

2023-12-01

Developing Next-Generation Ecoinformatics Tools for Advancing Global Change Science

Ifeanyi H. Nwigboji
University of Texas at El Paso

Follow this and additional works at: https://scholarworks.utep.edu/open_etd



Part of the [Ecology and Evolutionary Biology Commons](#)

Recommended Citation

Nwigboji, Ifeanyi H., "Developing Next-Generation Ecoinformatics Tools for Advancing Global Change Science" (2023). *Open Access Theses & Dissertations*. 4009.
https://scholarworks.utep.edu/open_etd/4009

This is brought to you for free and open access by ScholarWorks@UTEP. It has been accepted for inclusion in Open Access Theses & Dissertations by an authorized administrator of ScholarWorks@UTEP. For more information, please contact lweber@utep.edu.

DEVELOPING NEXT GENERATION ECOINFORMATICS TOOLS FOR
ADVANCING GLOBAL CHANGE SCIENCE

IFEANYI HUMPHREY Nwigboji

Doctoral Program in Environmental Science and Engineering

APPROVED:

Craig Tweedie, Ph.D., Chair

Amy Wagler, Ph.D.

Mauritz – Tozer Marguerite, Ph.D.

Anthony Darrouzet – Nardi, Ph.D.

Stephen L. Crites, Jr., Ph.D.
Dean of the Graduate School

Copyright ©

By

Ifeanyi Humphrey Nwigboji

2023

Dedication

This dissertation is dedicated to my wife and children who has been my most profound inspiration and to the Almighty God for walking every step of this journey with me.

DEVELOPING NEXT GENERATION ECOINFORMATICS TOOLS FOR
THE ADVANCEMENT OF GLOBAL CHANGE SCIENCE

by

IFEANYI HUMPHREY Nwigboji

DISSERTATION

Presented to the Faculty of Graduate School of
The University of Texas at El Paso
in Partial Fulfilment
of the Requirement
for the Degree of
DOCTOR OF PHILOSOPHY

Environmental Science and Engineering
THE UNIVERSITY OF TEXAS AT EL PASO

December 2023

Acknowledgements

I would like to express my deepest appreciation to my advisor, Dr. Craig E. Tweedie for his dedicated support, immense contributions, and overall insights in this field that made this an inspiring experience for me. This dissertation wouldn't have been possible without Dr. Tweedie's continuous encouragement, invaluable patients, feedback, and great mentorship. I look forward to years of future collaboration and guidance.

I am extremely grateful to my committee members, including Dr. Marguerite Mauritz-Tozer for her technical and probing questions, and excellent feedback. The meetings and conversations were vital in inspiring me to think outside of the box. To this, I say a huge THANK YOU. Dr. Amy Wagler for her valuable feedback and clear understanding and explanation of the various techniques in the Machine learning chapter, including some thought provoking, and probing questions. I am very grateful to have you with your vast knowledge and experience on my committee. Dr. Darrouzet – Nardi, Anthony for his insightful contributions and expertise knowledge on the overall project.

I had the pleasure of working and collaborating with a knowledgeable and enthusiastic lab group at SEL, whose contributions helped me in no small measure achieve my goals over the past years. They include Dr. Sergio Vargas - Zesati, Gesuri Ramirez, Kamal Nyaupane, Stephen Escarzaga, Maurico Barba, Sasha Peterson, Tabatha Fusion, Mariana Orejel, Ryan Cody, Muhammed Hanggito, and Daniel Cruz, among many others. Thank you all for all your motivation. I would be a remiss if I fail to mention the invaluable administrative and timely assistance of our program coordinator, Ms. Hamdan Lina throughout the course of my stay in the program.

To my family including my lovely wife – Amie Nwigboji, and children (Ikem and Iris) for being my support system. Their continuous encouragement and belief in me have kept my spirit and motivation high throughout this process. Thank you for all the emotional support as this couldn't have been done without you. To my parents, siblings, and friends, thank you for your unconditional support.

Lastly, to the Almighty God, for the gift of life, knowledge, understanding, wisdom, and all his provisions that made this dissertation a reality.

Abstract

Ecosystems are responding to a variety of human-induced, interlinked stressors that have emerged from changing climate, alteration to the global water cycle, sea-level rise, and land use and land cover change, among others. Quantifying these changes and their associated impacts on ecosystems requires a huge amount of long-term data. Due to advances in data collection techniques, such as remote sensing platforms, environmental sensors, synthesized datasets, and various software technologies, the volume and variety of long-term ecological data being collected has tremendously increased. Although there are several complex models and analyses that are increasingly parametrized with data from such sensors, there still exists a huge gap in managing, analyzing, visualizing, integrating, and sharing ecological data. The overarching goal of this dissertation is to develop ecoinformatics tools that will contribute to the advancement of global change science through: I) mitigating the challenges of new infrastructures for Big Data archiving, management and sharing, and analysis by developing a flexible system that supports multiple and novel data usage and visualization and II) attempt to utilize multi-sensor cross-correlation to detect rare soil moisture events in temporal data using some Machine Learning and Deep Learning (DL) models. To actualize the first objective, we developed web-based analytic tools capable of integrating spectral reflectance data from multiple instruments in the NASA Arctic-Boreal Vulnerability Experiment (ABoVE) study region using an open-source software – R shiny. *R-HyperSpectral* will help to dynamically view, interact, and discover optical properties of boreal and tundra plant communities. We also developed a multi-data fusion tool called *rDataFusion*, which is capable of aggregating heterogeneous data sets collected from a range of automated and semi-automated sensors and manual observations over a decade-long period. *rDataFusion* was developed using R shiny. Lastly, to achieve the second objective, we deployed several Machine and Deep Learning techniques for optimal rare soil moisture events detection in the Chihuahuan Desert, New Mexico. Specifically, the machine and deep learning techniques used for this study include both classification and regression methods, including a Decision Tree Classifier (DTC), Logistic Regression Classifier (LR), Random Forest Regression (RF), and the Long Short-Term Memory (LSTM) method of Artificial Neural Network (ANN). Of all these methods used, the DTC performed the best, with prediction accuracy of 88.8%, closely followed by LSTM model with 88%. The LR recorded a prediction accuracy of approximately 80%.

Lastly, through the tools we developed, data will be available for ecological and environmental science researchers to analyze and further understand ecosystem changes over multiple temporal scales and levels of biological organization and interaction. Furthermore, the analysis and prediction of rare soil moisture events in the dryland ecosystem unveils a pathway to understanding soil moisture events and the key drivers in drylands.

Table of Content

Acknowledgements.....	v
Abstract.....	vii
Table of Contents.....	ix
List of Tables.....	xi
List of Figures.....	xii
Chapter 1: Introduction.....	1
1.1 Background & Rationale.....	1
1.2 Project Description.....	4
1.3 Structure of the Dissertation.....	4
Chapter 2: R-HyperSpectral: An Interactive Tool for the Discovery and Analysis of Near Surface Hyperspectral Reflectance Data Acquired through the NASA ABoVE Campaign.....	6
2.1 Introduction.....	6
2.2 Materials and Method.....	9
2.3 Results.....	15
2.4 Discussion.....	27
2.5 Conclusion.....	29
Chapter 3: rDataFusion: A Project-Specific Multi-Data Fusion Tool for Discovering, Integrating, and Visualizing Heterogenous Long-term Data Set.....	31
3.1 Introduction.....	31
3.2 Materials and Methods.....	35
3.3 Results.....	47
3.4 Discussion.....	58
3.5 Conclusion.....	60
Chapter 4: Detecting Rare Soil Moisture Events in a Chihuahua Desert Dryland Ecosystem using Machine and Deep Learning.....	61
4.1 Introduction.....	61

4.2	Study Site.....	64
4.3	Methods.....	68
4.4	Results.....	80
4.5	Discussion.....	88
4.6	Conclusion.....	90
Chapter 5: General Discussion.....		92
5.1	Overview of this Dissertation.....	92
5.2	Future Research Direction.....	94
References.....		96
Vita.....		111

List of Tables

Table 2.1: Hyperspectral indices that are included in R-HyperSpectral, from a list generated by C. Laney and others from different websites. The Expression column gives the specific equation for each index, where R followed by a numeric value indicates reflectance for that specific wavelength. References are given where available.....	18
Table 3.1: showing the abbreviated variable names, full variable names, and the units for the climate data stream.....	38
Table 3.2: showing the abbreviated variable names, full variable names, and the units for the Flux data stream.....	38
Table 3.3: showing the abbreviated variable names, full variable names, and the units for the soil/ECTM data stream.....	43
Table 3.4: showing the abbreviated variable names, full variable names, and the units for the CS650 data stream.....	43
Table 4.1: Table of micrometeorological variables used in the study with their abbreviated, full variable names, and SI units.....	66
Table 4.2: DTC & LR modelling results before & after transforming the data with and without outliers.....	81
Table 4.3: Results of Random Forest Regression model.....	85
Table 4.4: Results from LSTM model.....	85

List of Figures

Figure 2.1: Map of the key five study sites that will be used in objective 2 of study- Barrow, Atqasuk, Imnaviat Creek, Toolik Lake and Ivotuk.....	11
Figure 2.2: Flow chart or visual map of R-HyperSpectral.....	15
Figure 2.3: Table of selected data after filters are applied.....	16
Figure 2.4: Data table showing metadata based on user's election.....	17
Figure 2.5(a-c): Table of calculated indices for all indices.....	21-23
Figure 2.5d: Table of calculated vegetation indices for all indices classified under broadband greenness.....	24
Figure 2.5e: Table of calculated vegetation indices for all indices classified under narrowband greenness.....	25
Figure 2.6: below shows the boxplot of the visualization by location of raw reflectance (where Atq, Brw, EACr, Imn, Sag, and Tol are Atqasuk, Barrow, Eagle Creek, Imnaviat, Sagwon, and Toolik)	26
Figure 3.1: Map showing the Jornada study site at the Chihuahuan Desert (Ramirez, 2011). The green, yellow, and brown colors represent the United States, Mexico, and Jornada site, part of the Map, respectively.....	37
Figure 3.2: UTEP-JER site showing the interconnection of different projects and the Robotic Tram System.....	44
Figure 3.3: A flow chart or visual map or rDataFusion.....	45
Figure 3.4: Select Data & View tab: This figure shows the processes of getting data into the application and view raw data table. The “chose date range “allows the user to select a date range of data that will be uploaded, the upload button allows the user to upload the data for the date range selected, the ‘chose data to inspect’ field allows users to choose the data stream they would want to upload.....	48
Figure 3.5: <i>Aggregate Raw Data</i> tab: This figure shows how raw data is aggregated from every minute to half-hourly or hourly data.....	49
Figure 3.6: <i>Flagged status of variables</i> tab: This figure allows users to filter data based on QC flags.....	49
Figure 3.7: The generated “semi” clean data table. Here, the data has passed the initial QAQC test but still contains outliers and gaps.....	50

Figure 3.8: These show time series graphs depicting outliers and extreme values. The figures show Air temperature, relative humidity, absolute humidity, and atmospheric pressure, respectively, from Jan. to Sept. 2020, with the red dots showing extreme values and the green color showing outliers.....	52-53
Figure 3.9: Statistics of the outliers and extreme values.....	54
Figure 3.10: Missing data points replaced from the selected data source.....	55
Fig. 3.11: Raw and clean data comparison.....	56
Figure 3.12: Merged data table of all the data sources.....	57
Figure 3.13: Time series visualization of temperature from the merged data table.....	57
Figure 4.1: Map showing the Jornada study site in the northern Chihuahuan Desert (Ramirez, G. 2011).....	67
Figure 4.2 UTEP-JER site showing a range of different sensing systems including the Robotic Tram System (Ramirez, G. 2011).....	67
Figure 4.3: Data points showing zero values, missing values, % of total values missing, total zero + missing values, its percentage, and the data type for each of the 30-minute from 2010 - 2020.....	68
Figure 4.4: Summary statistics of the data for exploratory data analysis.....	69
Figure 4.5: Box plot showing outlier distribution of the data points in the variables.....	69
Figure 4.6: Histogram to determine the frequency distribution of the data points.....	70
Figure 4.7: Histogram showing transformed variables (P_RAIN, NETRAD, SW-OUT, SW_IN, LEAF_WET, PPFD_IN, and PPFD_OUT) using Yeo-Johnson power transformation method.....	71
Figure 4.8: Box plot of variables before removing data points deemed as outliers.....	72
Figure 4.9: Box plot of variables after removing data points deemed as outliers.....	73
Figure 4.10: Correlation heat map of all the variables without data transformation.....	74
Figure 4.11: Correlation heat map after multi-correlated feature variables were removed using Variable Inflation Factor.....	75
Figure 4.12: Histogram of the dependent variable (SWC).....	77
Figure 4.13: DTC Modelling confusion matrix before & after transforming the data and with or without outliers.....	83
Figure 4.14: Logistic Regression confusion matrix for the different scenarios described above in order.....	84
Figure 4.15: DTC VIP without transforming or outliers removed.....	86
Figure 4.16: DTC VIP with transformed data, outliers and strongly correlated features removed.....	87

Figure 4.17: Visual interpretation of confusion matrix & confusion matrix of the best performing model in predicting rare soil moisture events.....88

Chapter 1: Introduction

1.1 Background and Rationale

The rapid development of earth-observation systems and other platforms such as high-performance computing have unveiled the vulnerable state of the biosphere being at a critical tipping point in terms of the preservation of biodiversity and ecosystem services (Barnosky et al., 2012). The wide range of habitats found in different ecosystems are under a looming threat from a variety of human-induced, interlinking climate stressors emerging from a changing climate, alteration to the global water cycle, sea-level rise, and loss of ice-dependent organisms due to the impacts of climate change on sea-ice extents among others (Liu et al., 2015; Baird et al., 2015).

However, discovering the changes in ecosystem functionalities and associated impacts requires huge amounts of data. Consequently, the volume and variety of data available for analysis continue to increase at a rapid pace because of increased availability of data from long-term ecological research, remote-sensing platforms, environmental sensors, synthesized datasets, and various software technologies (Porter, et al., 2009 & Hampton, et al 2017). The environmental science and ecological research communities are thus faced with the prospect of pursuing multidisciplinary scientific research across multiple scales, necessitating the need for synthetic research that can address critical environmental, ecological, data management, accessibility, and data fusion problems (Farley, et al., 2018, LaDeau, et al., 2017).

The growth of data in all these identified dimensions challenges traditional approaches to data management and analysis (Peters, et al., 2014; Laney, 2013). Methods that work well at small scales such as sharing spreadsheets by email, may not scale up. According to Williams et al., 2018; Dietze, 2017; Lynch, 2008 and Schnase et al., 2017, other notable challenges include complex relationships among extremely varied data sets that entail sophisticated data models, computationally intensive macroscale ecological forecasting, and the need for flexible system that support multiple and novel data use.

These identified challenges have motivated the research community to rapidly learn and implement concepts, techniques and software analytical tools needed to act on this new era of 'big data' and data intensive research. Big Data is defined by the first framework developed by the National Institute of Standards as data that either exceeds the capacity or capability of current

or conventional methods and systems (Farley et al., 2018). Because of this rapid rise in the availability of enormous amounts of data that expand both the temporal and spatial scale of observation, it becomes exigent to know the needs and expectations of ecological and environmental science researchers in the development of software technological tools for ecological and environmental data management, analysis, visualization, and sharing.

Furthermore, there is a dire need among the ecological and environmental science research communities for software tools that aid data storage, management, integration, fusing of data from different research labs, and accommodating different instrumentation through spectral band interpolation. Additionally, a software tool that is capable of visualizing diverse streams of data and providing a standard set of quality control and assurance across a given network is required for gap filling and facilitating intra-site data integration and spatial comparison (Laney, 2013).

It should also be noted that, due to these advances in data collection techniques, large (big) high resolution data sets with complex relationships have been produced across multiple spatial and temporal scales as mentioned earlier. Without a doubt, Machine Learning (ML) approaches are increasingly being used by researchers to model these complex relationships to detect anomalous behavior exhibited by these large data sets (Willcock, et al., 2018 & Olden et al, 2008). The strength and potential of ML has been demonstrated in the areas of precision agriculture to detect diseases, weeds, and pests (Hruska, et al., 2018), climate and weather applications such as automated warnings and notifying members of society of approaching weather extremes (Huntingford, et al., 2018) and real-time irrigation scheduling for sustainable and efficient water use for irrigated agriculture and healthy plant growth (Adeyemi, et al., 2018). While new solutions are emerging to these challenges, both within ecology and other disciplines; Centers like the National Center for Ecological Analysis and Synthesis (NCEAS) and National Socio-Environmental Synthesis center (SESYNC) and others have allowed for data sharing and synthesis (Hampton and Parker, 2011; Baron et al., 2017). There is a need for more ecoinformatics tools that will undoubtedly facilitate optimal use of data and ensure data sharing among wider user groups.

This dissertation will help address some of the data challenges mentioned above, with the overarching goal of *developing ecoinformatics tools that will contribute to the advancement of global change science* through; I) mitigating the challenges of new infrastructures for Big Data

archiving, management and sharing, and analysis by developing a flexible system that supports multiple and novel data usage and visualization and II) understanding complex relationships among variables in a data set. To achieve this goal, my objectives are to:

- I) Develop tools that will aid synthesis, analysis and discovery of spectral data collected from multiple sources, labs, and funding groups to enhance data reuse and availability to wider user group (Objective 1).
- II) Addressing challenges experienced by lab groups that collect multiple streams of data (climate, ecological, sensor, human observations/measurements) at one or several networked sites. To address what has been identified as a critical need in such a community (Laney et al. 2015), a new data fusion, quality checking, and visualization tool will be developed. This will hasten recognition that sensors are running as desired, ensure quality assurance and quality control (QA/QC), identify needs for and apply gap-filling or gap-fill data (Objective 2).
- III) Gain information about soil moisture anomalies or dynamics and how to detect them in the dryland ecosystem to better understand small and large-scale drought patterns (Objective 3).

These objectives and associated research questions described below will be answered in the three data chapters.

- Can we develop a tool that has the capability to calculate various or multiple spectral indices that are proxies to critical ecosystem properties and processes? (Chapter 2)
- How can we develop or create a workflow or template for project-specific multi-data fusion? (Chapter 3)
- Can we provide a standard set of quality control and assurance across the network, thereby facilitating site intercomparison and transfer lessons learned at one spot to another? (Chapter 3)
- What are rare events and why are they important? (Chapter 4)
- Why use ML in rare event detection? (Chapter 4)
- What processes and mechanisms control soil moisture? (Chapter 4)
- Can we improve our confidence or understanding of rare or extreme events by integrating data from multiple information sources to understand events that might be difficult to

measure directly, and given these inferential data sources, place a confidence interval around the probability that event has occurred? (Chapter 4)

1.2 Project description (Study Area)

The System Ecology Lab (SEL) located at The University of Texas at El Paso (UTEP) has been conducting ecological research for about 15 years at sites within the Alaskan Arctic and the Chihuahuan Desert. Researchers within the lab collect huge amounts of environmental and remote sensing data including hyperspectral reflectance data and other kinds of ecological data, with the purpose of understanding the biophysical factors controlling land-atmosphere exchange of carbon, water, and energy and how these factors contribute to global change. Five of the SEL sites (Utqiagvik (previously, Barrow), Atkasuk, Toolik Lake, Ivotuk, and Imnaviat Creek) will serve as reference sites for the development of R-HyperSpectral (Chapter 2), while the data from Jornada Experimental Range in southern New Mexico will serve as a case study site for the development of rDataFusion and to improve the detection of soil moisture dynamics (Chapter 3 & 4), respectively. The detailed description of these locations is presented in the various chapters.

1.3 Structure of the dissertation

This dissertation comprises a total of five chapters. Chapters 2-4 are data chapters that highlight the research conducted that make up this work and will be submitted for publication. Chapters 2 & 3 describe novel web-based ecological data management and visualization tools that look to help address the need of the ecological community in bringing to bear tools that are easy to use and highly accessible (Objective 1&2). Chapter 2 describes an open-source software tool that is designed to tackle the challenges of rapidly changing Arctic landscape, detailed field measurements of vegetation, particularly, hyperspectral reflectance measurement. The complexity of hyperspectral reflectance data contributes to the difficulty in managing, analyzing, visualizing, and sharing, yet few of these tools have been developed and those that have, have included a limited scope relative to the needs of the research that has utilized these tools. Chapter 2 presents a new tool that attempts to remedy this situation and allows users to view the hyperspectral reflectance scans and explore common spectral indices at multiple temporal scales. Chapter 3 describes a multi-data fusion and data integration tool that aggregates heterogeneous data sets collected from a range of automated and semi-automated sensors and manual

observations over a decade-long period. This chapter provides a tool that can help solve an integral problem that ecologists face in limited availability of custom analytic tool that aids researchers with improved capacities for aggregating different streams of data from a single intensive site by providing an open-source multi-data fusion tool that facilitates data management, sharing, and analysis. Chapter 4 describes a Machine learning analysis that can assist in gaining information about soil moisture anomalies or dynamics and how to detect them in the dryland ecosystem to better understand small and large-scale drought patterns (Objective 3). Chapter 5 summarizes the works in Chapters 2-4 and discusses how each of these chapters contributes to our further understanding of global change science.

Chapter 2: R-HyperSpectral: An Interactive Tool for the Discovery and Analysis of Near Surface Hyperspectral Reflectance Data Acquired through the NASA ABoVE Campaign

Abstract

Changes in climate variability over the past few decades have exacerbated many Arctic ecosystem changes that includes but is not limited to an increase in air temperature, sea ice loss, increased rates of coastal erosion, shifts in snow cover, permafrost thaw and degradation, species shifts, and land cover change. To better characterize and understand these changing northern ecosystems, NASA established the Arctic-Boreal Vulnerability Experiment (ABoVE) program to investigate links between changing land surface conditions and the vulnerability and resilience of the Arctic and Boreal ecosystems. To monitor changing Arctic landscapes, detailed field measurements of near surface vegetation optical properties along with corresponding airborne and satellite remote sensing observations, particularly, hyperspectral reflectance measurements have been widely used by NASA investigators. However, due to the size and complexity of data produced by such approaches, managing, analyzing, data sharing, and visualization has posed a great challenge for most of the ABoVE projects. Here, we present *R-HyperSpectral*, a web-based discovery and analytic tool capable of integrating spectral reflectance data from multiple instruments using open-source software – R shiny. R-HyperSpectral has been designed with the intent of permitting users to dynamically view, interact, and discover the optical properties of boreal and tundra plant communities. Users can view the hyperspectral reflectance scans and explore common spectral indices over multiple temporal scales to aid species-landscape characterization, ecological scaling, and change detection.

2.1 Introduction

Arctic ecosystems encompass a rich dynamic of plant and animal species with their vegetation covering a major portion of the earth's surface (Atkinson & Treitz, 2012). Arctic ecosystems are known for low soil temperatures, permafrost dominated landscapes, and a short growing season, with limited vegetation productivity (Langer, et al., 2023). They are generally considered to show sensitivity to disturbance (Reynolds & Tenhunen, 1996). These disturbances could lead to a change in vegetation caused by some external factors such as lightning induced fires, oil exploration, and climate change (Stow, et al., 2003, Foster, et al., 2022, Langer, et al., 2023 & IPCC, 2019).

Changes to Arctic ecosystems have been previously observed at plot to satellite scales using metrics such as vegetation phenology, vegetation biomass and cover, and shifts in species composition (IPCC, 2007 & 2019, Davidson, et al., 2016, Bjorkman, et al., 2018 & Myer-Smith, et al., 2020). Satellite remote sensing offers the possibilities to provide quality data for the assessment and monitoring of the patterns of vegetation that can be used to model or predict carbon fluxes, vegetation types, soil organic matter, soil moisture and nitrogen content, surface temperature, micro and macro topography, and thaw depth in the Arctic (Shaver, et.al., 2007, Davidson, et al, 2016). Furthermore, satellites can cover large areas at higher temporal frequency compared to human surveys. Remote sensing in the arctic is, however, subject to challenges, including satellite orbits that enable coverage at high northern latitudes, high spatial heterogeneity of landscape units, persistent cloud cover, terrain effects, low sun angles, and lack of standard band definitions among others (Stow, et al., 2003 & Gamon, et al., 2013).

Some studies have shown that repeat high resolution imagery from unoccupied aerial vehicles and near-surface sampling at plot to landscape scales can help overcome and/or improve understanding of satellite derived products (Gamon, et al., 2013 & Goswami, et al., 2011). Near-surface hyperspectral remote sensing studies can also help to understand fundamental ecosystem properties such as net ecosystem exchange of carbon fluxes and other biophysical factors (Boelman, et. al., 2003, Huemmrich, et al., 2013, and Zang, et al., 2013). The demand for high spatial and spectral resolution reflectance data is driving innovations in instrumentation and methodologies to process and analyze these data (Hilker, et al., 2010, Laney, 2013 & Wulder, et al., 2022).

Key challenges to the advancement of near-surface hyperspectral remote sensing of ecosystems include but are not limited to capacities to discover and access data, conduct quality checking and assurance to ensure clean and accurate data, manage and visualize data, among others (Gamon, et al., 2013). Additionally, approaches by different research groups appear to lack consistency, which likely limits the adoption of transcending solutions across ecosystems. To better work with very large, complex data sets, (big data) researchers may benefit from free, open-source tools that perform basic analysis, visualize data, and analyze workflow processes (Granell, et al., 2010, Hampton, et al., 2017, La Deau, et al., 2016, McCord, et al., 2021). Recently, there has been several efforts to develop and share spectral libraries using open-source

technologies to enhance data archival, visualization, and accessibility to a wider-user remote sensing community. This includes the Spectral Workbench (<https://spectralworkbench.org>), created and hosted by the not-for-profit Public Lab. Spectral Workbench is an open-source, web-based application, developed for the purpose of collecting, archiving, sharing, and analyzing spectral data (Padalia, et al., 2017). Also in this group is the Cornell Spectrum Imager (CSI) (Hovden, et al., 2013), which is a hyperspectral imaging software and a universal data analysis tool developed to extract hyperspectral signatures (Hovden, et al., 2013). There is also the United States Geological Survey (USGS) High Resolution Spectral Library that contains reflectance spectra, samples of minerals, rocks, plants, vegetation communities, micro-organisms, and man-made materials (<https://www.usgs.gov/centers/gggsc/science/usgs-high-resolution-spectral-library>). Others include ASTER, Spectral Library hosted by the Jet Propulsion Laboratory, is a library of both natural and man-made materials with over 2400 spectra (Baldrige, et al., 2008), the spectrum database SPECCHIO provides ready access to modelled data, spectral data, and existing spectral libraries (<https://www.specchio.ch>) (Bojinski, et al., 2003), among others.

While these existing spectral libraries are an immediate solution to improving capacities to access and utilize spectral data, most appear to have been developed with proprietary solutions that lack a user-friendly web interface and require local installation, which can limit data integration, discovery, analyses, and sharing spectral data and solutions to a wider-user group. It is also important to note that commercial solutions for maintaining spectral databases do exist (e.g. the USGS Spectral Library), but they are expensive, and/or are hard to update and apply to crowd-sourced data, and generally, do not encourage community-driven innovation and customization, which is known for catalyzing innovation in science and engineering (Laney, et al., 2013).

Importantly, none of the extant spectral libraries are uniquely dedicated to Arctic spectral data. These glaring limitations spurred us to develop *R-HyperSpectral*; a web-enabled, open-source multi-user collaborative tool for hyperspectral data with a focus on the Arctic. *R-HyperSpectral* accommodates spectral data from different instruments and can be used for visualizing, analyzing, and displaying metadata information. With a web-enabled application like *R-HyperSpectral*, our goal was to optimize access, discovery and use of spectral data from the Arctic by a wide-ranging remote sensing user community.

If spectral libraries that are modifiable, extensible, and have a full set of analytical and visualization tools are made available in a library common to ecologists, then data analysis may become more standardized and enhanced (Mineter, et al., 2003, Laney, et al., 2015). Instead of considering a new spectral library within the bounds of immediate research needs, *R-HyperSpectral* software which is currently connected via an API – Application Programming Interface (a set of functions and procedures allowing the creation of applications that access the features or data of an operating system, application, or other service) to EcoSIS (Ecological Spectral Information System), a useful tool for finding spectral data (<https://ecosis.org/>), is an ideal software to enhance Hyperspectral data analysis and visualization. This will help to scale the application to allow for data from other researchers within the ABoVE region and beyond. The overall goal of this paper is to develop a custom and shareable analytic and open-source multi-user collaborative tool for hyperspectral data with an initial focus on Arctic terrestrial and aquatic ecosystems. This application coined R-HyperSpectral, which will aid researchers collecting and sharing similar data sets, was built using a package in R called shiny. The rest of the paper is organized as follows: Section 2 describes the materials and methods. Section 3 presents our results. Discussions and the conclusion are given in Sections 4 and 5, respectively.

2.2 Materials and methods

2.2.1 Study Area

The System Ecology Lab (SEL) located at The University of Texas at El Paso (UTEP) has been conducting ecological research in the Alaskan Arctic for more than 20 years. Researchers within the lab collect a large volume and variety of environmental and remote sensing data including hyperspectral reflectance data to understand how biophysical factors control land-atmosphere exchange of carbon, water, and energy. SEL has several research sites in the Arctic, but for the purpose of this paper and developing *R-HyperSpectral* using SEL data collected for the International Tundra Experiment (ITEX) and the NASA ABoVE (Arctic Boreal Vulnerability Experiment) campaign as a case study, data was selected from five key locations that are research hotspots in Northern Alaska. These locations were a contributor to the NASA-ABoVE program saddled with multi-scale data collection, study of environmental change and its implications for socio-ecological system, gaining better understanding of the vulnerability of the Arctic and Boreal ecosystems to environmental change in western North-America, and providing

the scientific basis for informed decision making to guide societal responses at local to international levels (<https://above.nasa.gov/about.html>?). These localities include sites near Utqiagvik (previously, Barrow), Atqasuk, Toolik Lake, Ivotuk, and Imnaviat Creek), which are described below.

2.2.2 SEL- Utqiagvik, Atqasuk, Toolik lake, Imnaviat Creek, and Ivotuk

The SEL research sites located on the North Slope of Alaska cover a longitudinal gradient from high Arctic Coastal Plain tundra to low Arctic foothill tundra (Walker, et al., 2003). Except for Ivotuk, the primary sampling sites are divided into transects located within 1km² Arctic System Science Program (ARCSS) long-term monitoring grids that are adjacent to sites associated with the International Tundra Experiment (ITEX) (May et al., 2017). At Ivotuk, a remote site near the Brooks Range, sites are visited irregularly and there are no permanent ITEX sites. Sampling transects at each of the five locations are 50 m in length and span local moisture gradient and vegetation habitat types (Figure 1). Utqiagvik (Barrow) (71°18'N, 156°40'W; 7 m above sea level (a.s.l.) experiences a maritime high Arctic climate due to its proximity to the Arctic Ocean. Utqiagvik vegetation types range from dry heath to wet meadow, sedge-dominated tundra with a climate that is characterized by long, cold winters and short, cool summers during which the temperature can fall below 0°C on any given day (Oberbauer et al., 2007). Summers are generally cloudy, cool, wet, and windy (Brown et al., 1980; Oberbauer et al., 2007). The snow-free period is variable, but generally begins in early June and continues until early September. It is during this snow-free period that most of the data for this research is collected. The Atqasuk (70°29'N, 157°25'W, 21m a.s.l.) transect site; experiences a continental climate and is located approximately 100 km south of Utqiagvik. The transect spans plant habitat types ranging from dry heath, moist acidic tussock, to wet meadow sedge tundra. Compared to Utqiagvik, low clouds and fog typically dissipate by early afternoon in Atqasuk (Oberbauer et al., 2007). Atqasuk has long, cold winters and short, moderate summers. The summer temperature can fall below zero degree centigrade on any given day with daily maximums that may exceed 20 degrees centigrade (Oberbauer et al. 2007).

The Toolik lake site (68°37'N, 149°36'W, 736m a.s.l) covers dry heath to deciduous shrub-dominated tundra, to moist acidic tussock tundra and is located in the foothills of the Brooks Range. The climatic condition of Toolik lake is continental arctic with cold winters and

relatively warm summers (Chapin and Shaver, 1985; Oberbauer et al., 2007). There is a highly variable snow period, with snowmelt occurring between mid-May to early June. Fall snow initiation is also highly variable, starting as early to mid-September or as late as December. It is also worth noting that snow can fall anytime in the summer. Imnaviat Creek ($68^{\circ}37'W$, $149^{\circ}18'W$, 927 m a.s.l.) is located close to Toolik lake but is almost 200 m higher in elevation and spans plant habitat types ranging from dry heath to moist acidic tussock, to wet acidic tundra. Ivotuk, on the other hand, is part of the western Alaskan transect that starts in the north at Utqiagvik and goes south through Atqasuk and Oumalik (Walker, et al., 2003; Bratsch et al., 2016). It is dominated by deciduous shrubs (Walker, et al., 2003).

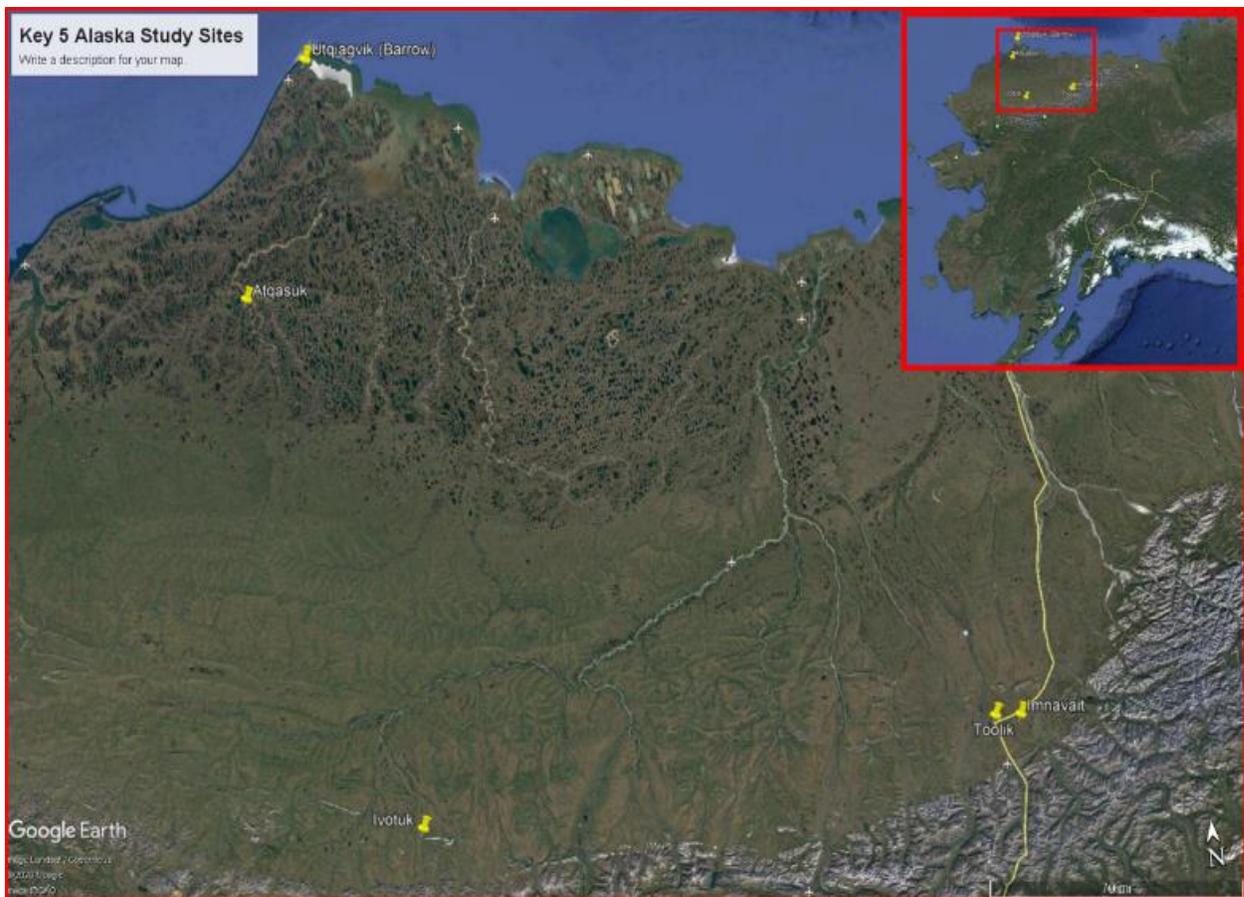


Figure 2.1: Map of the key five study sites - Barrow, Atqasuk, Imnaviat Creek, Toolik Lake, and Ivotuk.

