishing gradient problem as discussed in Section 3.2.2 above. The LSTM does show an improvement as seen in Figure 7.21, proving that it does address the RNN's vanishing gradient problem but nevertheless still performed poorly overall.

The SVM was not compatible with this application. It performed so poorly and experienced

large error as seen by the extremely high and sometimes negative R^2 values. It was so incompatible that the last case test (large dataset size, date inclusion, no normalisation applied) would not run. **Note**: this was an 'unoptimised' SVM model and results may have been different if hyperparameters had been changed or tuned.

On average, the application of normalisation did improve model performance, however, this is very variable, often due to the inclusion of date. The impact of the dataset size was also very variable and greatly dependent on the dataset's other two parameters (normalisation and date inclusion). This may indicate how different models dealt with the spatial increase and consequent inclusion of differing ET behaviours. In an isolated (i.e. highly biased to this project's application) summary: the NN, LSTM, RF performed better on the smaller dataset and the RNN performed better on the larger dataset. The SVM produced inconclusive results but could be argued to have performed better on the smaller dataset size due to the larger dataset size causing the model runtime error.

7.2.2 Data Variable Dataset Variation

Unlike originally hypothesised, the inclusion of the date variable worsened model performance and often introduced inaccuracies resulting in an unrealistically large or negative R^2 value. This is true of all models except the RF. At this stage, it was realised that the performance degradation caused by the date variable inclusion could be due to the data type used to 'feed-in' the date to the ML models. In the above tests, the date was specified as a string. Although in theory, during the training process of a ML model, the model should come to recognize the string to indicate the date, this may not be the best method. Another facet of Feature Engineering is changing a variables data type if it means providing the ML model with more relevant data.

Although there is a Python variable type specifically for dates, `datetime`, it is not supported by the pre-defined ML model libraries. Another means of data type manipulation would be to convert the date variable into a numeric form, in this case, Epoch. Unix Epoch is a timestamp that represents a date by the number of seconds that have elapsed since a specific point in time (1st January 1970) [71]. It is commonly used for numeric representation of a date especially in computing applications. Another commonly used alternative is to extract the day, month and year values in their numeric form [72].

To determine if these methods of feature engineering would improve model function they were applied to the best performing RF model. The model's configuration remained unchanged but was fed different training datasets. One that had just the Unix Epoch and another that had the Unix Epoch as well as the derived numeric day, month, and

year values. The functions used to manipulate the original training dataset to produce these two date variable manipulated datasets are found in Table 7.3 below.

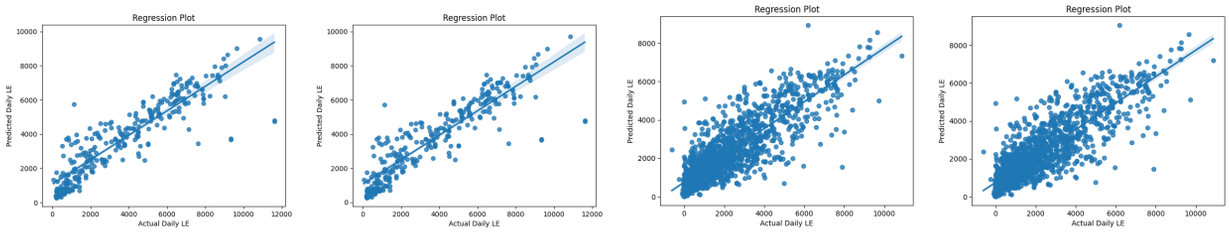
Date Variable Variation Functions	
Function	**Description**
df_epoch	Function that converts the date variable to the numeric form of Unix Epoch and removes the string date variable
df_epoch_and	Function that converts the date variable to the numeric form of Unix Epoch and separates the day, month and year variables and removes the string date variable

Table 7.3: Date Variable Variation Functions

This experiment was only applied to the original machine learning datasets that included a date variable. Therefore eight tests were performed in total: four for the Unix Epoch only dataset and four for the Unix Epoch and numeric derived date variables dataset. The two other parameterisations of the dataset, namely dataset size and normalisation were still varied to obtain all possible combinations. The results for the Unix Epoch-only dataset can be seen in Table 7.4 and Figure 7.24 below and the results for the Unix Epoch and numeric-derived date variables dataset can be seen in Table 7.5 and Figure 7.25 below.

Datset Size	**Date Included**	**Normalized**	**MAE**	**MSE**	**Normalised MAE**	**Normalised MSE**	R^2
Smaller	Yes	Yes	796.8031	1576293.0184	23.2469%	45988.8537%	0.7899
Smaller	Yes	No	794.2135	1568393.7041	23.1714%	45758.3887%	0.7910
Larger	Yes	Yes	766.0580	1133245.6660	33.9072%	50159.7204%	0.7181
Larger	Yes	No	767.8125	1135579.6097	33.9849%	50263.0254%	0.7175

Table 7.4: Unix Epoch Only Dataset Variation Results



(a) Smaller Dataset, Epoch Only Date & Normalisation Applied

(b) Smaller Dataset, Epoch Only Date & No Normalisation Applied

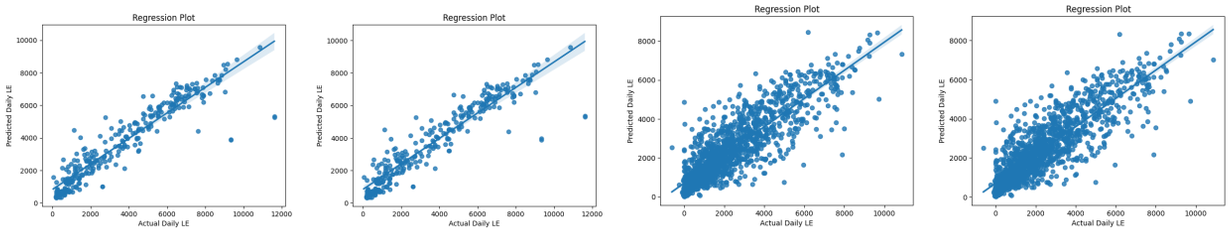
(c) Larger Dataset, Epoch Only Date & Normalisation Applied

(d) Larger Dataset, Epoch Only Date & No Normalisation Applied

Figure 7.24: RF Regression Plots with Unix Epoch Only Dataset Variation

Datset Size	Date Included	Normalized	MAE	MSE	Normalised MAE	Normalised MSE	R^2
Smaller	Yes	Yes	635.8123	1075575.2321	18.5500%	31380.2519%	0.8566
Smaller	Yes	No	636.8305	1079202.0422	18.5797%	31486.0652%	0.8562
Larger	Yes	Yes	719.8486	1021794.5197	31.8619%	45226.6697%	0.7458
Larger	Yes	No	720.5205	1024888.3853	31.8916%	45363.6104%	0.7451

Table 7.5: Unix Epoch and Derived Date Variables Dataset Variation Results



(a) Smaller Dataset, Epoch and Date Variables & Normalisation Applied

(b) Smaller Dataset, Epoch and Date Variables & No Normalisation Applied

(c) Larger Dataset, Epoch and Date Variables & Normalisation Applied

(d) Larger Dataset, Epoch and Date Variables & No Normalisation Applied

Figure 7.25: RF Regression Plots with Unix Epoch and Derived Date Variables Dataset Variation

Comparing Figure 7.24 with Figures 7.22a, 7.22c, 7.22e and 7.22g above shows that the Unix Epoch only dataset variation did not improve model performance, in fact it negligibly worsened performance as seen by the decimal decrease of the R^2 metric. However, comparing Figure 7.25 with Figures 7.22a, 7.22c, 7.22e and 7.22g, it is evident that the Unix Epoch and date derived variable dataset variation did improve performance. The previously best-performing model, in the case of smaller dataset size, date inclusion and no normalisation, experienced a R^2 increase from 0.8304 to 0.8562. However, during this variation, a 'new' best-performing model emerged in the case of smaller dataset size, date inclusion and normalisation with an R^2 of 0.8566. This model will be the one used in the upcoming South African Application.

7.3 South African Application

The best ML model was then tested on the limited South African ground truth dataset. To ensure program applicability, the dataset was processed in the same manner as described in Section 6.1. The results of this application can be seen by the regression plot in Figure 7.26 below.

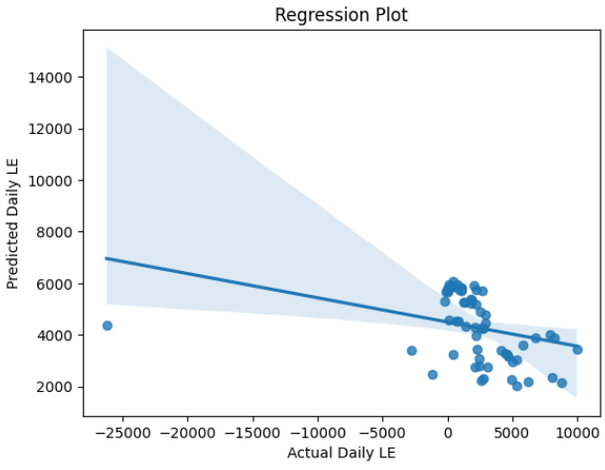


Figure 7.26: RF Regression Plot when Applied to South African Data

The performance metrics were as follows:

- MSE: 29743066.488962024
- MAE: 29743066.488962024
- Normalised MSE: 1388964.3299%
- Normalised MAE: 179.6997%
- R^2: -0.5060892970256223

It is evident that the model's performance was poor and would not provide a reliable ET estimation for a South African application. This was an expected outcome and can be attributed to the model's training data being taken not only from a different continent but from varying ecosystems with different and inconsistent ET behavioural patterns.

7.4 Acceptance Test Protocols

As outlined in Tables 4.1 and 4.2 above in the Requirements Analysis, this project needed to address certain user and functional requirements and specifications. Table 7.6 below addresses whether each of the project's ATPs were met or not and the reasons for this.

ATP#	Met	Description
UATP1	Yes	The satellite image extraction of a co-ordinate specified region can be visually verified, although not shown in this report it can be seen in Figure D.1 in Appendix D.
UATP3	Yes	The RF regression ML model attained a co-efficient of regression R^2 of 0.8566 which is >0.8 as required by specification S3.
FATP1&2	Yes	As explained by UATP1 the image extraction is correct and the quality is adequate seen in Figure D.1 in Appendix D.
FATP3	N/A	The plant indices were calculated correctly and were used in ML training. However the model was not trained without them to allow for comparison. Duncan, P et al. concluded that VIs contribute to ML model performance [19].
FATP4	Yes	The temporal resolution of Landsat 8 was proved to be eight days as shown by Figure 7.9. 44 images were extracted for a years duration ($365/8 \approx 44$).
FATP5	Yes	Each Python processing function works as intended as demonstrated in Section 7.1 above.
FATP6	Yes	The language and environment chosen were appropriate for this application but for larger scale ones, other environment options should be explored due to Google Colab's tendency to time-out as discussed in Section 5.1 above.

Table 7.6: Project Scope Acceptance Test Protocol

Chapter 8

Conclusion

IAPs are major contributors to water resource depletion posing a significant threat to South Africa. ET estimation and monitoring is used to track IAP water usage. This project aspired to create a ML model that could predict ET from RS derived variables in a South African context.

