Enhancing Open Source CLEWs Models with High-Resolution Land and Water Data

by

Yalda Saedi

M.Sc, University of Twente, 2016

B.Eng, Shahid Rajaee Teacher Training University, 2011

Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of
Master of Applied Science

in the
School of Sustainable Energy Engineering
Faculty of Applied Sciences

SIMON FRASER UNIVERSITY
Fall 2023

Declaration of Committee

Name:	**Yalda Saedi**
Degree:	**Master of Applied Science**
Title:	**Enhancing Open Source CLEWs Models with High-Resolution Land and Water Data**
Committee:	**Chair: Vivian Neal** Lecturer, Sustainable Energy Engineering **Taco Niet** Supervisor Assistant Professor, Sustainable Energy Engineering **Gordon McTaggart-Cowan** Committee Member Associate Professor, Sustainable Energy Engineering **Molly McVey** Committee Member Lecturer, Sustainable Energy Engineering **Amir Shabani** Examiner Lecturer, Sustainable Energy Engineering

Abstract

Climate change effects and increasing resource demands make it difficult for decision-makers to implement sustainable strategies to ensure access to water, energy, and food. The Climate, Land, Energy and Water systems (CLEWs) framework is widely used to analyze highly interconnected systems. CLEWs facilitates informed decision-making and supports sustainable planning by representing the interlinkages between these systems and their contribution to climate change. Literature highlights that there is a lack of functional tools to process detailed land and water data for developing the CLEWs model without increasing computational complexity. This thesis presents GeoCLEWs, an open source Python-based tool for reproducible processing of high-resolution land and water data to enhance regional and national CLEWs modelling. GeoCLEWs is openly accessible on GitHub and provides automated data collection, preparation, analysis, and statistics generation, which facilitate efficiently the CLEWs model-building process.

Keywords: open source modelling; water-energy-food nexus, integrated assessment model; agro-ecological assessment; CLEWs; sustainable development

Dedication

To my beloved husband, Farhad, your unwavering belief in me, endless love, and sacrifices have been my greatest source of strength throughout this academic journey. To my family, whose steadfast support and encouragement over the years have shaped my abilities and instilled in me the confidence to believe in myself and cultivate my skills. This thesis is dedicated to you all, with heartfelt gratitude for being the foundation of my success.

Lastly, to all women who strive for their advancement, breaking barriers and shaping a future of empowerment and equality.

Acknowledgements

I would like to express my deep appreciation to my supervisor, Dr. Taco Niet, for his invaluable guidance, unwavering support, and mentorship throughout this thesis. It has been an honour to meet him and be a part of his research team, learning valuable lessons that extend beyond the academic realm.

My warmest thanks go to the amazing members of the ΔE^{+} (Delta E Plus) Research Group. Their support, insights, and teamwork have truly made this research journey memorable, and it's been a joy working with such a talented and friendly group.

Additionally, I wish to express my gratitude for the valuable contributions of Dr. Gordon McTaggart-Cowan and Dr. Molly McVey. Their insightful comments and support have enriched the quality of this research.

I would like to extend my sincere appreciation to Catalyste+ and Mitacs for their financial support which played an important role in advancing my research.

Table of Contents

List of Tables

List of Figures

List of Acronyms

AEZ	Agro-Ecological Zone
CLEWs	Climate, Land, Energy, and Water systems
CLEWs UI	CLEWs User Interface
CWD	Crop Water Deficit
EVT	Crop Evapotranspiration
FAOSTAT	Food and Agriculture Organization of the United Nations Statistical Database
GADM	Global Administrative Area Database
GAEZ v4	Global Agro-Ecological Zones version4
GDF	GeoDataFrame
GHG	Greenhouse Gas
GIS	Geographic Information System
LCType	Land Cover Type
LERC	Limited Error Raster Compression
LUT	Land Utilization Type
OSeMOSYS	Open Source Energy Modelling System
RCPs	Representative Concentration Pathways
WEF	Water-Energy-Food
YLD	Agro-Climatic Potential Yield

Chapter 1.

Introduction

One of the most important global sustainability challenges is reaching a balance between supplying water, energy, and food (WEF) for an increasing number of people while dealing with climate change implications due to their interlinkages. Water, energy, and land are highly interconnected and changes in one system can cause cascading impacts on the others. Agricultural activities and energy production either from fossil-based or renewable sources are responsible for significant amounts of water withdrawals globally. Energy is required for water treatment, desalination, and pumping as well as land preparation, crop cultivation, and fertilizer production. Land resources are used for food and bioenergy production, power plant infrastructure, energy transmission, and watersheds. In addition, these interconnected systems are vulnerable to the consequences of climate change and alternation in each of them may contribute to climate change. These linkages are highly important to capture and analyze comprehensively to evaluate regional and national strategies to make informed sustainable decisions.