2.2.3 Data Collection (spectral measurements)

Spectral reflectance measurements were made using two different instruments in the snow-free summer months (June-September) during 2010-2019. The instruments include a SVC HR-1024i and a Dual Channel Unispec, PP Systems Amesbury MA, with spectral wavelength ranging from 300-2500nm and 300-1100nm, respectively. Most measurements were made using the Unispec instrument. The SVC instrument was used in 2017 only. These measurements were made at each meter along a 50-meter transect running east to west across the landscape capturing a diverse range of plant communities (Gamon et al., 2013). It is also important to note that the in situ data were collected using fiber optics pointed with a Nadir attitude at targets of interest and calibrated using a 99% reflective Labsphere® Spectralon® target panel (<http://www.labsphere.com>) as a reference, to account for changing light conditions during sampling. Spectral reflectance was calculated by dividing the surface radiance against the irradiance using the signal from the reflective Labsphere® above as standard for calibration.

Since the test data for this application is from two different measurement instruments and differing wavelengths, we used linear interpolation with an increment of 1 to interpolate the wavelengths of the two instruments. The wavelength for the data collected using the SVC instrument ranged from 338.1nm to 2516.3nm while that collected using Unispec ranged from 303nm to 1053nm. The wavelength range after interpolation was 350nm to 1900nm. The maximum value of the wavelength was reduced to 1900nm because objects found after that wavelength were mostly water and ice and are insignificant for the purpose of this project.

2.2.4 Why R-shiny

Owing to the rise of automated data gathering and collection tools, data size and complexity of analysis have placed a limitation between research disciplines and the required data analysis (Kasprzak, et al., archive & Donoho, 2017). A data analytics software tool that utilizes common task frameworks and can help interpret, quantify, and possibly close methodological variations across disciplines is highly needed (Dondo, 2017). Following the review of data analytics software tools, we discovered that data analytics software tools such as Minitab (Arend, 2010), MATLAB (Moler & MathWorks, 2012), GenStat (Payne, et al., 2007), and SPSS (Landau & Everitt, 2004), have attempted to proffer solution to this problem by creating a more user-friendly interfaces that either make coding easier to learn, or use drop down menus and radio button selections to bypass the command lines ((Kasprzak, et al., archive). These mentioned

analytical software have their limitations such as non-publication ready graphics, non-intuitive drop-down menus, restrictive interfacing with other software, pricing – including the cost of licensing the proprietary software. Others include the difficulties encountered by users when they attempt to run codes originating from the software on their platform (Kasprzak, et al., archive), among others.

R programming language have grown to become one of the most popular programming languages for statistics, environmental, and biological data analytics with over 14,000 free packages designed to address varying range of data analytics issues (Kasprzak, et al., archive). One of these packages is shiny. Shiny provides a framework for creating web-based interactive applications (Ramalho & Segundo, 2020). It can generalize R code to all levels of users, by simplifying the use of complex methodologies for people of different specialties, at the level of proficiency appropriate for the end users (Ramalho & Segundo, 2020). Shiny appears to have better graphics, and customizable visualization capacities that allows users to dynamically change visualizations by adjusting parameters in some controlled variables using buttons, selection list or by direct clicks on the graphs (Ramalho & Segundo, 2020). Since R shiny web application is free and open source, it eliminates the burden of purchasing proprietary software which can be inflexible and expensive and often takes resources away from research (LaZerte et al., 2017). Shiny applications can easily be interfaced with other software through API, as well as easy code adaptation (LaZerte et al., 2017).

The above important features of shiny and the flexibility of use, formed our decision in adopting shiny to build and develop R-HyperSpectral web application. The design of R-HyperSpectral utilized agile development approaches, including good architecture that allows the application logic to be broken down into smaller, independent parts, that are easier to maintain and verify. Also, included are testing the code, validating the data and app state, scaling, and performance. These are done to ensure quality assurance, validate the logic and the source code, including the test data, to minimize data quality issues, and the scalability and overall performance of the application. R-HyperSpectral drew functionality from similar applications like rHyperSpec (Laney, 2013), and followed best practices for shiny application development (Kasprzak, et al., archive).

The input data for R-HyperSpectral is currently available at EcoSIS, a useful tool for finding spectral data. There are currently at least 244, 966 spectral data at EcoSIS, contributed by over five organizations, including System Ecology Lab (SEL) with over 8,000 Spectra data contributions. The SEL spectra data serves as test case studies for the application and can be extended to data from other organizations in the future. During the development of R-HyperSpectral, we focused on key issues like usability, functionality, and users' perspective. These are based on feedback from presentations at lab meetings by peers, one on one with my research advisor and professors, and feedback from conferences. The plan is to launch it at scale as shiny apps are commonly run using multiple processes and servers.

2.2.5. Overview of R-HyperSpectral

R-HyperSpectral was built with R version 4.04, released in 2021 using the shiny package developed by RStudio, Inc. (<http://www.rstudio.com/shiny/>), which was first released in November of 2012. It aids in turning an analysis into an interactive web application without the knowledge of CSS, HTML, or JavaScript (<https://www.rstudio.com/products/shiny/>). Shiny is an R wrapper for JavaScript – an interpreted programming language. The implementation of JavaScript in web browsers aids in controlling the browser and its content, and in turn allows for interactivity with the user of the web browser (Laney, et al., 2013). *R-HyperSpectral* contains more than one thousand three hundred lines of code (>1300). Shiny web app is made up of two major parts, the User Interface (ui.R) and the server part (server.R) (<https://www.rstudio.com/products/shiny/>). The *ui.R* is made up of the user interface code. It also controls the loading of different libraries, for example, the following libraries are loaded on the background when R-HyperSpectral starts. They include: shinydashboard, shinycssloaders, lubridate, reshape2, plyr, readxl, highcharter, ggplot2, among others. The *server* (*server.R*) on the other hand contains all functions required for data selection, metadata, table of calculated indices, data visualization, and provide functionality behind user interface controls. The codes for the ui.R and server.R can run on personal computers or on servers. Figure 2.2 below shows the flow diagram or visual map of R-HyperSpectral.

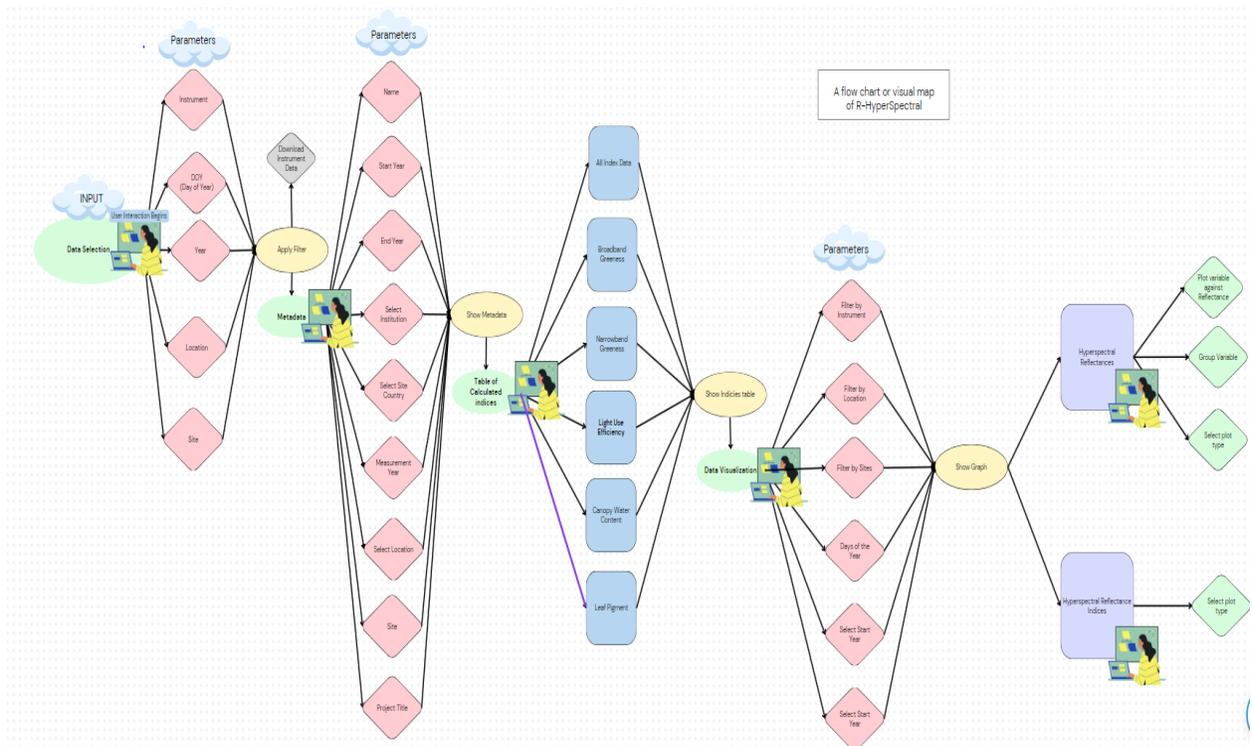


Figure 2.2: Flow chart or visual map of R-HyperSpectral

2.3 Results

2.3.1. R-HyperSpectral detailed Operation

When R-HyperSpectral is run, it opens on a web browser showing the *Introduction Page* that briefly gives users a general overview of the application, including its capabilities, limitations, data source description, system features, and how end users can interact with it. The *Data selection* tab allows users to select data based on the choice of instruments, locations, year, Day of the year (DOY), and/or site. Once the selection is complete, users are prompted to apply the filters they have selected, and a data table will be displayed. It is also possible to download the selected data at this point. Note that the data is connected to the application via an API from the EcoSiS website where hyperspectral reflectance data from System Ecology Lab (SEL) used in this study are archived and are publicly available.

R-HyperSpectral currently maps over 8,000 spectra spanning nearly eighteen observation sites across the ABoVE domain pertaining to multiple tundra vegetation species and communities. All these observations can be processed and analyzed within 7 minutes using R-HyperSpectral. This makes R-HyperSpectral a game changer as it presents a unique opportunity to the community as

a capable discovery tool for enhanced spectral analysis. It is important to highlight that this tool is not a replacement for desktop analysis using some sophisticated tools and approaches. R-HyperSpectral was built so that it can be scaled to accommodate more instruments, satellite sensors and platforms, allow for data from other regions other than the Arctic. This will require a slight modification of the underlying codes in an agile or version control environment like GitHub where the codes are currently housed (<https://github.com/SELDevTeam/R-HyperSpectral>). Figure 2.3 below shows a screen shot of R-HyperSpectral after a user clicks on Data selection tab. This allows the user to filter data by location, site, instruments, year, or Day of the year (DOY) and view a snapshot of the table of spectral data displayed based on the filters selected.

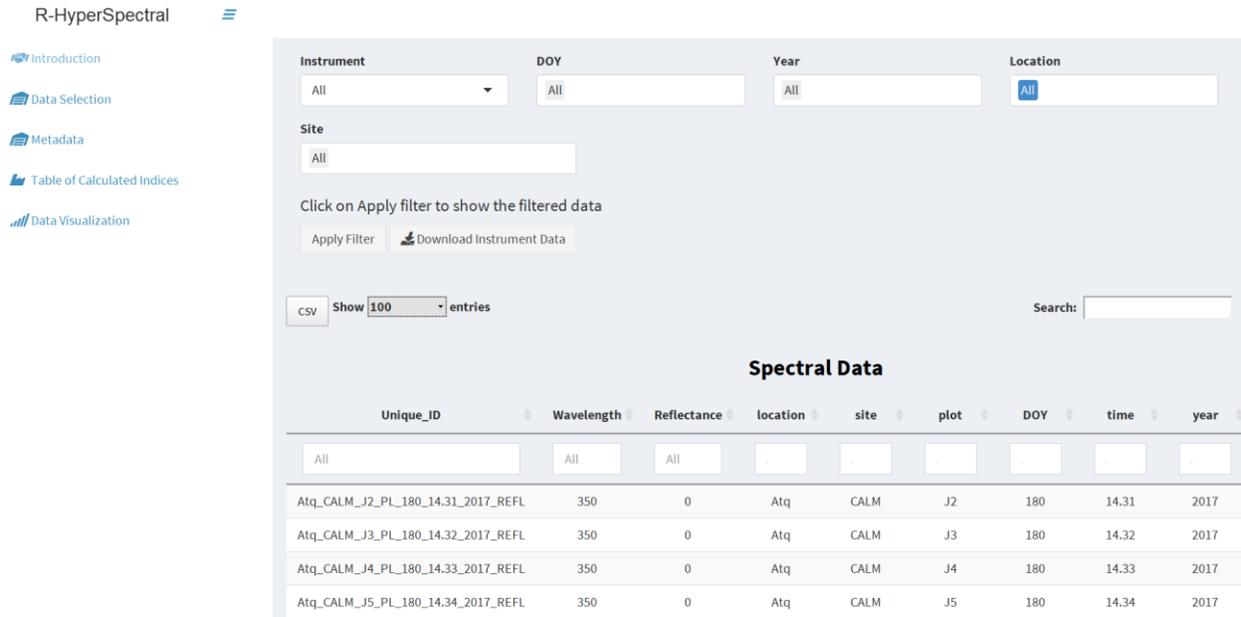


Figure 2.3: Table of selected data after filters are applied.

There is also a *Metadata* tab that links to the metadata of each project including the starting year of the project and the ending year. That metadata tab comes immediately after the data selection tab, so that users can interrogate the metadata to gain more understanding of the data. The metadata includes institution selection, the country where the site is located, measurement year, location, site name, and project title. Users can select from each of these based on available metadata and a metadata table will be displayed. Fig 3.2 below shows a metadata record based on user's selection.

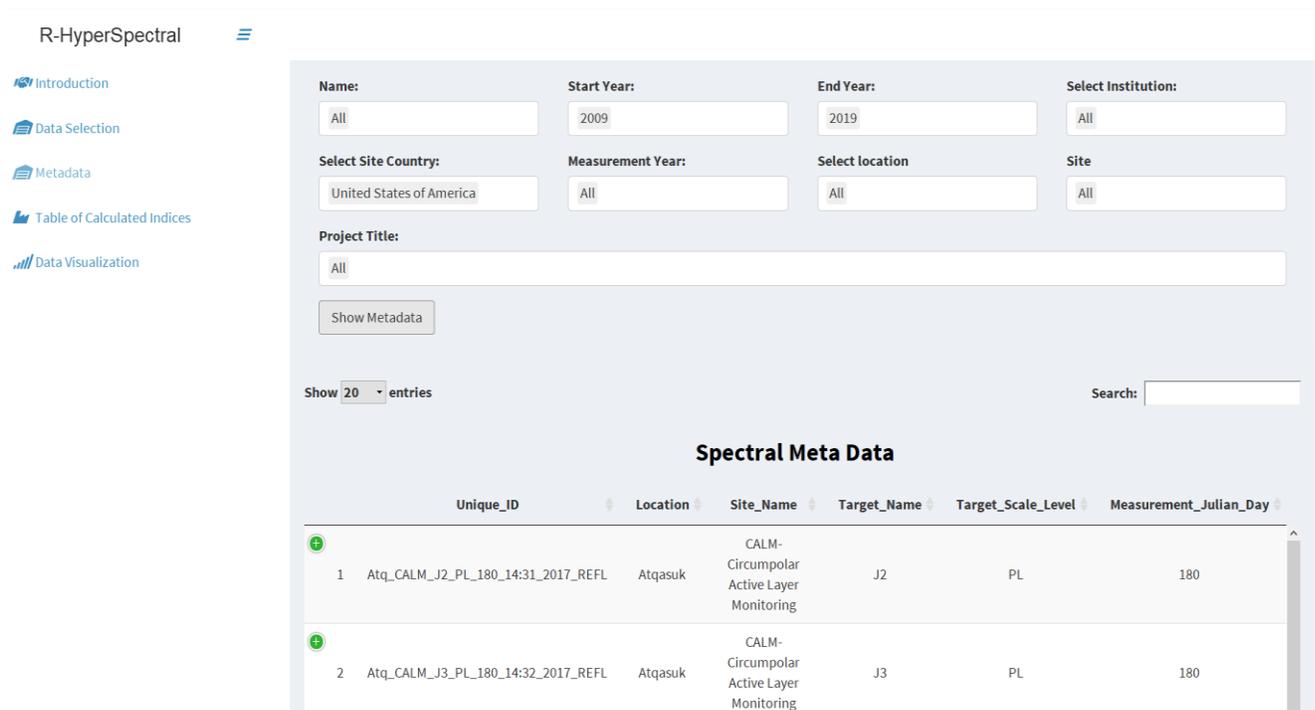


Figure 2.4 Data table showing metadata based on user's selection.

2.3.2. Table of Calculated Indices

The table of calculated Indices comes immediately after the metadata selection tab. The table of calculated indices displays all the spectral indices as calculated and printed by the R-HyperSpectral within a few seconds. Table 2.1 below shows all the spectral indices used in this application and their formular. The spectral indices are further classified into different categories based on their similarities. The All-Index data sub-tab prints all the SIs calculated by the application. The Broadband Greenness, Narrowband Greenness, Canopy Water Content, Leaf Pigment, and Light Use Efficiency are the sub-categories classified based on their similarities and functions. Figures 2.5a, 2.5b, 2.5c, 2.5d, and 2.5e below show the snapshot of the spectral indices as displayed from the application.

Table 2.1. Hyperspectral indices that are included in R-HyperSpectral, from a list generated by C. Laney and others from different websites. The Expression column gives the specific equation for each index, where R followed by a numeric value indicates reflectance for that specific wavelength. References are given where available.

Index Name	Index Abbreviation	Expression	Reference
Carotenoid 1	Cr1	$(1/R510)-(1/R550)$	Gitelson A.A. et al. 2002
Carotenoid 2	Cr2	$(1/R510)-(1/R700)$	Gitelson A.A. et al. 2002
Carter 1	Carter1	$R695/R760$	Carter, G.A. 1994
Carter 2	Carter2	$R695/R420$	Carter, G.A. 1994
Chlorophyll 1 A	Chl1a	$(R740^2)/(R675*R800)$	
Chlorophyll 1 B	Chl1b	$(R740^3)/(R675*R695*R800)$	
Chlorophyll 2 A	Chl2a	$(R685^2)/(R675*R800)$	
Chlorophyll 2B	Chl2b	$(R685^3)/(R675*R695*R800)$	
Curvature Index	Curvature	$(R675*R690)/(R683^2)$	Zarco-Tejada, P.J. et al. 2002
Datt 1	Datt1	$R860/(R708*R550)$	Datt, B. 1998
Gitelson 1	Gitelson1	$(R800-R700)/(R800+R700)$	Gitelson, A.A. & Merzlyak, M.N. 1994
Gitelson 2	Gitelson2	$(R750-R705)/(R750+R705)$	Gitelson, A.A. & Merzlyak, M.N. 1994
Gitelson 3	Gitelson3	$(R750-R445)/(R700-R445)$	Gitelson, A.A. et al. 2003
Gitelson 4	Gitelson4	$(1/R550)-(1/R750)$	Gitelson, A.A. et al. 2003
Gitelson 5	Gitelson5	$(1/R700)-(1/R800)$	Gitelson, A.A. et al. 2003
Greenness 1	Green1	$(R554/R675)$	
Modified Normalized Difference Vegetation Index	mndvi	$(R750-R705)/(R750+R705-2*R445)$	
Modified Simple Ratio	msr	$(R750-R445)/(R705-R445)$	Sims, D. A. & Gamon, J. A. 2002
Normalized Difference 1	Nd1	$(R682-R553)/(R682+R553)$	Gandia, S. et al. 2004
Normalized Difference 2	Nd2	$(R708-R546)/(R708+R546)$	

Normalized Difference 3	Nd3	$(R750-R705)/(R750+R705)$	Sims, D. A. & Gamon, J. A. 2002
	Ndswi_lin	$(R460-R1000)/(R460+R1000)$	
	Ndswi_log	$(\log(R1000)-\log(R460))/(\log(R1000)+\log(R460))$	
Normalized Difference Vegetation Index 1	Ndvi1	$(R800-R680)/(R800+R680)$	
Normalized Difference Vegetation Index 2	Ndvi2	$(R800-R667)/(R800+R667)$	
Normalized Difference Vegetation Index 3	Ndvi3	$(R750-R667)/(R750+R667)$	
Normalized Difference Vegetation Index 4	Ndvi4	$(R774-R677)/(R774+R677)$	
Optimized Soil Adjusted Vegetation Index	osavi	$1.16*(R800-R670)/(R800+R670+0.16)$	Rondeaux, G. et al 1996
Phytochrome 1	Phyt1	$R730/(R730+R652)$	
Phytochrome 2	Phyt2	$(R730-R652)/(R730+R652)$	
Phytochrome 3	Phyt3	$R724/(R724+R654)$	
Phytochrome 4	Phyt4	$(R724-R654)/(R724+R654)$	
Phytochrome 5	Phyt5	$R730/(R730+R666)$	
Phytochrome 6	Phyt6	$(R730-R666)/(R730+R666)$	
Photochemical Reflectance Index 1	Pri1	$(R531-R570)/(R531+R570)$	Gamon, J. et al. 1992
Photochemical Reflectance Index 2	Pri2	$(R530-R550)/(R530+R550)$	Sims, D. A. & Gamon, J. A. 2002
Photochemical Reflectance Index 3	Pri3	$(R531-R670)/(R531+R670)$	Gamon, J.A. et al. 1997
Photochemical Reflectance Index 4	Pri4	$(R531-R667)/(R531+R667)$	
Plant Sencence Reflectance Index	psri	$(R680-R500)/R750$	Merzlyak, M.N. et al. 1999
Reflectance Phytochrome	rphyto	$R730/(R730+R665)$	
RFFR 1	rffr1	$R730-R650$	
RFFR 2	rffr2	$(R730-R650)/(R685+R650)$	
RF Green	rfgreen	$R525-R550$	
RF Red	rfred	$R690-R650$	

RI	ri	$(R678-R667)/(R678+R667)$	
Structure Independent Pigment Index	sipi	$(R800-R450)/(R800-R650)$	Peñuelas, J. et al. 1995
Simple Ratio 01	sr01	R430/R762	
Simple Ratio 02	sr02	R550/R430	
Simple Ratio 03	sr03	R550/R650	
Simple Ratio 04	sr04	R672/R550	
Simple Ratio 05	sr05	R685/R655	
Simple Ratio 06	sr06	R690/R655	
Simple Ratio 07	sr07	R705/R715	
Simple Ratio 08	sr08	R705/R930	
Simple Ratio 09	sr09	R708/R545	
Simple Ratio 10	sr10	R750/R550	
Simple Ratio 11	sr11	R750/R700	
Simple Ratio 12	sr12	R750/R705	
Simple Ratio 13	sr13	R752/R690	
Simple Ratio 14	sr14	R775/R675	
Simple Ratio 15	sr15	R800/R650	
Simple Ratio 16	sr16	R800/R680	
Simple Ratio 17	sr17	R800/R750	
Simple Ratio 18	sr18	R860/R550	
Vogelman Red Edge 1	vog1	R740/R720	Vogelmann, J.E. et al. 1993
Vogelman Red Edge 2	Vog2	$(R734-R747)/(R715+R726)$	Vogelmann, J.E. et al. 1993
Vogelman Red Edge 3	Vog3	$(R734-R747)/(R715+R720)$	Vogelmann, J.E. et al. 1993
Water Band Index	wbi	R900/R970	Claudio, H. et al. 2006

Show **10** entries
 Search:

Unique_ID	instrument	location	year	site
Atq_CALM_E2_PL_159_11.54_2013_REFL	unispec	Atq	2013	CALM
Atq_CALM_E2_PL_164_12.14_2016_REFL	unispec	Atq	2016	CALM
Atq_CALM_E2_PL_165_14.29_2015_REFL	unispec	Atq	2015	CALM
Atq_CALM_E2_PL_168_10.46_2011_REFL	unispec	Atq	2011	CALM
Atq_CALM_E2_PL_170_13.58_2014_REFL	unispec	Atq	2014	CALM
Atq_CALM_E2_PL_172_12.10_2013_REFL	unispec	Atq	2013	CALM
Atq_CALM_E2_PL_174_11.06_2012_REFL	unispec	Atq	2012	CALM
Atq_CALM_E2_PL_176_14.34_2016_REFL	unispec	Atq	2016	CALM
Atq_CALM_E2_PL_180_15.00_2017_REFL	unispec	Atq	2017	CALM
Atq_CALM_E2_PL_181_16.40_2018_REFL	unispec	Atq	2018	CALM

Showing 1 to 10 of 916 entries
 Previous

...

Next

Figure 2.5a: Table of calculated indices for all indices (showing a snapshot).

All Index data Broadband Greenness Narrowband Greenness Canopy Water Content
 Leaf Pigment Light Use Efficiency

CSV Show 10 entries Search:

year	site	cri1	cri2	Carter1	Cart
2013	CALM	3.13088706236221	11.3655337708164	0.707835900148963	3.6142217760
2016	CALM	4.8038253412646	12.5707376328027	0.631183142679027	6.3935304896
2015	CALM	5.9234284240442	14.1348535133525	0.57947992369477	3.64718400710
2011	CALM	3.47646912671044	11.5775283157002	0.768855690424318	2.49528619528
2014	CALM	3.22049350321153	10.4992468526274	0.661901562245845	3.9005298117
2013	CALM	7.00655161429035	13.4610261240918	0.392955202310757	4.06893479458
2012	CALM	3.33779039276835	8.6469181370167	0.549547608789315	2.11350455679
2016	CALM	5.73847710794725	12.0970973712548	0.40466389385271	3.10714920990
2017	CALM	8.26523367052557	9.01914666805442	0.462745514585152	4.14879375300
2018	CALM	2.61001428223293	10.514473187139	0.598410708864974	2.6429306620

Showing 1 to 10 of 916 entries Previous 1 2 3 4 5 ... 92 Next

Figure 2.5b: Table of calculated indices for all indices (showing a snapshot). See table 2.1 for the full names and the expressions of each of the SI used.

All Index data Broadband Greenness Narrowband Greenness Canopy Water Content

Leaf Pigment Light Use Efficiency

CSV Show 10 entries Search:

ID	chl1a	chl1b	chl2a	chl2b	cu
79	1.31755418920255	1.76219329778487	0.607970968441745	0.552363633973668	1.025502
74	1.5307075114571	2.36268382811319	0.508663455692397	0.45259848378625	1.032697
09	1.63887299934977	2.7225494893395	0.475100356425452	0.424948075659362	1.014682
19	1.19414849518962	1.52172964357992	0.615831571348085	0.563563583441197	1.007126
37	1.41625707491892	2.05174994050004	0.54406397422184	0.488524383431165	1.036085
22	3.18117511003024	7.66133933556196	0.302997409169635	0.22520690767238	1.10802
28	1.819919782373	3.12368434618588	0.458206928176366	0.394621960487447	1.036240
07	2.52249918129652	5.748704732555	0.322335310226502	0.262594288747997	1.087423
13	3.40816451296596	8.76793841438729	0.321647989907392	0.254207247403994	1.085549
95	1.50984170055635	2.31112614982159	0.464071230639373	0.393825194564069	1.020262

Showing 1 to 10 of 916 entries Previous 1 2 3 4 5 ... 92 Next

Figure 2.5c: Table of calculated indices for all indices (Showing a snapshot). See table 2.1 for the full names and expressions of each of the SI used.

All Index data					
Broadband Greenness		Narrowband Greenness		Canopy Water Content	
Leaf Pigment		Light Use Efficiency			
CSV	Show 10 entries	Search: <input type="text"/>			
ndvi1	ndvi2	ndvi3	ndvi4		
0.286405378121625	0.339741442516984	0.305212256775207	0.262891163070122	0.210588	
0.35401600292472	0.352994473644416	0.373971917846978	0.330579969971342	0.264469	
0.381975738643437	0.299551826256463	0.396481070255301	0.361864425040181	0.250687	
0.269313363801553	0.275178184071893	0.289911029041464	0.240313679930293	0.175701	
0.32336943147247	0.348538771005592	0.33956604054932	0.299170647280801	0.246253	
0.594119667178263	0.393601559948276	0.61094873517185	0.58881243363056	0.48403	
0.418910068594671	0.301787698212302	0.438546217019046	0.402302168657309	0.324812	
0.547145471726429	0.351450873762548	0.559680983760283	0.5356959632468	0.425398	
0.506887445574395	0.179405794290955	0.499626281980501	0.522864426802492	0.342432	
0.437434495026958	0.400040881022622	0.477539581558735	0.396895761760789	0.37152	

Figure 2.5d: Table of calculated vegetation indices for all indices classified under broadband greenness. See table 2.1 for the full names and expressions of the SI used.

All Index data Broadband Greenness **Narrowband Greenness** Canopy Water Content

Leaf Pigment Light Use Efficiency

CSV Show **10** entries Search:

instrument	location	vog1	vog2	vog3
unispec	Atq	1.10419472427378	-0.0256788834484218	-0.026129763821804
unispec	Atq	1.14344329923344	-0.0287050426126082	-0.0295450522065592
unispec	Atq	1.20641042214274	-0.0450621705265974	-0.0466721424739088
unispec	Atq	1.07980791218843	-0.0210808957952468	-0.0213603473227207
unispec	Atq	1.12465640626414	-0.0265933794805215	-0.0272090959840037
unispec	Atq	1.23322252469644	-0.0398866495658012	-0.041938727365488
unispec	Atq	1.14853433012001	-0.025262123197903	-0.0261187709610759
unispec	Atq	1.25936697419724	-0.045668307134137	-0.048217076915024
unispec	Atq	1.47321945899643	-0.0919260375599776	-0.0970355283512125
unispec	Atq	1.10595231761458	-0.0275167179681364	-0.028152389000302

< >

Figure 2.5e: Table of calculated vegetation indices for all indices classified under narrowband greenness. See table 2.1 for the full names and expressions of each of the SI used.

2.3.3. Data Visualization

After filtering and viewing the table of calculated indices, users can then proceed to the data visualization tab. Here, users can have options to visualize hyperspectral reflectance data or SIs. Before clicking on the “show graph”, that allows users to see the visuals, they can filter the data by location, instruments, year, and site. It is also possible for users to choose a visualization type from a selection of plots that include options for a heatmap, boxplot, and/or regression scatter plot. These plots will enable users with more insights especially for the vegetation indices and appeals users interviewed from different user communities who have varied preferences for data

visualization. Fig.2.6 below shows a boxplot visualization for the selection of hyperspectral reflectance with all instruments, locations, sites, start and end year of 2017 and 2018, respectively, selected. This will allow the user community to see the values of reflectances from different locations, including the minimum, mean, and maximum values from each location. The expected production-ready plots will be made from desktop applications with customizations and annotation as required. The YouTube link below is the general overview of how users can interact with R-HyperSpectral: <https://youtu.be/3RoxSfByPhI>

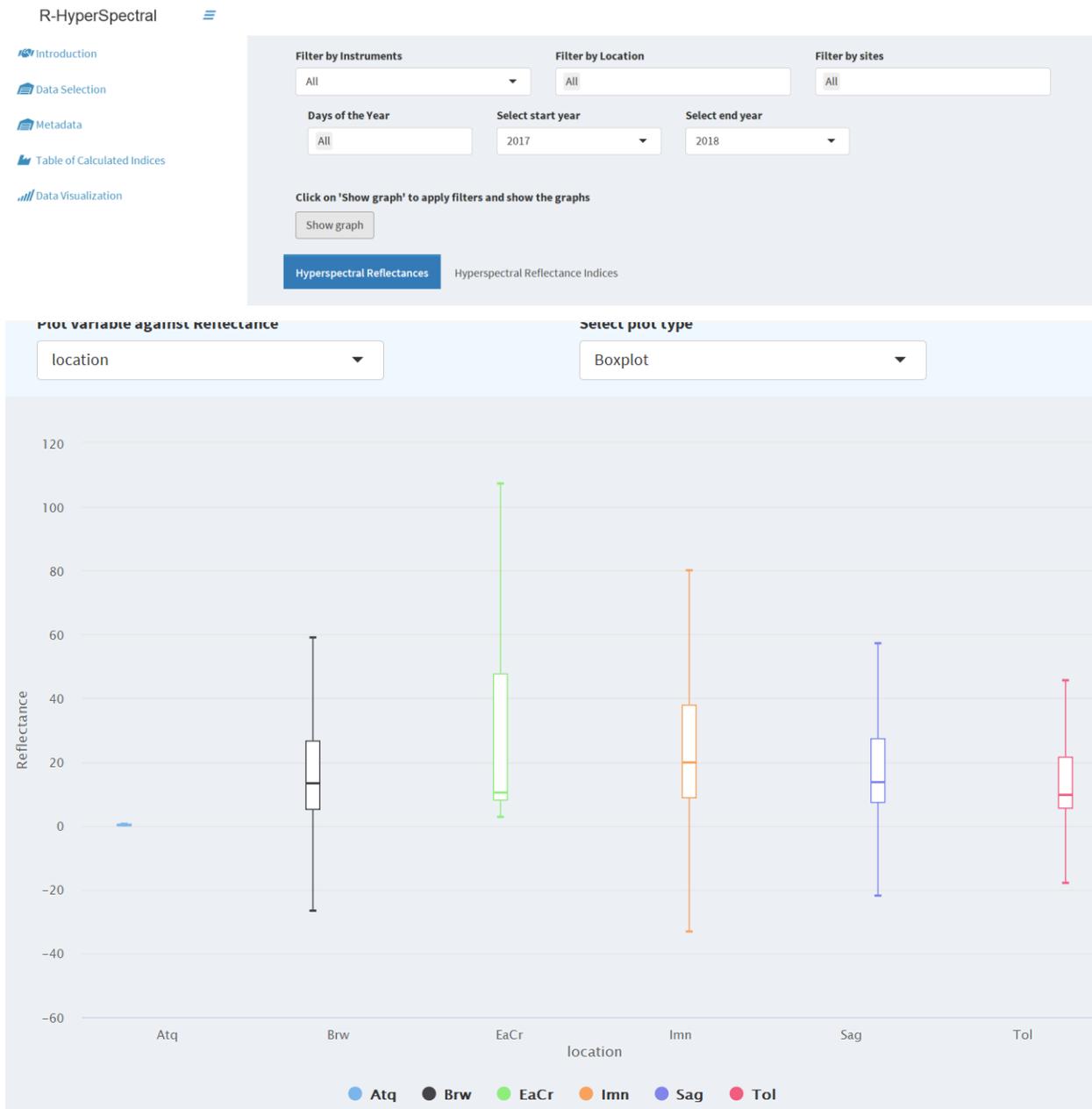


Figure 2.6: The boxplot of the visualization by location of raw reflectance (where Atq, Brw, EACr, Imn, Sag, and Tol are Atqasuk, Barrow, Eagle Creek, Imnaviat, Sagwon, and Toolik).