Although a good performance was achieved by the RF model when tested on the American data on which it was trained, the performance was not generalisable to the limited South African data. Given ET variability and measurement complexity, this was not an unexpected result. Nevertheless, the project contributes a foundation for further work in this field as detailed in Section 8.1 below as it sets up a process that can be followed when enough South African flux tower ground truth data has been accumulated.

ET estimation and monitoring in the interest of RS IAP tracking is still a fairly novel concept. Therefore, various implementations of this kind of project are underway globally. It is a worthwhile initiative that holds great promise with respect to ensuring an ecosystem's stabilisation and a country's water resource security.

8.1 Recommendations for Future Work

Along with the limitations outlined in the Introduction, several limitations only became apparent during the project's life cycle. These limitations should be noted and directly addressed in future work.

1. The small ML training dataset.
2. Relying on satellite imagery was limiting due to temporal and spatial resolution restrictions as well as atmospheric interference which results in infrequent data availability and reduced image data quality.
3. The limited quantity and quality of resources available in South Africa with specific reference to the lack of ground truth data because environmental sensory equipment is expensive.
4. The inherent difficulty associated with ET measurements, as explored in Section 2.4 above.

The dataset size placed considerable limitations on the project as well as ML performance. This dataset could be increased in several ways as described in Section 5.3 above. The dataset could be spatially increased, as attempted in this project, but restricted to regions with similar climatic characteristics to South Africa (i.e. with the same ET limiting factors). This would involve looking for these data collections in other countries besides America. The dataset could be temporally increased by incorporating other satellite data that precedes 2013 and has a different temporal resolution (e.g. Landsat 7). The temporal resolution does not need to be more frequent but will occur on different days thus increasing the number of days worth of satellite data available. During the data processing phase, less aggressive strategies could be adopted. For instance, instead of removing any entry that did not have a 'full day's worth' of data, a threshold could be set such that a day could be missing some data but still have enough to be included. Another common ecological data practice could be employed called patching, where missing data values are 'patched' instead of discarded, thus increasing the dataset size.

Landsat 8 was the chosen satellite for this project for reasons explained in Section 5.2.2 above, however as discussed above it places significant temporal restrictions on the project. An alternative would be to use the Sentinel -2 satellite which has improved spatial and temporal resolution. This could be used in conjunction with other satellites. Ultimately, an approach of data fusion is best suited for this project to obtain as many data points as possible, whether that be by combining satellite data or ground truth data.

Current initiatives are underway to begin to collect existing flux tower data for South Africa. This would be enormously useful as the country has such a unique climate and environment resulting in unique ET behaviour. If a ML model that was trained and tested on American data demonstrated such a good performance, then that should allow for extrapolation to South African data in a South African application producing accurate ML ET estimation.

Further work could also be devoted to optimising the ML models. Although the greatest limitation to their performance was insufficient data, a ML model's performance can be greatly improved through hyperparameter tuning, input variable re-selection and model configuration. Although not focused on in this project, it warrants attention. Another aspect pertaining to ML or more specifically Regression Machine Learning is that it seems that, from this project's conclusion, the type of regression is not strictly linear in nature and so the ML models should be altered to better suite non-linear regression tasks. Although nothing can be done to solve the difficulties of ET measurement, better calibrated ML models have the potential to produce accurate ET estimations.

This project's application is novel and innovative, and its objective is important, indicating the need for further exploration in future work.

Bibliography

[1] K. Zhang, J. S. Kimball, and S. W. Running, "A review of remote sensing based actual evapotranspiration estimation," *WIREs Water*, vol. 3, p. 834–853, Jul 2016.

[2] S. Amani and H. Shafizadeh-Moghadam, "A review of machine learning models and influential factors for estimating evapotranspiration using remote sensing and ground-based data," *Agricultural Water Management*, vol. 284, p. 108324, Apr 2023.

[3] I. Ghiat, H. R. Mackey, and T. Al-Ansari, "A review of evapotranspiration measurement models, techniques and methods for open and closed agricultural field applications," *Water*, vol. 13, no. 18, p. 2523, 2021.

[4] M. J. Goss and M. A. Oliver, *Encyclopedia of Soils in the environment*, vol. 5. Elsevier, 2 ed., 2023.

[5] P. Bai, "Comparison of remote sensing evapotranspiration models: Consistency, merits, and pitfalls," *Journal of Hydrology*, vol. 617, p. 128856, Nov 2022.

[6] "Radio waves and how satellites use them," Jul 2022. https://news.viasat.com/blog/scn/radio-waves-and-how-satellites-use-them.

[7] "Comparison of sentinel-2 and landsat," March 2019. https://www.usgs.gov/centers/eros/science/usgs-eros-archive-sentinel-2-comparison-sentinel-2-and-landsat.

[8] D. Sharma and N. Kumar, "A review on machine learning algorithms, tasks and applications," *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)*, vol. 6, no. 10, pp. 2278–1323, 2017.

[9] S. Rajbanshi, "Everything you need to know about machine learning," Mar 2021. https://www.analyticsvidhya.com/blog/2021/03/everything-you-need-to-know-about-machine-learning/.

[10] Ranbir, M. Kumar, G. Singh, J. Singh, N. Kaur, and N. Singh, "Machine learning-based analytical systems: Food forensics," *ACS omega*, vol. 7, no. 51, pp. 47518–47535, 2022.

[11] A. Bayen, T. Siauw, and D. E. Clough, *Python programming and numerical methods: A guide for engineers and scientists*. Academic Press, 2021.

[12] "Gradient descent in machine learning." https://www.javatpoint.com/gradient-descent-in-machine-learning.

[13] "Recurrent neural network (rnn)." https://machine-learning.paperspace.com/wiki/recurrent-neural-network-rnn.

[14] D. Kharkar, "About random forest algorithms.," Jul 2023. https://www.linkedin.com/pulse/random-forest-algorithms-dishant-kharkar.

[15] "Ameriflux," Jul 2023. https://ameriflux.lbl.gov/.

[16] "South africa - current climate," 2020. https://climateknowledgeportal.worldbank.org/country/south-africa/climate-data-historical#:~:text=Mean%20annual%20temperature%20for%20South,C%20(June%2C%20July).

[17] "What are the band designations for the landsat satellites?." https://www.usgs.gov/faqs/what-are-band-designations-landsat-satellites#:~:text=A%20panchromatic%20band%20was%20added,Thermal%20Infrared%20Sensor%20(TIRS).

[18] "The python standard library." https://docs.python.org/3/library/index.html.

[19] P. Duncan, E. Podest, K. J. Esler, S. Geerts, and C. Lyons, "Mapping invasive herbaceous plant species with sentinel-2 satellite imagery: Echium plantagineum in a mediterranean shrubland as a case study," *Geomatics*, vol. 3, p. 328–344, Apr 2023.

[20] L. Royimani, O. Mutanga, J. Odindi, T. Dube, and T. N. Matongera, "Advancements in satellite remote sensing for mapping and monitoring of alien invasive plant species (aips)," *Physics and Chemistry of the Earth, Parts A/B/C*, vol. 112, p. 237–245, Aug 2019.

[21] L. Henderson, *Invasive alien plants in South Africa*, vol. 21. Agricultural Research Council, 2020.

[22] W. C. Government, *Overview of Provincial Revenue and Expenditure*. 2023.

[23] A. Rebelo, K. Esler, and D. le Maitre, "New project aims to map alien invasive tree," *The Water Wheel Januray/February 2023*, vol. 22, no. 1, pp. 26–28, 2023.

[24] "Alien and invasive plant species – notable category 1b species," Aug 2020. https://www.envass.co.za/alien-and-invasive-plant-species-notable-category-1b-species/.

[25] J. Scotcher, K. Johnson, and R. Heath, "Code of good practice for managing alien and invasive species in the south african forestry industry." Forestry South Africa, 2021. https://www.forestrysouthafrica.co.za/wp-content/uploads/2022/02/FSA-Code-of-Good-Practice-AIS-Revised-November-2021-FINAL.pdf.

[26] D. Le Maitre, B. van Wilgen, C. Gelderblom, C. Bailey, R. Chapman, and J. Nel, "Invasive alien trees and water resources in south africa: Case studies of the costs and benefits of management," *Forest Ecology and Management*, vol. 160, p. 143–159, May 2002.

[27] K. Rutledge, M. McDaniel, S. Teng, H. Hall, T. Ramroop, E. Sprout, J. Hunt, D. Boudreau, and H. Costa, "Invasive species," 2023. https://education.nationalgeographic.org/resource/invasive-species/.

[28] D. Lieurance, A. Kendig, and C. Romagosa, "The stages of invasion: How does a nonnative species transition to an invader?," *EDIS*, vol. 2022, Aug 2022.

[29] B. van Wilgen, D. Richardson, and S. I. Higgins, "Integrated control of invasive alien plants in terrestrial ecosystems," *Land Use and Water Resources Research*, vol. 1, pp. 1–8, Jan 2001. https://ideas.repec.org/a/ags/luawrr/47853.html.

[30] B. Mtengwana, T. Dube, Y. P. Mkunyana, and D. Mazvimavi, "Use of multispectral satellite datasets to improve ecological understanding of the distribution of invasive alien plants in a water-limited catchment, south africa," *African Journal of Ecology*, vol. 58, p. 709–718, Jun 2020.

[31] B. W. van Wilgen, J. Measey, D. M. Richardson, J. R. Wilson, and T. A. Zengeya, "Biological invasions in south africa: An overview," *Biological Invasions in South Africa*, p. 3–31, Mar 2020.

[32] "South africa - floral surprises on africa's southern tip," Sep 2023. https://www.fauna-flora.org/countries/south-africa/.

[33] P. B. Holden, A. J. Rebelo, and M. G. New, "Mapping invasive alien trees in water towers: A combined approach using satellite data fusion, drone technology and expert engagement," *Remote Sensing Applications: Society and Environment*, vol. 21, p. 100448, Jan 2021.

[34] W. Meijninger and C. Jarmain, "Satellite-based annual evaporation estimates of invasive alien plant species and native vegetation in south africa," *Water SA*, vol. 40, p. 95, Jan 2014.

[35] A. Rebelo, "Mapwaps - mapping woody invasive alien plant species and their impacts in strategic water source areas," 2023.

[36] D. Stoll, "Water crisis in south africa: Causes, effects, and solutions," Oct 2022. https://earth.org/water-crisis-in-south-africa/#:~:text=What%20Led%20to%20A%20Water,lack%20of%20rain)%20water%20scarcity.

[37] G. R. Moncrieff, J. A. Slingsby, and D. C. Le Maitre, "Propagating uncertainty from catchment experiments to estimates of streamflow reduction by invasive alien plants in southwestern south africa," *Hydrological Processes*, vol. 35, Mar 2021.

[38] E.-M. Steenkamp, "Impacts and control of invasive alien plants in south africa," Dec 2020. https://www.hortgro.co.za/news/impacts-and-control-of-invasive-alien-plants-in-south-africa/.

[39] J. Chamier, K. Schachtschneider, D. Le Maitre, P. Ashton, and B. Van Wilgen, "Impacts of invasive alien plants on water quality, with particular emphasis on south africa," *Water SA*, vol. 38, Apr 2012.