2.4. Discussion

The development of R-HyperSpectral attempts to present a custom and shareable analytic and open-source multi-user collaborative tool for hyperspectral data with an initial focus on Arctic terrestrial and aquatic ecosystems using the SEL data as a test case study. This will not only help improve the utilization of an extensive spectral library by the NASA ABoVE researchers, but also enable the user community to have access to an application that permits an advanced search, and filtering of spectral records, calculate a range of spectral indices, and enhance advanced visualizations. R-HyperSpectral will aid the Arctic research community to have access to data for improved rapid change detection across the ABoVE domain. R-HyperSpectral allows users to view, interact, and discover optical properties of Arctic tundra plant communities at different locations such as Barrow, Aquasak, Toolik Lake, etc. Users can view the hyperspectral reflectance scans, explore common spectral indices, and visualize their results from a choice of plotting methods. The features of R-HyperSpectral include the data selection tab that enables users select hyperspectral reflectance data based on the instruments, location, sites, year, among others. There is also the metadata tab that shows the metadata including the researcher's name, start and end year, institution, and project title, just to mention a few. Others include data visualization that allows users to visualize either raw reflectance or hyperspectral indices, and table of calculated indices. Furthermore, the table of calculated indices gives users the opportunity to view the hyperspectral reflectance scans and explore common spectral indices at temporal scales.

The main difference of R-HyperSpectral compared to other comparable tools is the generation and visualization of SIs within a very short time, usually less than a minute, and further groups them based on categories that calculate similar properties. Spectral indices are designed to enhance spectral contributions from desired vegetation characteristics, while minimizing the interference of other factors (Xue, et al., 2017). Most importantly, SIs are calculated from a discrete number of reflectance bands linked to specific plant traits and are applied in developing proxies for traits of interest and reflectance properties (Sims & Gamon, 2003). Spectral Indices (SI) obtained from remote sensing of canopies are quite simple and effective algorithms for quantitative and qualitative evaluations of vegetation cover, vigor, and growth dynamics, among

other applications (Xue, et al., 2017). The importance of remote sensed information of terrestrial vegetation growth, vigor, and dynamics in providing useful insights for application in environmental monitoring, biodiversity conservation, and agriculture cannot be overemphasized (Xue, et al., 2017). Hence, the importance of having an application that can easily calculate, produce, and visualize these SIs. Depending on the SI, information on various aspects of vegetation growth and development such as chlorophyll content, canopy structure, leaf area, and water content can be monitored in the Arctic region. Table 2.1 above shows all the spectral indices that can be produced by R-HyperSpectral.

In continuation, the metadata tab allows users to gather all the relevant and important information about the who, what, when, where, and by whom for spectral data collection, which allows to access pertinent information for tracing data provenance, and context. Metadata is essential for both discovery and validation of data (Leipzig, et al., 2021). Metadata has emerged as important component for supporting the federal and institutional recommendations, that mandates to build a sustainable research infrastructure to support FAIR data initiatives and produce reproducible research (Margolis, et al., 2014, Brito, et al., 2020, & Wilkinson, et al., 2016) with standards supporting research life cycle, especially in ecological research. Hence, the inclusion in R-HyperSpectral. The development of this web-enabled spectral library application using open-source technology will enable wider data archival and sharing among wider ecological and remote sensing communities, thereby supporting the initiative above. All these components of the app are strategically included to enhance data sharing, accessibility, availability, and further allow users understand how these data were collected, who collected them, the measurement instruments used and their calibration among other vital information about the data, the metadata field reveals.

The data for this application is housed in EcoSIS (Ecological Spectral Information System) and was used as a case study for building and developing the web app. EcoSIS is a useful tool for finding spectral data (<https://ecosis.org/>) and is connected to the application via an API, this opens the application to be continuously enriched with more hyperspectral data through a slight modification of the underlying codes. This will further open up the possibility for researchers collecting similar data to leverage the application to view, archive, share, and analyze their spectral data. Importantly, the SEL data in EcoSIS is for testing purposes, in the future, data from

other regions will be included in the application. This will help bridge the gap for the need for open access to data and open-source software in the remote sensing community and the field of ecology at large. To the best of our knowledge, there is no other tool with potentially greater access or functionality for similar use at present. R-HyperSpectral will ensure that researchers in the community that require similar software for their data will have a template to build on or modify R-HyperSpectral to meet their need thereby increasing the capacity of R-HyperSpectral and sharing same to the community. If R-HyperSpectral is modified by the researchers, they would be a proper channel in place to verify that the code works as intended, with a thorough validation of the scientific issues the code is intended to solve.

Future work would require improvement that enhances more user experience that is particularly user driven with better visualization options. Expanding the software to accommodate data from other research groups collecting similar data but from a different region or instruments would be a central focus in the future. At the moment, R-HyperSpectral inputs its data from EcoSIS which contains spectral data only, expanding the utility of the application to further include data from more instruments, such as the Airborne Visible and Infrared Imaging Spectrometer [AVIRIS] by (Carlson, et al., 2007), including data from satellite platforms and sensors such as Landsat and MODIS provides a foundation for improving the analysis, and modeling capabilities needed to understand and predict ecosystem responses and societal implications. Secondly, by linking R-HyperSpectral to these platforms and sensors, and even other regions, other than the Arctic, would help the research community with increased access to data and help speed up ecological research in these areas. It would be important to initiate these future improvements using a code repository like GitHub (<https://github.com>) for versioning and tracking purpose. See the GitHub link provided above for access to R-HyperSpectral underlying codes.

The future plan is to utilize Rstudio Connect in combination with Amazon Web Services (AWS) in hosting the application to ensure easy access and availability as well as help eliminate any issues related to data security as AWS provides a secure platform for hosting such applications.

2.5. Conclusion

We have demonstrated that R-HyperSpectral as a software application has the capability to analyze, visualize, and integrates diverse hyperspectral data streams and give the research community access to data for improved rapid change detection across the ABoVE domain. It can

also calculate spectral indices and classify indices based on their similarities. With R-HyperSpectral written in R Programming language that several ecologists and researchers find more convenient to use in their data analytics task, it will be relatively easy for future improvements to be initiated by ecologists instead of expert programmers. Furthermore, since R-HyperSpectral is web-enabled, researchers, ecologists, and academicians that work with hyperspectral data may be encouraged to incorporate it in the analysis, visualization, and contribute towards improving the overall working of the application for efficient and effective spectral data discovery.

Chapter 3: *rDataFusion*: A Project-Specific Multi-Data Fusion Tool for Discovering, Integrating, and Visualizing Heterogenous Long-term Data Sets

Abstract

To understand ecosystem change over a range of spatial and temporal scales and levels of biological organization and interaction, multiple streams of ecological data need to be collected, integrated, and analyzed. However, due to the size and complexity of these data streams and many other challenges (e.g., personnel turnover, methodological changes, and gaps in observing records), managing, analyzing, sharing, and visualization of these data has posed a significant challenge. To resolve these challenges, we developed a multi-data fusion tool called *rDataFusion*, which is capable of aggregating heterogeneous data sets collected from a range of automated and semi-automated sensors and manual observations over a decade-long period. *rDataFusion* is developed using a free, open-source software package in R called *shiny*. *rDataFusion*, has the capability to integrate and filter data from two instrument nodes and different data streams that include micro-meteorological variables (e.g., temperature, relative humidity), soil conditions (e.g., temperature and soil moisture), and ecosystem trace gas and energy fluxes. After initial compilation and filtering, users can visualize data in near real-time to check that all sensors are running properly, and/or ensure preliminary flagging for data that is deemed out of range or problematic in some way. *rDataFusion*, also, has the capacity for exploratory data analysis through quality control and quality assurance processes that allow for identifying missing values, outliers, and gap-filling missing or problematic data, visualize data to allow for preliminary summaries and interpretations, and compare data across time or by site. The overarching goal is to develop a customizable analytic tool that aids researchers with improved capacities for aggregating different streams of data from a single intensive site by providing an open-source multi-data fusion tool that facilitates data management, sharing, and analysis and serves as a template for other research groups with similar challenges.

3.1. Introduction

The changes that are occurring in ecology create challenges with respect to gathering, managing, analyzing, and visualizing large volumes of data collected from different instruments and sensors across different research groups. One particularly daunting challenge lies in dealing with the scope and enormous variability and veracity of these datasets (Michener, et al., 2012 & Farley, et

al., 2018). The diversity in the spatiotemporal scale of a given study and the different ways in which these studies are carried out result in large number of small, distinctive datasets that accumulate from thousands of scientists that collect relevant ecological and environmental datasets (Heidorn, 2008 & Michener, et al., 2012). Such heterogeneity can be attributed, in part, to methodological specialization to address specific scientific hypotheses, but also to a lack of standard protocols for organizing, managing, storing, discovering, integrating, accessing, and curating data from different small, lowly funded academic research groups (Laney, et al., 2022). Unfortunately, only a small fraction of ecological data collected is readily discoverable and accessible due to lack or no ecological or environmental data integration tool to assist in organizing and managing data collected by these academic researchers (Michener, et al., 2012).

Data loss is a big challenge for both “long tail science”, where many small research groups that do not have access to a deal of funding sources other than through public funding each produce a small body of knowledge that collectively make a large contribution to the science; and ‘big science’, where a few large research groups produce large and complex data sets as a product of strong public support and funding (Howe et al. 2011., Laney, et al., 2015, & Latif, et al., 2019). The need for a data integration software and information system tool that can improve efficiency and ameliorate the difficulty of data management and analysis tasks cannot be overemphasized (Recknagel, 2011, Laney, 2013). It is important to note that new networks such as DataOne, National Ecological Observatory Networks (NEON), and the US Long Term Ecological Research Network (LTER), among others, have been forming to promote ecological information management standards and tools (Collins et al. 2011, Michener et al. 2012, National Ecological Observatory Network 2013, McCord et al., 2021).

Of these networks mentioned above, DataOne has been integral in creating data management approach using the data life cycle. The DataOne data life cycle describes the process of data management procedures starting from planning, collecting, assuring, describing, preserving, discovering, integrating, and lastly, analyzing (McCord, et al., 2021). The DataOne lifecycle acts as a standard organizational structure of how data moves through different stages (Michener, 2015 & McCord, et al., 2021). It further provides an avenue that can prevent data loss through broader collaboration, shared repositories, use and reuse, and data integration. Interestingly, the DataOne data life cycle was developed in a time when ecological data integration, management,

and sharing was still nascent, and failed to capture some of the critical data management and integration processes that pose a great challenge to long tail ecological scientist (McCord, et al., 2021). It is also pertinent to note that, the old method in which ecologist rely on spread sheets, emails, hard drives, among others, for data sharing and management is not sustainable and could potentially lead to loss and under-utilization of data (Michener et al., 2012). Interestingly, statistical computing software tools support robust data management, by integrating quality assurance and quality control (QAQC), including data and metadata management and integration and data analysis workflows systems to allow for properly documented data to be easily accessed for data and metadata completeness and quality (Laney, 2013).

However, the challenge of effective data management, integration, visualization, curation, use, re-use, and analysis for large-scale synthesis studies in ecology is still profound among the long tail scientist or small-scale research labs (Laney, 2013 & Michener et al. 2012). Due to advances in technology, ecological data are collected through several automated means, including field instrumentations, satellite and aerial platforms, automated and semi-automated sensor networks (Collins, S.L. et al., 2006; Porter, J.H. et al., 2009 & Michener et al., 2012). All these data from disparate sources when combined will generate terabytes to petabytes of data annually.

Integrating and managing data from these sources for a small lab group is time consuming and labor intensive as it requires understanding the methodological differences, transforming data into a common representation, and converting data to a compatible semantics before analysis can begin (Michener, et al., 2012). Furthermore, ecological data show high variability with highly complex interactions that makes it tedious and difficult to analyze, manage, and integrate (Michener, et al., 2012). It is also pertinent to note that long tail ecological scientists have traditionally used Excel to manipulate and convert their data for integration and analysis; however, these methods and processes are always error-prone and is not reproducible because of lack of provenance regarding its operations, good practices of documentations, and metadata (Michener, et al., 2012).

Statistical software and scripting tools have been shown to be highly effective in data analytics and management in Ecology and other fields and disciplines such as Health Sciences, Social Sciences, Engineering, Agriculture and Medicine, among others. In the field of Agriculture, important data management decisions makings have been aided by the use of statistical software

tools (Perakis, et al., 2020 & Krisnawijaya, et al., 2022). It is the same story in the field of Medicine and Health Sciences, including Community Medicine, where data analytics software tools have aided researchers to bridge the gap between data generation and analysis to extrapolate meaningful results and conclusions (Joshi, et al., 2021). These tools aid the researchers even without an in-depth knowledge of statistics to analyze and manage their research data (Joshi, et al., 2021).

Statistical software and scripting tools such as R and a package in R called shiny – a free, open-source statistical software application and programming language that has the capacity to execute complex statistical analysis, produce sophisticated and customizable scientific visualizations and can interface with databases, clouds, and data archives (Laney, 2013; Wanyanhan, et al, 2022; Chang, et al, 2020, & Chang, W., 2018) is a very good example. These tools bring to bear a set of approaches that explicitly encode the semantics of observational data to automate or semi-automate the process of data integration, management, and analysis (Wanyanhan, et al, 2022 & Laney, 2013). These tools allow for data analytic and management approaches to be built from bottom up to streamline the process (Laney, 2013). They can help to foster the culture of data sharing, integration, synthesis, and automate the documentation of data workflows (Farley, et al., 2018). They also help provide solutions to previously unanswered research questions and provide avenues for training a new generation of ecological data scientists (Farley, et al., 2018).

It is important to reiterate challenges ecologists face with respect to gathering, managing, analyzing, and visualizing large volumes of data collected from different instruments and sensors across different research group due to lack or low deployment and use of software tools for long tail science. The daunting task of dealing with the scope and enormous variability and veracity of these datasets (Michener, et al., 2012) persists among long tail ecological scientists. The varied diversity in scales studied and the different ways in which these studies are carried out result in large number of small, distinctive datasets that accumulate from thousands of scientists that collect relevant ecological and environmental datasets (Heidorn, 2008 & Michener, et al., 2012) portend a great challenge to manage and analyze by long tail scientists without adequate deployment of software analytical tools. These challenges, among others, spurred us to develop a data integration and management tool called rDataFusion.

In this study, we introduce rDataFusion; aimed at developing a customizable analytic tool that aids researchers with improved capacities for aggregating different streams of data from a single intensive site by providing an open-source multi-data fusion tool that facilitates data management, sharing, integration, curation, visualization, and analysis and serves as a template for other research groups with similar challenges. The System Ecology Lab (SEL) manages a site at the Jornada Experimental Range (JER) where it amasses a heterogeneous amount of data from different instruments and sensors over a long period of time, and is faced with the challenge of aggregating, analyzing, curating, visualizing, and managing these huge amounts of data. rDataFusion offers capacities to visualize data in near real-time to check that all sensors are running properly, ensure preliminary flagging for data that is deemed out of range, filter data based on quality flags, and align data with that from other sensors. It also employs statistical algorithms to filter extreme values and outliers, gap-fill with sensors at the site, visualize data to allow for some preliminary summaries and interpretations. The rest of this paper is organized as follows: Section 2 shows the study sites and methods, section 3 discusses the results, while section 4 presents the discussion and conclusion.

3.2 Materials and Method

3.2.1 Study Site

The study site was built in 2009 and became operational in 2010, with the overarching goal of studying global change science in arid ecosystems, with special focus on changes and feedback cycles in land cover, hydrology, and land – atmosphere exchange of water, carbon, and energy. Since its inception in 2009, the site has been supported by more than ten research grants and has included over 100 grad students, post docs, and technicians who have explored a range of research topics and questions. This has included several theses and dissertations including, “Furthering our understanding and scaling patterns and controls of land – atmosphere carbon, water and energy exchange in the Chihuahuan desert shrubland with novel cyberinfrastructure” (Jaime, 2014), “Towards new data and information management solutions for data – intensive ecological research” (Laney, 2013), “Assessing data quality in a sensor network for environmental monitoring” (Ramirez, 2011), “Spatiotemporal variability of plant phenology in drylands: A case study of the Northern Chihuahuan desert” (Luna, 2016), “Development of low

cost network of webcams for monitoring plant phenology in Chihuahuan desert” (Gonzalez, 2011), among others.

The SEL-Jornada site lies within the United States Department of Agriculture (USDA) Agricultural Research Services (ARS) in southern New Mexico (32° 34' 59'' N, 106° 37' 34'' W; 1417 m a.s.l, Figure 1). The site is a shrubland with a mixture of *Larrea tridentatae* (Creosote) and *Prosopis glandulosa* (Honey Mesquite) that is typical of the northern Chihuahuan Desert (Laney, 2013). Other notable species available in the area are *Flourensia cernua* (Tarbush), *Muhlenbergia porter* (Bush Muhly) and *Dasyochloa pulchella* (Fluffgrass). It is characterized by a shallow sandy to gravelly soil that is generally less than 1m in depth. The study site slopes westward by approximately 2° from east to west. The long-term average rainfall at the JER Headquarters (approximately 13km from SEL-Jornada) was 245.1 mm from 1915 to 1995, with a standard deviation of 85.0 mm (Wainwright, 2006., Laney, 2013). A large portion of annual precipitation occurs mostly during the summer monsoon season (40-50% on average, 5.7 cm regionally from 1910 to 2010) (Petrie et al 2014). It is also important to note that this region exhibits an out of phase interaction between spring and summer growing seasons, where precipitation events induce pulses of vegetation productivity, nutrient cycling, and fluxes of water and carbon differently between spring and summer seasons (Petrie, et al., 2014).



Figure 3.1: Map showing the Jornada study site at the Chihuahuan Desert (Ramirez, 2011). The green, yellow, and brown colors represent the United States, Mexico, and Jornada site, part of the Map, respectively.

3.2.2 Data Collection

Site infrastructure include instrumentation such as: Extended Open Path Eddy Covariance System – A 10m tall tower hosting an open path eddy covariance system was designed to measure the land – atmosphere flux exchange, and provides digital output of carbon dioxide, density, sensible heat, temperature, humidity, net radiation, horizontal wind speed, and direction among others, (Laney, 2013 Streams of data like Climate, Soil Moisture, Flux, and cs650 for this data integration tool were all collected from the study site using different instruments and sensors. These datasets provide the capacity for studying seasonal variabilities of ecosystems changes and to accurately estimate the land – atmosphere fluxes of carbon dioxide, water vapor, and energy. Tables 3.1, 3.2, 3.3, and 3.4, below show variable abbreviations, full variable names and the SI units for Climate, soil moisture, Flux, and cs650 data streams, respectively.

Table 3.1: showing the abbreviated variable names, full variable names, and the units for the climate data stream.

S/N	Abbreviated Variable Name	Full Variable Name	SI Units
1	t_hmp	Air Temperature	°C
2	rh_hmp	Relative Humidity	Percent
3	e_hmp	Absolute Humidity	kPa
4	atm_pressure	Atmospheric pressure	kPa
5	hor_wnd_spd	Horizontal Wind Speed	M/S
6	hor_wnd_dir	Horizontal Wind Direction	Degree
7	Precip_tot	Total Precipitation	mm
8	Par	Photosynthetically Active Radiation	umol/m/s
9	albedo	albedo	unitless
10	LEAF_WET	Leaf Wetness	mV
11	NetRs	Net Solar Radiation	W/m ²
12	NetRI	Net Radiation	W/m ²
13	UpTot	total downwelling; upward facing sensor	W/m ²
14	DnTot	total upwelling; downward facing sensor	W/m ²
15	CO2_raw	Carbon dioxide	mmol/m ³
16	H2O_raw	Water	mmol/m ³

Table 3.2: showing the abbreviated variable names, full variable names, and the units for the Flux data stream.

S/N	Abbreviated Variable Name	Full Variable Name	SI Units
1	HS	sensible heat flux using sonic temperature	W/m ²
2	H	Sensible heat flux using the fine wire thermocouple	W/m ²
3	Fc_wpl	Carbon dioxide flux (LI-7500)	mg/(m ² s)
4	LE_wpl	W/m ² Latent heat flux (LI-7500)	W/m ²
5	Hc	Sensible heat flux computed from Hs and LE_wpl	W/m ²
6	tau	Momentum flux	kg/(m s ²)
7	u_star	Friction velocity	m/s
8	Ts_mean	Average Sonic Temperature	C
9	stdev_Ts	Standard deviation sonic temperature	C

10	cov_Ts_Ux	Covariance of sonic temperature and horizontal wind (x-axis)	m C/s
11	cov_Ts_Uy	Covariance of sonic temperature and horizontal wind (y-axis)	m C/s
12	cov_Ts_Uz	Covariance of sonic temperature and vertical wind	m C/s
13	CO2_mean	Average Carbon dioxide	mg/m ³
14	stdev_CO2	Standard deviation of carbon dioxide	mg/m ³
15	cov_CO2_Ux	Covariance of carbon dioxide (LI-7500) density and horizontal wind (x-axis)	mg/(m ² s)
16	cov_CO2_Uy	Covariance of carbon dioxide (LI-7500) density and horizontal wind (y-axis)	mg/(m ² s)
17	cov_CO2_Uz	Covariance of carbon dioxide (LI-7500) density and vertical wind	mg/(m ² s)
18	H2O_Avg	Average water vapor (LI-7500) density	g/(m ² s)
19	stdev_H2O	Standard deviation of water vapor (LI-7500) density	g/m ³
20	cov_H2O_Ux	Covariance of water vapor (LI-7500) density and horizontal wind (x-axis)	g/m ³
21	cov_H2O_Uy	Covariance of water vapor (LI-7500) density and horizontal wind (y-axis)	g/(m ² s)
22	cov_H2O_Uz	Covariance of water vapor (LI-7500) density and vertical wind	g/(m ² s)
23	fw_Avg	Average finewire temperature	C
24	stdev_fw	Standard deviation of finewire temperature	C
25	cov_fw_Ux	Covariance finewire temperature and horizontal wind (x-axis)	m C/s
26	cov_fw_Uy	Covariance finewire temperature and horizontal wind (y-axis)	m C/s
27	cov_fw_Uz	Covariance finewire temperature and vertical wind	m C/s
28	Ux_Avg	Average Horizontal wind (x-axis)	m/s

28	stdev_Ux	Standard deviation of average horizontal wind (x-axis)	m/s
30	cov_Ux_Uy	Covariance of horizontal wind (x-axis and y-axis)	(m/s) ²
31	cov_Ux_Uz	Covariance of horizontal wind (x-axis) and vertical wind	(m/s) ²
32	Uy_Avg	Average horizontal wind (y-axis)	m/s
33	stdev_Uy	Standard deviation of horizontal wind (y-axis)	m/s
34	cov_Uy_Uz	Covariance horizontal wind (y-axis) and vertical wind	(m/s) ²
35	Uz_Avg	Average vertical wind	m/s
36	stdev_Uz	Standard deviation of vertical wind	m/s
37	press_Avg	Average Barometric pressure (LI-7500)	kPa
38	Atm_press_mean	Average Barometric pressure (CS105)	kPa
39	T_hmp_mean	Average temperature from HMP45C	C
40	H2O_hmp_mean	Mean HMP45C vapor density	kg/m ³
41	Rh_hmp_mean	Mean HMP45C relative humidity	percent
42	Rho_a_mean	Mean air density	kg/m ³
43	Wnd_dir_compass	Resultant wind direction using compass coordinate system	degrees
44	Wnd_dir_csat3	Resultant wind direction the CSAT3's right-handed coordinate system	degrees
45	Wnd_spd	Wind speed	m/s
46	Rslt_wnd_spd	Resultant wind speed	m/s
47	Std_wnd_dir	Standard deviation of wind direction	degrees
48	Fc_irga	Carbon dioxide flux (LI-7500), without Webb et al. term	mg/(m ² s)
49	LE_irga	Latent heat flux (LI-7500), without Webb et al. term	W/m ²
50	C02_wpl_LE	Carbon dioxide flux (LI-7500), Webb et al. term due to latent heat flux	mg/(m ² s)
51	C02_wpl_H	Carbon dioxide flux (LI-7500), Webb et al. term due to sensible heat flux	mg/(m ² s)

52	H20_wpl_LE	Latent heat flux (LI-7500), Webb et al. term due to latent heat flux	W/m ²
53	H20_wpl_H	Latent heat flux (LI-7500), Webb et al. term due to sensible heat flux	W/m ²
54	N_Tot	Number of samples in the statistics (fluxes, variances, mean, etc)	samples
55	Csat_warnings	Number of times any CSTA3 flag warning was set high	samples
56	Irga_warnings	Number of times any L1-7500 flag warning was set high	samples
57	del_T_f_Tot	Number of times delta temperature warnings from CSAT3	samples
58	sig_lck_f_Tot	Number of poor signal lock warnings from CSAT3	samples
59	amp_h_f_Tot	Number of amplitude high warnings from CSAT3	samples
60	amp_l_f_Tot	Number of amplitude low warnings from CSAT3	samples
61	chopper_f_Tot	Number of chopper warnings from LI-7500	samples
62	detector_f_Tot	Number of chopper detector from LI-7500	samples
63	pll_f_Tot	Number of choppers PII from LI-7500	samples
64	sync_f_Tot	Number of chopper synchronization warnings from LI-7500	samples
65	agc_Avg	Average AGC from LI-7500	unitless
66	agc_thrshld_excded_Tot	Number of times LI-7500 AGC exceeded a user set threshold	samples
67	lws_1_Avg	LSW1	mV
68	lws_2_Avg	LSW1	mV
69	Rn_nr_Avg	Average net radiation	W/m ²
70	albedo_Avg	Average Albedo	unitless
71	Rs_downwell_Avg	SW_OUT, this is not downwelling, it's downward pointing sensor	W/m ²
72	Rs_upwell_Avg	SW_IN, this is not upwelling, it's upward pointing sensor	W/m ²

73	Rl_downwell_Avg	LW_OUT, this is not downwelling, it's downward pointing sensor	W/m ²
74	Rl_upwell_Avg	LW_IN, this is not upwelling, it's upward pointing sensor	W/m ²
75	T_nr_Avg		W/m ²
76	Rl_down_meas_Avg	not temperature corrected (can use for net calculations because temp term cancels)	W/m ²
77	Rl_up_meas_Avg	not temperature corrected (can use for net calculations because temp term cancels)	W/m ²
78	par_Avg		umol/m/s
79	hfp01_1_Avg	Soil heat flux (5cm/15cm) 15 open... (channel 3H/L. Field label: 10O)	W/m ²
80	hfp01_2_Avg	Soil heat flux (5cm/15cm) 10 open... (channel 6H/L. Field label: 15O)	W/m ²
81	hfp01_3_Avg	Soil heat flux (5cm/15cm) 15 bush... (channel 10H/L. Field label: 15B)	W/m ²
82	hfp01_4_Avg	Soil heat flux (5cm/15cm) 10 bush... (channel 8L (=16 SE). Field label: 10B) *single-ended	W/m ²
83	precip_Tot	Total Precipitation	mm
84	hor_wnd_spd_mean	Average 03002 horizontal wind speed	m/s
85	hor_wnd_spd_mean_rslt	Average 03002 resultant wind speed	m/s
86	hor_wnd_dir_mean_rslt	Average resultant horizontal wind direction	Deg
87	hor_wnd_dir_stdev	Standard deviation of wind direction	Deg
88	panel_temp_Avg	Average CR3000 panel temperature	C
89	batt_volt_Avg	Average battery voltage	V

Table 3.3: Abbreviated variable names, full variable names, and the units for the soil/ECTM data stream.

S/N	Abbreviated Variable Name	Full Variable Name	SI Units
1	VWC1	Volumetric Water Content	%
2	VWC2	Volumetric Water Content	%
3	VWC3	Volumetric Water Content	%
4	VWC4	Volumetric Water Content	%
5	VWC5	Volumetric Water Content	%
6	VWC6	Volumetric Water Content	%
7	VWC7	Volumetric Water Content	%
8	VWC8	Volumetric Water Content	%
9	Temp1	Soil temperature	C
10	Temp2	Soil temperature	C
11	Temp3	Soil temperature	C
12	Temp4	Soil temperature	C
13	Temp5	Soil temperature	C
14	Temp6	Soil temperature	C
15	Temp7	Soil temperature	C
16	Temp8	Soil temperature	C

Table 3.4: Abbreviated variable names, full variable names, and the units for the CS650 data stream.

S/N	Abbreviated Variable Name	Full Variable Name	SI Units
1	CS650 VWC 1 AVG	Soil moisture 1 average	%
2	CS 650 VWC 2 AVG	Soil moisture 2 average	%
3	CS650 VWC 3 AVG	Soil moisture 3 average	%
4	CS650 VWC 4 AVG	Soil moisture 4 average	%
5	CS650 VWC 5 AVG	Soil moisture 5 average	%
6	CS650 EC 1 AVG	Soil Conductivity 1 average	ds/m
7	CS650 EC 2 AVG	Soil Conductivity 2 average	ds/m
8	CS650 EC 3 AVG	Soil Conductivity 3 average	ds/m
9	CS650 EC 4 AVG	Soil Conductivity 4 average	ds/m
10	CS650 EC 5 AVG	Soil Conductivity 5 average	ds/m
11	CS650 P 1 AVG	Soil permittivity 1 average	F/m
12	CS650 P 2 AVG	Soil permittivity 2 average	F/m
13	CS650 P 3 AVG	Soil permittivity 3 average	F/m
14	CS650 P 4 AVG	Soil permittivity 4 average	F/m
15	CS650_P_5_AVG	Soil permittivity 5 average	F/m
16	CS650_PA_1_AVG	Soil period 1 average	uS
17	CS650_PA_2_AVG	Soil period 2 average	uS
18	CS650_PA_3_AVG	Soil period 3 average	uS

19	CS650_PA_4_AVG	Soil period 4 average	uS
20	CS650_PA_5_AVG	Soil period 5 average	uS
21	CS650_VR_1_AVG	Soil Voltratio 1 average	ratio
22	CS650_VR_2_AVG	Soil Voltratio 2 average	ratio
23	CS650_VR_3_AVG	Soil Voltratio 3 average	ratio
24	CS650_VR_4_AVG	Soil Voltratio 4 average	ratio
25	CS650_VR_5_AVG	Soil Voltratio 5 average	ratio
26	CS650_T_1_AVG	Soil Temperature 1 average	Celsius
27	CS650_T_2_AVG	Soil Temperature 2 average	Celsius
28	CS650_T_3_AVG	Soil Temperature 3 average	Celsius
29	CS650_T_4_AVG	Soil Temperature 4 average	Celsius
30	CS650_T_5_AVG	Soil Temperature 5 average	Celsius

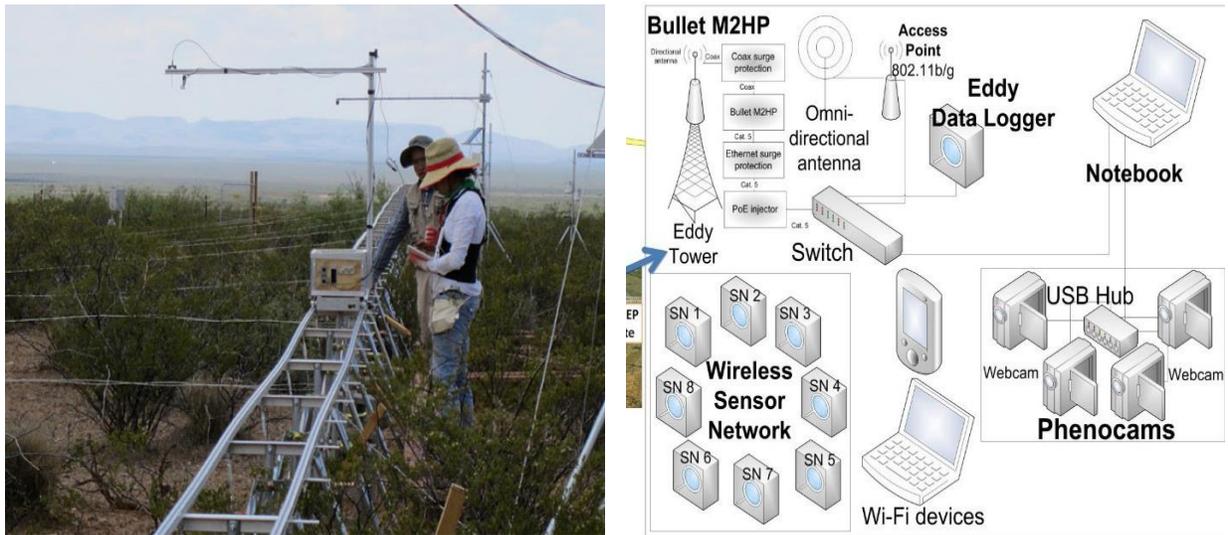


Figure 3.2: UTEP-JER site showing the interconnection of different projects (right) and a photo of the Robotic Tram System (left) (Ramirez, 2011).

3.2.3 Overview of rDataFusion

We built rDataFusion with R version 4.04, released in 2021 using the shiny package developed by RStudio, Inc. (<http://www.rstudio.com/shiny/>), which was first released in November of 2012. Shiny is an R wrapper for JavaScript – an interpreted programming language. JavaScript implementation in web browsers helps control the browser and its content and allows for interactivity with the user of the web browser (Laney, et al., 2013). Shiny is a free, open-source statistical software application and programming language that has the capacity to execute

complex statistical analysis, produce sophisticated and customizable scientific visualizations and can interface with databases, clouds, and data archives (Laney, 2013; Wanyanhan, et al, 2022; Chang, et al, 2020, & Chang, W., 2018). rDataFusion contains approximately two thousand five hundred lines of code (2500). The app starts by loading different libraries mainly from the user interface including shinydashboard, DT, xts, gt, highcharter, ggplot2, dplyr, lubridate, inputTS, gridExtra, CaTools, shinyjs, zoo, glue, among others. The *server* part (*server.R*) contains all functions required to read data files, select data and view, aggregate raw data, flagged status of variable, distribution of flagged status, clean data through QA/QC process, perform outlier and extreme value detection, relace missing values, compare raw and clean data variables, merge data, and display new tables and graphs as described in section 3.1 below, and provide functionality behind user interface controls. The user interface (*ui.R*) is made up of the user interface code. The codes for the server and ui can run on personal computers or on servers. Figure 2.3 below shows a flow chart or visual map of rDataFusion.

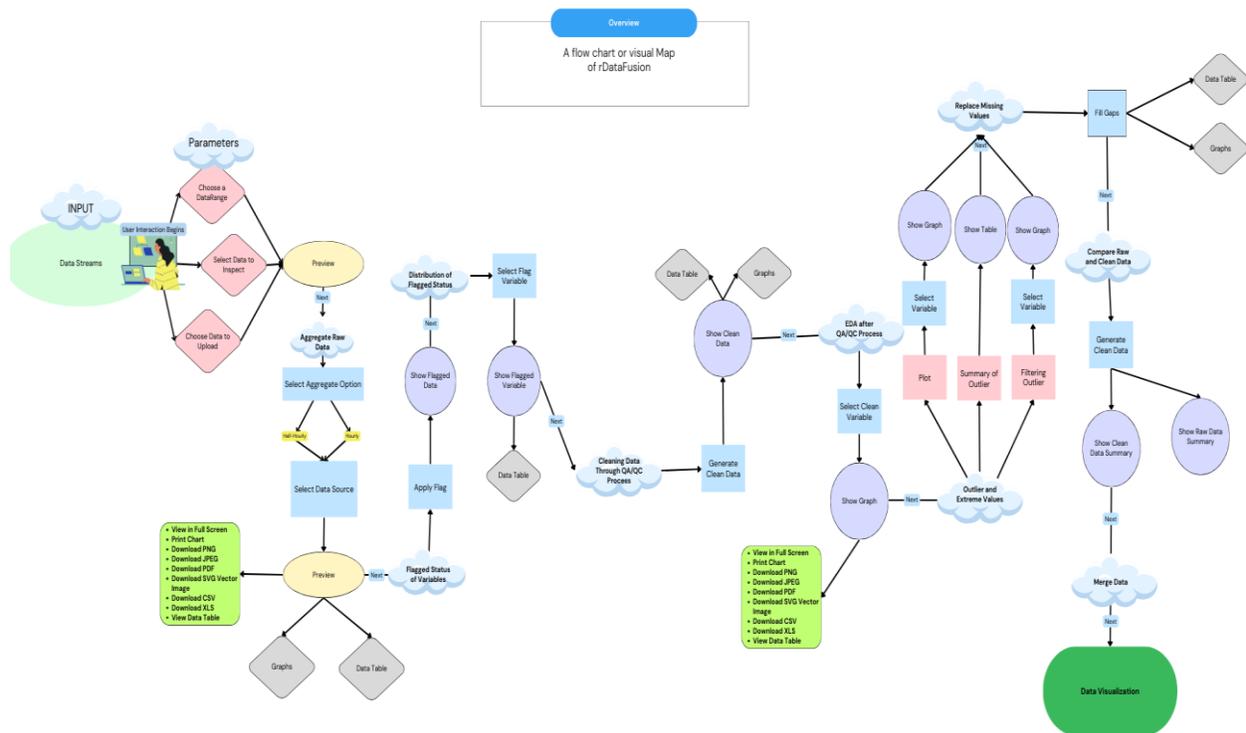


Figure 3.3: A flow chart or visual map or rDataFusion

3.2.4 Why R-shiny

Owing to the rise of automated data gathering and collection tools, data size and complexity of analysis have placed a limitation between research disciplines and the required data analysis (Kasprzak, et al., archive & Donoho, 2017). A data analytics software tool that utilizes common task frameworks and can help interpret, quantify, and possibly close methodological variations across disciplines is highly needed (Dondo, 2017). Following the review of data analytics software tools, we discovered that data analytics software tools such as Minitab (Arend, 2010), MATLAB (Moler & MathWorks, 2012), GenStat (Payne, et al., 2007), and SPSS (Landau & Everitt, 2004), have attempted to proffer solution to this problem by creating a more user-friendly interfaces that either make coding easier to learn, or use drop down menus and radio button selections to bypass the command lines (Kasprzak, et al., archive). These mentioned analytical software have their limitations such as non-publication ready graphics, non-intuitive drop-down menus, restrictive interfacing with other software, pricing – including the cost of licensing the proprietary software. Others include the difficulties encountered by users when they attempt to run codes originating from the software on their platform (Kasprzak, et al., archive), among others.

R has grown to become one of the most popular programming languages for statistics, environmental, and biological data analytics with over 14,000 free packages designed to address varying range of data analytics issues (Kasprzak, et al., archive). One of these packages is shiny. Shiny provides a framework for creating web-based interactive applications (Ramalho & Segundo, 2020). It can generalize R code to all levels of users, by simplifying the use of complex methodologies for people of different specialties, at the level of proficiency appropriate for the end users (Ramalho & Segundo, 2020). Shiny appears to have better graphics, and customizable visualizations capacities that allows users to dynamically change visualizations by adjusting parameters in some controlled variables using buttons, selection list or by direct clicks on the graphs (Ramalho & Segundo, 2020). Since R shiny web application is free and open source, it eliminates the burden of purchasing proprietary software, which can be inflexible and expensive and often takes resources away from research (LaZerte et al., 2017). Shiny applications can easily be interfaced with other software through API, as well as easy code adaptation (LaZerte et al., 2017).

The above important features of shiny and the flexibility of use, formed our decision in adopting shiny to build and develop rDataFusion web application. The design of rDataFusion utilized agile development approaches, including good architecture that allows the application logic to be broken down into smaller, independent parts, that are easier to maintain and verify. Also, included are testing the code, validating the data and app state, scaling, and performance. These are done to ensure quality assurance, validate the logic and the source code, including the test data, to minimize data quality issues, and the scalability and overall performance of the application. rDataFusion followed best practices for shiny application development (Kasprzak, et al., archive).

3.3 Results

3.3.1 rDataFusion Detailed Operation

rDataFusion application when run, opens in a web browser, showing the *About Page* that briefly gives user the general overview of the application, including its capabilities, limitations, data source description, features of the system, and how end users will interact with it. It also contains the *Reset App* tab that users can utilize to refresh the application when it is taking more time than expected to run. The *Next* button helps users to navigate between tabs. The *Select Data & View* tab comes with various options to make users' experience seamless, ranging from *Choose a date Range* that enables users to select the range of date of all the raw data they want to process and analyze based on the provision of the application. Then comes the *Select one data source at a time to inspect* that gives users the option to choose either of the four data sources - Climate, Soil/ECTM, Flux, and CS650 data. Once a choice of data source is made, the *Upload Raw Datasets* allows users to upload the selected data source into the application. The data to be uploaded needs to be formatted to a data table with properly labeled columns. The user then goes on to preview the raw data in a data table using the *preview* button. Figure 3.4 below shows the *Select Data & View Page with chose a date range, select data to inspect, upload raw data set, and preview selected data.*

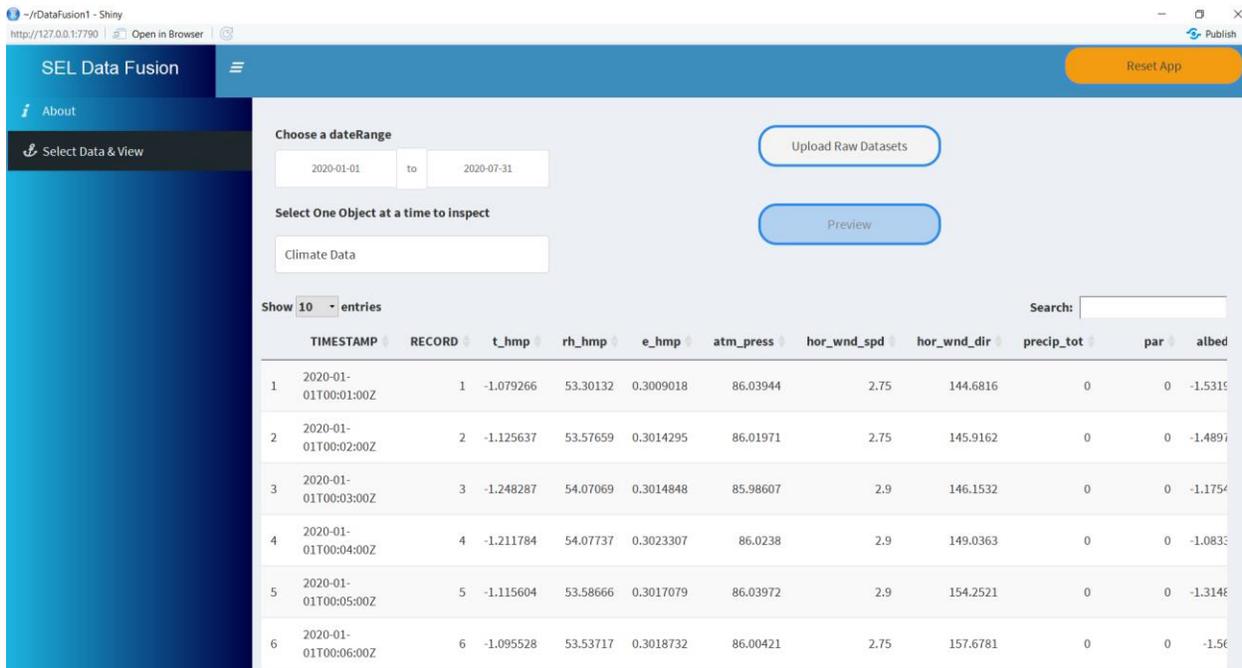


Figure 3.4: Select Data & View tab: This figure shows the processes of getting data into the application and view raw data table. The “chose date range “allows the user to select a date range of data that will be uploaded, the upload button allows the user to upload the data for the date range selected, the ‘chose data to inspect’ field allows users to choose the data stream they would want to upload. See table 3.1, 3.2, 3.3, and 3.4 for the full names and the SI units of all the variables used.

Furthermore, since the raw data is aggregated every minute, the *Aggregate raw data* tab enables users to transform the raw data from every minute to half hourly or hourly data for all the data streams within the date range selected. Data can further be previewed in a table and graphed interactively, as seen in fig. 3.5. below. The next step is the *Flagged Status of Variables* tab that allows users to flag data based on QC filtering data based on these conditions: outside high range, outside low range, data rejected due to QC, missing data – given as NAs or shown using non-physical values such as 999 or 9999 or similar and passed L1QC test. A *csv* file that contained most of the variables minimum and maximum values based on the instrument specification was used to apply flag to filter out data points deemed to be outside high and low ranges. We generated this *csv* file by carefully researching each of the variable’s measurement instruments applicable range of value specifications. These steps are the initial stages of data cleaning processes that *rDataFusion* offers so that users can have clean data to work with. This tab further allows users to show color coded flagged data that reflects various conditions of the

data – either missing value, data rejected due to QC, and so on. Figure 3.6 below shows the flagged status of variables.

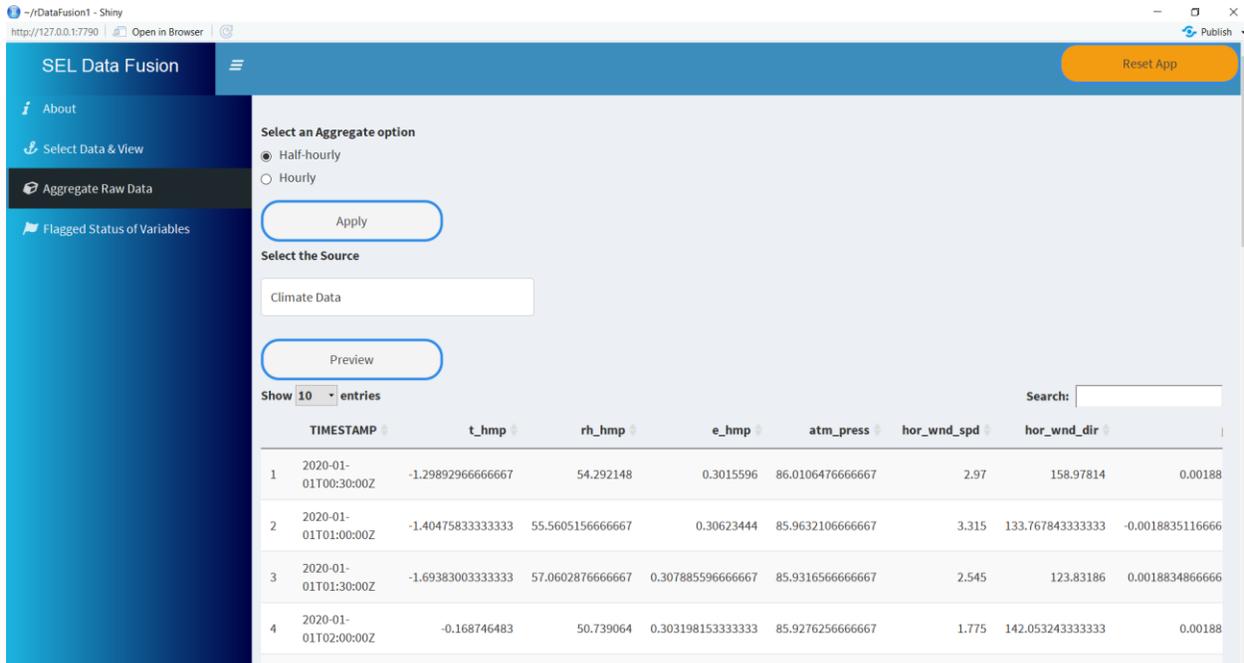


Figure 3.5: *Aggregate Raw Data* tab: This figure shows how raw data is aggregated from every minute to half-hourly or hourly data. See table 3.1, 3.2, 3.3, 3.4 for the full names and the SI units of all the variables used

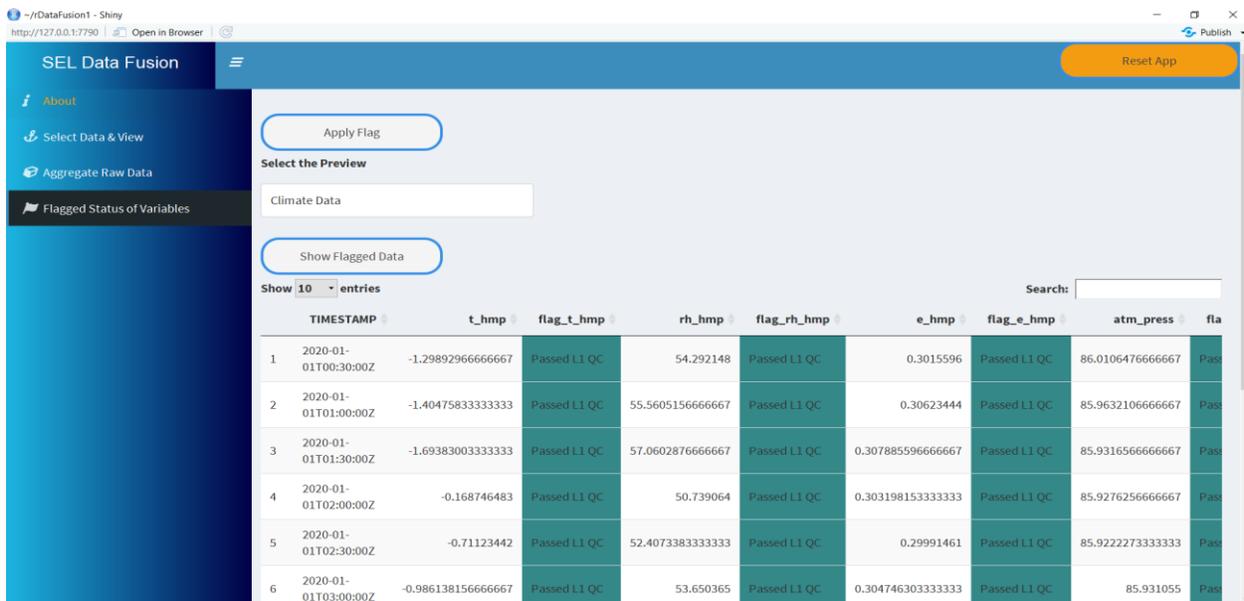


Figure 3.6: *Flagged status of variables* tab: This figure allows users to filter data based on QC flags. See table 3.1, 3.2, 3.3, and 3.4 for the full names and the SI units of all the variables used.

In addition to Flagged status of variables, there is the *distribution of flagged status* tab that show the flagged status of variables individually, for example, when users chose air temperature, they will be able to see different flagged status with a clean flagged variable summary table that show the number of data points that fall into each of the flagging categories. In the example of air temperature mentioned above, we observed that, for the time range selected – 01/01/2020 to 06/30/2020, there are no data points that are out of low or high range, rejected due to QA/QC, or suspected as bad data. We do observe that there are two data points that are missing and eight thousand six hundred and eighty-six data points that passed QC test. There is also a tab for *Cleaning Data through QA/QC* process that enables users to generate clean data sets. These are data sets that have passed through the initial cleaning process – where data that is deemed out of range or problematic based on the sensor specifications have been filtered out. These data sets still contain gaps, outliers, and extreme values. Fig.3.7 below shows the generated “semi” clean data table. Semi clean implies data has passed QC test but still contains gaps and outliers.

The screenshot shows a Shiny web application interface. On the left is a dark blue sidebar with navigation options: 'Select Data & View', 'Aggregate Raw Data', 'Flagged Status of Variables', 'Distribution of Flagged Status', 'Cleaning Data through QA/QC Process' (highlighted), 'EDA after QA/QC Process', and 'Outlier and Extreme Values'. The main content area has a 'Generate Clean Data' button at the top. Below it, a 'View Data' section contains a text input field with 'Climate Data'. A 'Show Clean Data' button is also present. Below the buttons, there is a 'Show 10 entries' dropdown and a search bar. The main part of the interface is a data table with the following columns: 'TIMESTAMP', 't_hmp', 'rh_hmp', 'e_hmp', 'atm_press', 'hor_wnd_spd', 'hor_wnd_dir', and 'p'. The table contains 7 rows of data, each representing a timestamp and its corresponding values for the seven variables.

	TIMESTAMP	t_hmp	rh_hmp	e_hmp	atm_press	hor_wnd_spd	hor_wnd_dir	p
1	2020-01-01T00:30:00Z	-1.29892966666667	54.292148	0.3015596	86.0106476666667	2.97	158.97814	0.001883
2	2020-01-01T01:00:00Z	-1.40475833333333	55.5605156666667	0.30623444	85.9632106666667	3.315	133.767843333333	
3	2020-01-01T01:30:00Z	-1.69383003333333	57.0602876666667	0.307885596666667	85.9316566666667	2.545	123.83186	0.00188348666666
4	2020-01-01T02:00:00Z	-0.168746483	50.739064	0.303198153333333	85.9276256666667	1.775	142.053243333333	0.001883
5	2020-01-01T02:30:00Z	-0.71123442	52.4073383333333	0.29991461	85.9222273333333	1.705	164.109048433333	0.005650
6	2020-01-01T03:00:00Z	-0.986138156666667	53.650365	0.304746303333333	85.931055	0.74	168.98905	0.007534
7	2020-01-01T03:30:00Z	-1.59428120333333	55.8793046666667	0.303463636666667	85.927551	1.86	192.424743333333	0.00753437533333

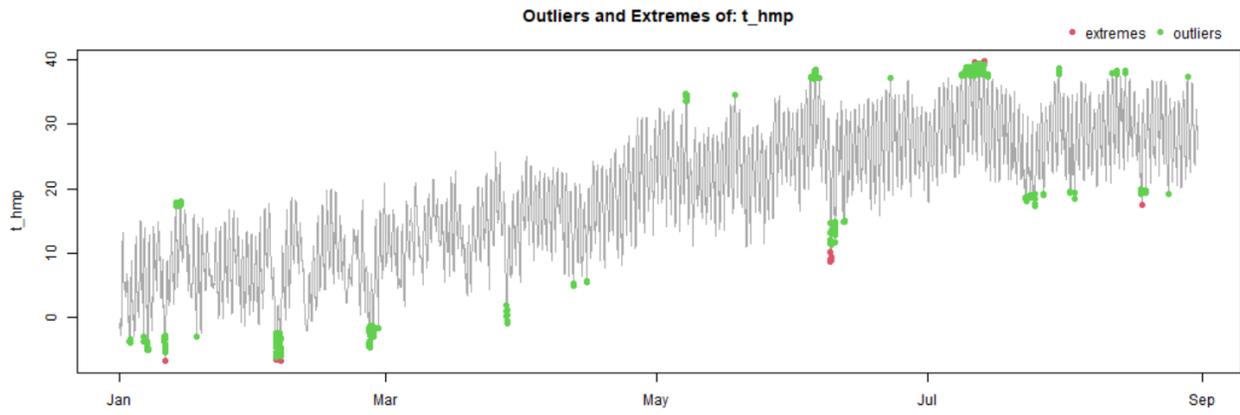
Figure 3.7: The generated “semi” clean data table. Here, the data has passed the initial QAQC test but still contains outliers and gaps. See table 3.1, 3.2, 3.3, and 3.4, for the full names and the SI units of all the variables used.

Subsequently, the *outlier and extreme value detection* tab gives users the opportunity to filter out outliers and extreme values in the data. There are different techniques for evaluating outliers and

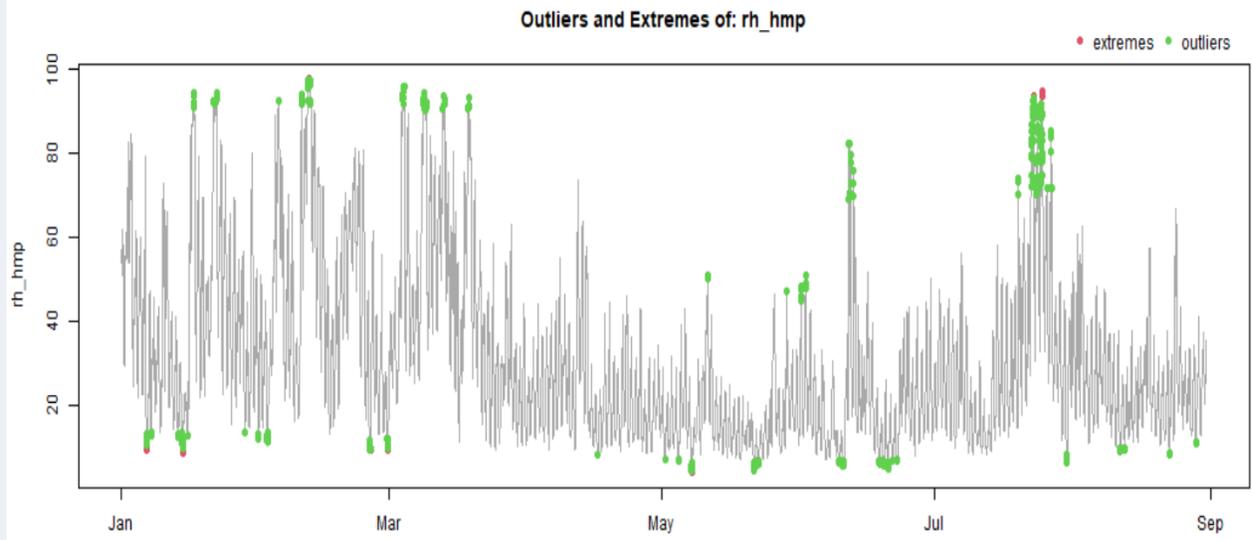
extreme values based on scientific and practical applications, because clear outliers or bad data may contain valuable information about the process or the data gathering process (Faybishenko, et al., 2021) Detailed description of these methods is beyond the scope of this work, however, the first step in the evaluation of outliers and extreme values is to access whether the data are within a reasonable range (natural overall, seasonal, and instrumental). As an example, rainfall will only have positive values when there is precipitation and zeros when there is none; solar radiation is expected to be only positive values and zeros at night, etc. The outlier and extreme value” detection tab filters datasets based on datapoints considered to be an outlier and extreme value, utilizing statistical algorithms known as Median Absolute Deviation (MAD) that shows the graphs of data points with outliers and extreme values clearly labelled. This is based on the Hampel approach that uses a sliding window to go over the data vector and calculate the median and standard deviation expressed as median absolute deviation (Faybishenko, et al., 2021). Considering seasonal fluctuations and trends in our time series meteorological data, we utilized the *runquantile* function from R Package *caTools* which uses a moving window to calculate the quantiles over a vector of the variable. A 6-month moving window was used, with probabilities of 0.999 and 0.001 used for upper and lower extreme values, and 0.975 and 0.025, for the upper and lower outlier values, respectively, following (Faybishenko, et al., 2021). The rolling window length and thresholds are usually anchored on the outlier detection goals.

In addition, the outlier detection part of this app has a sub-tab that populates the summary of outliers in a table. It also allows users with the option to filter out the data points considered as outliers and extreme values or not. Fig. 3.8 below show time series graphs depicting outliers and extreme values, and the statistics of the outliers and extreme values are summarized in fig. 3.9 below.

Show Graph



Show Graph



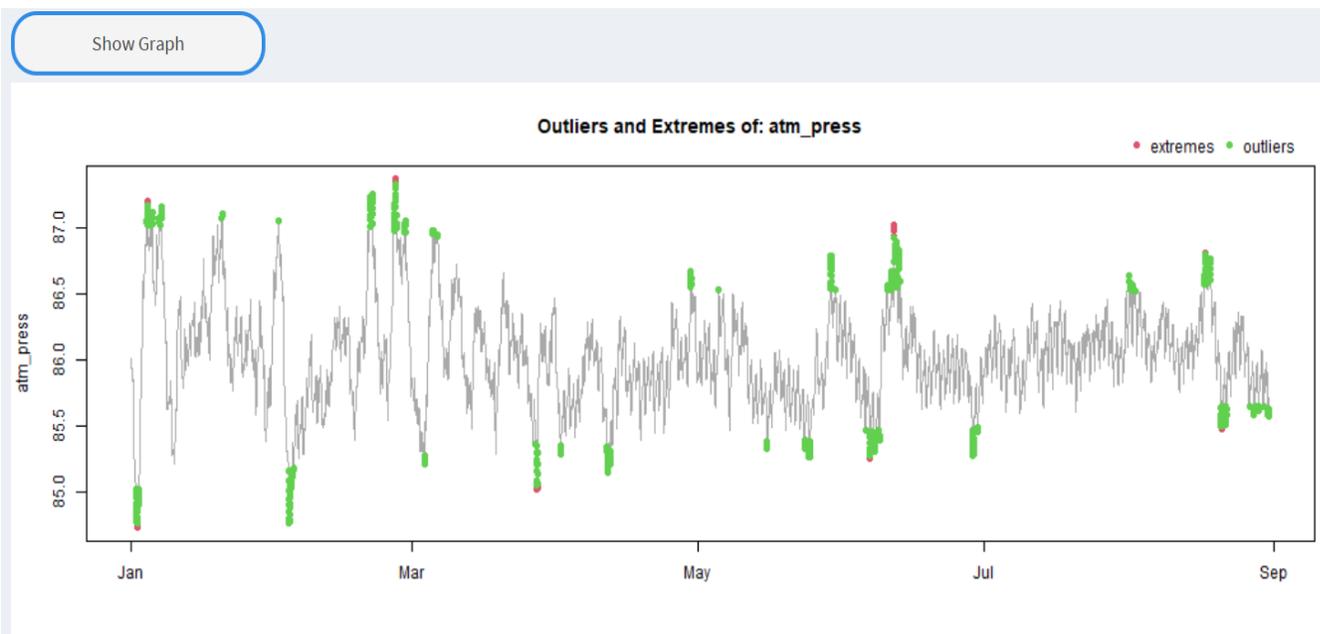
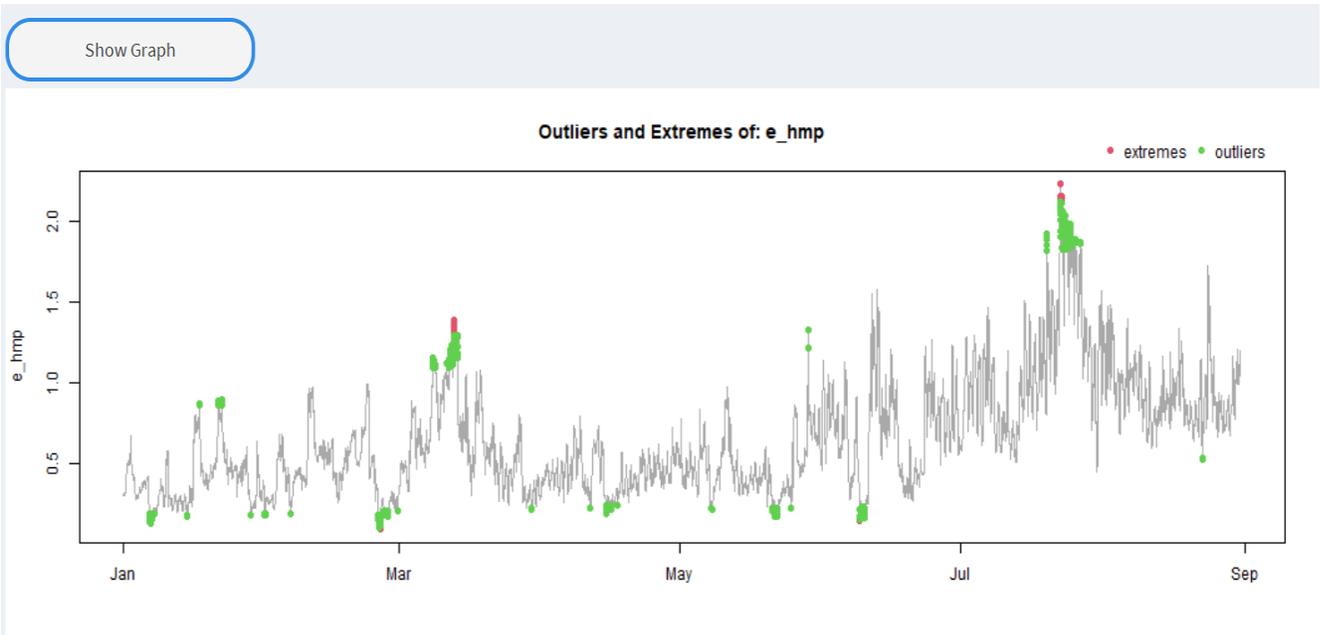


Figure 3.8: These show time series graphs depicting outliers and extreme values. The figures show Air temperature, relative humidity, absolute humidity, and atmospheric pressure, respectively, from Jan. to Sept. 2020, with the red dots showing extreme values and the green color showing outliers. See table 3.1, 3.2, 3.3, and 3.4 for the full names and SI units of all the variables used.

Plot Summary of Outlier Filter Outlier

Select the Source

Climate Data

Show Table

CSV Show 10 entries Search:

	variable	upper_extreme	upper_extreme_percent	lower_extreme	lower_extreme_percent	upper_outlier	upper_outlier_percent	lower_outlier	lower_outlier_percent
1	t_hmp	4	0.04	9	0.09	102	1	126	1.24
2	rh_hmp	8	0.08	10	0.1	199	1.96	186	1.83
3	e_hmp	8	0.08	13	0.13	174	1.71	240	2.36
4	atm_press	13	0.13	11	0.11	227	2.23	216	2.12
5	hor_wnd_spd	10	0.1	3	0.03	227	2.23	162	1.59
6	hor_wnd_dir	14	0.14	14	0.14	248	2.44	255	2.51
7	par	4	0.04	2400	23.58	92	0.9	118	1.16
8	albedo	13	0.13	10	0.1	245	2.41	249	2.45
9	lws_2	9	0.09	9	0.09	183	1.8	154	1.51
10	NetRs	5	0.05	5	0.05	142	1.4	212	2.08

Showing 1 to 10 of 14 entries Previous 1 2 Next

Figure 3.9: Statistics of the outliers and extreme values. See table 3.1, 3.2, 3.3, and 3.4, for the full names and SI units of all the variables used.

Modern data collection instruments (data loggers) collect and record time series or other types of data, but collected data are always prone to distinct types of errors (Faybishenko, et al., 2021). These errors include errors of omission such as improper recording of metadata or data due to lack of proper documentation, anomalies in the data collection field, human errors, and commission with examples as: error in data entry and malfunctioning of instruments, among others (Faybishenko, et al., 2021). It is also important to note that collected time series data could be imperfect due to different time frequency of measurement, varied units of measurements in the same time series, changes in sensors due to calibration or other sensor malfunctioning, and abnormal values (Faybishenko, et al., 20210). Interestingly, the QAQC processes we introduced into our data helped in cleaning up these issues and in return introduced gaps in the data as some of the data points were filtered out.

In our app, there is a tab for *replacing missing values or Gap Fill*. Missing values are observed when no values are stored for the variable in an observation. Missing value mechanisms and patterns are different for different data types. Imputation or gap filling of missing values is a

challenging problem because of the non-generic nature of the techniques and are different for different kinds of data. The *imputeTS* package in R provides a univariate time series imputation, with different time series imputation algorithms included. In the app, we utilized the mean of the neighboring data points to fill the gaps, and only filled gaps that are within 2- hours-time range. Figure 3.10 below shows the climate data source with all the missing values filled.