[40] Y. Adam, N. S. Ngetar, and S. Ramdhani, "The assessment of invasive alien plant species removal programs using remote sensing and gis in two selected reserves in the ethekwini municipality, kwazulu-natal," *South African Journal of Geomatics*, vol. 6, p. 90, May 2017.

[41] "11 application of remote sensing in environmental monitoring-enhancing sustainability," Jul 2023. https://www.spatialpost.com/remote-sensing-environmental-monitoring/#:~:text=Remote%20sensing%20is%20widely%20used,and%20disease%20outbreaks%20in%20crops.

[42] D. Analytics, "What are the benefits and challenges of using satellite data for environmental analytics?," Aug 2023. https://www.linkedin.com/advice/0/what-benefits-challenges-using-satellite-data-environmental.

[43] Q. Wang, Y. Tang, Y. Ge, H. Xie, X. Tong, and P. M. Atkinson, "A comprehensive review of spatial-temporal-spectral information reconstruction techniques," *Science of Remote Sensing*, vol. 8, p. 100102, Dec 2023.

[44] "Sentinel 2 - colour vision for copernicus." https://www.esa.int/Applications/Observing_the_Earth/Copernicus/Sentinel-2.

[45] "What is the landsat satellite program and why is it important?." https://www.usgs.gov/faqs/what-landsat-satellite-program-and-why-it-important#:~:text=The%20Landsat%20Program%20is%20a,was%20later%20renamed%20Landsat%201.

[46] S. Dzikiti, N. Z. Jovanovic, R. D. Bugan, A. Ramoelo, N. P. Majozi, A. Nickless, M. A. Cho, D. C. Le Maitre, Z. Ntshidi, and H. H. Pienaar, "Comparison of two remote sensing models for estimating evapotranspiration: Algorithm evaluation and application in seasonally arid ecosystems in south africa," *Journal of Arid Land*, vol. 11, p. 495–512, Sep 2019.

[47] "Openet - what is evapotranspiration?," 2023. https://openetdata.org/what-is-evapotranspiration/.

[48] Z. Cai, Y. Tang, and Q. Zhan, "A cooled city? comparing human activity changes on the impact of urban thermal environment before and after city-wide lockdown," *Building and Environment*, vol. 195, p. 107729, Feb 2021.

[49] R. Bugan, C. L. García, N. Jovanovic, I. Teich, M. Fink, and S. Dzikiti, "Estimating evapotranspiration in a semi-arid catchment: A comparison of hydrological modelling and remote-sensing approaches," *Water SA*, vol. 46, Mar 2020.

[50] O. Gwate, S. K. Mantel, A. R. Palmer, L. A. Gibson, and Z. Munch, "Measuring and modelling evapotranspiration in a south african grassland: Comparison of two improved penman-monteith formulations," *Water SA*, vol. 44, Jul 2018.

[51] "The water balance," Sep 2023. https://geography-revision.co.uk/gcse/physical-gcse/the-water-balance/.

[52] A. Pedro, "Fluxnet15 - how to convert latent heat flux to actual evapotranspiration?," May 2021. https://earthscience.stackexchange.com/questions/20733/fluxnet15-how-to-convert-latent-heat-flux-to-actual-evapotranspiration.

[53] A. Ramoelo, N. Majozi, R. Mathieu, N. Jovanovic, A. Nickless, and S. Dzikiti, "Validation of global evapotranspiration product (mod16) using flux tower data in the african savanna, south africa," *Remote Sensing*, vol. 6, no. 8, p. 7406–7423, 2014.

[54] A. K. Maini and V. Agrawal, *Satellite Technology: Principles and applications*. Wiley, 2014.

[55] "Transforming energy into imagery: How satellite data becomes stunning views of earth," Mar 2020. https://www.nesdis.noaa.gov/news/transforming-energy-imagery-how-satellite-data-becomes-stunning-views-of-earth#:

~:text=Satellite%20imagers%20use%20remote%20sensing,wavelengths%20along%20the%20electromagnetic%20spectrum.

[56] Mehtapriyankapm, "Supervised, unsupervised and reinforcement learning," Nov 2020. https://medium.com/@mehtapriyanka1pm/supervised-unsupervised-and-reinforcement-learning-246781f26730.

[57] S. Raschka and V. Mirjalili, *Python machine learning: Machine learning and deep learning with python, scikit-learn, and tensorflow 2*. Packt Publishing Ltd, 2019.

[58] K. Mali, "Everything you need to know about linear regression!," Sep 2023. https://www.analyticsvidhya.com/blog/2021/10/everything-you-need-to-know-about-linear-regression/#h-evaluation-metrics-for-linear-regression.

[59] S. Allwright, "Mse vs mae, which is the better regression metric?," July 2022. https://stephenallwright.com/mse-vs-mae/.

[60] B. Müller, J. Reinhardt, and M. T. Strickland, *Neural networks: An introduction*. Springer, 1990.

[61] S. Shanmuganathan, "Artificial neural network modelling: An introduction," *Artificial Neural Network Modelling*, p. 1–14, Feb 2016.

[62] S. Hochreiter, "The vanishing gradient problem during learning recurrent neural netsnbsp; and problem solutions," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 06, no. 02, p. 107–116, 1998.

[63] "What is the main difference between rnn and lstm?," Nov 2022. https://www.theiotacademy.co/blog/what-is-the-main-difference-between-rnn-and-lstm/#:~:text=LSTM%20networks%20differ%20from%20RNN,can%20be%20of%20any%20length.

[64] S. E. R, "Understand random forest algorithms with examples (updated 2023)," Oct 2023. https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/.

[65] F. Tabsharani, "What is a support vector machine?: Definition from whatis," Aug 2023. https://www.techtarget.com/whatis/definition/support-vector-machine-SVM#:~:text=A%20support%20vector%20machine%20(SVM)%20is%20a%20type%20of%20supervised,data%20set%20into%20two%20groups.

[66] M. J. Nelson and A. K. Hoover, "Notes on using google colaboratory in ai education," *Proceedings of the 2020 ACM Conference on Innovation and Technology in Computer Science Education*, Jun 2020.

[67] C. Edwards, "Which five american states have the most similar weather to south africa?," Apr 2019. https://www.thesouthafrican.com/news/weather/which-five-american-states-have-the-most-similar-weather-to-south-africa/#.

[68] "Cc by 4.0 deed — attribution 4.0 international." https://creativecommons.org/licenses/by/4.0/.

[69] "What is feature engineering?," May 2023. https://www.geeksforgeeks.org/what-is-feature-engineering/.

[70] A. Bhandari, "Feature engineering: Scaling, normalization, and standardization (updated 2023)," Jul 2023. https://www.analyticsvidhya.com/blog/2020/04/feature-scaling-machine-learning-normalization-standardization/#:~:text=One%20key%20aspect%20of%20feature,is%20on%20the%20same%20scale.

[71] B. Jack, "What is unix time and when was the unix epoch?," Feb 2021. https://www.makeuseof.com/what-is-unix-time-and-when-was-the-unix-epoch/.

[72] A. Long, "Machine learning with datetime feature engineering: Predicting healthcare appointment no-shows," Feb 2020. https://towardsdatascience.com/machine-learning-with-datetime-feature-engineering-predicting-healthcare-appointm~:text=Basically%20you%20can%20break%20apart,begins%20in%20the%20prior%20year.

[73] L. Gibson, C. Jarmain, Z. Su, and F. Eckardt, "Review: Estimating evapotranspiration using remote sensing and the surface energy balance system – a south african perspective," *Water SA*, vol. 39, no. 4, 2013.

[74] Z.-L. Li, R. Tang, Z. Wan, Y. Bi, C. Zhou, B. Tang, G. Yan, and X. Zhang, "A review of current methodologies for regional evapotranspiration estimation from remotely sensed data," *Sensors*, vol. 9, no. 5, p. 3801–3853, 2009.

[75] T. Suwanlertcharoen, T. Chaturabul, T. Supriyasilp, and K. Pongput, "Estimation of actual evapotranspiration using satellite-based surface energy balance derived from landsat imagery in northern thailand," *Water*, vol. 15, no. 3, p. 450, 2023.

Appendix A

Literature Review Tables

A.1 Satellite Options Table

Satellite	Description	Spatial Resolution (m)	Spectral Resolution (# of bands)	Temporal Resolution (days)	Operational	Citation
ASTER	Sensor abroad NASA's terra satellite and used for creating maps that convey LST, emissivity, elevation and reflectance data.	15 - 90	14	16	2000 - Now	[40] [73] [20]
AVHRR	Sensor aboard NOAA series and captures multispectral earth data for applications such as weather forecasting, climate change monitoring and vegetation analysis.	1100 - 4000	5	0.5	1978 - Now	[40] [74] [20]
IKONOS	First commercial Earth observation satellite and is used for national security, military applications and is often used by government.	1 - 4	5	3	1999 - Now	[40] [20]
Landsat 1	First satellite in the Landsat series pioneered by NASA and the US Geological Survey and provides multispectral earth land surface imagery.	60	4	18	1972 - 1978	N/A
Landsat 2	Improved resolution from Landsat 1 as a result of enhanced sensors. Provides more detailed Earth observation data.	80	4	18	1975 - 1982	N/A
Landsat 3	Used for continued data collection. Has further applications in forestry and agriculture.	80	4	18	1978 - 1983	N/A
Landsat 4	Improved spatial and spectral resolutions with the introduction of the Thematic Mapper (TM) sensor.	30 - 120	7	16	1982 - 1993	N/A
Landsat 5	Used for continued data collection. One of the longest serving satellites.	30 - 120	7	16	1984 - 2013	[74] [20] [75]
Landsat 6	Failed to reach orbit.	15 - 30	8	16	1993	N/A