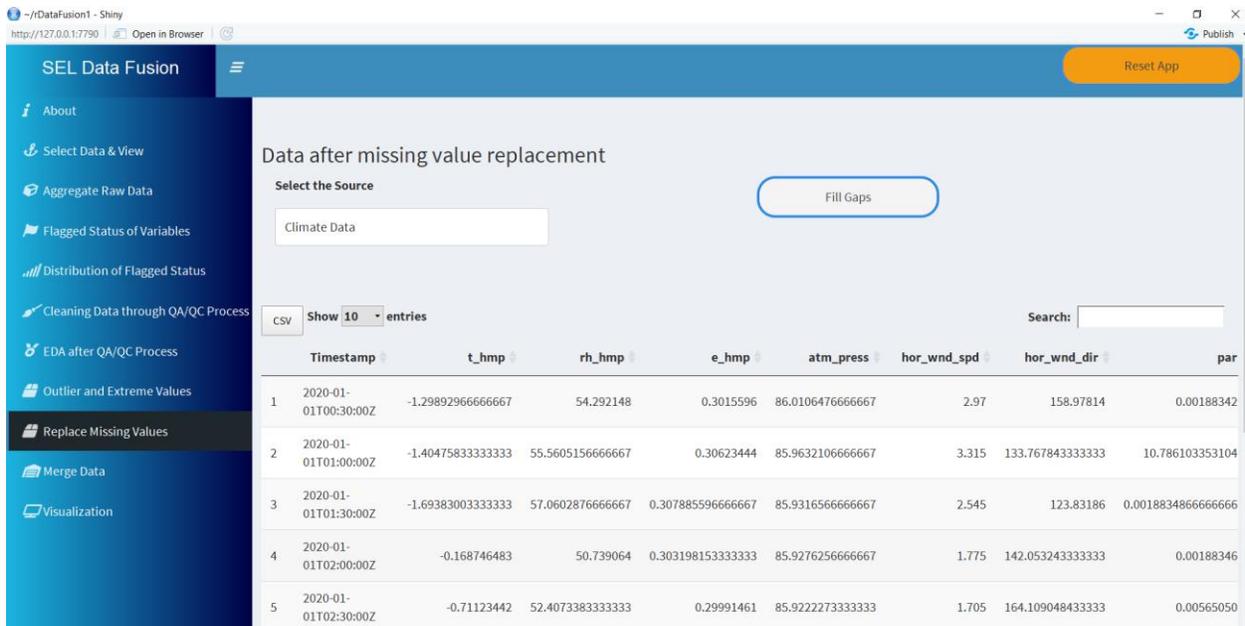
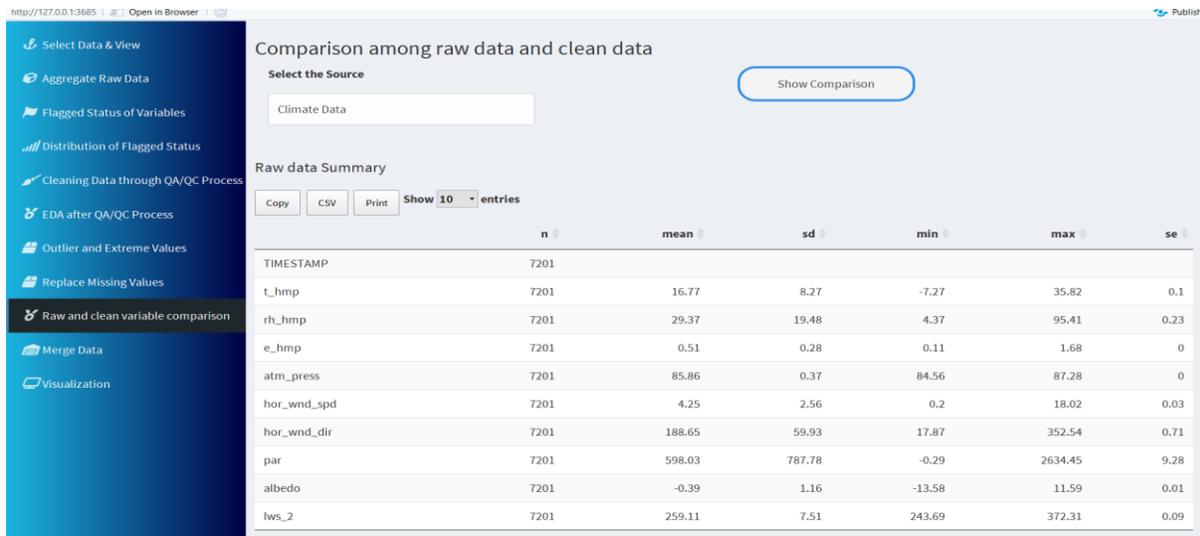


Figure 3.10: Missing data points replaced from the selected data source. See table 3.1, 3.2, 3.3, 3.4, for the full names and the SI units of all the variables used.

Furthermore, to ensure that the data cleaning processes such as QAQC, outlier and extreme values detection, filtering out outliers and extreme values, and missing values replacement were effective, we compared raw data (before the cleaning processes) and the clean data (after the cleaning processes). For the raw data, we summarized the data to see the mean, standard deviation, minimum, maximum, and standard error and compare it with the clean data, but included the number of gaps filled data points, as well as the number of data points removed as outliers. Figure 3.11 below shows the raw and clean data comparison.



Clean data Summary

Copy CSV Print Show 10 entries

	n	mean	sd	min	max	se	# Removed	# Gap Filled
t_hmp	7053	16.8	7.93	-2.31	33.76	0.09	196	48
rh_hmp	6942	28.49	17.59	7.02	80.81	0.21	350	91
e_hmp	6922	0.5	0.26	0.15	1.45	0	319	40
atm_press	6885	85.86	0.33	85.04	86.66	0	327	11
hor_wnd_spd	7077	4.11	2.23	0.2	10.94	0.03	339	215
hor_wnd_dir	7115	188.81	55.63	76.88	291.08	0.66	359	273
par	4651	920.92	806.69	-56.9	2427.78	11.83	2780	230
albedo	7188	-0.38	0.72	-3.08	0.94	0.01	363	350
lws_2	6970	258.34	2.27	254.36	264.61	0.03	303	72
NetRs	7180	222.48	292.68	-9.89	867.36	3.45	290	269

Figure 3.11: Raw and clean data comparison. See table 3.1, 3.2, 3.3, and 3.4 for all the full names and the SI units of all the variables used.

Lastly, all the clean data sources were merged into one data table using the *merge data tab*. This will allow users to have all the data in one place, download all the data in an excel or csv file. Users can visualize the data by each variable using the data *visualization tab*. Users can chose to visualize either one, three, and six months of data. They can also visualize one year of data and all these depends on the length of data selected to upload to the application at the data selection and view stage. Users can also view charts or graphs in full screen, print charts, download images either in PNG, JPEG, PDF, and SVG vector formats. It is expected that production-ready plots will be made from desktop applications with customizations and annotations as required. It is also important to note that the data behind the charts can also be downloaded in csv or xls

formats or view the data table. Figures 3.12 and 3.13 below show the merged data table and visualized temperature data from climate data source, respectively. A YouTube video that shows users how to navigate through rDataFusion can be viewed at <https://youtu.be/ByGdTZRiXTQ>

Timestamp	t_hmp	rh_hmp	e_hmp	atm_press	hor_wnd_spd	hor_wnd_dir	par	albedo
2020-01-01T00:30:00Z	-1.298929666666667	54.292148	0.3015596	86.0106476666667	2.97	158.97814	0.001883429	-1.475406993333333
2020-01-01T01:00:00Z	-1.404758333333333	55.5605156666667	0.30623444	85.9632106666667	3.315	133.7678433333333	11.1107662798002	-1.425687516666667
2020-01-01T01:30:00Z	-1.693830033333333	57.0602876666667	0.307885596666667	85.9316566666667	2.545	123.83186	0.00188348666666667	-1.002255083333333
2020-01-01T02:00:00Z	-0.168746483	50.739064	0.303198153333333	85.9276256666667	1.775	142.0532433333333	0.001883467	-1.293018666666667
2020-01-01T02:30:00Z	-0.71123442	52.4073383333333	0.29991461	85.9222273333333	1.705	164.1090484333333	0.005650501	-1.000849936666667
2020-01-01T03:00:00Z	-0.986138156666667	53.650365	0.304746303333333	85.9311055	0.74	168.98905	0.007534197	-1.583095896666667
2020-01-01T03:30:00Z	-1.594281203333333	55.8793046666667	0.303463666666667	85.927551	1.86	192.4247433333333	0.00753437533333333	-2.040060626666667
2020-01-01T04:00:00Z	-2.212395366666667	58.8508863333333	0.305175906666667	85.9008283333333	1	175.6388733333333	0.022603928	-1.183494186666667

Figure 3.12: Merged data table of all the data sources. See table 3.1, 3.2, 3.3, and 3.4 for the full names and SI units of all the variables used.

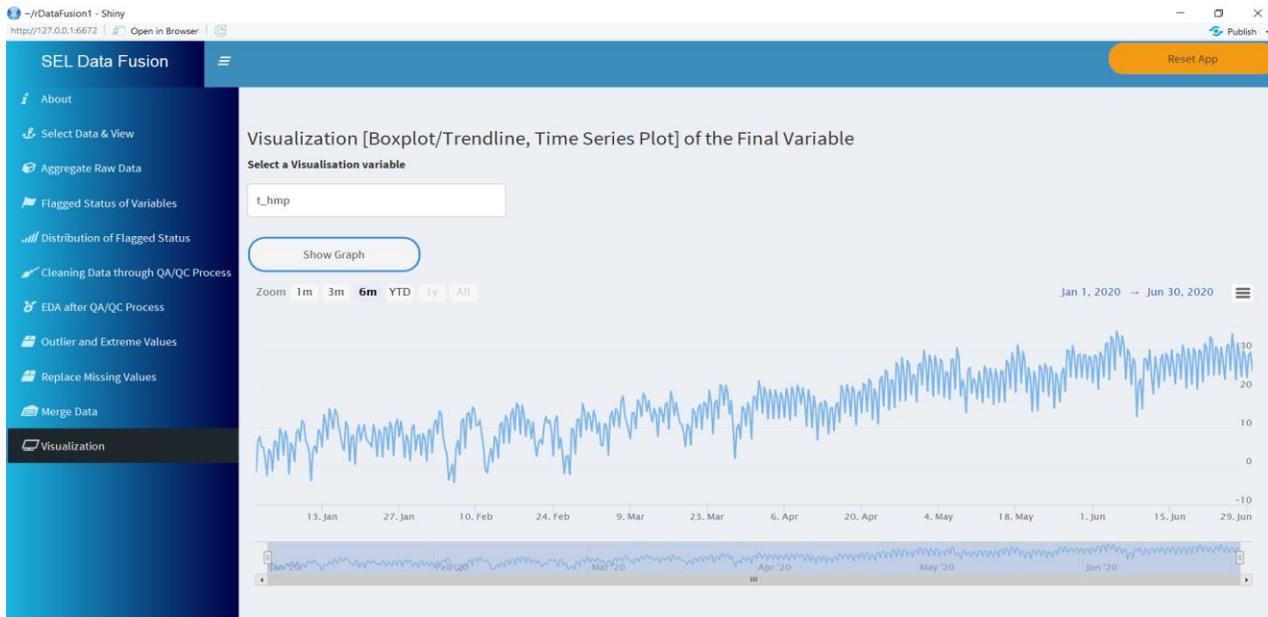


Figure 3.13: Time series visualization of temperature from the merged data table. See Table 3.1, 3.2, 3.3, and 3.4 for the full names and SI units of all the variables used.

3.4. Discussion

Data driven research is often faced with the enormous task of data preparation and cleaning with an estimated 80% effort spent towards achieving this goal within the academia and even in the industry (Huber, et al., 2021 & Press, 2016). This has not only led to a very reduced amount of time spent in the actual data analysis which brings real value in data, but also discouraged some researchers from attempting their data driven research because of data cleaning and preparation difficulties (Huber et al., 2021). New ways of doing data-driven research, data analytics, and integration have been provided by technological advances over the past few years coupled with a great deal of computing resources (Hey, et al., 2009). The development of these resources has given researchers the leverage to shift their focus to activities that allow them extract knowledge from their data. With these challenges in mind, and the need to adequately utilize new technological advances in data analytics, the development of *rDataFusion* aimed at developing a customizable analytic tool that aids researchers with improved capacities for aggregating different streams of data from a single intensive site by providing an open-source multi-data fusion tool that facilitates data management, sharing, analysis, and serves as a template for other research groups with similar challenges was born.

Furthermore, the extensive need for standardized, systematic, and long-term monitoring to understand ecological and environmental changes that lead to global changes cannot be overemphasized (Weigel, et al., 2020). *rDataFusion* is carefully built and designed to meet the needs of ecologist or environmental scientists or researchers who are collecting data from disparate sources, instruments, or sensors but have no way of aggregating, managing, or visualizing their data. The QA/QC process, filtering outliers, gap filling and other carefully thought features of the app were added to ensure clean data since collected time series data are often irregular, with different units of measurements in the same time series, time stamp duplicates, periodic failure or malfunctioning of sensors, changes due to calibrations or missing data (Faybishenko, et al., 2021). Other key features of *rDataFusion* include the capacity to visualize data in near real-time to check that all sensors are running properly, ensure preliminary flagging for data that is deemed out of range, and filter data based on flags.

For the ecological, environmental science and research community that utilize data analytical and integration tool, growing data and advances in analytics are essential in gathering the knowledge needed to address issues such as global change, biodiversity loss among other urgent ecological or environmental issues (Raban & Gordon, 2020). Data management tools are necessary and have a crucial role to play in addressing these challenges. Therefore, a data tool like rDataFusion that supports efficient and effective data management, integration, and visualization will help in addressing the challenges.

Besides, rDataFusion is embodied with terrific features that starts with the “About page” – a brief description of the application and the various data sources utilized in the app, with an automatic connection of the application to an online data archive that feed data to the app. There is also the “select data & view” tab that not only allows users to select date range, but also give them leverage to upload raw datasets into the app with “a click and select” one data stream at a time to preview data in a data table. Since the raw data is collected every minute, we added a tab to convert the data in half hourly or hourly intervals. After data conversion, users can select the data source to preview and graph the selected data variables using the preview graph tab for preliminary checks.

Similarly, there is a “flagged status of variables” tab that allows users to apply preliminary flagging to filter variables based on the minimum and maximum values of the instruments used for measurement of that variable. This would allow data variables to be filtered based on level one quality control to either, “passed”, “outside low range”, “outside high range”, and “rejected”. Next, is the “distribution of flagged status” that shows the distribution of flagging status of specific variables. It also shows the summary of flagged variable based on the number of data points that passed QC test, rejected due to QC test, outside high range, outside low range, missing data, and suspect data. Immediately following, is the tab for “cleaning data due to QA/QC” that allows users “to generate and show clean data table”; data that has passed QC test. All these features are essential in modern data analytic and integration tool for effective cleaning, utilization, and extraction of desired knowledge from data and further reduce the time spent in data processing and cleaning as rDataFusion is built with these challenges in mind.

Future work would include expanding rDataFusion software to include data from other research groups and include metadata information. Currently, rDataFusion fetches input data from SEL

data archive, future work would enable rDataFusion software to be connected to a standard database for data input. Future work will allow users customize the merging of the data based selected criteria. Future improvements using a code repository like GitHub (<https://github.com>) for versioning and tracking purpose as the codes for this application is currently in GitHub and here is the link: <https://github.com/SELDevTeam/rDataFusion1>

3.5. Conclusion

Through rDataFusion, we have demonstrated that it is possible to aggregate data from different sources by long tail ecological scientist using a programming language that is common to ecologists. This relied on established web-based practices that in principle allow for the design of uniform, programming language, and research infrastructure independent practices, that enable seamless integration of data for analysis. This has become viable in practice through the design of rDataFusion as a roadmap that can help promote open access to data and open-source software development, which is important as it helps researchers who need software for data analytics to be able to find one online or generate one and share with the community.

Chapter 4: Detecting Rare Soil Moisture Events in a Chihuahua Desert Dryland Ecosystem using Machine and Deep Learning

Abstract

Detecting rare soil moisture events in dryland ecosystems is important for understanding how rising temperatures and shifting precipitation regimes are altering the frequency and severity of drought across all dryland regions. However, predicting rare soil moisture events might require sensor cross-correlation. Hence, the need for cross-correlation of sensors. The objective of this study is to assess if multi-sensor cross-correlation can predict rare soil moisture events in time series data using Machine Learning and Deep Learning (DL) models. We deployed several Machine and Deep Learning techniques and cross correlated these for optimal rare soil moisture events detection in a Chihuahuan Desert shrubland on the Jornada Experimental Range in southern New Mexico. Specifically, the machine and deep learning techniques used for this study utilized both classification and regression methods, including Decision Tree Classifier (DTC), Logistic Regression Classifier (LR), Random Forest Regression (RF), and a Long Short-Term Memory (LSTM) method used in Artificial Neural Network (ANN). Of the methods used, the DTC performed the best, with prediction accuracy of 88.8%, closely followed by a LSTM model with 88%. The LR recorded a prediction accuracy of approximately 80%. The Variable Importance Plot (VIP) showed that soil temperature and soil heat flux are the most prominent factors in predicting soil moisture dynamics in this dryland ecosystem at 54% and 38%, respectively, when a DTC modelling method was used. Similarly, with the random forest regression model, the VIP plot showed that soil heat flux made the highest contribution to determine rare soil moisture event with a feature value of 50%, closely followed by soil temperature with a 35% contribution. This result will further aid in understanding drought severity in these regions and help ecologists to manage ecohydrological and agricultural processes to ensure human well-being and sustainable environmental management. This will further help ecologists to understand both small and large- scale drought patterns in the region.

4.1 Introduction

Rare or anomalous events in time series data describe both the unusual magnitude and time interval of an event whose value deviates from the remaining measured data points (Nikou Gunnemann-Gholizadeh, 2018). According to the IPCC 2021 report (Seneviratne, et al., 2021), it is evident that rare events such as hot temperature extremes or droughts are intensifying in many

regions of the world and will continue to do so in the future. Other events that are likely to intensify are heat waves, floods, soil moisture anomalies, snow-cover-induced albedo anomalies among others. In drylands much research focuses on understanding how rare or extreme hydrometeorological events affect dryland ecosystems and their functioning (Mahecha et al., 2017, Frank et al., 2015, & Niu et al., 2014). Some of this research has focused on the manifestation of extreme anomalies of phenology (Ma et al., 2015), exploration of soil moisture anomalies occurring coincidentally with unusual climate patterns that catalyze anomalous vegetation responses (Nicolai-Shaw et al., 2017), biogeochemical fluxes (Frank et al., 2015), and the global inter-annual variability in atmospheric carbon uptake due to extreme anomalies in gross primary production (GPP) (Mahecha et al., 2017).

Although there are ever more sophisticated climatic models to project or detect future changes, the effects of climatic fluctuations on complex ecosystem attributes such as soil moisture dynamics remains poorly understood, both empirically and theoretically (Woodward et al., 2016). For example, soil moisture dynamics are continually perturbed by changes in catchment geomorphology and land-use, local physio-chemical parameters, and changes in the timing of rare events relative to normal seasonal cycles (Garner et al., 2015 & Death et al., 2015). Furthermore, although rare events may be viewed simply as one end of a gradient of fluctuations, anthropogenic influences are increasingly altering their intensity, frequency, and duration, with potentially dramatic consequences for water availability (Ledger & Milner., 2015). Separating the biological impacts of rare events from the effects of inherent and chronic background fluctuations in soil moisture dynamics remains an important challenge.

Soil moisture dynamics influence both global water and energy budgets, and can control the redistribution of rainfall into infiltration, runoff, percolation in soil and evapotranspiration (Ali et al., 2015). Hence, it is regarded as a space-effective driver of hydrological and vegetation processes. Rare soil moisture conditions that are represented by saturation or the permanent wilting point can promote flood events or indicate droughts respectively. For meteorological processes, soil moisture is the “memory of precipitation” because it stores rainwater and re-emits it to the atmosphere via evaporation, sometimes following considerable delay (Ali, et al., 2015). Due to these characteristics and to the important effect soil moisture has on surface energy exchange, soil moisture content may be one of the best metrics to observe in order to understand

how climate change dynamics are impacting dryland ecosystems. This hastens the need for detecting rare events in soil moisture in dryland ecosystems to better understand small and large-scale drought patterns.

The soil water holding capacity of any particular soil, especially in dryland ecosystems, contribute greatly to soil moisture availability within the soil profile (Weir, P. et al., 2023). Typically, soil moisture measurements attempt to quantify the amount of moisture stored within the soil. To determine high soil moisture values, which are rare events in drylands, we are only interested in the rarest high soil measurements. It is these low frequency soil moisture events with higher magnitude (rare events) that we intend to predict using predictive approaches. We are focusing on predicting rare high soil moisture events in dryland ecosystem because of its importance in understanding climate change dynamics in dryland ecosystems. Soil moisture also plays an important role as a space-effective driver for both hydrological and vegetation processes, growth, and development.

Scientists from multiple disciplines, have over the past few decades, increasingly adopted predictive approaches to solving scientific problems such as the description of soil moisture anomalies, global climate change, emerging diseases, biodiversity loss, food security and many more (Willcock et al, 2018). Consequently, ecologists are challenged by the need to understand and predict complex ecological processes and patterns. To address these challenges, Machine Learning (ML), a fast-growing field that is now well embodied within the discipline of ecoinformatics is concerned with identifying structures in complex and often non-linear data to generate accurate predictive models or algorithms (Olden et al, 2008, Ghahramani, 2015). Simply stated, ML is a process that is used to fit a model to a data set, through training or learning. The learned model is then used against an independent data set to determine how well the learned model can generalize against the unseen data, a process called testing (Ghahramani, 2015).

Advances in both data collection techniques and FAIR Data Principles (Wilkinson et al. 2016), the availability of large high-resolution data sets spanning multiple spatiotemporal scales has increased dramatically. Accordingly, ML approaches are increasingly used by researchers to model complex relationships and predict anomalous behavior in these large data sets has been reputed as robust alternatives to traditional ecological modelling approaches (Olden et al, 2008).

This is because ecological data are known to be non-linear and multi-dimensional with many interactions (Olden, et al., 2008). Approaches that assume linearity, therefore, are unable to cope with complex interaction effects (Knudby et al., 2010). Expectedly, ML methods show greater strength in accuracy and general capacity to predict and explain anomalous patterns especially in ecological data (Olden, 2008, Pichler & Hartig, 2023, Scowen, et al., 2021).

In this study, we will attempt to utilize multi-sensor cross-correlation to predict rare events in soil moisture using data derived from a well-studied dryland study site using Machine Learning and Deep Learning (DL) -a subset of machine learning that uses structures and patterns like the human brain to analyze complex patterns and relationships in data (Pichler & Hartig, 2023). For this purpose, we will be adopting Decision Tree Classifier (DTC), Logistic Regression, Random Forest, and Long Short-Term Memory (LSTM) of neural networks. These methods were chosen because of their ease of implementation and interpretation (Cruz, et al., 2006). The rest of this chapter is organized as follows: Section 2 shows the study sites, section 3 highlights the methods, section 4 discusses the results, while section 5 presents the conclusion.

4.2 Study Site:

This study utilized data collected in a shrubland study site managed by researchers at the University of Texas at El Paso and located on the Jornada Experimental Range (32° 34' 59" N, 106° 37' 34" W; 1417m asl, Figure 2.1), which is owned and managed by the United States Department of Agriculture – Agricultural Research Service, in Southern New Mexico and northern Chihuahuan Desert. The study site was initiated in 2009 and became operational in mid-2010, with the overarching goal of studying global change impacts on dryland ecosystems, with special focus on the biophysical controls and feedback associated with land – atmosphere exchange of water, carbon, and energy. Since its inception in 2009, the site has been supported by numerous grants and has included over 100 undergraduate and graduate students, post docs, and technicians who have explored a range of research topics and questions. Notable studies include “Furthering our understanding and scaling patterns and controls of land – atmosphere carbon, water, and energy exchange in the Chihuahuan desert shrubland with novel cyberinfrastructure” (Jaime, H. A., 2014), Towards new data and information management solutions for data – intensive ecological research” (Laney, C.M., 2013), “Assessing data quality in a sensor network for environmental monitoring” (Ramirez, G., 2011), “Spatiotemporal

variability of plant phenology in drylands: A case study of the Northern Chihuahuan desert” (Luna, R.N., 2016), “Development of low cost network of webcams for monitoring plant phenology in Chihuahuan desert” (Gonzalez, L., 2011), among others.

A range of biophysical data such as soil temperature, soil water content, net radiation, relative humidity, wind speed, wind direction, soil heat flux, surface reflectance, land-atmosphere carbon and energy exchange, and vegetation phenology have been collected since 2010. Data included in this study spanned May 2010 to March 2020. Table 2.1 below shows all the data variables used with their abbreviated and full variable name and corresponding international system of units (SI units). The relative humidity and air temperature were measured at 5m height with an HMP155A probe (Vaisala Corporation, Helsinki, Finland), Net radiation was measured with a CNR4 sensor from the Campbell Scientific. Soil heat flux was measured using four self-calibrating Hukseful USA HFP01 sensor plates in four representative positions of the landscape. Soil instruments are categorized into two sub-systems of soil profiles installed to capture underground soil temperature and heat under Mesquite shrubs and bare soil. The depth of these profiles are 2cm, 10cm, 15cm, and 20cm with the data collected and stored in a Campbell Scientific CR3000 data logger. It is also pertinent to note that the TE525 (Texas Electronics, Dallas, TX, USA) tipping-bucket rain gauge was used to measure the amount of rain on the site (Jamie, H. 2014). All these Biometeorological variables were measured every second and averaged and stored every 30 minutes.

The study site is a shrubland with a mixture of *Larrea tridentata* (Creosote) and *Prosopis glandulosa* (Honey Mesquite) that is typical of the northern Chihuahuan Desert (Laney, 2013). Other notable species include *Flourensia cernua* (Tarbush), *Muhlenbergia porteri* (Bush Muhly) and *Dasyochloa pulchella* (Fluffgrass). The site is characterized by a shallow sandy to gravelly soils that are generally less than 1m in depth and are underlain by caliche. The study site slopes westward by approximately 2° from east to west. The long-term average annual rainfall at the JER Headquarters (approximately 13km from SEL-Jornada) was 245.1 mm from 1915 to 1995, with a standard deviation of 85.0 mm (Wainwright, 2006., Laney, 2013). Petrie, et al. (2014) stated that a sizable portion of annual precipitation occurs mostly during the summer monsoon season (40-50% on average, 5.7 cm regionally from 1910 to 2010). It is also important to note that this region exhibits an out of phase interaction between spring and summer growing seasons,

where precipitation events induce ecosystem pulses of vegetation productivity, nutrient cycling, and fluxes of water and carbon between spring and summer seasons (Petrie, et al., 2014).

Table 4.1: Table of micrometeorological variables used in the study with their abbreviated, full variable names, and SI units.

S/N	Abbreviated Variable Name	Full Variable Name	SI Units
1	TS	Soil Temperature	$^{\circ}\text{C}$
2	TA	Air Temperature	$^{\circ}\text{C}$
3	SW-OUT	Shot Wave Outgoing Radiation	W/m^2
4	SW-IN	Short Wave Incoming Radiation	W/m^2
5	NETRAD	Net Radiation	W/m^2
6	SWC	Soil Water Content	%
7	PA	Atmospheric Pressure	kPa
8	P RAIN	Precipitation (Rain)	mm
9	RH	Relative Humidity	%
10	LEAF WET	Leaf Wetness	mV
11	WD	Wind Direction	degrees
12	WS	Wind Direction	m/s
13	PPFD_IN	Photosynthetic Photon Flux Density Incoming	$\mu\text{mol}/\text{m}^2/\text{s}$
14	PPFD_OUT	Photosynthetic Photon Flux Density Outgoing	$\mu\text{mol}/\text{m}^2/\text{s}$
15	G	Soil heat flux	W/m^2
16	LW-OUT	Long Wave Outgoing Radiation	W/m^2
17	LW-IN	Long Wave Incoming Radiation	W/m^2



Figure 4.1: Map showing the Jornada study site in the northern Chihuahuan Desert (Ramirez, G. 2011)

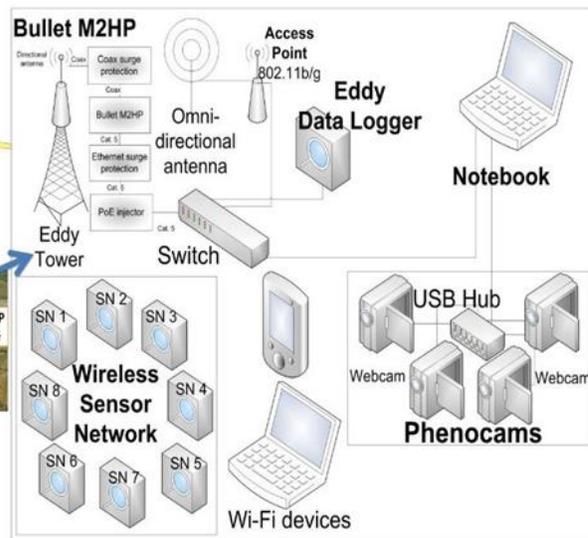


Figure 4.2: UTEP-JER site showing a range of different sensing systems including the Robotic Tram System (Ramirez, G. 2011).

4.3 Methods

4.3.1 Exploratory Data Analysis

To understand the data and ensure accuracy of predictions, we adopted different approaches in exploring the data through exploratory data analysis. All variables were analyzed using python and its relevant libraries to check for missing values, null values, percentage of missing values, and the data type of each variable as seen in figure 4.3 below. In addition, we employed descriptive statistics to explore features of the datasets by generating summary statistics such as minimum, mean, standard deviation, first quartile and the maximum values of the numerical variables. See figure 4.4 below.

	Zero Values	Missing Values	% of Total Values	Total Zero Missing Values	% Total Zero Missing Values	Data Type
LW_OUT	0	65696	36.6	65696	36.6	float64
LW_IN	0	65696	36.6	65696	36.6	float64
NETRAD	0	49026	27.3	49026	27.3	float64
SW_OUT	0	49026	27.3	49026	27.3	float64
SW_IN	0	49025	27.3	49025	27.3	float64
TS	0	43247	24.1	43247	24.1	float64
WD	0	30379	16.9	30379	16.9	float64
RH	0	27789	15.5	27789	15.5	float64
TA	0	27789	15.5	27789	15.5	float64
WS	0	27789	15.5	27789	15.5	float64
SWC	0	27695	15.4	27695	15.4	float64
LEAF_WET	64732	27688	15.4	92420	51.5	float64
PPFD_OUT	0	27688	15.4	27688	15.4	float64
G	0	26031	14.5	26031	14.5	float64
PA	0	21363	11.9	21363	11.9	float64
PPFD_IN	582	14176	7.9	14758	8.2	float64
P_RAIN	167637	8657	4.8	176294	98.3	float64

Figure 4.3: Data points showing zero values, missing values, % of total values missing, total zero + missing values, its percentage, and the data type for each of the 30-minute from 2010-2020

	LW_OUT	LW_IN	NETRAD	SW_OUT	SW_IN	TA	RH	WD	WS	G	LEAF_WET	PA	PPFD_IN	PPFD_OUT	P_RAIN	SWC	TS
count	113611	113611	130281	130281	130282	151518	151518	148928	151518	153276	151619	157944	165131	151619	170650	151612	136060
mean	415.16	317.6	88.99	44.44	232.26	17.26	36.49	178.33	3.46	-0.19	5.84	86.08	495.69	59.65	0.02	4.09	8.3
std	79.97	52.42	214.71	58.26	321.05	9.55	21.76	63.93	2.11	15.98	19.1	0.41	677.02	81.26	0.26	3.01	3.44
min	253.78	185.31	-159.38	-5.97	-8.26	-24.69	2.82	5.31	0.2	-35.46	0	84	-0.56	0.91	0	-1.06	0
25%	353.61	276.14	-71.96	2.03	-1.89	10.06	19.21	125.26	2.02	-12.32	0	85.82	0.51	1	0	2.19	5.41
50%	405.29	315.13	-32.05	3.77	5	17.9	31.32	174.51	3.1	-5.27	0.15	86.09	14.46	1.32	0	3.28	8.25
75%	457.91	359.52	244.02	82.33	458.58	24.54	50.4	231.27	4.26	11.04	1.35	86.35	974.23	116.83	0	5.57	11.07
max	698.53	448.91	908.46	473.02	1173.94	40.25	98.91	352.54	18.74	88.22	100	87.76	2649.49	874.65	46.48	18.41	20.43

Figure 4.4: Summary statistics of the data for exploratory data analysis. See table 4.1 for the full names and the SI units of all the variables above.

Following, all missing values as seen in fig. 4.3 were removed to ensure only available data points were used for further analysis. Data were then visualized to check for outliers on both the dependent and independent variables (figure 4.5), including the alignment of the datapoints to determine if some of the datapoints required normalization, by plotting the histogram of each of the variables, as seen in figs. 4.6 below.

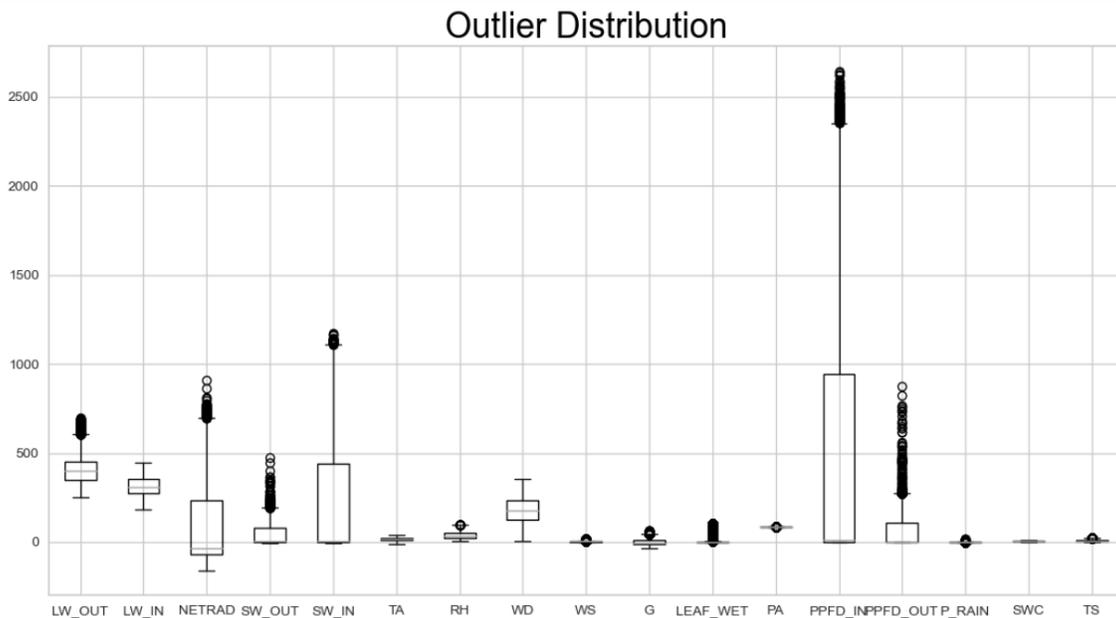


Figure 4.5: Box plot showing outlier distribution of the data points. See table 4.1 for all the full names and the SI units of all the variables above.