Satellite	Description	Spatial Resolution (m)	Spectral Resolution (# of bands)	Temporal Resolution (days)	Operational	Citation
Landsat 7	Improved capabilities with the introduction of the Enhanced Thematic Mapper Plus (ETM+).	15 - 30	8	16	1999 - Now	[73, 20, 75]
Landsat 8	Improved capabilities with the use of the Operational Land Imager (OLI) and Thermal Infrared Sensor (TIRS).	15 - 30	9 or 11	16	2013 - Now	[30, 75]
Landsat 9	Improved capabilities with the use of the Operational Land Imager 2 (OLI-2) and Thermal Infrared Sensor 2 (TIRS-2).	15 - 100	11	16	2021 - Now	N/A
MERIS	The Medium Resolution Imaging Spectrometer was on-board the ESA's Envisat satellite, operating in the visible and NIR electomagnetic regions it was used for water colour, water quality and land cover observance.	300	15	3	2002 - 2012	N/A
MODIS	Instrument aboard NASA's Terra & Aqua satellite to observe Earth's surface and atmosphere. Monitors land & cloud cover and sea surface temperature.	250 - 1000	36	1 - 2	1999 - Now	[49, 46, 34]
QuickBird	Commercial satellite launched by DigitalGlobe for various applications: urban planning, disaster response and environmental monitoring.	0.65 - 2.24	5	1 - 3.5	2001 - 2015	[20]
RapidEye	A constellation of satellites launched by Planet Labs Inc (private Earth imaging company) for various applications: environmental monitoring, resource management and land cover classification.	5	5	1 - 5.5	2008 - 2020	N/A
Senitel 1	Synthetic Aperture Radar (SAR) satellite launched by the ESA, used for radar imaging of the Earth.	5	1	6 - 12	2014 - Now	[33]
Senitel 2	Multispectral Imager (MSI) satellite, launched by the ESA and designed for land observation - its bands capture data used commonly for vegetation and agriculture analysis.	10 - 60	13	5	2015 - Now	[33, 19, 30]
Senitel 3	ESA launched satellite that is dedicated to ocean monitoring and houses: an Ocean and Land Colour Instrument (OLCI) and a Sea and Land Surface Temperature Radiometer (SLSTR).	300 - 1000	21	1 - 27	2016 - Now	N/A
Senitel 4	ESA launched satellite that is dedicated to air quality observation through the monitoring of atmosphere composition by means of an Ultra-Violet Visible Near-Infrared Spectrometer (UVN).	0.5nm - 0.12nm	3	0.1	2019 - Now	N/A
Senitel 5	Also dedicated to air quality monitoring, instead using an Ultra-Violet Visible Near-Infrared Shortwave-Infrared Spectrometer (UVNS).	5.5 - 7	7	16	2017 - Now	N/A
Senitel 5P	ESA launched satellite that is dedicated to atmospheric observance and houses a Tropospheric Monitoring Instrument (TROPOMI).	8 - 50	7	1	2018 - Now	N/A

Satellite	Description	Spatial Resolution (m)	Spectral Resolution (# of bands)	Temporal Resolution (days)	Operational	Citation
SPOT 5	Satellite Pour l'Observation de la Terre (SPOT) was launched by the French National Centre for Space Studies (CNES) in collaboration with the company Airbus Defence and Space. It housed High-Resolution Geometric (HRG) and High-Resolution Stereoscopic (HRS) instruments that are used for various applications: environmental monitoring, urban planning and disaster management.	2.5 - 20	4	2 - 3	2002 - 2015	[49, 34, 20]
SPOT 6	Improvement to SPOT 5 with an improved version of the High-Resolution Geometric (HRG) instrument. Used for: resource management, change detection and land mapping.	1.5 - 6	4	1	2012 - Now	N/A
WorldView 1	A commercial satellite launched by DigitalGlobe. Used for urban planning, agriculture, mapping and disaster response.	0.5 - 2	5	1.1	2007 - Now	[40, 20]
WorldView 2	Improved capabilities with the introduction of the Multispectral sensors. Used for land cover mapping and environmental monitoring.	0.46 - 1.8	9	1.1	2009 - Now	N/A
WorldView 3	Improved capabilities with the additional or more multispectral band: SWIR. Used for: agriculture, environmental monitoring, and defense.	0.31 - 3.7	17	1	2014 - Now	N/A
WorldView 4	Was intended as a data collection continuation but experienced a control failure and declared lost.	0.31 - 1.24	13	1	2016 - 2019	N/A

Table A.1: Satellite Options

A.2 Vegetation Indices (VIs) Table

Vegetation Indice	Acronym	Measurement Method	Description	Unit
Atmospherically Resistant Vegetation Index	ARVI	Derived from multispectral bands: $\frac{NIR-(2\cdot Red)+Blue}{NIR+(2\cdot Red)+Blue}$	Provides an indication of vegetation with the removal of atmospheric disturbance (i.e. atmospheric scattering effects).	non-dimensional
Digital Elevation Model	DEM	Calculated by triangulating elevation values from stereo pairs of images captured by a multispectral satellite, utilizing parallax and geometric principles.	Provides a topographical indication of the height of vegetation at different locations.	m

Vegetation Indice	**Acronym**	**Measurement Method**	**Description**	**Unit**
Difference Vegetation Index	DVI	Derived from multispectral : $NIR - Red$	Indicates variations in canopy structure and leaf area highlighting changes in vegetation conditions.	non-dimensional
Enhanced Vegetation Index	EVI	Derived from multispectral bands: $2.5 \cdot \frac{NIR-Red}{NIR+(6 \cdot Red)-(7.5 \cdot Blue)+1}$	Index with improve sensitivity in high biomass regions and reduce atmospheric influences and background soil brightness.	non-dimensional
Generalised Difference Vegetation Index	GDVI	Derived from multispectral bands: $NIR - Green$	Provides an indication of plant density and health with less sensitivity to soil background reflectance variation.	non-dimensional
Land Surface Temperature	LST	Calculated by calibrating thermal infrared bands, correcting for atmospheric effects, converting radiance to brightness temperature, estimating land surface emissivity, and applying the Stefan-Boltzmann Law: $LST = \frac{K_2}{ln(\frac{K_2}{emissivity \cdot Radiance+1})}$	Indicates temperature, amount of radiation, water stress and other thermal characteristics.	celcius
Land Use and Land Cover	LULC	Determined using classification algorithms that assign pixels to specific categories based on their spectral signatures, with supervised algorithms like Maximum Likelihood Classifier or unsupervised algorithms like K-Means Clustering.	Indicates how land is being used (i.e. activities, modifications or interventions) and the nature of its coverage (i.e. natural or artificial).	categorical classes
Lead Area Index	LAI	Calculated by establishing an empirical relationship between NDVI and LAI based on ground truth data.	Catgeorizes plant characteristics (i.e. canopies, shape, species).	non-dimensional
Normalised Difference Greeness Index	NDGI	Derived from multispectral bands: $\frac{Green-Red}{Green+Red}$	Indicates temporal change in vegetation.	non-dimensional
Normalised Difference Red Edge Index	NDRE	Derived from multispectral bands: $\frac{NIR-RedEdge}{NIR+RedEdge}$	Provides a more advanced indication of crop health variations.	non-dimensional
Normalised Difference Vegetation Index	NDVI	Derived from multispectral bands: $\frac{NIR-Red}{NIR+Red}$	Indicates plant health and vegetation cover as it focuses on reflectance values.	non-dimensional
Normalised Difference Water Index	NDWI	Derived from multispectral bands: $\frac{Green-NIR}{Green+NIR}$	Used to detect the presence of bodies of water, also by using relectance values.	non-dimensional

Vegetation Indice	**Acronym**	**Measurement Method**	**Description**	**Unit**
Photosynthetically Active Radiation	PAR	Measured by a silicon photovoltaic detector	Indicates the amount of light available for photosynthesis (i.e. light in the 400 to 700 nm wavelength range).	µmol/(m^2·s)
Refractive Index	RI	Derived from multispectral bands: $\frac{Red-Green}{Red+Green}$	Indicates the behaviour of light as it passes through differing materials.	non-dimensional
Soil Adjusted Vegetation Index	SAVI	Derived from multispectral bands: $\frac{NIR-Red}{NIR+Red+L} \cdot (1+L)$	Provides an indication of plant density and health with less sensitivity to soil brightness in regions of sparse vegetation.	non-dimensional
Simple Ratio	SR	Derived from multispectral bands: $\frac{NIR}{Red}$	Indicates plant health and density.	non-dimensional
Surface Albedo	A	Calculated using the formula: $\frac{\text{Reflectance in Short Wave Bands}}{\text{Incoming Solar Radiation}}$	Provides a measurement of reflectivity (i.e. how much incoming solar radiation is reflected back into space).	non-dimensional
Visible Atmospherically Resistant Index	VARI	Derived from multispectral bands: $\frac{Green-Red}{Green+Red-Blue}$	Emphasises vegetation in the visible portion of the spectrum and minimises atmposheric disturbances.	non-dimensional

Table A.2: Vegetation Indices (VIs) [19, 2]

Appendix B

AmeriFlux

B.1 AmeriFlux Portal

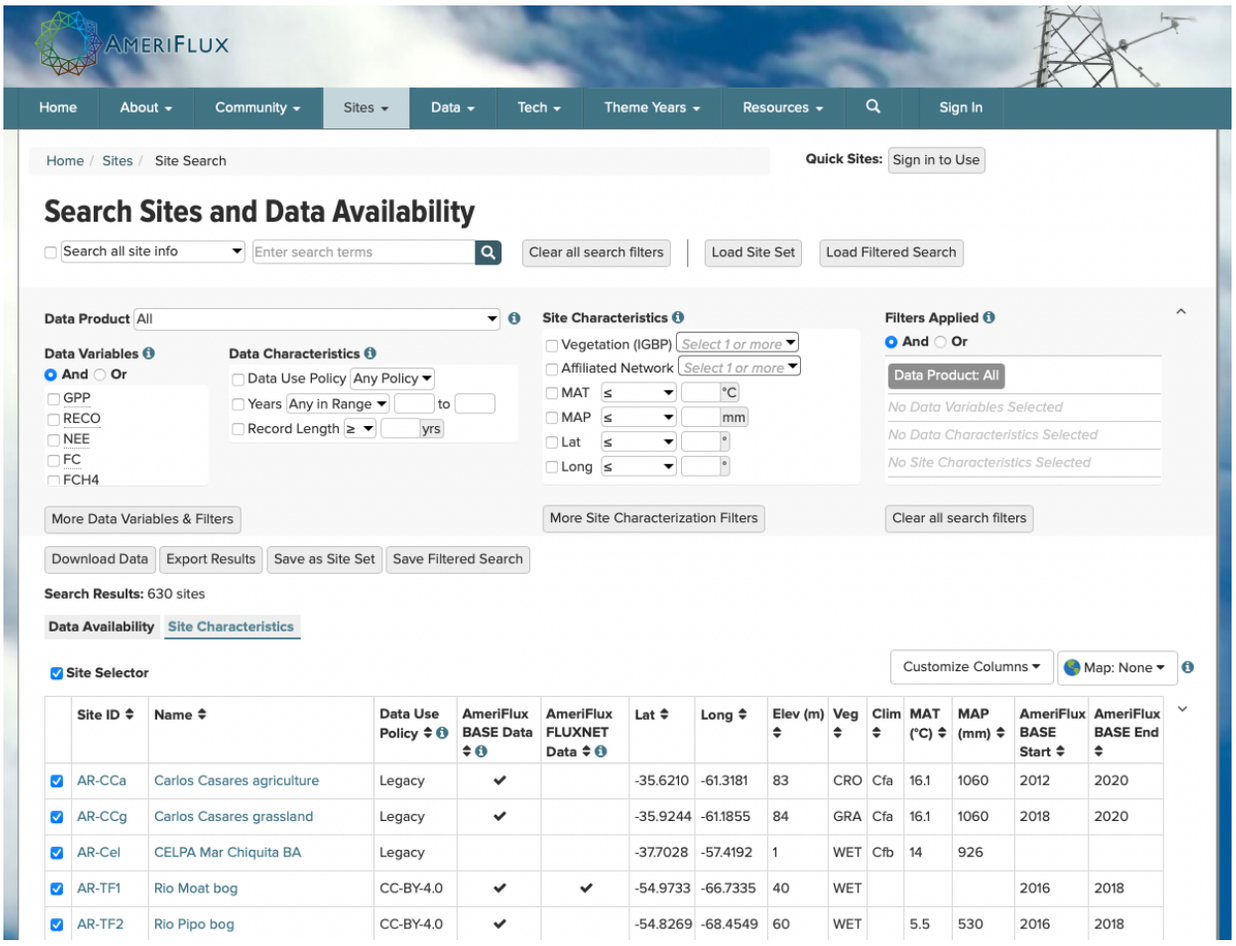


Figure B.1: Site and Data Availability Search Website Page [15]

Figure B.1 above shows the site and data availability search website page. This page shows all available datasets (i.e. 630) before filtering has been applied. The filters include those on the data variables offered, data characteristics and site characteristics. Once the desired filters have been applied, these datasets can be accessed via the 'Download Data' button which will direct you to the next webpage as shown by Figure B.2 below. The AmeriFlux site overview dataset that summarises the selected flux tower sites' characteristics can be download via the 'Export Results' button.