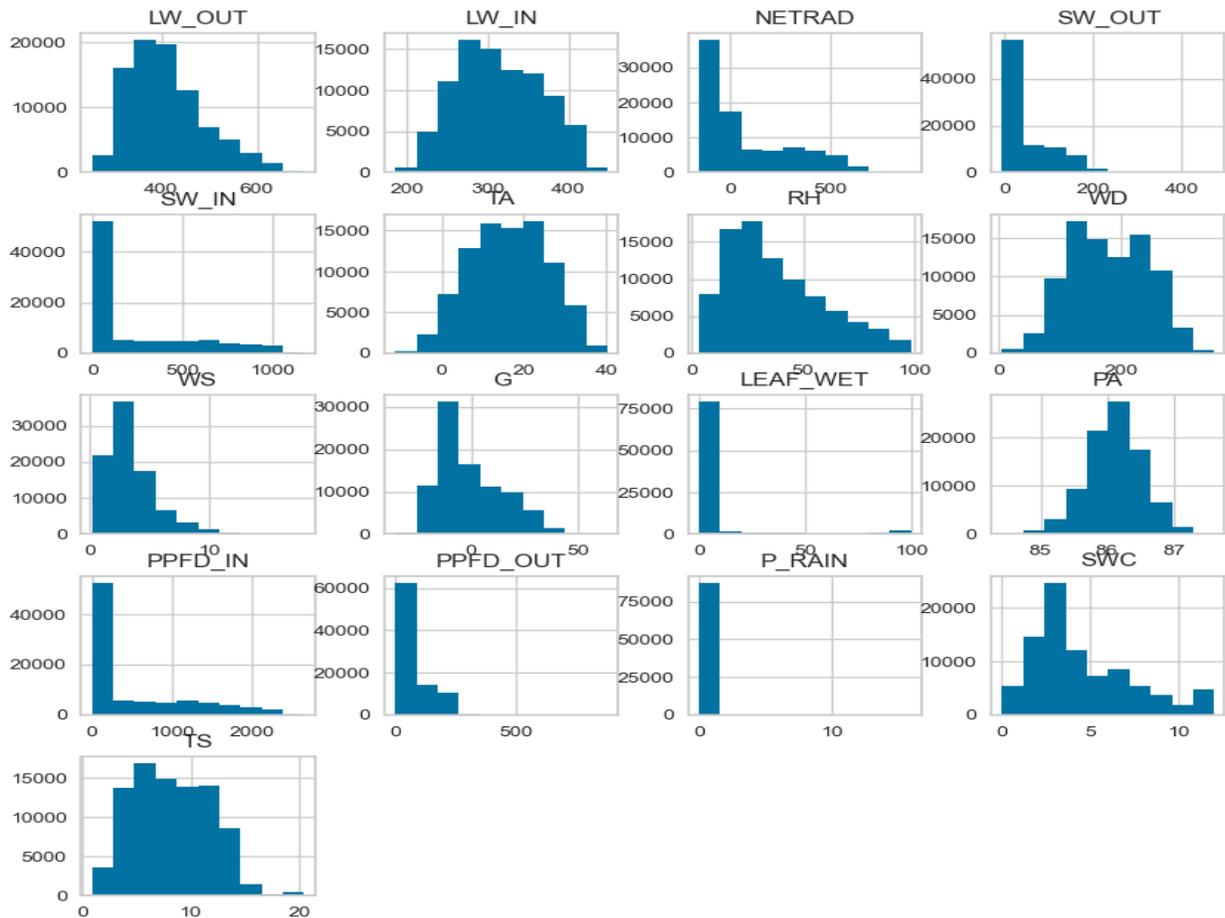


Figure 4.6: Histogram to determine the frequency distribution of the data points for each variable. See table 4.1 for the full names and SI units of all the variables above.

4.3.2 Variable Transformation

Normally distributed data is essential for the application of large-scale statistical analysis. To statisticians, the adequacy and normal distribution of data are very crucial. However, data analysts are usually forced to deal with unusual data, including transforming data from non-normal to normally distributed data (Hamasha, et al., 2022). Therefore, it is important to transform the non-normal data points to something close to normal as different statistical methods require distinct levels of normality (Hamasha, et al., 2022). There are different non-normal to normality transformation methods available based on the nature of the data being transformed such as Log transformation, Box-Cox transformation, Yeo-Johnson, Reciprocal, and

Square-Root transformation methods. Since most of our variables are numeric, consisting of both negative and zero values, we must apply a transformation method that is consistent with our data, hence, we utilized some numeric variable transformation techniques like standardization - implemented in scikit-learn – a python ML Library, and the Yeo-Johnson transformation method, allowing the values to be centered around the mean with a unit standard deviation. The Yeo-Johnson power transformation that allows data to be more Gaussian-like and removes skew in the data distribution was also utilized (Hamasha, et al., 2012). The variables in our dataset that were transformed using the above methods include: P_RAIN, NETRAD, SW-OUT, SW_IN, LEAF_WET, PPFD_IN, and PPFD_OUT as seen in fig. 4.7 below.

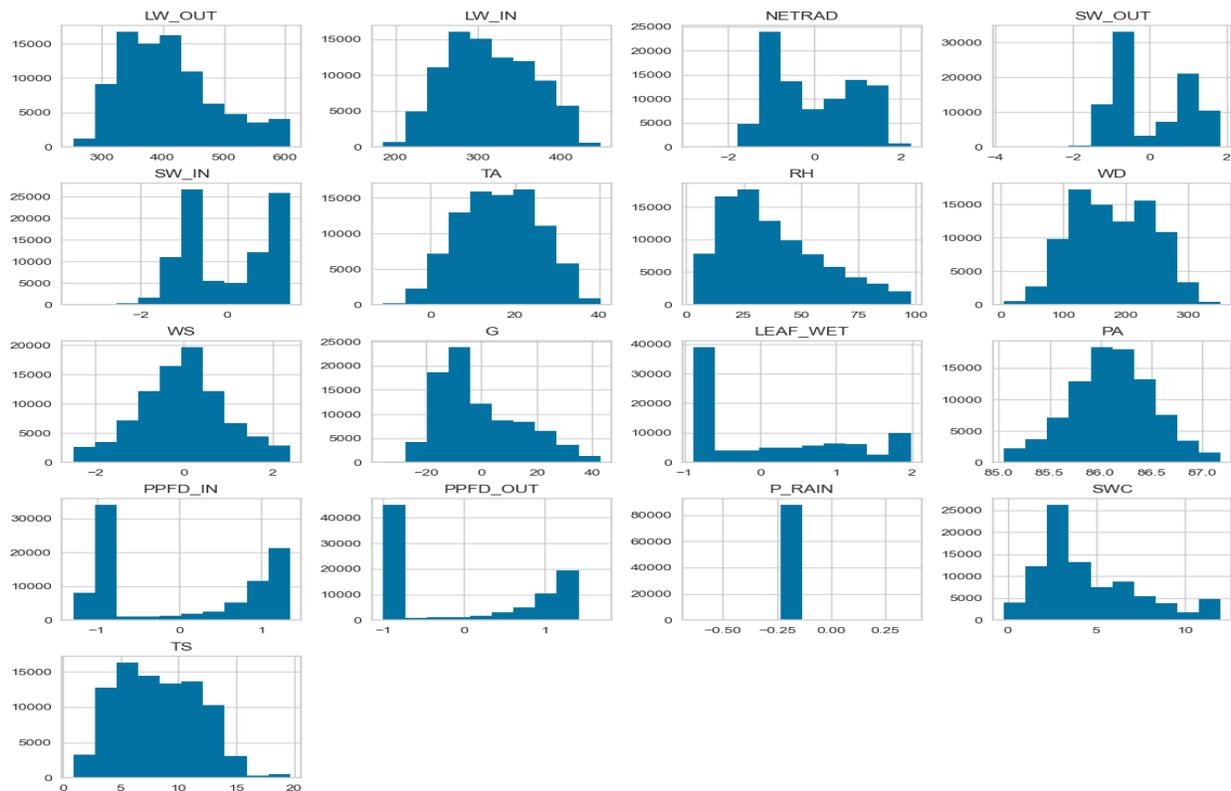


Figure 4.7: Histogram showing transformed variables (P_RAIN, NETRAD, SW-OUT, SW_IN, LEAF_WET, PPFD_IN, and PPFD_OUT) using Yeo-Johnson power transformation method. See table 4.1 for the full names and SI units of all the variable above.

4.3.3 Removing Outliers, Checking Multicollinearity, and Removing strongly correlated feature variables.

Upon further analysis of the data, we found some outlying values in the data points and removed them, but also tested our final model with and without outliers. We used an Interquartile range technique and calculated the difference between the third (Q3) – the 75th percentile and the first quartile (Q1) – 25th percentile to return values at a given quantile within a specified range. The data points that fall below $Q1 - 1.5IQR$ or above $Q3 + 1.5IQR$ were deemed outliers. It is also pertinent to note that, for the purposes of analysis and prediction accuracy, we also decided to model the data with the datasets deemed as outliers, in order to compare the results with or without outliers after the final analysis. Figures 4.8 and 4.9 below show the datasets before and after removing outliers, respectively.

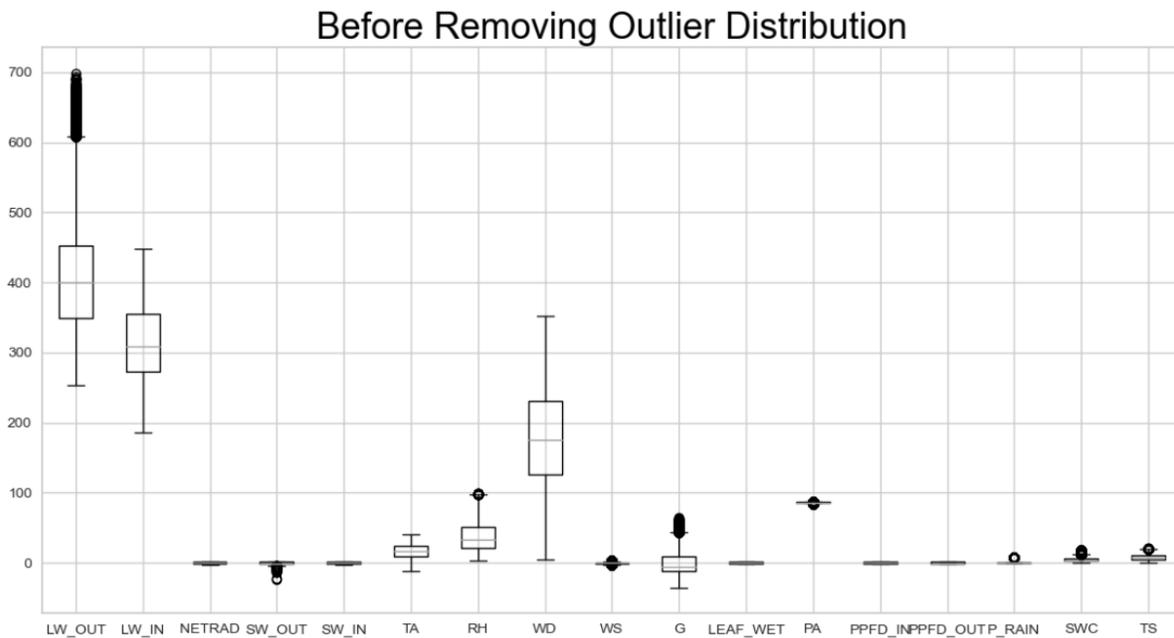


Figure 4.8: Box plot of variables before removing data points deemed as outliers. See table 4.1 for the full names and the SI units of all the variables above.

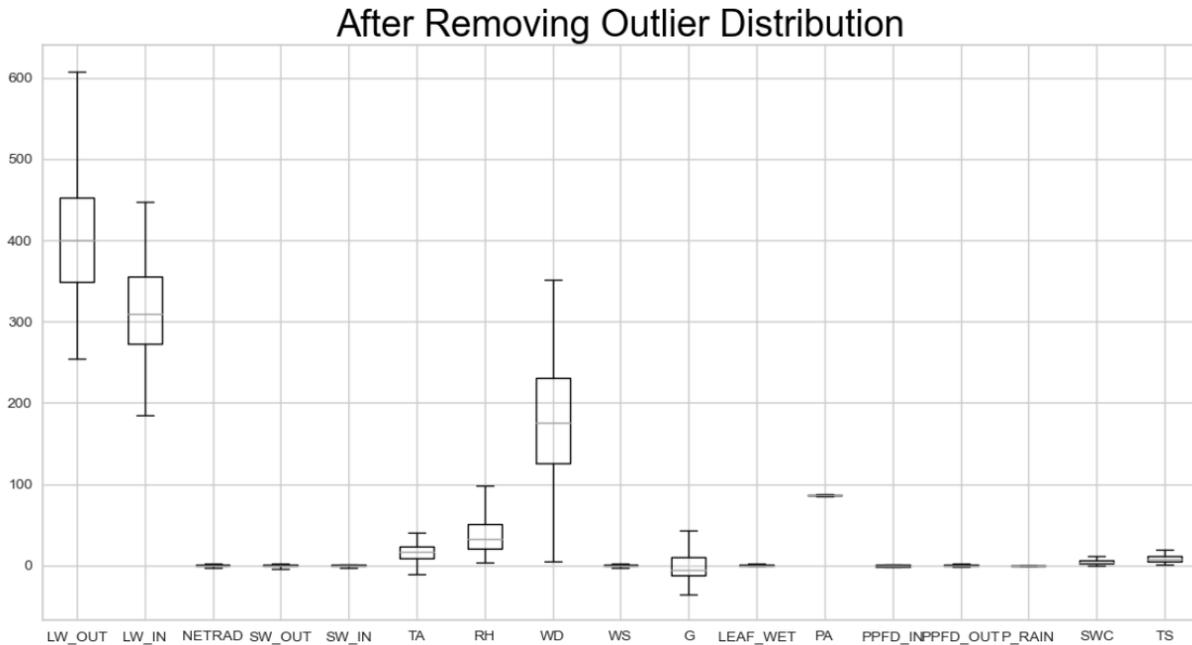


Figure 4.9: Box plot of variables after removing data points deemed as outliers. See table 4.1 for the full names and the SI units of all the variables above.

Furthermore, we explored the strength of the relationship between these variables by creating a correlation heatmap (figure 4.10). The correlation heatmap helps to provide a quick and intuitive way to visualize the correlation matrix and identify relationship between variables. Upon closer examination of the correlation heatmap, we found several variables are multi-correlated.

However, multi-collinearity portends a significant issue for linear models. Collinearity can cause unstable parameter estimation, unreliable models, and weak prediction ability (Cheng, J. et al., 2022). In other to address this problem, the Variance Inflation Factor (VIF) was introduced for feature selection. VIF is a tool that helps to identify the degree of multicollinearity. VIF measures how much the behavior (variance) of an independent variable is influenced, or inflated, by its interaction/correlation with the other independent variables (Cheng, J. et al., 2022). Following best practice (Cheng, et al, 2022), any feature variable with VIF values above the threshold of 5 was removed. The correlation heatmap below shows the remaining feature variables after applying VIF to the variables (Figure 4.11). To compare the prediction accuracy between the transformed and untransformed, we also performed the final modelling both with and without removing the multi-correlated feature variables.

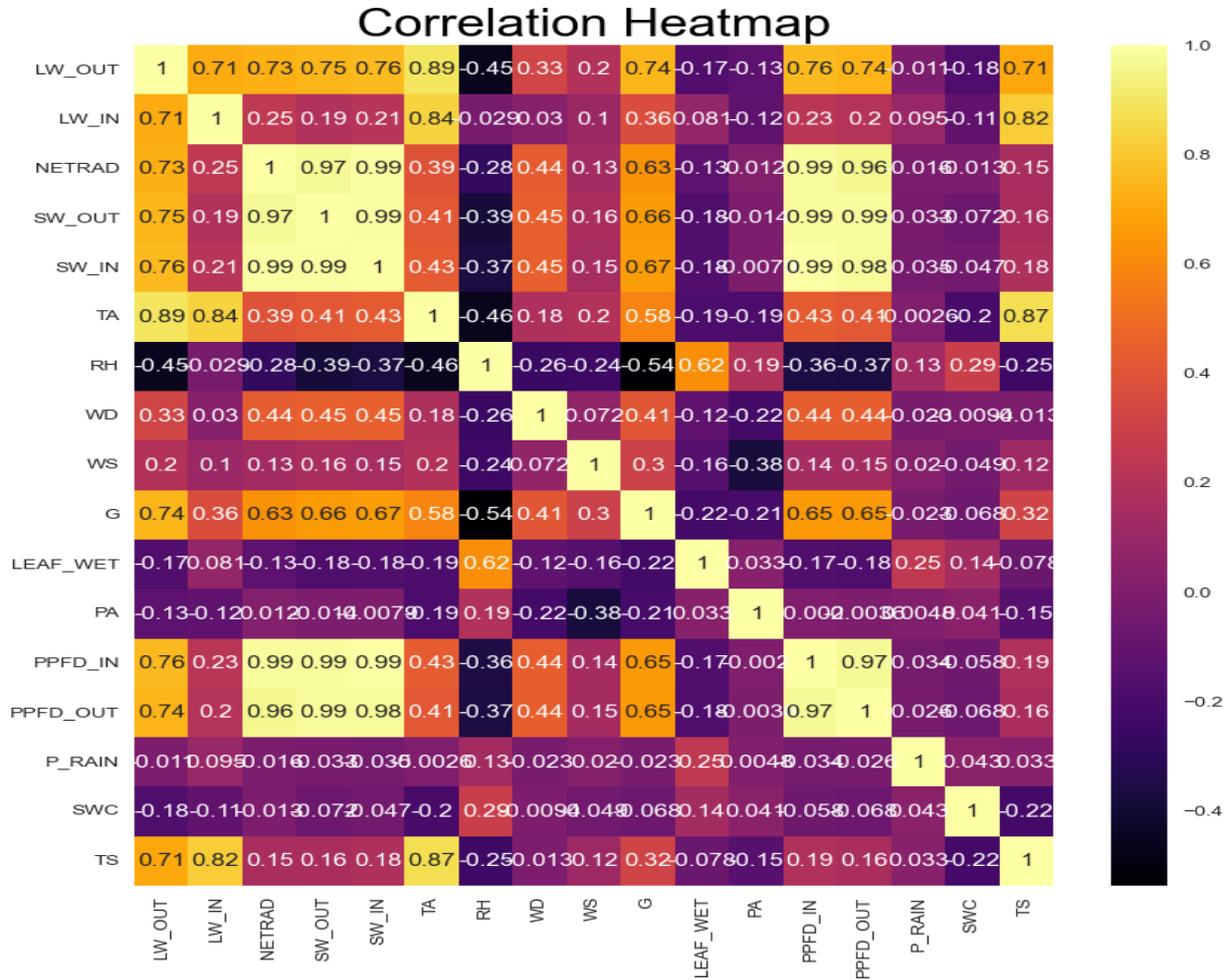


Figure 4.10: Correlation heat map of all the variables without data transformation. See table 4.1 for the full names and the SI units of all the variables above.

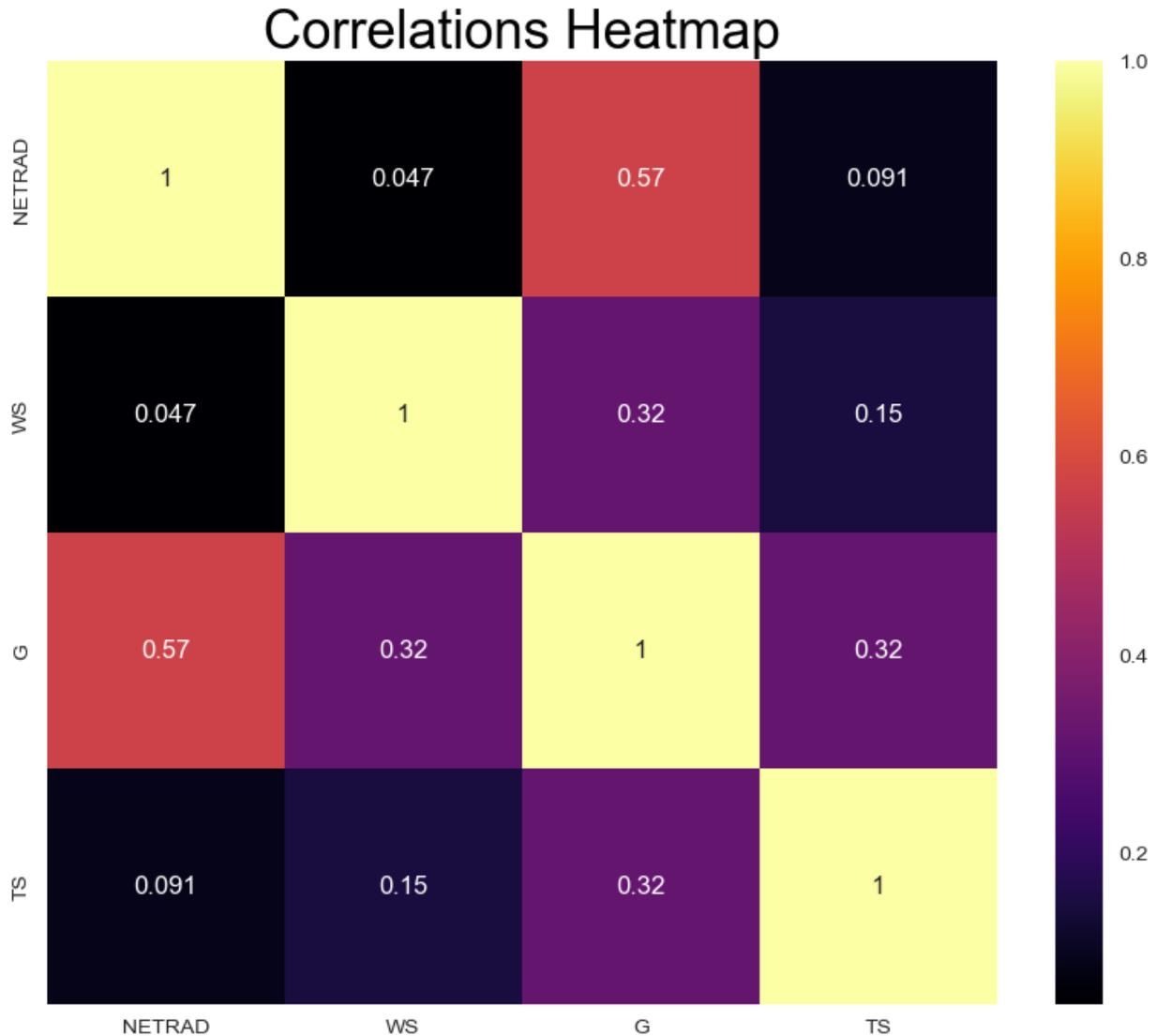


Figure 4.11: Correlation heat map after multi-correlated feature variables were removed using Variable Inflation Factor. See table 4.1 for the full names and the SI units of all the variables above.

4.3.4 Modelling Techniques:

Machine learning (ML) methods utilize a variety of statistical, probabilistic, and optimization methods to learn from the data, and detect useful patterns from large, unstructured, and complex dataset (Uddin, et., 2019 & Mitchel, 1997). In this study, we are mostly adopting supervised learning approaches for our modelling. In supervised learning (SL), a labelled training dataset is first used in training the algorithm. This trained algorithm is then fed on the unlabeled test dataset to categorize data into similar groups. SL is well suited to problems focused on either classification or regression-based issues. For this reason, we have identified Decision Tree Classifier (DTC),

Logistic Regression (LR), Random Forest (RF), and a deep learning algorithm – Long Short – Term Memory (LSTM) as different algorithms for modelling our problem - soil moisture dynamics. In each of these methods, we first evaluate the prediction capabilities of the model on the training sets of the data, before proceeding to evaluate the prediction capabilities of the models on the data set aside for testing called the ‘test set’. For the classification models - logistic regression and Decision tree classifier, we summarized the dependent variable - SWC, and found that the minimum value is -0.2114, while 4.56, 2.86, 2.52, 6.33, and 12.049 were the mean, standard deviation, first quartile, third quartile, and maximum, respectively. Since this is a binary classification, it procedurally dictates that a certain threshold needs to be assigned to distinguish between the two classes. In this regard, we made the histogram (Figure 4.12) below, to see the frequency distribution of the dependent variable (SWC) and utilized percentile – a statistical measure that indicates the value below which a certain percentage of observation in the dataset will fall. It is used to describe the distribution of a set of data by dividing it into one hundred equal parts. Based on the frequency distribution from the histogram, we took the 80th percentile as the threshold; this means that any value of SWC greater than or equal to 80th percentile was classified as 1 – to symbolize a rare event, while any value of SWC that falls below 80th percentile was classified as 0 – an event that is not rare. The value of SWC corresponding to the 80th percentile is 6.9%. Based on the 80th percentile categorization, we found that 17591 SWC data points fell into this category of ‘1’ (rare events) and 70362 fell to the category of 0 (not rare events). Note: the sole purpose for the use of percentile is to convert the dependent variable to a categorical variable for binary classification. Hence, in our models, we are predicting the accuracy of Soil Water Content events equal to or greater than the high soil moisture threshold event, that is, the less frequently observed soil moisture with higher magnitudes in terms of the measured values.

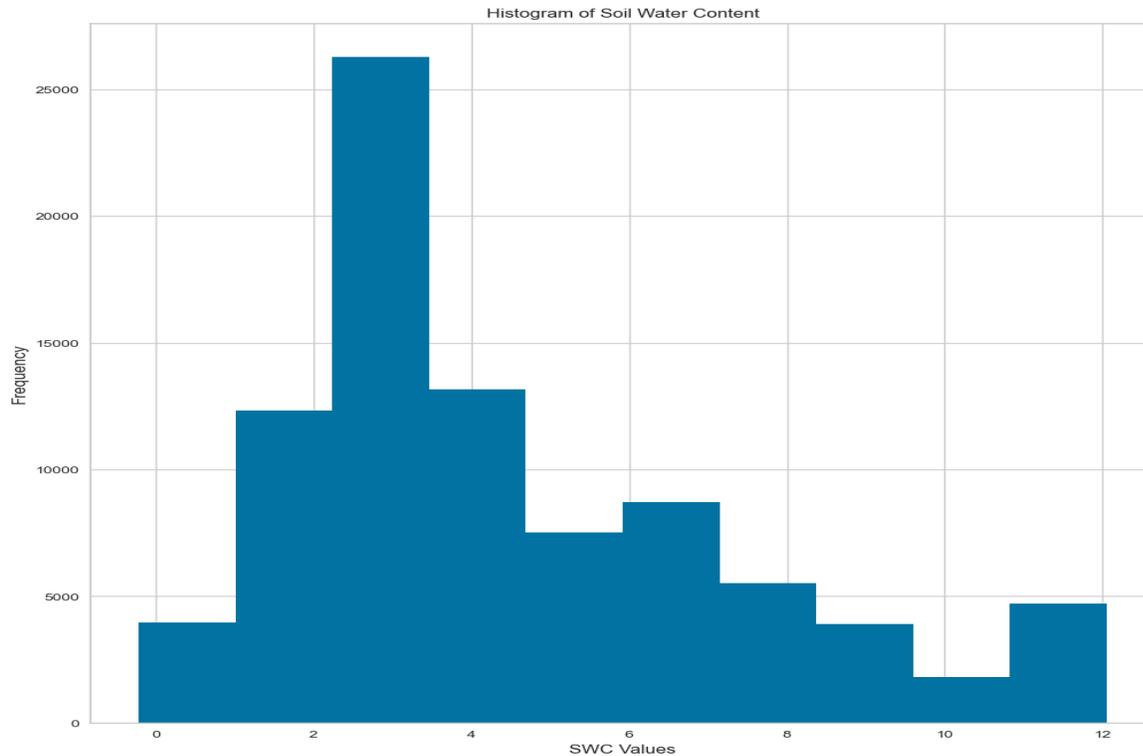


Figure 4.12: Histogram of the dependent variable (SWC).

4.3.5 Decision Tree Classifier, Logistic Regression, Random Forest, and Long Short – Term Memory

4.3.6 Decision Tree Classifier

DTC is a kind of nonlinear supervised classification model which has tree-like structures as a classifier. Here, the connection point between branches represents the condition for discrimination and the leaf nodes at the end of the branches represent the categories the records belong to. In using DTC, we continuously select different branches and repeat the process until reaching the leaf node (Zhang et., 2021). The nodes of a DT normally have multiple levels where the topmost or first node is known as the root node. All nodes having at least one child or internal node represent the test data or input variables. The classification algorithm branches towards the appropriate child node where the process of test and branching repeats until it reaches the leaf node depending on the test outcome (Uddin, et al., 2019). DTs have been found to be relatively easy to interpret and quick for users to learn in other fields (Cruz, et al., 2006).

To identify each node, an attribute and a split condition on a given attribute minimizes the mixing of class labels, resulting in pure subsets. We utilized the GINI index to evaluate the

goodness of fit. The GINI index determines the purity of a specific class after splitting along a particular attribute (Zhang et., 2021). The best split increases the purity of the sets resulting from the split. We also utilized a test size of 0.25, a random state of 50, with a maximum depth of 6, maximum features of 4, and minimum sample leaf of 10 to build our decision tree classifier. Continuously, we ran the algorithm with all these parameters but without filtering outliers from the data to compare with results derived from data sets where outliers were excluded, as described above. Additionally, we ran the algorithm without invoking any kind of variable transformation technique or check to removing multi-collinearity but filtered out datapoints considered as outliers. This will further help evaluate and compare the overall accuracies of each of these techniques.

4.3.7 Logistic Regression

Logistic Regression is one of the most frequently used methods in supervised learning for the purpose of prediction. To predict the class labels of instances, the logistic regression assigns each instance as a probability. The probability of an anomalous instance ($y_j = 1$) is denoted by $P(y_j = 1|X_j)$, and $P(y_j = 0|X_j) = 1 - P(y_j = 1|X_j)$, for $y_j = 0$. The estimated probability is learned based on the sigmoid basis function: $P(y_j = 1|X_j) = f(g(X_j, \beta)) = \frac{1}{1+e^{-g(x_j, \beta)}}$, ----- (1)

where $0 \leq f(g(X_j, \beta)) \leq 1$ and $g(X_j, \beta) = \beta_0 + \sum_{i=1}^T B_i \cdot x_{t_i}^j$ is a linear expression including the explanatory features and the regression coefficients β . Equation 1 is called the linear logistic regression (Nikou Gunnemann-Gholizadeh, 2018; Bishop, 2006; Goodfellow et al., 2016). Being that it is a probability, the outcome lies between 0 and 1. We utilized the processes outlined in section 3.4 above to distinguish between the two classes, thereby classifying soil water content values that meet the 80th percentile threshold as either rare or not a rare event depending on whether they are greater than or equal to or less than 6.9% which is the SWC value at 80th percentile.

We used the same data transformation procedures as DTC for logistic regression to model the data. We also modelled the data without any transformation to compare the results before and after. A variable importance plot helped to determine the contribution of each of the features selected in the overall performance of the model (Figures 4.14 & 4.15).

4.3.8 Random Forest

Random Forest (RF) was also used to model the rare events in soil moisture. RF – is an ensemble regression consisting of many decision trees similar to a forest having numerous trees. Its foundation is based on bagging and random subspace methods (Ganesh, et al., 2021). With bagging, several learner trees are created and ensembled to ensure prediction accuracy is obtained. Following, independent bootstrap samples are created from the original training sample data for use in training the learner trees. Each bootstrap sample is created by drawing examples from the original training data. It is allowed to replace the examples while creating the bootstrap samples. In general, the bootstrap samples may be around 2/3 of the training data, without any duplicate examples.

In the Random Forest regression Model, bagging reduces variance and overfitting in the ensemble, making it important for the learner trees to be correlated. Importantly, the samples from the original training datasets which were not selected for training the regression tree during bagging are collated to constitute an out of bag (OOB) dataset. The regression tree's performance in terms of mean square error is calculated based on OOB, which is usually one-third of the training dataset (REF). For the random forest regression model on non-dichotomous target variable – Soil Water Content (SWC), we transformed some of the variables that were badly skewed using the Yeo-Johnson power transformation method. We also removed outliers as well as checked and removed multi-correlated feature variables using variable inflation factor (VIF) values above 5.

Several accuracy metrics were implemented in the model. This included the coefficient of determination (R^2), consideration for the total number of samples, the mean absolute error (MAE), the mean squared error (MSE), as well as root mean squared error (RMSE). In this study, accuracy metrics help to understand how well the model performed in predicting the dynamics of soil moisture in dryland region.

4.4.9 Long Short-Term Memory (LSTM)

Neural network (NN) methods have an extraordinarily strong learning potential and the capacity to represent nonlinear relationships between the inputs and outputs of a system (Adeyemi et al., 2018). NN has been applied to some specific water resources management problems such as crop yield prediction (Guo et al., 2014 & Gandhi et al., 2016), rainfall-runoff modelling (Sarkar et al.,

2012 & Khan et al., 2006) and the prediction of soil moisture to aid irrigation (Capraro et al., 2008 & Tsang et al., 2016). Specifically, the LSTM, a class of Recurrent Neural Networks (RNN), has been successfully applied in the control of nonlinear dynamic systems (Wang et al., 2017 & Wang, 2017). The LSTM model – which is a branch of NN requires minimal input data for pre-processing and can preserve vital information across multiple time steps (Chauhan, S. & Vig, L., 2015). It has shown excellent performance in modelling water table depth according to Zhang et al. (2018), where they applied time series data on water dispersion, evaporation, precipitation, and temperature as inputs to the model. The authors reported R^2 scores ranging between 0.789 and 0.952 for the LSTM models, outperforming other models. The excellent water table depth prediction demonstrated by the LSTM models highlights their ability to preserve and learn previous information from long-term time series data. This ability is particularly desirable in modelling soil moisture dynamics in dryland ecosystems and underpins the rationale for including it in this study.

The LSTM is the output of a NN feeding back to the input that allows modelling data sequence or chains of information. Since the vanishing gradient problem of the training algorithm of RNN occurs because of backpropagation through time, the LSTM method uses several gated units to overcome this challenge (Gonzalez, J. & Yu, W., 2018). In this model, a total of 64 memory cells were used, with a l2 kernel regularizer that has a value of 0.01. There is also the input shape, the timesteps – that specifies the timesteps in the input sequence, and the input dimension that entails the number of features in each time step, and further complied with binary cross-entropy, RMSprop with a value of 0.01, and accuracy - as the loss function, optimizer, and the metrics, respectively. Subsequently, the model is trained with input training data and corresponding output training data. in this study, the model was trained with 50 epochs, a batch size of 16, and a validation split of 0.2.

4.4 Results

4.4.1 Analysis of the results from different models

Biogeophysical data collected from 2010 to 2020 at our Chihuahuan Desert site were analyzed using four different machine learning models to understand and predict rare soil moisture events. Firstly, we modelled the data with DTC without transforming the data or filtering out outliers, and a prediction accuracy of 88.8% was achieved, with both precision and recall values of 93%.

However, when datasets deemed to be outliers were removed and skewed features were filtered out using a variance inflation factor (VIF), we noticed a reduction in the accuracy of the prediction. Prediction accuracy reduced to 81%, and the precision also reduced to 82%, while the recall increased to 99%. Moreso, when the model was built without removing outliers but excluded multi-correlated feature variables, the predication accuracy remained the same at 81%, precision, and recall slightly varied to 83% and 97%, respectively.

When the logistic regression model was run without filtering outliers, transformed, or removed strongly correlated features, including skewed features, the prediction accuracy was 80.23%, with precision and recall values of 82% and 97%, respectively. When run with outliers and skewed factors removed, the prediction accuracy did not see a significant change at approximately 80%, with precision and recall of 80% and 100%, respectively. Accordingly, when the model was built with outliers included and multi-correlated feature variables excluded, the prediction accuracy was reduced to 79%, while the precision and recall were pegged at 81% and 96%, respectively. These prediction results show that DTC, LR, and LSTM were more than 80% accurate in predicting that Soil Water Content with values equal or greater than 6.9% are rare soil moisture events. Table 4.1 below shows the results from DTC and LR.

Table 4.2: DTC & LR modelling results before & after transforming the data with and without outliers.