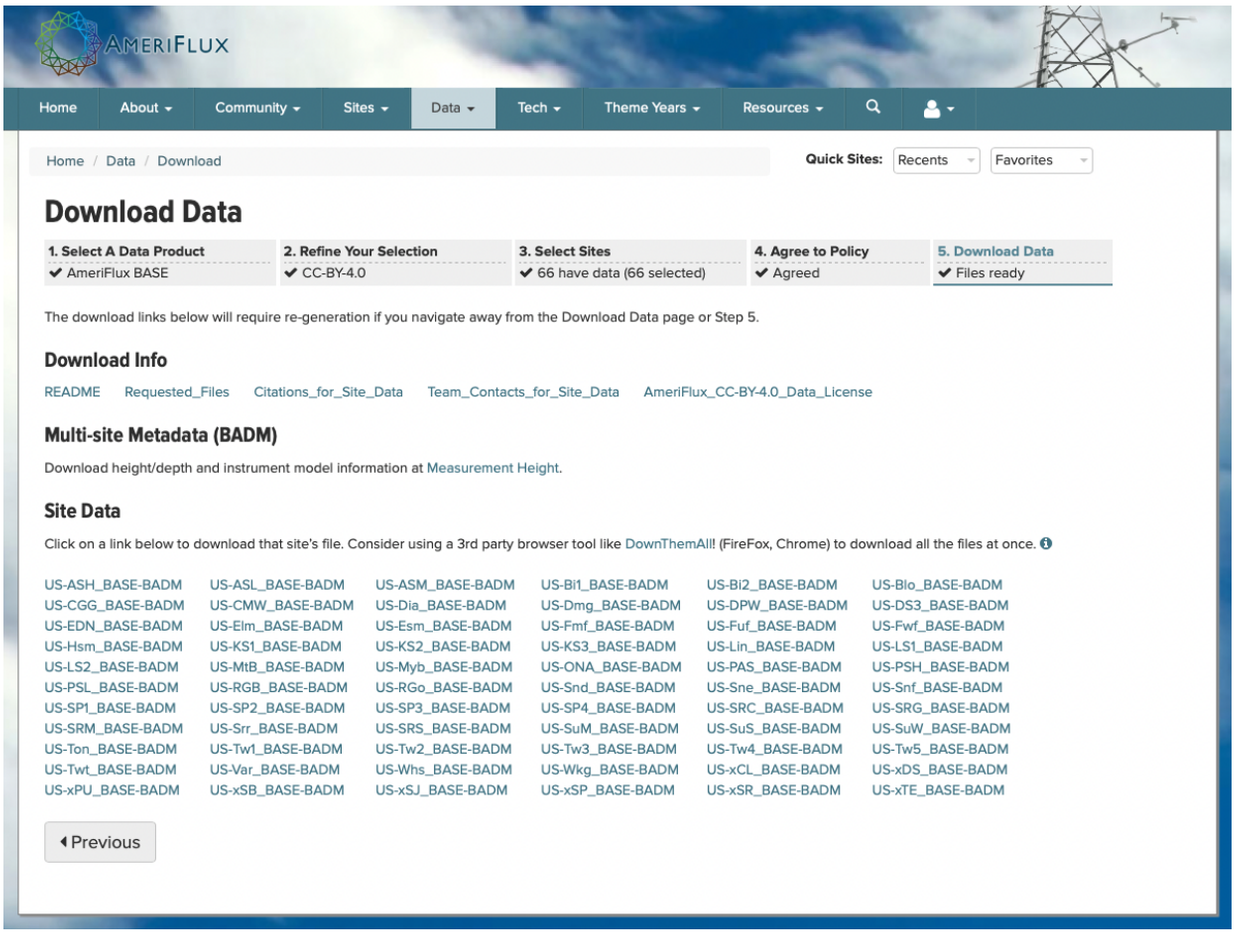


Figure B.2: Data Download Website Page [15]

Figure B.2 above shows the data download webpage. The filters have now been applied adn can be seen in the top bar: AmeriFlux product, data use policy and the number of resulting or selected sites. Each individual AmeriFlux flux tower dataset can be found on the bottom of the webpage under the **Site Data** title, each is a hyperlink that once pressed will download to the local Personal Computer (PC).

B.2 AmeriFlux Data

B.2.1 AmeriFlux Individual Flux Tower Data

Table B.1 below provides an exhaustive list of all possible data variables provided by an individual AmeriFlux flux tower dataset. The subset of these variables that an individual tower does offer is dependent on the type and purpose of the tower.

Type	Variable	Description	Units
Timekeeping	TIMESTAMP_START	ISO timestamp start of averaging period (up to a 12-digit integer as specified by the data's temporal resolution)	YYYYMMDDHHMM
	TIMESTAMP_END	ISO timestamp end of averaging period (up to a 12-digit integer as specified by the data's temporal resolution)	YYYYMMDDHHMM
	TIMESTAMP	ISO timestamp (up to a 12-digit integer as specified by the data's temporal resolution)	YYYYMMDDHHMM
Aquatic	COND_WATER	Conductivity (i.e., electrical conductivity) of water	µS cm-1
	DO	Dissolved oxygen in water	µmol L-1
	PCH4	Dissolved methane (CH4) in water	nmolCH4 mol-1
	PCO2	Dissolved carbon dioxide (CO2) in water	µmolCO2 mol-1
	PN2O	Dissolved nitrous oxide (N2O) in water	nmolN2O mol-1
	PPFD_UW_IN	Photosynthetic photon flux density, underwater, incoming	µmolPhotons m-2 s-1
	TW	Water temperature	deg C
Biological	DBH	Diameter of tree measured at breast height (1.3m) with continuous dendrometers	cm
	LEAF_WET	Leaf wetness, range 0-100	%
	SAP_DT	Difference of probes temperature for sapflow measurements	deg C
	SAP_FLOW	Sap flow	mmolH2O m-2 s-1
	T_BOLE	Bole temperature	deg C
	T_CANOPY	Temperature of the canopy and/or surface underneath the sensor	deg C
Footprint	FETCH_70	Distance at which cross-wind integrated footprint cumulative probability is 70%	m
	FETCH_80	Distance at which cross-wind integrated footprint cumulative probability is 80%	m
	FETCH_90	Distance at which cross-wind integrated footprint cumulative probability is 90%	m
	FETCH_FILTER	Footprint quality flag (i.e., 0, 1): 0 and 1 indicate data measured when wind coming from direction that should be discarded and kept, respectively	nondimensional
	FETCH_MAX	Distance at which footprint contribution is maximum	m
Gases	CH4	Methane (CH4) mole fraction in wet air	nmolCH4 mol-1
	CH4_MIXING_RATIO	Methane (CH4) in mole fraction of dry air	nmolCH4 mol-1
	CO	Carbon Monoxide (CO) mole fraction in wet air	nmolCO mol-1
	CO2	Carbon Dioxide (CO2) mole fraction in wet air	µmolCO2 mol-1
	CO2_MIXING_RATIO	Carbon Dioxide (CO2) in mole fraction of dry air	µmolCO2 mol-1
	CO2_SIGMA	Standard deviation of carbon dioxide mole fraction in wet air	µmolCO2 mol-1
	CO2C13	Stable isotopic composition of CO2 - C13 (i.e., d13C of CO2)	‰ (permil)
	FC	Carbon Dioxide (CO2) turbulent flux (no storage correction)	µmolCO2 m-2 s-1
	FCH4	Methane (CH4) turbulent flux (no storage correction)	nmolCH4 m-2 s-1
	FN2O	Nitrous oxide (N2O) turbulent flux (no storage correction)	nmolN2O m-2 s-1
	FNO	Nitric oxide (NO) turbulent flux (no storage correction)	nmolNO m-2 s-1
	FNO2	Nitrogen dioxide (NO2) turbulent flux (no storage correction)	nmolNO2 m-2 s-1
	FO3	Ozone (O3) turbulent flux (no storage correction)	nmolO3 m-2 s-1
	H2O	Water (H2O) vapor in mole fraction of wet air	mmolH2O mol-1
	H2O_MIXING_RATIO	Water (H2O) vapor in mole fraction of dry air	mmolH2O mol-1
	H2O_SIGMA	Standard deviation of water vapor mole fraction	mmolH2O mol-1
	N2O	Nitrous Oxide (N2O) mole fraction in wet air	nmolN2O mol-1

Type	Variable	Description	Units
Gases	N2O_MIXING_RATIO	Nitrous Oxide (N2O) in mole fraction of dry air	nmolN2O mol-1
	NO	Nitric oxide (NO) mole fraction in wet air	nmolNO mol-1
	NO2	Nitrogen dioxide (NO2) mole fraction in wet air	nmolNO2 mol-1
	O3	Ozone (O3) mole fraction in wet air	nmolO3 mol-1
	SC	Carbon Dioxide (CO2) storage flux	µmolCO2 m-2 s-1
	SCH4	Methane (CH4) storage flux	nmolCH4 m-2 s-1
	SN2O	Nitrous oxide (N2O) storage flux	nmolN2O m-2 s-1
	SNO	Nitric oxide (NO) storage flux	nmolNO m-2 s-1
	SNO2	Nitrogen dioxide (NO2) storage flux	nmolNO2 m-2 s-1
	SO2	Sulfur Dioxide (SO2) mole fraction in wet air	nmolSO2 mol-1
	SO3	Ozone (O3) storage flux	nmolO3 m-2 s-1
Heat	FH2O	Water vapor (H2O) turbulent flux (no storage correction)	mmolH2O m-2 s-1
	G	Soil heat flux	W m-2
	H	Sensible heat turbulent flux (no storage correction)	W m-2
	LE	Latent heat turbulent flux (no storage correction)	W m-2
	SB	Heat storage flux in biomass	W m-2
	SG	Heat storage flux in the soil above the soil heat fluxes measurement	W m-2
	SH	Sensible heat (H) storage flux	W m-2
	SLE	Latent heat (LE) storage flux	W m-2
Meteorological Atmosphere	PA	Atmospheric pressure	kPa
	PBLH	Planetary boundary layer height	m
	RH	Relative humidity, range 0-100	%
	T_SONIC	Sonic temperature	deg C
	T_SONIC_SIGMA	Standard deviation of sonic temperature	deg C
	TA	Air temperature	deg C
	VPD	Vapor Pressure Deficit	hPa
Meteorological Precipitation	D_SNOW	Snow depth	cm
	P	Precipitation	mm
	P_RAIN	Rainfall	mm
	P_SNOW	Snowfall	mm
	RUNOFF	Run off	mm
	STEMFLOW	Excess precipitation that drains from outlying branches and leaves and is channeled through the stems to the ground	mm
	THROUGHFALL	Excess precipitation that passes directly through a canopy or drips from wet leaves to the ground	mm
Meteorological Radiation	ALB	Albedo, range 0-100	%
	APAR	Absorbed PAR	µmolPhoton m-2 s-1
	EVI	Enhanced Vegetation Index	nondimensional
	FAPAR	Fraction of absorbed PAR, range 0-100	%
	FIPAR	Fraction of intercepted PAR, range 0-100	%
	LW_BC_IN	Longwave radiation, below canopy incoming	W m-2
	LW_BC_OUT	Longwave radiation, below canopy outgoing	W m-2
	LW_IN	Longwave radiation, incoming	W m-2
	LW_OUT	Longwave radiation, outgoing	W m-2
	MCRI	Carotenoid Reflectance Index (Gitelson et al., 2002)	nondimensional
	MTCI	Meris Terrestrial Chlorophyll Index (Dash and Curran, 2004)	nondimensional
	NDVI	Normalized Difference Vegetation Index	nondimensional
	NETRAD	Net radiation	W m-2
	NIRV	Near Infrared Vegetation Index (Badgley et al., 2017)	W m-2 sr-1 nm-1
	PPFD_BC_IN	Photosynthetic photon flux density, below canopy incoming	µmolPhoton m-2 s-1
	PPFD_BC_OUT	Photosynthetic photon flux density, below canopy outgoing	µmolPhoton m-2 s-1
	PPFD_DIF	Photosynthetic photon flux density, diffuse incoming	µmolPhoton m-2 s-1
	PPFD_DIR	Photosynthetic photon flux density, direct incoming	µmolPhoton m-2 s-1
	PPFD_IN	Photosynthetic photon flux density, incoming	µmolPhoton m-2 s-1
	PPFD_OUT	Photosynthetic photon flux density, outgoing	µmolPhoton m-2 s-1
	PRI	Photochemical Reflectance Index	nondimensional
	R_UVA	UVA radiation, incoming	W m-2
	R_UVB	UVB radiation, incoming	W m-2

Type	Variable	Description	Units
Meteorological Radiation	REDCI	Red Edge Chlorophyll Index	nondimensional
	REP	Red Edge Position (Dash and Curran, 2004)	nm
	SPEC_NIR_IN	Radiation (near infra-red band), incoming (hemispherical)	W m-2 nm-1
	SPEC_NIR_OUT	Radiation (near infra-red band), outgoing	W m-2 sr-1 nm-1
	SPEC_NIR_REFL	Reflectance (near infra-red band)	nondimensional
	SPEC_PRI_REF_IN	Radiation for PRI reference band (e.g., 570 nm), incoming (hemispherical)	W m-2 nm-1
	SPEC_PRI_REF_OUT	Radiation for PRI reference band (e.g., 570 nm), outgoing	W m-2 sr-1 nm-1
	SPEC_PRI_REF_REFL	Reflectance for PRI reference band (e.g., 570 nm)	nondimensional
	SPEC_PRI_TGT_IN	Radiation for PRI target band (e.g., 531 nm), incoming (hemispherical)	W m-2 nm-1
	SPEC_PRI_TGT_OUT	Radiation for PRI target band (e.g., 531 nm), outgoing	W m-2 sr-1 nm-1
	SPEC_PRI_TGT_REFL	Reflectance for PRI target band (e.g., 531 nm)	nondimensional
	SPEC_RED_IN	Radiation (red band), incoming (hemispherical)	W m-2 nm-1
	SPEC_RED_OUT	Radiation (red band), outgoing	W m-2 sr-1 nm-1
	SPEC_RED_REFL	Reflectance (red band)	nondimensional
	SR	Simple Ratio	nondimensional
	SW_BC_IN	Shortwave radiation, below canopy incoming	W m-2
	SW_BC_OUT	Shortwave radiation, below canopy outgoing	W m-2
	SW_DIF	Shortwave radiation, diffuse incoming	W m-2
	SW_DIR	Shortwave radiation, direct incoming	W m-2
	SW_IN	Shortwave radiation, incoming	W m-2
	SW_OUT	Shortwave radiation, outgoing	W m-2
	TCARI	Transformed Chlorophyll Absorption in Reflectance Index	nondimensional
Meteorological Soil	SWC	Soil water content (volumetric), range 0-100	%
	SWP	Soil water potential	kPa
	TS	Soil temperature	deg C
	TSN	Snow temperature	deg C
	WTD	Water table depth	m
Meteorological Wind	MO_LENGTH	Monin-Obukhov length	m
	TAU	Momentum flux	kg m-1 s-2
	U_SIGMA	Standard deviation of velocity fluctuations (towards main-wind direction after coordinates rotation)	m s-1
	USTAR	Friction velocity	m s-1
	V_SIGMA	Standard deviation of lateral velocity fluctuations (cross main-wind direction after coordinates rotation)	m s-1
	W_SIGMA	Standard deviation of vertical velocity fluctuations	m s-1
	WD	Wind direction	Decimal degrees
	WD_SIGMA	Standard deviation of wind direction (Yamartino, 1984)	decimal degree
	WS	Wind speed	m s-1
	WS_MAX	maximum WS in the averaging period	m s-1
	ZL	Monin-Obukhov Stability parameter	nondimensional
Products	GPP	Gross Primary Productivity	µmolCO2 m-2 s-1
	NEE	Net Ecosystem Exchange	µmolCO2 m-2 s-1
	RECO	Ecosystem Respiration	µmolCO2 m-2 s-1
Quality Control Flag	FC_SSITC_TEST	Results of the quality flagging for FC according to Foken et al 2004, based on a combination of Steady State and Integral Turbulence Characteristics tests by Foken and Wichura (1996) (i.e., 0, 1, 2)	nondimensional
	FCH4_SSITC_TEST	Results of the quality flagging for FCH4 according to Foken et al 2004, based on a combination of Steady State and Integral Turbulence Characteristics tests by Foken and Wichura (1996) (i.e., 0, 1, 2)	nondimensional
	FN2O_SSITC_TEST	Results of the quality flagging for FN2O according to Foken et al 2004, based on a combination of Steady State and Integral Turbulence Characteristics tests by Foken and Wichura (1996) (i.e., 0, 1, 2)	nondimensional
	FNO_SSITC_TEST	Results of the quality flagging for FNO according to Foken et al 2004, based on a combination of Steady State and Integral Turbulence Characteristics tests by Foken and Wichura (1996) (i.e., 0, 1, 2)	nondimensional

Type	Variable	Description	Units
Quality Control Flag	FNO2_SSITC_TEST	Results of the quality flagging for FNO2 according to Foken et al 2004, based on a combination of Steady State and Integral Turbulence Characteristics tests by Foken and Wichura (1996) (i.e., 0, 1, 2)	nondimensional
	FO3_SSITC_TEST	Results of the quality flagging for FO3 according to Foken et al 2004, based on a combination of Steady State and Integral Turbulence Characteristics tests by Foken and Wichura (1996) (i.e., 0, 1, 2)	nondimensional
	H_SSITC_TEST	Results of the quality flagging for H according to Foken et al 2004, based on a combination of Steady State and Integral Turbulence Characteristics tests by Foken and Wichura (1996) (i.e., 0, 1, 2)	nondimensional
	LE_SSITC_TEST	Results of the quality flagging for LE according to Foken et al 2004, based on a combination of Steady State and Integral Turbulence Characteristics tests by Foken and Wichura (1996) (i.e., 0, 1, 2)	nondimensional
	TAU_SSITC_TEST	Results of the quality flagging for TAU according to Foken et al 2004, based on a combination of Steady State and Integral Turbulence Characteristics tests by Foken and Wichura (1996) (i.e., 0, 1, 2)	nondimensional

Table B.1: AmeriFlux Individual Flux Tower Data Variables

B.2.2 AmeriFlux Site Overview Data

Table B.2 below provides a list of all data variables provided by an AmeriFlux site overview dataset.

Variable	Description
Site ID	Unique identification code for a flux tower site (e.g. AR-CCa)
Name	Name of flux tower site (e.g. Carlos Casares Agriculture)
Principal Investigator	Person responsible for site implementation and research, including their email adress
Data Use Policy	The data use policy that governs the flux tower data collected (CC-BY-4.0 or Legacy)
AmeriFlux BASE Data	Does the flux tower offer an AmeriFlux BASE Data product (Yes or No)
AmeriFlux FLUXNET Data	Does the flux tower offer an AmeriFlux FLUXNET Data product (Yes or No)
Vegetation Abbreviation (IGBP)	Abbreviation of a flux tower's surrounding vegetation
Vegetation Description (IGBP)	Description of a flux tower's surrounding vegetation
Climate Class Abbreviation (Koeppen)	Abbreviation of a flux tower environment's climate classification
Climate Class Description (Koeppen)	Description of a flux tower environment's climate classification
Mean Average Precipitation (mm)	Average precipitation or rainfall of the region in which a flux tower is situated
Mean Average Temperature (°C)	Average temperature of the region in which a flux tower is situated
Country	Country in which a flux tower is situated
Latitude (degrees)	Latitude co-ordinate of a flux tower's location
Longitude (degrees)	Longitude co-ordinate of a flux tower's location
Elevation (m)	Elevation of the flux tower
Years of AmeriFlux BASE Data	List of years that an AmeriFlux BASE Data product has been operational
AmeriFlux BASE DOI	Digital Object Identifier (DOI) and an AmeriFlux BASE product
Years of AmeriFlux FLUXNET Data	List of years that an AmeriFlux FLUXNET Data product has been operational
AmeriFlux FLUXNET DOI	Digital Object Identifier (DOI) and an AmeriFlux FLUXNET product
Site Start	Start year of a flux tower's operation
Site End	End year of a flux tower's operation
BASE variables available	List of variables recorded by an AmeriFlux BASE Data product
FLUXNET variables available	List of variables recorded by an AmeriFlux FLUXNET Data product

Table B.2: AmeriFlux Site Overview Data Variables

Appendix C

Code

C.1 GitHub Repository

Git, hosted on GitHub, was used extensively throughout this project for code storage, version control and facilitates future collaborations. The GitHub Repository is accessible here or via the QR code in Figure C.1. It has been made public in the spirit of sharing scientific data to facilitate progress.

Figure C.1: Git Repository QR Code

Appendix D

Results

D.1 Machine Learning Extended Results

The five ML models were tested on eight combinations with regards to dataset size, date inclusion or data normalisation. Each of these eight combinations were 'run' three times to ensure accurate performance summary. This results in 24 tests per ML model. The summarised results are in Chapter 7 above but the detailed and all inclusive results are shown in Tables D.1, D.2, D.3, D.4 and D.5 below.