Algorithm Name	Categories	Precision	Recall	F1-score	Accuracy
Decision Tree Classifier	0	0.93	0.93	0.93	0.8876
	1	0.72	0.72	0.72	
Decision Tree Classifier (After Transforming & removed)	0	0.82	0.99	0.89	0.8123
	1	0.68	0.10	0.17	

outliers)					
Logistic Regression	0	0.82	0.97	0.89	0.8023
	1	0.51	0.12	0.20	
Logistic Regression (data transformed, outliers & strongly correlated features excluded)	0	0.80	1.00	0.89	0.7986
	1	0.00	0.00	0.00	
Logistic Regression (with outliers but excluded strongly correlated features)	0	0.81	0.96	0.88	0.7896
	1	0.40	0.12	0.18	

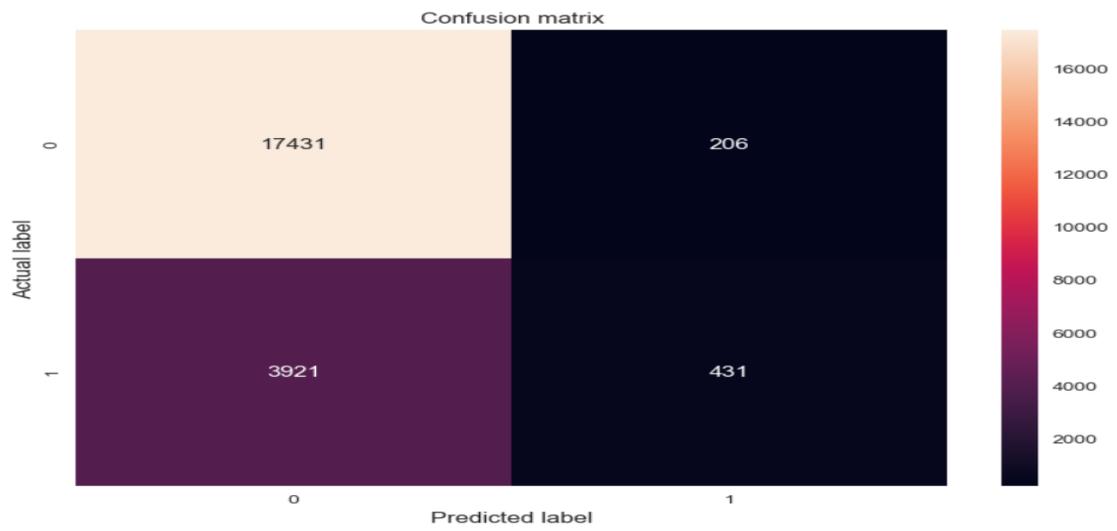
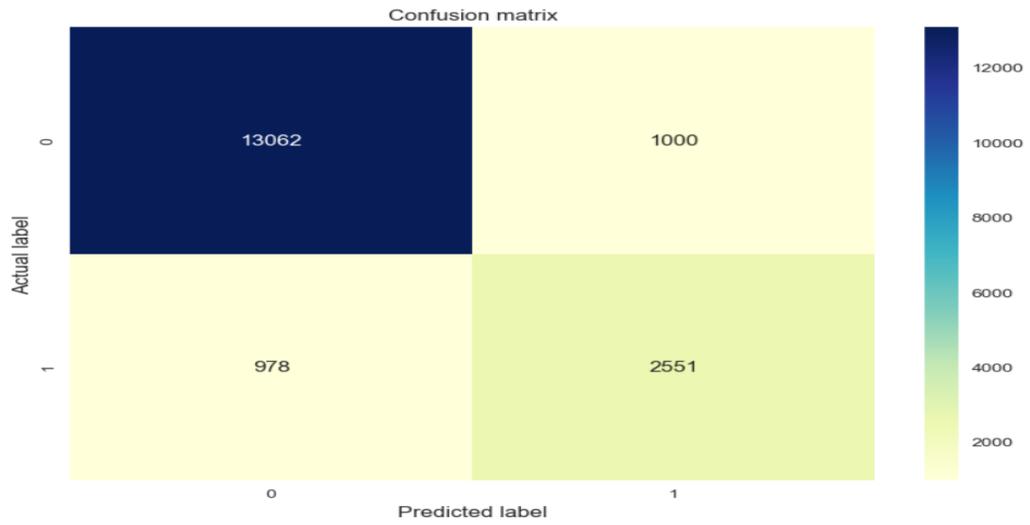


Figure 4.13: DTC Modelling confusion matrix before & after transforming the data and with or without outliers

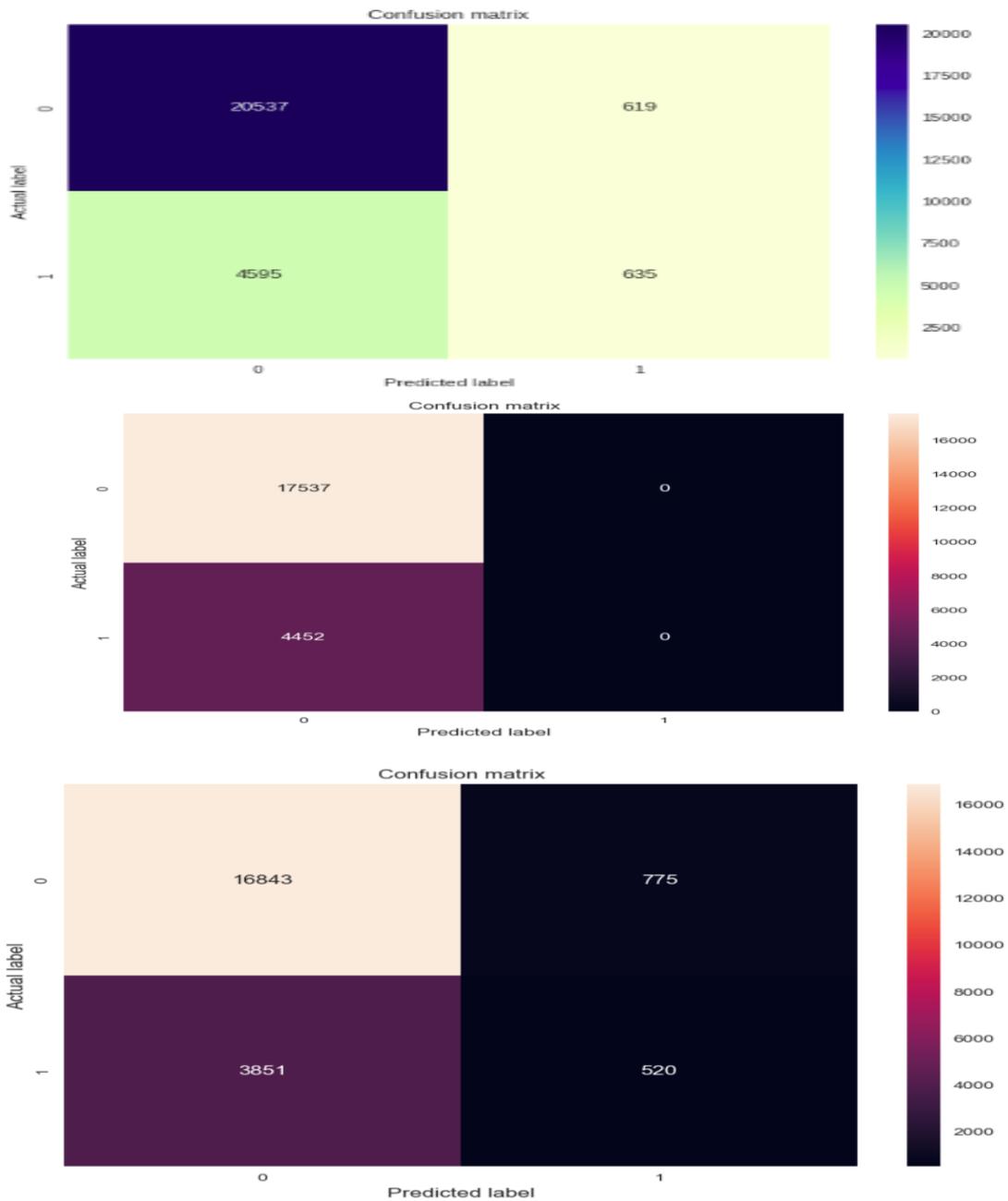


Figure 4.14: Logistic Regression confusion matrix for the different scenarios described above in order.

The result of the random forest regression model on the non-dichotomous target variable – Soil Water Content (SWC), see table 4.2 below, shows an R^2 of 21%, it implies that the model is only able to explain 21% variation in the dataset.

Table 4.3: Results of Random Forest Regression Model

Algorithm	Metrics	Results
Random Forest Regression Model	Mean Squared Error	6.53
	Root Mean Squared Error	2.55
	Mean Absolute Error	1.96
	R^2	0.21

Furthermore, for a deep learning model, LSTM, at 50 timesteps, batch size of 16, sigmoid as activation, loss function as binary cross entropy, and Adam as optimizer, we applied the model without filtering out outliers or transforming as well as not removing strongly correlated feature variables, the model achieved an accuracy of 88% with a loss of 0.2664. When we transformed the data using the Yeo-Johnson transformation method, removed outliers and strongly correlated features, we arrived at an accuracy of 81% and loss of 0.4120. Alternatively, when we kept the outliers but transformed and removed strongly correlated features, the prediction accuracy remained the same at 82% with a loss of 0.4098. Table 4.3 below shows the results in a clear form.

Table 4.4: Results from LSTM model.

Algorithm name	Epoch	Loss	Accuracy
LSTM	50/50	0.2664	0.8813
LSTM (With transformation, outliers & correlated feature excluded)	50/50	0.4120	0.8141

LSTM (Outliers included but transformed & correlated features excluded)	50/50	0.4098	0.8168
---	-------	--------	---------------

The Variable Importance Plot (VIP) for the DTC model, showed that Soil temperature and soil heat flux are the most important features for predicting rare soil moisture events with 16% and 14%, respectively, without the transformation or removing outliers from the data and over 54% and approximately 38% contributions, respectively, when the data was transformed, outliers filtered out, strongly correlated features removed, including skewed feature variables. When the data was transformed and outliers were removed with strongly correlated features, but without removing skewed features, the contribution of soil temperature and soil heat flux to the prediction retained their importance relative to other feature variables at 30% and 26%, respectively. With the random forest regression model, the VIP interestingly showed that soil heat flux made the highest contribution to determine rare soil moisture event with feature value of 50%, closely followed by Soil temperature with a 35% contribution.

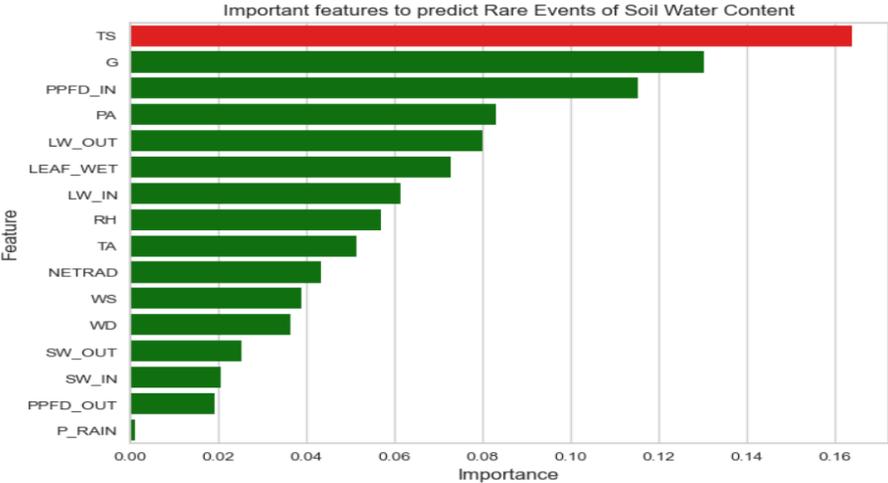


Figure 4.15: DTC VIP without transforming or outliers removed. See table 4.1 above for the full names and the SI units of all the variables in this plot.

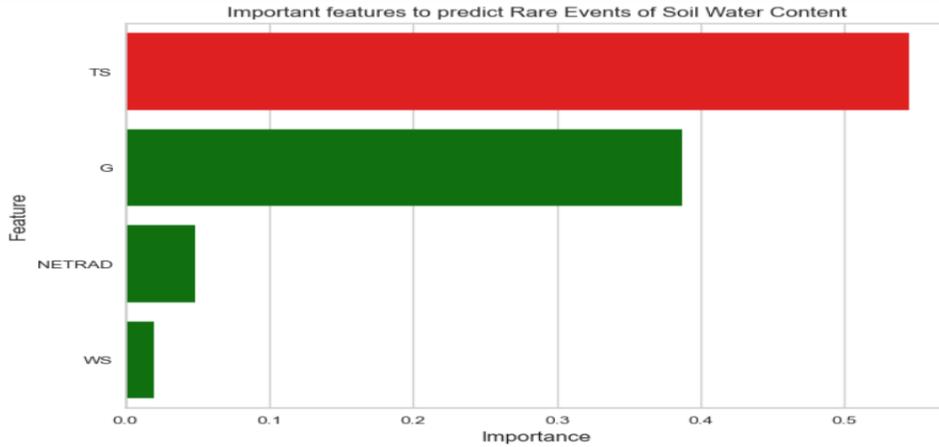


Figure 4.16: DTC VIP with transformed data, outliers and strongly correlated features removed. See table 4.1 above for the full names and the SI units of all the variables used in the plot above.

To measure the effectiveness of the model and to assess how well the best model captured and predicted rare soil moisture dynamics, we analyzed the confusion matrix of the best model (DTC) shown below in Figure 4.12. A confusion matrix is a performance measurement for machine learning classification methods, where output can be two or more classes (Das, et al., 2022). It is a table of four different combinations of actual and predicted values, where each row represents an instance in actual values, while each column represents the predicted values, or vice versa (Kulkarni, et al., 2020 & Das, et al., 2022). The four different combinations include: true positive (TP), true negative (TN), false positive (FP), and false negative (FN). The TP means both the actual and predicted values are positive, that is the number of actual positive examples are classified accurately (Kulkarni, et al., 2020 & Das, et al., 2022). We had 13062 examples predicted as rare soil moisture events that are actual rare events from the DTC. FP means that the actual value is negative, but the model predicted value is positive, that is the number of actual negative values classified as positive. In our model, that value stands at 978, i.e. 978 unrare soil moisture events were classified as rare events. FN on the other hand, is the number of actual positive examples classified as negative and we have 1000 false negatives in our best performing model. Lastly, TN, means both the actual and predicted values are negative, that is, the number of negative examples classified accurately. In our model, we had 2551 classified as unrare soil moisture events that are actually unrare events. In the analysis above, we can see that the number of TP and TN are very high compared to FP and FN as predicted by the model. This further attests to the accuracy of our predicted results, which depends largely on

the TP and TN. In other words, the misclassification is low since the number of FP and FN are very low.

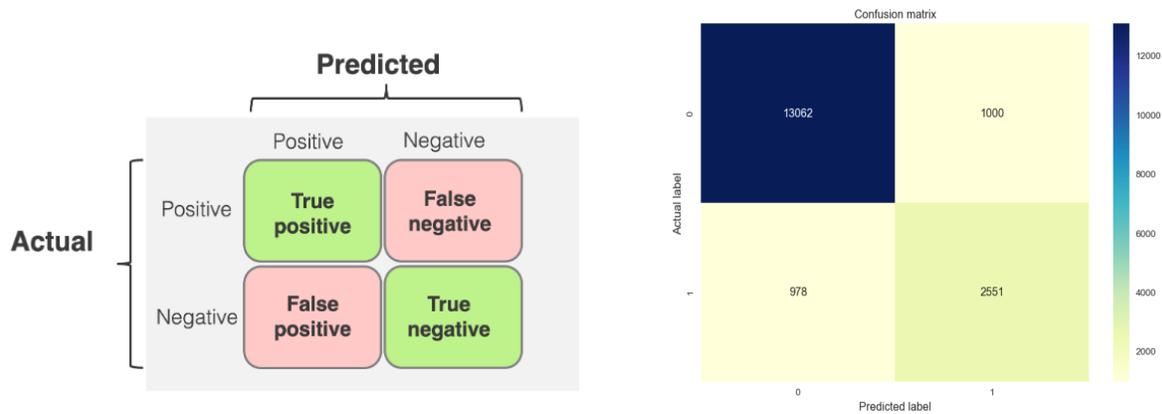


Figure 4.17: Visual interpretation of confusion matrix & confusion matrix of the best performing model in predicting rare soil moisture events

4.4.2 Discussion

Predicting rare soil moisture events in dryland ecosystems is essential as rising temperatures and shifting precipitation events are increasing the frequency and severity of drought across all dryland regions (Chenoweth, et al., 2022). Understanding drought severity in these regions is essential to manage ecohydrological and agricultural processes to ensure human well-being and sustainable environmental management (Bradford, et al., 2019). This study utilized multi-sensor cross-correlation to predict rare events, that is soil moisture events that are less frequently occurring with higher magnitudes in terms of its measurement values compared to others, in time series data in a northern Chihuahuan Desert shrubland situated on the Jornada Experimental Range in Southern New Mexico using predictive approaches. We chose a predictive approach to model this problem based on its capacities as outlined in the introduction section above.

Results affirmed the importance of soil temperature and soil heat flux as the best predictors of soil moisture variability based on the ML models applied. Soil Moisture has been shown to have a tremendous influence on soil temperature by controlling the partitioning between sensible and latent heat (Seneviratne, et al., 2016). This leads to more absorbed radiative energy being re-distributed into latent heat flux (Jin, S.N. & Mullens, T., 2014). Therefore, soil moisture interacts with soil temperature in controlling the exchange of water and heat energy between the land surface and the atmosphere through evaporation and plant transpiration. Results from the VIP

illustrating that soil temperature and soil heat flux have a greater influence on modeling soil moisture than rainfall is understandable considering that this region experiences prolonged periods of relatively high temperatures without rainfall.

Land-atmosphere exchange of water and heat is an interactive process, where temperature gradients affect the soil moisture potential and both liquid and vapor movement in the soil (Julie, A. et al., 2021). The effects of soil moisture on the surface energy exchange may have a profound impact on climate change dynamics in the Chihuahuan Desert region. Hence, from our models findings that showed 88.8% accuracy soil moisture events that are equal or above 6.9% in value or greater than or equal to 80th percentile are rare soil moisture events, with soil temperature and heat flux being the most contributing variables to soil moisture dynamics, ecologists may be able to understand small and large-scale drought patterns within the Chihuahuan Desert dryland region, and link their affect to global or changing climate.

It is important to note that the prediction accuracies were comparable among the models we utilized, with variation occurring depending on the state of the data. For example, we generally obtained higher prediction accuracy when there was no transformation of the data, no outliers were filtered out, and strongly correlated and skewed features were retained. The prediction accuracy slightly changed when we either filtered outliers, transformed the data, and/or removed strongly correlated features and strongly correlated features. This is noticed across the three Models of DTC, LR, and LSTM. This may be connected to the size of the data, as it has been established that ML algorithms, especially classification algorithms, perform poorly with limited sized data sets (Althnian, A., et al., 2021, Pitchler & Hartig, 2023). This is attributed to the fact that limited datasets will likely lead to less details, hence, the model cannot generalize patterns in training data (Althnian, A., et al., 2021). In addition, it may also lead to overfitting of the model (Althnian, A., et al., 2021, Scowen, et al., 2021). Overfitting occurs when the model becomes too complex and fits the training data too closely, leading to deficient performance on new unseen (test) data (Althnian, A., et al., 2021, Pitchler & Hartig, 2023). The above postulation could aptly explain the slight reduction in accuracy of our models when the size of our data was reduced following the removal of outliers, multi-correlated features, and skewed features.

In the Random Forest Regression Model, the accuracy metrics used, such as R^2 , showed a different prediction strength in modeling rare soil moisture events compared to the other models.

For instance, the RF R^2 value of 21% showed that the model only explained 21% of the variation in the dataset. The mean squared error, which measures the amount of error in the model, was valued at 6.53, which could be viewed to be high as values closer to zero are usually more appropriate. This deficient performance of the RF regression model could be attributed to the inability of RF to extrapolate training data, challenging the model to make effective predictions based on the average of the previously observed labels. To state it differently, in a regression problem, the range of prediction a RF model can make is bound by the highest and the lowest labels in the training data (Jonsson, E., & Fredrikson S., 2021). This behavior becomes problematic in situations where the training and prediction inputs differ in their range and/or distribution, which is the case for our data, as the distribution of each of the variables in the training sets are different. Overall, our model's performance were better for classification problems compared to regression as seen from the results of the random forest regression analysis and other modelling approaches, we deployed.

Future directions would be to incorporate the different depth of soil moisture measurement into the models. This cross-correlation analysis could further be extended to analyze common climatic indices such as drought, wind patterns, wildfires, heatwaves, etc., to back-cast soil moisture in the past, which could help determine long-term change and event durations.

4.5 Conclusion

This paper has presented different modelling approaches such as DTC, LR, and RF as well as the deep learning approach LSTM to model rare soil moisture events between 2010 and 2020 at a dryland study site in the northern Chihuahuan Desert of Southern New Mexico. The performances of each modelling approach were evaluated based on the state of the data – filtering or retaining outliers, transforming, or not transforming the data, excluding or not excluding strongly correlated, and skewed features variables. All the models but the regression-based Random Forest achieved a prediction accuracy of over 80%, with DTC and LSTM reaching a prediction accuracy of 88.4% and 88%, respectively. The precision and recall which measures the accuracy of positive predictions and the completeness of positive predictions, respectively, were all far above 85%. ML and DL algorithms are powerful predictive modelling and data analysis tools for modelling complex ecosystem attributes such as rare events in soil moisture and how it affects drought patterns in the region. Hence, this prediction will further aid

ecologists in understanding drought severity in these regions and help them to manage ecohydrological and agricultural processes to ensure human well-being and sustainable environmental management.

Chapter 5: General Discussion

The overarching goal of this dissertation is to *develop ecoinformatics tools that will contribute to the advancement of global change science* through; I) mitigating the challenges of new infrastructures for Big Data archiving, management and sharing, and analysis by developing a flexible system that supports multiple and novel data usage and visualization and II) understanding complex relationships among variables in a data set. In this chapter, we will discuss each of the objectives that addressed the above-mentioned overarching goal. Future research directions are also presented.

5.1 Overview of the dissertation

The first objective of this dissertation was to develop a tool that will aid synthesis, analysis, and discovery of spectral data collected from multiple instruments and lab groups to enhance data reuse and availability to wider user community. We achieved this by successfully developing web based analytic tools capable of integrating spectral reflectance data from multiple instruments actively used in the NASA ABoVE domain. R-HyperSpectral will help to dynamically view, interact, and discover optical properties of boreal and tundra plant communities. Users can view the hyperspectral reflectance scans and explore common spectral indices at temporal scales.

In the Arctic, multiple streams of detailed field measurements of vegetation optical properties along with corresponding airborne and satellite remote sensing observations, particularly, hyperspectral reflectance measurements are being collected and used by NASA ABoVE investigators. These streams of data aid them to understand Arctic change, ecosystem vulnerabilities, including permafrost thaw and degradation, snow cover and sea ice loss, and air temperature rise, among others. However, due to the size and complexity of the data being collected, managing, analyzing, sharing, and visualizing has become challenging. With these challenges in mind, researchers have adopted some other ways of sharing and managing their data such as storing it in local archives or exchanging data through emails, which could make the data prone to destruction, data loss, as well as bring about accessibility issues.

To mitigate the susceptibility of data under – utilization, loss, destruction, coupled with the challenges of managing, analyzing, sharing, reuse, and visualization of these streams of

hyperspectral reflectance data, we developed R-HyperSpectral. R-HyperSpectral will not only aid the synthesis and integration of diverse hyperspectral data streams but also, give the research community access to data for improved rapid change detection across the ABoVE domain.

Similarly, the second objective of the dissertation is motivated by the urgency to address what has been identified as a critical need by the ecological community. Addressing lab groups that collect multiple streams of data (climate, ecological, sensor, human observations/measurements) at one or several networked sites but have no means of managing, analyzing, visualizing, and sharing these data sets have persisted in the environmental and ecological science communities for decades. To achieve this objective and help address these challenges faced by the community, we developed *rDataFusion* – a multi-data fusion tool capable of aggregating heterogeneous data sets collected from a range of automated and semi-automated sensors and manual observations over a decade-long period.

rDataFusion, has the capability to integrate and filter data from two instrument nodes and different data streams that include micro-meteorological variables (e.g., temperature, relative humidity), soil conditions (e.g., temperature and soil moisture), and ecosystem trace gas and energy fluxes. After initial compilation and filtering, users can visualize data in near real-time to check that all sensors are running properly, and/or ensure preliminary flagging for data that is deemed out of range or problematic in some way. *rDataFusion*, also, has the capacity for exploratory data analysis through quality control and quality assurance processes and allow for identifying missing values, outliers, and gap-filling missing or problematic data, visualize data to allow for preliminary summaries and interpretations, and compare data across time or by site.

With Chapters 2 and 3, we identified an issue within the ecological community and proffered a solution by building a web-based information and data management system that allows researchers to analyze and dynamically visualize data. This also streamlines documentation of data and incorporates metadata management, including data aggregation. These new data management tools or system have the capability to improve data management with others in the ecological research community faced with similar challenges, and the open-source code will permit users to modify and/or add new or extended modules that can be further shared and innovated. This could potentially in the future become a new data management solution for the

majority of the ecological or environmental research community and allow for easy access and discovery of data for research.

Lastly, the objective of data chapter 4 was to test and gain spatiotemporal information about soil moisture anomalies or dynamics and how to predict them in the dryland ecosystems to better understand small and large-scale drought patterns. We achieved this objective by utilizing multi-sensor cross-correlation to predict rare soil moisture events in temporal data using some Machine Learning and Deep Learning (DL) models. This is we did by deploying several Machine and Deep Learning techniques and cross correlated these sensors for optimal rare soil moisture events detection in the Northern Chihuahuan Desert, in Southern New Mexico. Specifically, the machine and deep learning techniques used for this study include both classification and regression methods, including Decision Tree Classifier (DTC), Logistic Regression Classifier (LR), Random Forest Regression (RF), and the Long Short-Term Memory (LSTM) method of Artificial Neural Network (ANN). Our models predicted that soil moisture events that are equal or greater than 6.9% are rare events, with prediction accuracy of over 88%.

In conclusion, through the tools we developed, data will be available for ecological and environmental science researchers to analyze and further understand ecosystem changes over a range of spatial and temporal scales and levels of biological organization and interaction. It will also aid Arctic ecosystem researchers to understand ecosystem changes such as coastal erosion, permafrost thaw and degradation, air temperature rise, snow cover and sea ice loss orchestrated by Changes in climate variability over the past decades. Furthermore, the analysis and prediction of rare soil moisture events in the dryland ecosystem unveils a pathway to understanding soil moisture events and the key driver of soil moisture in drylands. The understanding of ecosystem and climate variability either in the desert or in the arctic, and drivers of soil moisture coupled with rare soil moisture events detection in dryland are all case studies that collectively contribute to the advancement of the global change sciences.

5.2 Future research directions

We have developed web-based data analytic tools that could aid ecologists better manage their data. The study opens the possibilities for future research directions. Modifying machine and deep learning models to accommodate soil moisture measurements at different depths, including seasonal variations will enhance knowledge of how rare soil moisture events persist in dryland

ecosystems. Seasonal variation of soil moisture availability has a profound impact on dryland ecological niches through linkages between soil moisture and vegetation composition and functioning. Soil moisture content may be different in different locations. Creating ML models that clearly incorporate these locational differences to see if there are emergent properties between sites that could serve as predictors and seasonal variations will further enhance the understanding of both small and large-scale drought patterns in the region.

Furthermore, improving the detailed understanding of the processes and mechanisms that control soil moisture in dryland ecosystems would be a viable research opportunity in the future. Estimating soil moisture availability and the key drivers in dryland ecosystems is critical for advancing earth system science, planning, and management, and understanding the impacts of drought on the ecosystems. This will help develop more consistent methods to approximate soil moisture dynamics needed for ecosystem functioning and resource management.

Another future direction would be to link R-HyperSpectral to the ABoVE Spectral Library (ASTRAL)- a web mapping tool for spectral information to introduce a geospatial component. Another potential future research direction would be to find ways to improve on the data aggregation and management components of the rDataFusion tool to develop and/or include metadata for datasets. Expanding rDataFusion to include information about data provenance, content, structure, and permission, so that data will be discoverable for future use, will further enhance the tool's potential. Importantly, expanding on most of the research questions such as "How can we develop or create a workflow or template for project-specific multi-data fusion?" to include data from other researchers and open it up to other projects that are similar and would be neat to explore. Similarly, expanding and creating standard frameworks that do not only take into consideration the set of quality control and assurance standards across the network, but also facilitate intra and inter - site data integration and spatial comparison will enhance site intercomparison, transfer lessons learned at one spot to another.

Reference

- Adeyemi, O., Grove, I., Peets, S., Domun, Y., and Northon, T. (2018). Dynamic Neural Network Modelling of Soil Moisture Content for Predicting Irrigation Scheduling. *Sensors* 18, 3408; doi:10.3390/s18103408
- Ali, I., Greifender, F., Stamenkovic, J., Neumann, M., Notarnicola, C. (2015). Review of machine learning approaches for biomass and soil moisture retrievals from remote sensing data. *Remote Sens.*, 7, 16398-16421.
- Althnian, A., AlSaeed, D., Al-Baity H., Samha A., Bin Dris A., Alzakari, N., Elwafa, A.B., & Kurdi, H. (2021). Impact of Dataset Size on Classification Performance: An Empirical Evaluation in the Medical Domain.. *Appl. Sci.* 2021, 11, 796.
<https://doi.org/10.3390/app11020796>
- Arend, D. (2010). Minitab 17 Statistical Software. Minitab, Inc.,
- Atkinson, D. M. and Treitz, P. (2012). Arctic Ecological Classification Derived from Vegetation Community and Satellite Spectral Data. *Remote Sens.* 2012, 4, 3948-3971; doi:10.3390/rs4123948
- Baird, D. J., Van den Brink, Chariton, A. A., Dafforn, K. A., & Johnson E. L. (2015). Research Front. *Marine and Freshwater Research*. <http://dx.doi.org/10.1071/MF15330>
- Baldrige, A.M., Hook, S.J., Grove, C.I., & Rivera, G. (2008). The ASTER spectral library version 2.0. *Remote Sensing of Environment* 113 (2009) 711–715
- Baldrige, A.M., Hook, S.J., Grove, C.I., & Rivera, G. (2009). The ASTER spectral library version 2.0. *Remote Sensing of Environment* 113 (2009) 711-715
- Barnosky, A. D., Hadly, E. A., Bascompte, J., Berlow, E. L., Brown, J. H., Fortelius, M., Getz, W. M., Harte, J., Hastings, A., Marquet, P. A., Martinez, N.D., Mooers, A., Roopnarine, P., Vermeij, G., Williams, J.W., Gillespie, R., Kitzes, J., Marshall, C., Matzke, N., Mindell, D. P., Revilla, E., and Smith, A. B. (2012). Approaching a state shift in Earth’s biosphere. *Nature* 486, 52–58. doi:10.1038/NATURE11018
- Baron, J. S., Specht, A., Garnier, E., Bishop, P., Campbell, C. A., Davis, F. W. and Winter, M. (2017). Synthesis centers as critical research infrastructure. *BioScience*, 67(8), 750-759. <https://doi.org/10.1093/biosci/bix053>
- Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.
- Bjorkman A.D., Myers-Smith I.H., Elmendorf S.C., Normand S., R uger N., Beck P.S.A., Blach-Overgaard A., Blok D., Cornelissen J.H.C., Forbes B.C., Georges D., Goetz S.J., Guay K.C.,

- Henry G.H.R., Lambers J. H., Karger D.N., Hollister R.D., Manning P., Kattge J., Rixen C., Prevéy J.S., Thomas H.J.D., Schaepman-Strub G., Wilmsking M., Vellend M., Carbone M., Wipf S., Lévesque E., Hermanutz L., Petraglia A., Molau U., Tomaselli M., Vowles T., Soudzilovskaia N.A., Spasojevic M.J., Anadon-Rosell A., Angers-Blondin S., Alatalo J.M., Alexander H.D., Björk R.G., Buchwal A., Beest M.T., Berner L., Cooper E.J., Dullinger S., Buras A., Christie K., Grau O., Frei E.R., Eskelinen A., Elberling B., Heijmans M.M.P.D., Harper K.A., Hallinger M., Grogan P., Iversen C.M., Iturrate-Garcia M., Hülber K., Hudson J., Kaarlejärvi, E., Jørgensen R.H., Johnstone J.F., Jaroszynska F., Lantz T., Little C.J., Speed J.D.M., Michelsen A., Klady R., Kuleza S., Kulonen A., Lamarque L.J., Oberbauer S.F., Olofsson J., Onipchenko V.G., Rumpf S.B., Milbau A., Nabe-Nielsen J., Nielsen S.S., Ninot J.M., Tape K.D., Suding K.N., Treier U.A., Trant A., Shetti R., Semenchuk P., Street L.E., Collier L.S., Boulanger-Lapointe N., Zamin T., Hik D.S., Gould W.A., Tremblay M., Tremblay J.P., Weijers S., Venn S., Wookey P.A., Bahn M., Magnusson B., Tweedie C., Jørgensen J., Klein J., Hofgaard A., Jónsdóttir I.S., Cornwell W.K., Craine J., Cerabolini B.E.L., Chapin F.S., Bond-Lamberty B., Campetella G., Blonder B., Bodegom P.M. van, Onoda Y., Niinemets Ü., Milla R., Green W., Enquist B.J., Díaz S., de Vries F.T., Dainese M., Schamp B., Sandel B., Reich P.B., Poschlod P., Poorter H., Penuelas J., Ozinga W.A., Ordoñez J.C., Weiher E. & Sheremetev S. (2018), Plant functional trait change across a warming tundra biome., *Nature* 562(7725): 57-62. Doi: 10.1038/s41586-018-0563-7
- Boelman, N. T., Stieglitz, M., Rueth, H. M., Sommerkorn, M., · Kevin L. Griffin, K. L., Shaver, G. R., and Gamon, J. A. (2003). Response of NDVI, biomass, and ecosystem gas exchange to long-term warming and fertilization in wet sedge tundra. *Ecosystem Ecology Oecologia* (2003) 135:414–421 DOI 10.1007/s00442-003-1198-3
- Bojinski, S., Schaepman, M., Schlapfer, D., & Itten, K. (2003). SPECCHIO: a spectrum database for remote sensing applications. *Computer & Geosciences* 29: 27-38.
- Bradford, J.B., Schlaepfer, D.R., Lauenroth, W.K., Palmquist, K.A., Chambers, J.C., Maestas, J.D., & Campbell, S.B. (2019). Climate driven shift in soil temperature and moisture regimes suggest opportunities to enhance assessment of dry resilience and resistance. *Frontiers in Ecology and Evolution*. Volume 7 - 2019 | <https://doi.org/10.3389/fevo.2019.0035>
- Bratsch, S. N., Epstein, H. E., Buchhorn, M., and Walker, A. D. (2016). Differentiating among Four Arctic Tundra Plant Communities at Ivotuk, Alaska Using Field Spectroscopy. *Remote Sens.* 2016, 8(1), 51; <https://doi.org/10.3390/rs8010051>
- Brito, J.J., Li, J., Moore, J.H., Greene, C.S., Nogoy, N.A., Garmire, L.X., & Mangul, S. (2020). Recommendations to enhance rigor and reproducibility in Biomedical research. *Gigascience*, 9.
- Brown, J., Everett, K. R., Webber, P. J., MacLean, S. F. J., and Murray, D. F. (1980). The coastal tundra at Barrow. Pages 1– 29 in J. Brown, P. C. Miller, L. L. Tieszen, and F. L. Bunnell, editors. *An arctic ecosystem: the coastal tundra at Barrow, Alaska*. Dowden, Hutchinson, and Ross, Stroudsburg, Pennsylvania, USA.