Neural Network (NN)							
Dataset Size	**Date Included**	**Normalized**	**MAE**	**Loss**	**Normalised MAE**	**Normalised Loss**	R^2
1st Attempt							
Smaller	No	Yes	1641.28173828125	4317814.0	48.137606546542 81%	126638.36233918858%	0.4086114764213562
Smaller	No	No	2194.22705078125	6983893.0	64.35509271851105%	204832.53152454522%	0.043452441692352295
Smaller	Yes	Yes	4539.16455078125	27799226.0	132.31392191778 29%	810330.750777856%	-2.8631434440612793
Smaller	Yes	No	2577.775634765625	8546276.0	75.1406123845496%	249118.8153020797%	-0.18764054775238037
Larger	No	Yes	1198.623291015625	2501998.75	53.05346491847871%	110743.47036651419%	0.3777409791946411
Larger	No	No	1305.438720703125	2845316.25	57.781329539627755%	125939.38978396222%	0.2923562526702881
Larger	Yes	Yes	4591.021484375	25096194.0	203.20779604982332%	1110807.7558197386%	-5.241543292999268
Larger	Yes	No	3332.158203125	13354848.0	147.4879886428043%	591112.2912180916%	-2.3214144706726074
2nd Attempt							
Smaller	No	Yes	1669.3128662109375	4424285.0	48.959739258960695%	129761.07977829452%	0.3940286636352539
Smaller	No	No	2199.21875	6969453.5	64.50149564701893%	204409.03143098013%	0.04543018341064453
Smaller	Yes	Yes	2883.720703125	14507573.0	84.05872747669046%	422887.04444701277%	-1.0160572528839111
Smaller	Yes	No	2322.3115234375	7488373.0	67.69398688751895%	218281.5778825866%	-0.04062819480895996
Larger	No	Yes	1209.0732421875	2546531.5	53.516000664323 53%	112714.57897716573%	0.36666548252105713
Larger	No	No	1353.3648681640625	3001684.75	59.902636710992866%	132860.57243683375%	0.2534666657447815
Larger	Yes	Yes	2675.567138671875	8794313.0	118.4259545034257%	389253.8879603159%	-1.187187671661377
Larger	Yes	No	1665.5126953125	5762831.0	73.71892404758117%	255074.42962380752%	-0.43324363231658936
3rd Attempt							
Smaller	No	Yes	1625.7470703125	4256189.5	47.681985968387856%	124830.96031585656%	0.41705185174942017
Smaller	No	No	2192.582275390625	6818554.0	64.30685264566758%	199983.25821383775%	0.06609803438186646
Smaller	Yes	Yes	2403.0380859375	9293380.0	70.04711772642578%	270896.4484357914%	-0.2914625406265259
Smaller	Yes	No	2338.593994140625	7220006.5	68.16861113458906%	210458.85550072513%	-0.0033344030380249
Larger	No	Yes	1233.84765625	2532744.25	54.61256579633178%	112104.32770833088%	0.3700944185256958
Larger	No	No	1331.5709228515625	2854962.75	58.93799308881508%	126366.36317349353%	0.2899571657180786
Larger	Yes	Yes	2598.6982421875	10769665.0	115.02358335518909%	476686.9195217564%	-1.678467035293579
Larger	Yes	No	2214.13916015625	8703359.0	98.00222900595065%	385228.0819507342%	-1.164566993713379

Table D.1: Neural Network (NN) Results

Reccurent Neural Network (RNN)							
Dataset Size	**Date Included**	**Normalized**	**MAE**	**Loss**	**Normalised MAE**	**Normalised Loss**	R^2
1st Attempt							
Smaller	No	Yes	1935.8814697265625	5512736.0	56.77800364002557%	161684.56053185454%	0.24494904279708862
Smaller	No	No	2197.226318359375	6811084.5	64.4430590677588%	199764.18318015343%	0.06712090969085693
Smaller	Yes	Yes	2308.3681640625	7542367.5	67.28754633155853%	219855.48514615122%	-0.048131585121154785
Smaller	Yes	No	2328.4892578125	7082316.5	67.87406413622364%	206445.2746523014%	0.015799760818481445
Larger	No	Yes	1162.2869873046875	2226995.75	51.44514741902317%	98571.28739431946%	0.4461355209350586
Larger	No	No	1036.5958251953125	1901509.375	45.88180511663701%	84164.60924369424%	0.527085542678833
Larger	Yes	Yes	1653.6815185546875	4075989.75	73.19525249391641%	180411.46107420741%	-0.013718128204345703
Larger	Yes	No	1516.8084716796875	3834282.75	67.13697760046328%	169713.02567655622%	0.04639559984207153
2nd Attempt							
Smaller	No	Yes	1937.141845703125	5562730.5	56.81496955601998%	163150.86306502682%	0.23810148239135742
Smaller	No	No	2215.397216796875	7206339.5	64.97599837926148%	211356.72651490013%	0.012984871864318848
Smaller	Yes	Yes	2362.577880859375	7248145.5	68.86772703556328%	211279.09045964584%	-0.007244706153869629
Smaller	Yes	No	2346.1884765625	7146932.0	68.38998573842468%	208328.77769036748%	0.006820499897003174
Larger	No	Yes	1122.1708984375	2082334.75	49.669529066422555%	92168.30211437464%	0.48211342096328735
Larger	No	No	1175.921875	2175256.25	52.0486548274248%	96281.19361028848%	0.45900338888168335
Larger	Yes	Yes	1709.1314697265625	4186175.75	185288.5138808104%	185288.5138808104%	-0.0411219596862793
Larger	Yes	No	1446.701416015625	4134494.5	64.03389905519357%	183000.99835831887%	-0.02826857566833496
3rd Attempt							
Smaller	No	Yes	1850.6883544921875	6424478.5	54.27935117471441%	188425.3087611756%	0.12007230520248413
Smaller	No	No	1915.9560546875	5323498.5	56.193605625318234%	156134.36149753715%	0.27086782455444336
Smaller	Yes	Yes	2344.47265625	7209021.0	68.33997060627576%	210138.6347700231%	-0.0018078088760375977
Smaller	Yes	No	2356.884521484375	7191592.5	68.70176902735139%	209630.60445687943%	0.0006142258644104004
Larger	No	Yes	1123.6243896484375	2136103.75	49.73386349538517%	94548.22562877972%	0.4687408208847046
Larger	No	No	1181.760498046875	2443626.0	52.307083964678704%	108159.77566649207%	0.39225852489471436
Larger	Yes	Yes	1531.1500244140625	4262705.5	67.77176341730235%	188675.873727126%	-0.06015515327453613
Larger	Yes	No	1471.9168701171875	3938887.75	65.14998550170941%	174343.05225738054%	0.020379841327667236

Table D.2: Recurrent Neural Network (RNN) Results

Long Short-Term Memory (LSTM)							
Dataset Size	**Date Included**	**Normalized**	**MAE**	**Loss**	**Normalised MAE**	**Normalised Loss**	R^2
1st Attempt							
Smaller	No	Yes	1983.490966796875	5687969.0	58.17435576190408%	166824.01770805137%	0.22094827890396118
Smaller	No	No	2075.360107421875	6082140.0	60.868811224380835%	178384.76810665592%	0.16696065664291382
Smaller	Yes	Yes	2359.7333984375	7240822.5	68.78481208043135%	211065.6294054443%	-0.006227135658264160
Smaller	Yes	No	2306.180419921875	7338249.5	67.223774903102850%	213905.56797265605%	-0.019766092300411504
Larger	No	Yes	1066.5833740234375	1958038.625	47.20911402317098%	86666.70694546818%	0.5130264759063721
Larger	No	No	1421.308837890625	3121130.5	62.909972745032405%	138147.48030420623%	0.2237599492073059
Larger	Yes	Yes	1739.51171875	4265615.5	76.99426887307389%	188804.67614909622%	-0.06087899208068848
Larger	Yes	No	1618.7723388671875	4033558.5	71.65010235895252%	178533.3690579787%	-0.0031652450561523438
2nd Attempt							
Smaller	No	Yes	1807.422607421875	4846433.0	53.01039809930034%	142142.37535627998%	0.33620905876159670
Smaller	No	No	2038.1121826171875	5894364.0	59.776356524433396%	172877.43381050433%	0.19267934560775757
Smaller	Yes	Yes	2499.765380859375	10573124.0	72.86666031062413%	308200.217840143%	-0.4693032503128052
Smaller	Yes	No	2354.73876953125	7206153.5	68.63922164595455%	210055.04886630565%	-0.001409292221069336
Larger	No	Yes	1075.128173828125	2041166.625	47.587323957911444%	90346.11853775119%	0.49235212802886963
Larger	No	No	1362.80419921875	2986077.75	60.32044038852648%	132169.7753922668%	0.2573481798171997
Larger	Yes	Yes	1786.056640625	4401619.5	79.05444023661904%	194824.48528917972%	-0.0947037935256958
Larger	Yes	No	1619.82861328125	4035092.75	71.69685517778291%	178601.2780325175%	-0.0035468339920043945
3rd Attempt							
Smaller	No	Yes	1842.8148193359375	4982958.0	54.04842608206162%	146146.55488285472%	0.3175099492073059
Smaller	No	No	1913.122802734375	5269573.5	56.110508394303174%	154552.78024157273%	0.2782537341117859
Smaller	Yes	Yes	2311.616455078125	7243367.0	67.38223206480292%	211139.80005304993%	-0.006580710411071777
Smaller	Yes	No	2307.34814453125	7325875.0	67.2578133745127%	213544.8587257331%	-0.01804649829864502
Larger	No	Yes	1073.45361328125	2061699.5	47.51320455785187%	91254.94466490326%	0.48724550008773804
Larger	No	No	1326.149169921875	3015192.5	58.698015456970374%	133458.4524764794%	0.2501072287559509
Larger	Yes	Yes	1530.8138427734375	4443701.0	67.7568833452974%	196687.0966706716%	-0.10516965389251709
Larger	Yes	No	1808.312744140625	4471760.0	80.03953990549645%	197929.04414766934%	-0.11214816570281982

Table D.3: Long Short-Term Memory (LSTM) Results

Random Forest (RF)							
Dataset Size	**Date Included**	**Normalized**	**MAE**	**MSE**	**Normalised MAE**	**Normalised MSE**	R^2
1st Attempt							
Smaller	No	Yes	806.8974605453657	1479407.0690002071	23.665719037252448%	43389.93955070903%	0.7973733365249996
Smaller	No	No	808.8887756410801	1483816.165769674	23.724122869057748%	43519.25517066864%	0.7967694455567489
Smaller	Yes	Yes	789.3202010289027	1224131.9681320917	23.008210052462573%	35682.71205779803%	0.8298875231237407
Smaller	Yes	No	787.9831147315441	1220223.929830529	22.969234789511248%	35568.79508719698%	0.8304306067883398
Larger	No	Yes	777.4493993732239	1151251.3127554106	34.41146584144092%	50956.68638465122%	0.7136782931428409
Larger	No	No	776.4131248010092	1155595.767465603	34.36559825562392%	51148.980641975235%	0.7125978064808898
Larger	Yes	Yes	766.4599892327493	1133985.4571129913	33.925052562362694%	50192.465070516366%	0.7179723940077988
Larger	Yes	No	763.5368214049998	1127794.2726386497	33.79566730599167%	49918.430859121225%	0.7195121711933017
2nd Attempt							
Smaller	No	Yes	806.8974605453657	1479407.0690002071	23.665719037252448%	43389.93955070903%	0.7973733365249996
Smaller	No	No	808.8887756410801	1483816.165769674	23.724122869057748%	43519.25517066864%	0.7967694455567489
Smaller	Yes	Yes	789.3202010289027	1224131.9681320917	23.008210052462573%	35682.71205779803%	0.8298875231237407
Smaller	Yes	No	787.9831147315441	1220223.929830529	22.969234789511248%	35568.79508719698%	0.8304306067883398
Larger	No	Yes	777.4493993732239	1151251.3127554106	34.41146584144092%	50956.68638465122%	0.7136782931428409
Larger	No	No	776.4131248010092	1155595.767465603	34.36559825562392%	51148.980641975235%	0.7125978064808898
Larger	Yes	Yes	766.4599892327493	1133985.4571129913	33.925052562362694%	50192.465070516366%	0.7179723940077988
Larger	Yes	No	763.5368214049998	1127794.2726386497	33.79566730599167%	49918.430859121225%	0.7195121711933017
3rd Attempt							
Smaller	No	Yes	806.8974605453657	1479407.0690002071	23.665719037252448%	43389.93955070903%	0.7973733365249996
Smaller	No	No	808.8887756410801	1483816.165769674	23.724122869057748%	43519.25517066864%	0.7967694455567489
Smaller	Yes	Yes	789.3202010289027	1224131.9681320917	23.008210052462573%	35682.71205779803%	0.8298875231237407
Smaller	Yes	No	787.9831147315441	1220223.929830529	22.969234789511248%	35568.79508719698%	0.8304306067883398
Larger	No	Yes	777.4493993732239	1151251.3127554106	34.41146584144092%	50956.68638465122%	0.7136782931428409
Larger	No	No	776.4131248010092	1155595.767465603	34.36559825562392%	51148.980641975235%	0.7125978064808898
Larger	Yes	Yes	766.4599892327493	1133985.4571129913	33.925052562362694%	50192.465070516366%	0.7179723940077988
Larger	Yes	No	763.5368214049998	1127794.2726386497	33.79566730599167%	49918.430859121225%	0.7195121711933017

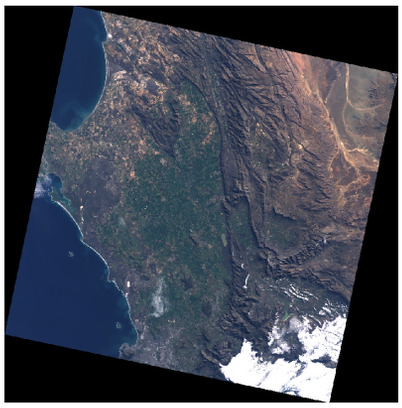
Table D.4: Random Forest (RF) Results

State Vector Machine (SVM)							
Dataset Size	**Date Included**	**Normalized**	**MAE**	**MSE**	**Normalised MAE**	**Normalised MSE**	R^2
1st Attempt							
Smaller	No	Yes	2279.264538514051	7812280.703596083	66.8491807417378%	229128.5438505345%	-0.07000730648771891
Smaller	No	No	2196.095172546354	8518991.490360333	64.40988337901263%	249855.86019238344%	-0.16680179379580418
Smaller	Yes	Yes	29914159.978777215	1195008826537548.2	871979.8067216489%	34833789962149.89%	-166065355.23891053
Smaller	Yes	No	77494276.91264307	8640780421276447.0	2258911.654288833%	251873562453844.7%	-1200772953.120427
Larger	No	Yes	1413.5611172266686	3778283.1359228874	62.56704312775866%	167234.3707203597%	0.0603229160470059
Larger	No	No	1308.5601759337203	3308838.783825618	57.9194913931578%	146455.82448998635%	0.17707597133364072
Larger	Yes	Yes	214464188.79266086	5.006109936310042e+16	9492614.069546815%	2215804413904137.0%	-12450434806.742098
Larger	Yes	No	N/A	N/A	N/A	N/A	N/A
2nd Attempt							
Smaller	No	Yes	2279.264538514051	7812280.703596083	66.8491807417378%	229128.5438505345%	-0.07000730648771891
Smaller	No	No	2196.095172546354	8518991.490360333	64.40988337901263%	249855.86019238344%	-0.16680179379580418
Smaller	Yes	Yes	29914159.978777215	1195008826537548.2	871979.8067216489%	34833789962149.89%	-166065355.23891053
Smaller	Yes	No	77494276.91264307	8640780421276447.0	2258911.654288833%	251873562453844.7%	-1200772953.120427
Larger	No	Yes	1413.5611172266686	3778283.1359228874	62.56704312775866%	167234.3707203597%	0.0603229160470059
Larger	No	No	1308.5601759337203	3308838.783825618	57.9194913931578%	146455.82448998635%	0.17707597133364072
Larger	Yes	Yes	214464188.79266086	5.006109936310042e+16	9492614.069546815%	2215804413904137.0%	-12450434806.742098
Larger	Yes	No	N/A	N/A	N/A	N/A	N/A
3rd Attempt							
Smaller	No	Yes	2279.264538514051	7812280.703596083	66.8491807417378%	229128.5438505345%	-0.07000730648771891
Smaller	No	No	2196.095172546354	8518991.490360333	64.40988337901263%	249855.86019238344%	-0.16680179379580418
Smaller	Yes	Yes	29914159.978777215	1195008826537548.2	871979.8067216489%	34833789962149.89%	-166065355.23891053
Smaller	Yes	No	77494276.91264307	8640780421276447.0	2258911.654288833%	251873562453844.7%	-1200772953.120427
Larger	No	Yes	1413.5611172266686	3778283.1359228874	62.56704312775866%	167234.3707203597%	0.0603229160470059
Larger	No	No	1308.5601759337203	3308838.783825618	57.9194913931578%	146455.82448998635%	0.17707597133364072
Larger	Yes	Yes	214464188.79266086	5.006109936310042e+16	9492614.069546815%	2215804413904137.0%	-12450434806.742098
Larger	Yes	No	N/A	N/A	N/A	N/A	N/A

Table D.5: Support Vector Machine (SVM) Results

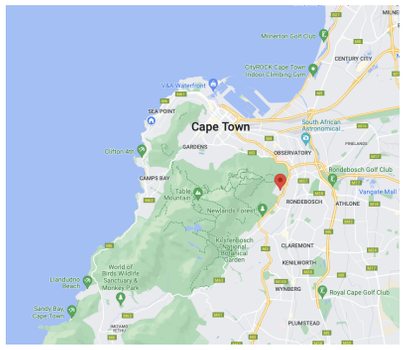
D.2 Acceptance Test Protocol

User Requirement (UR) 1 and Function Requirement (FR) 1 and 2, with corresponding Acceptance Test Protocol (ATP) UATP1 and FATP1&2 as described in Table 4.1 and 4.2, were met but their results were not included in the above report. Figure D.1 shows that a satellite image can be extracted for a co-ordinate specified location and visually verified. The quality can be deemed adequate also through visual confirmation. The co-ordinates chosen were (-33.957447294917145, 18.46125274152161), the location of the University of Cape Town.



(a) Satellite Image Displayed using IPython.display

(b) Satellite Image Displayed using Folium



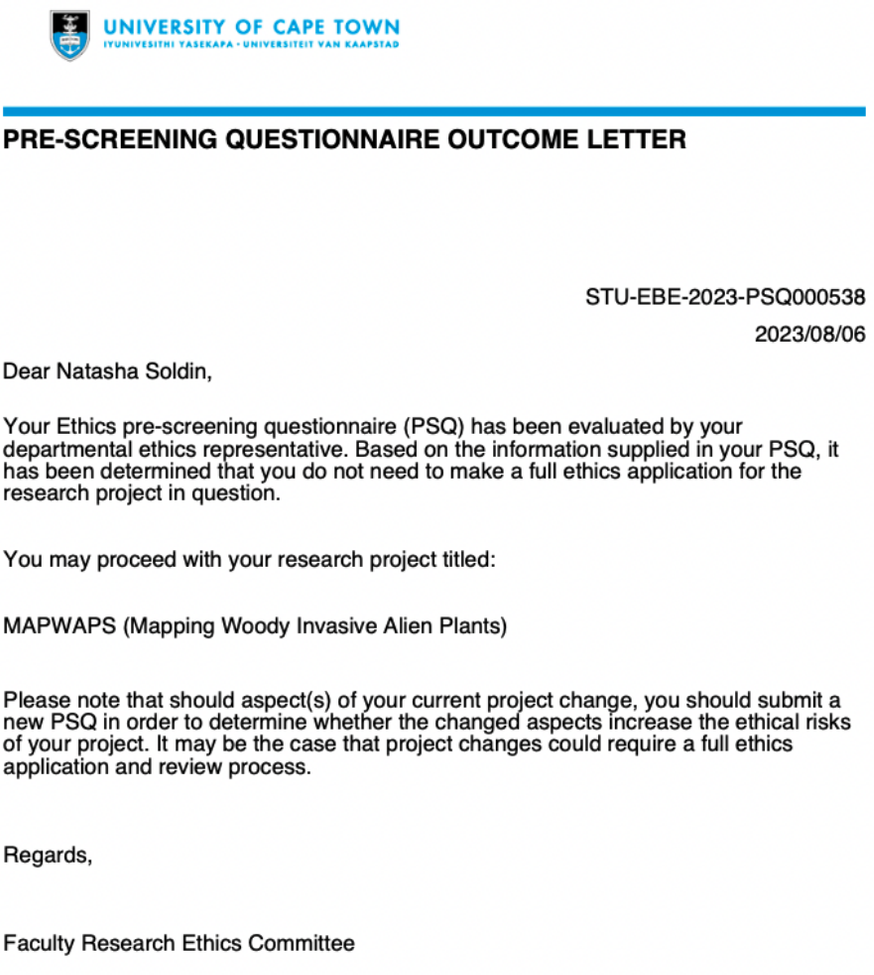
(c) Google Map Confirmation

Figure D.1: Satellite Image extracted at Specified Co-ordinates

Appendix E

Ethics Clearance

Ethics clearance was required from the university before project commencement. Figure E.1 below shows the certificate granting ethics clearance for this project.



UNIVERSITY OF CAPE TOWN
IYUNIVESITHI YASEKAPA • UNIVERSITEIT VAN KAAPSTAD

PRE-SCREENING QUESTIONNAIRE OUTCOME LETTER

STU-EBE-2023-PSQ000538

2023/08/06

Dear Natasha Soldin,

Your Ethics pre-screening questionnaire (PSQ) has been evaluated by your departmental ethics representative. Based on the information supplied in your PSQ, it has been determined that you do not need to make a full ethics application for the research project in question.

You may proceed with your research project titled:

MAPWAPS (Mapping Woody Invasive Alien Plants)

Please note that should aspect(s) of your current project change, you should submit a new PSQ in order to determine whether the changed aspects increase the ethical risks of your project. It may be the case that project changes could require a full ethics application and review process.

Regards,

Faculty Research Ethics Committee

Figure E.1: Ethics Clearance Certificate