- Capraro, F., Patiño, D., Tosetti, S. and Schugurensky, C. (2008) Neural network-based irrigation control for precision agriculture. In Proceedings of the 2008 IEEE International Conference on Networking, Sensing and Control (ICNSC), Sanya, China, 6–8 April 2008; pp. 357–362
- Carlson, K. M., Asner, G. P., Hughes, R. F., Ostertag, R., & Martin. R. E., (2007). Hyperspectral remote sensing of canopy biodiversity in Hawaiian lowland rainforests. *Ecosystems* 10:536-549.
- Chang, W. (2018). shinythemes: Themes for Shiny. R package version 1.1.2. <https://CRAN.R-project.org/package=shinythemes>.
- Chang, W., Cheng, J., Allaire, J., Xie, Y., & McPherson J. (2020). shiny: Web Application Framework for R. R package version 1.4.0.2. <https://CRAN.R-project.org/package=shiny>
- Chapin, F. S., III, and Shaver, G. R., (1985). Individualistic response of tundra plant species to environmental manipulations in the field. *Ecology* 66:564–576.
- Chauhan, S., Vig, L. (2015). Anomaly detection in ECG time signals via deep long short-term memory networks. In Proceedings of the 2015 IEEE International Conference on Data Science and Advanced Analytics, Paris, France, 19–21 October 2015
- Chenoweth, D.A., Schlaepfer, D.R., Chambers, J.C., Brown, J.L., Urza, A.K., Hanberry, B., Board, D., Crist, M., Bradford, J.B. (2022). Ecologically relevant moisture and temperature metrics for assessing dryland ecosystem dynamics. US Forest Services. https://www.fs.usda.gov/rm/pubs_journals/2022/rmrs_2022_chenoweth_d001.pdf
- Collins, S. L., Bettencourt, L.M., Hagberg, A., Brown, R. F., Moore, D. I., Bonito, G., Delin, K. A., Jackson, S. P., Johnson, D.W., Burleigh, S.W., Woodrow, R.R., & McAuley, J.M. (2006). New opportunities in ecological sensing using wireless sensor networks. *Frontiers in Ecology and the Environment* 4:402-407.
- Cruz, J.A. & Wishart, D.S. (2006). Application of Machine Learning in Cancer Prediction and Prognosis. *Cancer Informatics*. Doi:[10.1177/117693510600200030](https://doi.org/10.1177/117693510600200030)
- Das, C., Sahoo, A.K. & Pradhan, C. (2022). Multicriteria recommender system using different approaches. *ScienceDirect (Journal and Books)*. Chapter 12. <https://doi.org/10.1016/B978-0-323-85117-6.00011-X>

- Davidson, S.J., Santos, M.J., Sloan, L.V., Watts, J.D., Phoenix, G. K., Oechel, W.C., and Zona, D. (2016) Mapping Arctic Tundra Vegetation Communities Using Field Spectrometry and Multispectral Satellite Data in North Alaska, USA. *Remote Sens.* 2016, 8, 978.
- Death, R.G., Fuller, I.C., Macklin, M.G. (2015). Resetting the river template: The potential for climate-related extreme floods to transform river geomorphology and ecology. *Freshwater Biol.* 60, 2477 – 2496. (doi:10.1111/fwb.12639)
- Dietze, M.C. (2017). Prediction in ecology: a first-principles framework. *Ecological applications*, 27(7), 2018, pp. 2048-2060
- Donoho, D. (2017). 50 Years of Data Science. *Journal of Computational and Graphical Statistics*, 26(4):745–766, ISSN 1061-8600. doi: 10.1080/10618600.2017.1384734.
- Farley, S. S., Dawson, A., Goring, S. J. and Williams, J. W. (2018). Situating Ecology as a Big-Data Science: Current Advances, Challenges, and Solutions. *BioScience* 563, August 2018/Vol. 68 No. 8. doi:10.1093/biosci/biy068
- Faybishenko, B., Versteeg, R., Pastorello, G., Dwivedi, D., Varadharajan, C., & Agarwal, D. (2021). Challenging problems of quality assurance and quality control (QA/QC) of meteorological time series data. *Stochastic Environmental Research and Risk Assessment* 36:1049–1062.
- Foster, A.C., Wang, J.A., Frost, G.V., Davidson, S.J., Hoy, E., Turner, K.W., Sonnentag, O., Epstein, H., Berner, L.T., Armstrong, A.H., Kang, M., Rogers, B.M., Campbell, E., Miner, K.R., Orndahl, K.M., Bourgeau – Chavez, L.L., Lutz, D.A., French, N., Cheng, D., Du, J., Shestakova, T.A., Shuman, J.K., Tape, K., Virkkala, A., Porter, C., & Goetz, S. (2022). Disturbance in North American Boreal Forest and Arctic Tundra: Impacts, Interactions, and Responses. *Environmental Research Letters*, 17 113001
- Frank, D., Reichstein, M., Bahn, M., Frank, D., Mahecha, M. D., Smith, P., Thonike, K., van der Velde, M., Vicca, S., Babst, F., Beer, C., Buchmann, N., Canadell, J. G., Ciais, P., Cramer, W., Ibrom, A., Miglietta, F., Poulter, B., Rammig, A., Seneviratne, S. I., Walz, A., Wattenbach, M., Zavala, M. A., and Zscheischler, J. (2015). Effects of climate extremes on the terrestrial carbon cycle: concepts, processes and potential future impacts, *Glob. Change Biol.*, 21, 2861–2880.
- Gamon, J. A., Cheng, Y., Claudio, H., MacKinney, L. and Sims, D. A. (2006). A mobile tram system for systematic sampling of ecosystem optical properties. *Remote Sensing of Environment* 103:246-254.

- Gamon, J.A., Huemmrich, K.F., Stone, R.S., Tweedie, C.E. (2013). Spatial and temporal variation in primary productivity (NDVI) in the coastal Alaskan tundra: Decreased vegetation growth following snowmelt. *Remote Sensing of Environment*, 129 (2013) 144-153
- Gandhi, N. and Petkar, O. (2016). Armstrong, L.J. Rice crop yield prediction using Artificial Neural Networks. In *Proceedings of the IEEE International Conference on Technological Innovations in ICT for Agriculture and Rural Development*, Chennai, India, 15–16 July 2016; pp. 105–110.
- Ganesh, N., Jain, P., Choudhury, A., Dutta, P., Kalita, K., & Barsocchi, P. (2021). Random Forest Regression based Machine Learning Model for Accurate Estimation of Fluid Flow in Curved Pipes. *MDPI: Processes* 9 (11), 2095; <https://doi.org/10.3390/pr9112095>
- Garner, G., Van Loon, A.F., Prudhomme, C., Hannah, D.M. (2015) Hydro climatology of extreme river flows. *Freshwater Biol.* 60, 2461 – 2476. (doi:10.1111/ fwb.12667)
- Ghahramani, Z, G., (2015). Probabilistic machine learning and artificial intelligence. *Nature*, 521 (2015), pp.452-459.
- Gonzalez, J., & Yu, W. (2018). Non-linear System Modelling using LSTM network. Elsevier; *IFAC PapersOnLine*, Vol.51, Issue 13, Pages 485 – 489.
- Gonzalez, L., (2011). Development Of a Low-Cost Network of Webcams for Monitoring Plant Phenology in a Chihuahuan Desert Shrubland. Master’s Thesis, The University of Texas at El Paso.
- Goodfellow, I., Bengio, Y., Courville, A., (2016). *Deep Learning*. MIT Press, – <http://www.deeplearningbook.org>
- Goswami, S., Gamon, J. A. and Tweedie, C. E. (2011). Surface hydrology of an arctic ecosystem: Multiscale analysis of a flooding and draining experiment using spectral reflectance. *Journal of Geophysical Research-Biogeosciences* 116:Paper #G00I07, 01-14.
- Granell, C., Díaz, L. and Gould, M. (2010). Service-oriented applications for environmental models: Reusable geospatial services. *Environmental Modelling & Software* 25:182-198.
- Guo, W.W. and Xue, H. (2014). Crop yield forecasting using artificial neural networks: A comparison between spatial and temporal models. *Math. Probl. Eng.* 2014, 7.

- Hamasha, M.M., Ali, H., Hamasha, S., & Ahmed, A. (2022). Ultra-fine transformation of data for normality. Elsevier, 2405-8440. Vol. 8, Issue 5. DOI: <https://doi.org/10.1016/j.heliyon.2022.e09370>
- Hampton, S.E. and Parker, J. N. (2011). Collaboration and productivity in scientific synthesis. *BioScience* 61: 900–910.
- Hampton, S.E., Jones, M. B., Wasser, L.A., Schildhauer, M. P., Supp, S.R., Brun, J., Hernandez, R. A., Boettiger, C., Collins, S. L., Gross, L. J., Fernandez, D.S., Budden, A., Ethan, W. P., Teal, T. K., Labou, S.G., and Aukema, J.E. (2017). Skills and Knowledge for Data-Intensive Environmental Research. *BioScience*. June 2017/Vol.67 No. 6
- Heidorn, P.B. (2008). Shedding Light on the Dark Data in the Long Tail of Science. *Library Trends* 57, 280 – 299.
- Hey, T., et al., (2009). *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Microsoft. Microsoft Research, Redmond, WA. https://www.microsoft.com/en-us/research/wp-content/uploads/2009/10/Fourth_Paradigm.pdf
- Hilker, T., Nestic, Z., Coops, N. C. and Lessard, D. (2010). A new, automated, multi-angular radiometer instrument for tower-based observations of canopy reflectance (Amspec II). *Instrumentation Science & Technology* 38:319-340
- Hovden, R., Cueva, P., Mundy, J.A., & Muller, D.A. (2013). The Open-sources Cornell Spectrum Imager. *Microscopy Today*, Volume 21, Issue 1, Pages 40–44, <https://doi.org/10.1017/S1551929512000995>
- Hoven, R., Cueva, P., Mundy, J.A., & Muller, D.A. (2012). *The Open-Source Cornell Spectrum Imager*. Published online by Cambridge University Press: 21 December 2012.
- Howe, B., G. Cole, E. Souroush, P. Koutris, A. Key, N. Khoussainova, & L. Battle. (2011). Database-as-a-service for long-tail science. *Scientific and Statistical Database Management*. Springer. Pg 480-489.
- Hruska, J., Adao, T., Padua, L., Marques, P., Cunha, A., Peres, E., Sousa, A., Morais, R., and Sousa, J. J. (2018). Machine learning classification methods in hyperspectral data processing for agricultural applications. *JCGDA '18*, April 20-22, 2018
- [Huber, R., D’Onofrio, C., Devaraju, A., Klump, J., Loescher, H.W., Kindermann, S., Guru, S., Grant, M., Moris, B., Wyborn, L., Evans, B., Goldfarb, D., Genazzio, M.A., Ren, X., Magagna, B., Thirmann, H., & Stocker M. \(2021\). Integrating data and analysis technologies within leading environmental research infrastructures: Challenges and approaches. *Ecological Informatics*, Vol. 61. 101245.](#)

- Huemmrich, K.F.; Gamon, J.A.; Tweedie, C.E.; Entcheva Campbell, P.K.; Landis, D.R.; Middleton, E.M. Arctic tundra vegetation functional types based on photosynthetic physiology and optical properties. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* 2013, 6, 265–275
- Huntingford, C., Jeffers, E.S., Bonsall, M.B., Christensen H. M., Lees, T., and Yang, H., (2019). Machine learning and artificial intelligence to aid climate research and preparedness. *Environ. Res. Lett.* 14(2019) 124007
- Intergovernmental Channel on Climate Change (IPCC). (2007). Working Group, I: The Physical Science Basis of Climate Change. In ARA 4 Report, Observations 4: Changes in Snow, Ice and Frozen Ground; 2007. Available online: <http://ipcc-wg1.ucar.edu/wg1/wg1-report.html>
- Intergovernmental Channel on Climate Change (IPCC). (2019): Summary for Policymakers. In: IPCC Special Report on the Ocean and Cryosphere in a Changing Climate [H.- O. Pörtner, D.C. Roberts, V. Masson-Delmotte, P. Zhai, M. Tignor, E. Poloczanska, K. Mintenbeck, M. Nicolai, A. Okem, J. Petzold, B. Rama, N. Weyer (eds.)]. In press.
- IPCC, (2012). *Managing the risks of extreme events and disasters to advance climate change adaptation*. Cambridge, UK: Cambridge University Press.
- Jaimés Hernández, A. (2014) *Furthering Our Understanding and Scaling Patterns and Controls of Land-Atmosphere Carbon, Water and Energy Exchange in a Chihuahuan Desert Shrubland with Novel Cyberinfrastructure*. Ph.D. Dissertation, The University of Texas at El Paso, Texas.
<http://search.proquest.com/docview/1561147292>
- Jiang, W., Chen, H., Yang, L., & Pan, X. (2022). MoreThanANOVA: A user-friendly Shiny/R application for exploring and comparing data with interactive visualization. *PLoS ONE* 17(7).
- Jin, S.N. & Mullens T. (2014). A Study of the Relations between Soil Moisture, Soil Temperatures and Surface Temperatures Using ARM Observations and Offline CLM4 Simulations. *Climate*, 2(4), 279-295; <https://doi.org/10.3390/cli2040279>
- Jonsson, E., & Fredrikson S. (2021). An Investigation of how well Random Forest Regression can predict demand of Groceries. <https://www.diva-portal.org/smash/get/diva2:1594694/FULLTEXT01.pdf>

- Joshi, K.P., & Jamadar, D.C. (2021). Statistical Software Applications and Statistical Method used in Community Medicine and Public Health Research Studies. *National Journal of Community Med.* 12(3):53-56. DOI:10.5455/NJCM.20210329094615
- Julie A., Howe A., & Peyton S. (2021). Principles and application of soil microbiology. (Third Edition).
- Kasprzak, P., Mitchell, L., Kravchuk, O., & Timmins, A. (in archive). Six Years of Shiny in Research Collaborative Development of Web Tools in R.
- Khan, M.S. and Coulibaly, P. (2006). Bayesian neural network for rainfall-runoff modeling. *Water Resour. Res.* 42, 1–18.
- Knudby, A., LeDrew, E., and Brenning, A., (2010) Predictive mapping of reef fish species richness, diversity and biomass in Zanibar using IKONOS imagery and machine-learning techniques. *Remote Sensing of Environment* 114(6): 1230-1241. DOI: 10.1016/j.rse.2010.01.007
- Krisnawijaya, N.N.K., Tekinerdogan, B., Catal, C., & Van der Tol., Rik. (2022). Data analytics for agricultural systems: A systematic literature review. *Computer and Electronics in Agriculture.* Vol. 195, 106813.
- Kulkarni, A., Chong, D., & Batarseh, F.A. (2020). Foundations of data imbalance and solutions for a data democracy. Elsevier. *Data Democracy at the Nexus of Artificial Intelligence, Software Development, and Knowledge Engineering.* Pages 86-105.
<https://doi.org/10.1016/B978-0-12-818366-3.00005-8>
- LaDeau, S.L., Han, B. A., Rosi-Marshall, E.J. and Weathers, K. C. (2017). The next decade of big data in ecosystem science. *Ecosystems* 20: 274–283.
- LaDeau, S.L., Han, B.A., Rosi-Marshall, E.J., & Weathers, K.C. (2016). The Next Decade of Big data in Ecosystem Science. *Springer Link.*, 20, 274 – 283
- Landau, S. & Everitt, B. (2004) *A Handbook of Statistical Analyses Using SPSS.* Chapman & Hall/CRC, Boca Raton, ISBN 978-1-58488-369-2
- Landau, S. & Everitt, B. (2004) *A Handbook of Statistical Analyses Using SPSS.* Chapman & Hall/CRC, Boca Raton, ISBN 978-1-58488-369-2

- Laney, C., Borgman, and Bourne, P. (2022). Why it takes Village to Manage and Share Data. HDSR, Issue 4.3. <https://doi.org/10.1162/99608f92.42eec111>
- Laney, C.M., (2013). Toward new data and information management solutions for data-intensive ecological research. Ph.D. Dissertation, The University of Texas at El Paso, Texas. <https://scholarworks.utep.edu/dissertations/AAI3609494/>
- Laney, C.M., Pennington, D.D., & Tweedie, C.E. (2015). Filling the gaps: sensor network use and data sharing practices in ecological research. *Frontiers in Ecology and the Environment*. Vol. 13, Issue 7, Pgs. 363-368.
- Langer, M., von Deimling, T.S., Westermann, S., Rolph, R., Rutte, R., Antonova, S., Rachold, V., Schultz, V., Oehme, A., & Grosse, G., (2023). Thawing permafrost poses environmental threat to thousands of sites with legacy industrial contamination. *Nature Communications*, 14, 1721
- Latif, A., Limani, F., & Tochtermann, K. (2019). A Generic Research Data Infrastructure for Long Tail Research Data Management. *Data Science Journal*, 18(1), p.17.
DOI: <http://doi.org/10.5334/dsj-2019-017>
- LaZerte, S.E., Reudink, M.W., Otter, K.A., Kusack, J., Bailey, J.M., Woolverton, A., Paetkau, M., de Jong, A., & Hill, D.J. (2017). Feedr and animalnexus.ca: A paired R package and user-friendly Web application for transforming and visualizing animal movement data from static stations. *Ecology and Evolution*, 7(19):7884–7896, 2017. ISSN 2045-7758. doi: 10.1002/ece3.3240.
- Ledger, M., Milner, A. (2015). Extreme events in running waters. *Freshwater Biol.* 60, 2455 – 2460. (doi:10.1111/fwb.12673).
- Leipzig, J., Nust, D., Hoyt, C.T., Ram, K., & Greenberg, J. (2021). The role of metadata in reproducible computational research. *ScienceDirect, Patterns*. Vol. 10. Issue, 9.
- Liu, J., Mooney, H., Hull, V., Davis, S.J., Gaskell, J., Hertel, T., Lubchenco, J., Seto, K. C., Gleick, P., Kremen, C., and Li, S. (2015). Systems integration for global sustainability. *Science* 347, 1258832. doi:10.1126/SCIENCE. 1258832
- Luna, R.N., (2016). “Spatiotemporal variability of plant phenology in drylands: A case study from the northern Chihuahuan Desert”. *ETD Collection for University of Texas, El Paso*. AAI10250930.
<https://scholarworks.utep.edu/dissertations/AAI10250930>
- Lynch, C. (2008). Big data: How do your data grow? *Nature* 455: 28–29

- Ma, X., Huete, A., Moran, S., Ponce-Campos, G., and Eamus, D. (2015). Abrupt shifts in phenology and vegetation productivity under climate extremes, *J. Geophys. Res.-Biogeo.*, 120, 2036–2052, <https://doi.org/10.1002/2015JG003144>
- Mahecha, D.M., Gans, F., Sippel, S., Donges, J.F., Kaminski, T., Metzger, S., Migliavacca, M., Papale, D., Rammig, A., and Zscheischler, J. (2017). Detecting impacts of extreme events with ecological in situ monitoring networks. *Biogeosciences*, 14, 4255-4277.
- Margolis, R., Derr, L., Dunn, M., Huerta, M., Larkin, J., Sheehan, J., Guyer, M., & Green, E.D. (2014). The National Institute of Health big data to knowledge (BD2K) initiative: Capitalizing on Biomedical big data. *Journal of American Medical Association*, 21. Pp. 957-958
- May, J.I., Healey, N. C., Ahrends, H. E., Hollister, R. D., Tweedie, C. E., Welker, J. M., Gould, W. A., and Oberbauer, S. E. (2017). Short-Term Impacts of Air Temperature on Greening and Senescence in Alaskan Arctic Plant Tundra Habitats. *Remote Sens.* 2017, 9, 1338; doi:10.3390/rs9121338
- McCord, S.E., Webb, N.P., Van Zee, J.W., Burnett, S.H., Christensen, E.M., Courtright, E.M., Laney, C.M., Lunch, C., Maxwell, C., Karl, J.W., Slaughter, A., Stauffer, N.G., & Tweedie, C. (2021). Provoking a Cultural Shift in Data Quality. *BioScience*, Vol.71, Issue 6, pg. 647 – 657
- Michener, W.K., & Jones, M.B. (2012). Ecoinformatics: supporting ecology as a data-intensive science. *Review Special Issue: Ecology and Evolutionary Informatics*. Vol. 27, Issue 2, P85 – 93
- Mineter, M. J., Jarvis, C. H., and Dowers, S. (2003). From stand-alone programs towards grid aware services and components: a case study in agricultural modelling with interpolated climate data. *Environmental Modelling & Software* 18:379-391
- Mitchel. T.M. (1997). *Machine Learning*. McGraw-Hill, Inc., New York
- Moler & MathWorks (2012). *MATLAB 8.0 and Statistics Toolbox 8.1*. The MathWorks, Inc.,
- Myer-Smith, I.H., Kerby, J.T., Phoenix, G.K., Bjerke, J.W., Epstein, H.E., Assmann, J.J., John, C., Andreu-Hayles, L., Angers-Blodin, S., Beck, P.S.A., Berner, L.T., Bhatt, U.S., Bjorkman, A.D., Blok, D., Bryn, A., Christiansen, C.T., Cornelissen, C.T., Cunliffe, A.M., Elmendorf, S.C., Forbes, B.C., Goetz, S.J., Hollister, R.D., de Jong, R., Loranty, M.M., Macias -Fauria, M., Maseyk, K., Noramand, S., Olofsson, J., Parker, T.C., Parmentier, F.W., Post, E., Shaepman-Strub, G., Stordal, F., Sullivan, P.F., Thomas, H.J.D., Tommervik, H., Treharne, R., Tweedie, C.E., Walker, D.A., Wilking, M., & Wipf, Sonja. (2020). Complexity Revealed in the Greening of the Arctic. *Nature Climate Change*, **10**, 106 -117

- Nicolai-Shaw, N., Zscheischler, J., Hirschi, M., Gudmundsson, L., and Seneviratne, S. I. (2017). A drought event composite analysis using satellite remote-sensing based soil moisture, *Remote Sens. Environ.*, <https://doi.org/10.1016/j.rse.2017.06.014>
- Nikou, G. (2018). Machine Learning Methods for Detecting Rare Events in Temporal Data. A dissertation submitted to the technical university of Munich.
- Niu, S., Luo, Y., Li, D., Cao, C., Xia, J., Li, J., and Smith, M. B. (2014). Plant growth and mortality under climatic extremes: An overview, *Environ. Exp. Bot.*, 98, 13–19.
- Oberbauer, S. F., Tweedie, C. E., Welker, J. M., Fahnestock, J. T., Henry, H. R. G., Webber, P.J., Hollister R. D., Walker, M. D., Kuchy, A., Elmore, E. and Starr, G. (2007). Tundra CO₂ Fluxes in response to experimental warming across latitudinal and moisture gradients. *Ecological Monographs*, 77(2), 2007, pp. 221-238
- Olden, J. D., Lawler, J. J., and Poff, L. N., (2008). Machine Learning Methods without tears: A Primer for Ecologists. *The Quarterly Review of Biology*, June 2008, Vol. 83, No. 2 The University of Chicago Press.
- Padalia, H., Pandey, K., & Arumugam, R.A. (2017). Development of a web-enabled spectral data archival, visualization and analysis of architecture for tropical phytodiversity inventory. *Tropical Ecology* 58(2): 307-314.
- Payne, R., Murray, D., Harding, S., Baird, D., & Soutar, D., GenStat. VSN International.
- Perakis, F., Lampathaki, F., Nikas, K., Georgiou, Y., Marko, O., & Maselyne J. (2020). Cybele – Fostering precision agriculture & livestock farming through secure access to large-scale HPC enabled virtual industrial experimentation environments fostering scalable big data analytics. *Comput. Networks*, 168, pp.1-10
- Peters, D. P. C., K. M. Havstad, J. Cushing, C. Tweedie, O. Fuentes, and N. Villanueva-Rosales. (2014). Harnessing the power of big data: infusing the scientific method with machine learning to transform ecology. *Ecosphere* 5(6):67. <http://dx.doi.org/10.1890/ES13-00359>
- Petrie, M.D., Collins, S.I., Gutzler, D.S., and Moore, D.M. (2014). Regional trends and local variability in monsoon precipitation in the northern Chihuahuan Desert, USA. *Journal of Arid Environments* 103 (2014) 63-70
- Pichler, M., & Hartig, F. (2023). Machine learning and deep learning – A review for ecologist. *Methods in Ecology and Evolution*. Vol. 14, issue 4, Pg. 994-1016

- Porter, J.H., Nagy, E., Kratz, T.K., Hanson, P., Collins, S.L. and Arzberger, P. (2009). New eyes on the world: Advanced sensors for ecology. *BioScience* 59: 385–397.
- Press, G. (2016). Cleaning Big Data: most time-consuming, least enjoyable data science task, survey says. *Forbes* 23, 3. <https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/>.
- Raban, D.R., Gordon, A., 2020. The evolution of data science and big data research: a bibliometric analysis. *Scientometrics* 122, 1563–1581. <https://doi.org/10.1007/s11192-020-03371-2>.
- Ramalho, L.F., & Segundo, W.R.C (2020). R-Shiny as an Interface for Data Visualization and Data Analysis on the Brazilian Digital Library of Theses and Dissertations (BDTD). *MDI*. 8(2), 24; <https://doi.org/10.3390/publications8020024>
- Ramirez, G., (2011). Assessing data quality in a sensor network for environmental monitoring. Masters Thesis, The University of Texas at El Paso.
- Recknagel, F. (2011). *Ecological Informatics: A discipline in the making*. Elsevier: *Ecological Informatics*, Vol.6, Issue. Pg 1-3. DOI: <https://doi.org/10.1016/j.ecoinf.2010.12.002>
- Reynolds, J. F. and Tenhunen, J. D. (1996). Ecosystem response, resistance, resilience, and recovery in Arctic landscapes: Introduction. In J. F. Reynolds, & J. Tenhunen (Eds.), *Landscape function and disturbance in Arctic Tundra*. *Ecological Studies*, vol. 120 (pp. 3 – 18). Heidelberg: Springer.
- Safriel, U. and Adeel, Z. Dryland systems. In: Hassan R, Scholes R, Ash N, editors. *Ecosystems and Human well-being: Current State and Trends*. Washington DC: Island Press; 2005. pp. 623–62. Volume. [[Google Scholar](#)]
- Sarkar, A. and Kumar, R. (2012) Artificial Neural Networks for Event Based Rainfall-Runoff Modeling. *J. Water Resour. Prot.* 4, 891–897.
- Schnase, J. L., Duffy, D. Q., Tamkin, G. S., Nadeau, D., Thompson, J. H., Grieg, C. M., McInerney, M. A. and Webster, W. P. (2017). MERRA analytic services: Meeting the big data challenges of climate science through cloud-enabled climate analytics-as-a-service. *Computers, Environment, Urban Systems* 61: 198–211.
- Scowen, M., Athanasiadis, I.N., Bullock, J.M., Eigenbrod, F., & Willock, S. (2021). The current and future uses of machine learning in ecosystems services research. *Science of the Total Environment*. Vol.799, 10 149263

Seneviratne, S.I., Luthi, D., Litschi, M., and Schar, C. (2006). Land-Atmosphere Coupling and Climate Change in Europe. *Nature*, 443, 205–209. 10.1038/nature05095

Seneviratne, S.I., X. Zhang, M. Adnan, W. Badi, C. Dereczynski, A. Di Luca, S. Ghosh, I. Iskandar, J. Kossin, S. Lewis, F. Otto, I. Pinto, M. Satoh, S.M. Vicente-Serrano, M. Wehner, and B. Zhou. (2021). Weather and Climate Extreme Events in a Changing Climate. In *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*

Shaver, G.R., Street, L.E., Rastetter, E.B., Van Wijk, M.T., Williams, M. (2007). Functional convergence in regulation of net CO₂ in heterogeneous tundra landscapes in Alaska and Sweden. *J. Ecol.* 2007, 95, 802–817

Sims, D. A., and Gamon, J.A., (2003). Estimation of Vegetation Water Content and Photosynthetic Tissue Area from Spectral Reflectance: A Comparison of Indices Based on Liquid Water and Chlorophyll Absorption Features. *Remote Sensing of Environment* 84: 526–37.

Stow, D. A., Hope, A., McGuire, D., Verbyla, D., Gamon, J., Huemmrich, F., Houston, S., Racine, C., Sturm, M., Tape, K., Hinzman, L., Yoshikawa, K., Tweedie, C. E., Noyle, B., Silapaswan, C., Douglas, D., Griffith, B., Jia, G., Epstein, H., Walker, D., Daeschner, S., Peterson, A., Zhou, L. and Myneni, R. (2003). Remote sensing of vegetation and land-cover change in Arctic Tundra Ecosystems. *Remote Sensing of Environment* 89, 281-308

Tsang, S.W. and Jim, C.Y. (2016) Applying artificial intelligence modeling to optimize green roof irrigation. *Energy Build.* 127, 360–369.

Uddin, S., Khan, A., Hossain, M.E. & Moni, M.A. (2019). Comparing different supervised machine learning algorithms for disease predictions. Springer: *BMC Medical Information and Decision Making*; 19 article number 281. <https://link.springer.com/article/10.1186/s12911-019-1004-8>?

Wainwright, J. (2006). *Climate and Climatological Variations in the Jornada Basin. The Jornada Basin long-term ecological research site.* Oxford University Press, USA.

Walker, D.A., Epstein, H.E., Jia, G., Balsler, A., Copass, C., Edwards, E.J., Gould, W.A., Hollingsworth, J., Knudson, J.A., Maier, H.A. (2003). Phytomass, LAI, and NDVI in northern Alaska: Relationships to summer warmth, soil pH, plant functional types, and extrapolation to the circumpolar Arctic. *J. Geophys. Res.* **2003**, 108.

- Wang, Y. (2017). A new concept using LSTM Neural Networks for dynamic system identification. In Proceedings of the 2017 American Control Conference, Seattle, WA, USA, 24–26 May 2017; pp. 5324–5329.
- Wang, Y., Kirubakaran, V., Biao, H. (2017). A Long-Short Term Memory Recurrent Neural Network Based Reinforcement Learning Controller for Office Heating Ventilation and Air Conditioning Systems. *Processes*, 5, 1–18.
- Wanyanhan, J., Cheng, H., Yang, L., & Pan, X. (2022). moreThanANOVA: A user-friendly Shiny/R application for exploring and comparing data with interactive visualization. *PLOS ONE*.
- Weigel, Tobias, et al., (2020). Making data and workflows findable for machines. *Data Intell.* 2 (1–2), 40–46. https://doi.org/10.1162/dint_a_00026.
- Weir, P., & Dahlhaus, P. (2023). In search of pragmatic soil moisture mapping at the field scale: A review. *Smart Agricultural Technology*, Vol. 6. 100330
- Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J.J., Appleton, G., Axton, A., Baak, M., Blomberg, N., Boiten, J.W., da Silva Santos, L.B., & Bourne, P.E., Bouwman, J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C.T., Finkers, R., Gonzalez – Beltran, A., Gray, A.G.J., Growth, P., Goble, C., Grethe, J.S., Heringa, J., Hoen, P.A.C., Hooft, R., Kuhn, T., Kok, R., Lusher, S.J., Martone, M.E., Mons, A., Packer, A.L., Person, B., Rocca-Serra, P., Roos, M., Shaik, R., Sasone, S.A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M.A., Thomson, M., Van der Lei, J., Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J., & Barend, M. (2016). The FAIR guiding principles for scientific data management and stewardship. *Sci. Data.*, 3. P. 160018
- Willcock, S., Martinez-Lopez, J., Hooftman, A. D. P., Bagstad, K. J., Balbi, S., Marzo, A., Prato, C., Sciandrello, S., Signorello, G., Voigt, B., Villa, F., Bullock, J. M., and Athanasiadis J. N. (2018). Machine learning for ecosystem services. *Ecosystem Services* 33 (2018) 165–174
- Woodward, G., Bonada, N., Brown, L. E., Death, R.G., Durance, I., Gray, C., Hladyz, S., Ledger, M.E., Milner, A.M., Ormerod, S.J., Thompson, R.M., and Pawar, S. (2016). The

effects of climatic fluctuations and extreme events on running water ecosystems. *Phil. Trans. R. Soc. B* 371: 20150274.

Wulder, M.A., Roy, P.A., Radeloff, V.C., Loveland, T.R., Anderson, M.A., Johnson, D.M., Healey, S., Zhu, Z., Scambos, T.A., Pahlevan, M., Hansen, M., Gorelick, N., Crawford, C.J., Masek, J.G., Hermosilla, T., White, C.J., Belward, A.S., Schaaf, C., Woodcock, E.C., Huntington, J.L., & Cook, B.D. (2022). Fifty years of Landsat Science and Impacts. *Remote Sensing of Environment*. Vol. 280, 113195

Xue, J., & Su, B. (2017). Review of Significant Remote Sensing Vegetation Indices: A Review of Developments and Applications. *Journal of Sensors* Volume 2017, Article ID 1353691, 17 pages <https://doi.org/10.1155/2017/1353691>

Zhang, J., Zhu, Y., Zhang, X., Ye, M., Yang, J. (2018). Developing a Long Short-Term Memory (LSTM) based model for predicting water table depth in agricultural areas. *J. Hydrol.* 561, 918–929.

Zhang, W.; Miller, P.A.; Smith, B.; Wania, R.; Koenigk, T.; Döscher, R. (2013). Tundra shrubification and tree-line advance amplify arctic climate warming: Results from an individual-based dynamic vegetation model. *Environ. Res. Lett.* 8, 1–10

Zhang, C., Hu, C., Xie, S., & Cao, S. (2021). Research on the application of Decision Tree and Random Forest Algorithm in the main transformer fault evaluation. *Journal of Physics: Conference series*. 1732, 012086

Vita

Ifeanyi H. Nwigboji earned his Bachelor of Science degree in Industrial Physics and Electronics from Ebonyi State University Abakaliki, Nigeria in 2008, and worked as a high school Physics teacher at Community Secondary School Echialike in Ebonyi State from 2009 to 2012. In 2013, he proceeded for his Masters degree and obtained a Master of Science degree in Material Physics from Southern University and A & M College, Baton Rouge, Louisiana in 2015. His Masters degree at Southern was funded by the prestigious Ebonyi State government Scholarship board. Ifeanyi received another Master of Science degree in Computational Science from the University of Texas at El Paso (UTEP) in 2017.

Ifeanyi joined the doctoral program in Environmental Science and Engineering at UTEP in the fall of 2017. He has published five papers in a peer reviewed journals and three conference proceedings from his previous work. He has also presented his research in several national and international conferences, including American Physical society (APS), and the American Geophysical Union (AGU) among others. Ifeanyi's dissertation was supervised by Dr. Craig Tweedie. Ifeanyi currently works as IT Manager at the Bank of the West, now Bank of Montreal and will continue to work in that capacity following his graduation.

This dissertation was typed by Ifeanyi H, Nwigboji.