The nexus approach implements an integrated assessment of systems to capture interlinkages leading to improving analysis of sustainable planning and management pathways [1]. WEF nexus are interrelated and unattainable to comprehend in isolation. The WEF nexus approach aims to provide insights to decision-makers for resource planning as well as prevent issues led by individual resource analysis by clarifying the interconnections of WEF resources at the local, and global levels [2]. The integrated nexus approach promotes a more holistic understanding of their interrelationships [3] enabling optimizing positive interactions and managing trade-offs while the assessment of WEF in isolation can lead to missing cross-system impacts. Resilience and sustainability in the WEF system are influenced by the interconnection among various subsystems. Variations to one subsystem have the potential to quickly and extensively propagate through systems, setting off a series of feedback [4], a detailed understanding of systems' cross-sectorial dependencies is required to enhance nexus assessment.

The 'Climate, Land, Energy, and Water systems' (CLEWs) framework [5] stands out as one of the most comprehensive tools within the nexus approach [6] representing interlinkages within the resource systems, nexus analysis, and climate change to support sustainable policy planning. CLEWs models simultaneously assess the interlinkages between water, energy and land systems and analyze their contribution to climate change enabling decision-makers to assess the impact of various development and climate mitigation strategies on the interdependent nexus through the mathematical optimization [7].

The amount of available land and water resources is finite and requires sustainable management across multiple uses to address demand; CLEWs models provide information to implement sustainable planning which involves optimizing the allocation of land to different types of uses, minimizing the environmental impacts, and addressing the water competition. To create a representation of land and water systems, the CLEWs model requires spatial information on land cover, soil suitability, water availability, and crop agro-climatic characteristics. It utilizes precipitation, crop water deficit, and evapotranspiration to create crop water balances during the growth cycle. Insufficient water can lower crop yield and monitoring the amount of water evaporated from plants and soil surfaces is highly important to effectively irrigate land crops and manage water use. Effective agriculture management and land cover preservation involve cultivating proper crops based on crop agro-climatic potential yield, water deficit, evapotranspiration, climate condition, soil characteristics and precipitation.

Despite extensive research conducted on integrated nexus assessment within the CLEWs framework, there still exist noticeable gaps in the literature pertaining to detailed assessment of land and water systems. Gaps in the literature include incomplete evaluation due to coarse resolution data and the challenges of computation complexity pertaining to the finer spatial resolution [8]. Also highlighted is a lack of functional CLEWs-compatible tools for geoprocessing high-resolution land and water data [9] necessitating modellers to possess a high level of domain-specific technical expertise, knowledge of programming, and utilization of Geographic Information System (GIS) processing tools. Currently, many CLEWs models employ low-resolution spatial data and do not include reproducible processing methods. Improving the spatial resolution of CLEWS models, through a reproducible and easy-to-use data workflow, is critical to achieving detailed cross-regional analysis [10].

This thesis presents GeoCLEWs v1.0.0, a new Python-based tool for reproducible processing of detailed land and water data to facilitate the development of CLEWs model using high-resolution spatial data. GeoCLEWs addresses the identified literature gaps through improving the level of detail included in CLEWs models. It is a versatile open source script that offers a wide range of useful features for both developers and modellers. GeoCLEWs is designed to collect, analyze, and process high-resolution data from the most updated Global Agro-ecological Zones database (GAEZ v4) [11], in an automated and time-efficient manner. It provides detailed land and water data processing for any arbitrary geographic region. GeoCLEWs leverages open source tools and open datasets to encourage collaboration and accessibility. The tool is openly licensed under an MIT License and is available on GitHub [12], which includes a transparent script and supplementary documentation such as running instructions. GeoCLEWs automates the process of detailed geospatial data collection, analysis and processing enhancing efficiency and accessibility for advanced CLEWs modelling.

GeoCLEWs utilizes high-resolution spatial datasets from GAZE v4 including soil diversity, crop attainability, level of land management technologies, watering system, and future simulated datasets to generate detailed crop agro-ecological potential yield, crop water deficit, crop evapotranspiration, precipitation, and land cover statistics on national and regional scale, which enhance interlinkage identification within CLEWs framework. These agro-ecological characteristics significantly vary across different geographical regions and comprehensive spatial assessment generated by GeoCLEWs enables exploring synergies and trade-offs within administrative regions. However, the currently available methods utilizing open source datasets offer an approximation for an entire country using historical records. The workflow of the existing approach to develop land and water systems within CLEWs framework are illustrated in Figure 1, which involves significant manual calculations and the use of general ratios to obtain rough estimations. In addition, automated workflow of GeoCLEWs for data collection, preparation standardization, and processing has substantially reduced the manual efforts required in existing methods including manual data collection, mathematical calculations, and GIS processing. Figure 2 illustrates GeoCLEWs workflow generating automatically detailed statistics on regional and national scale.

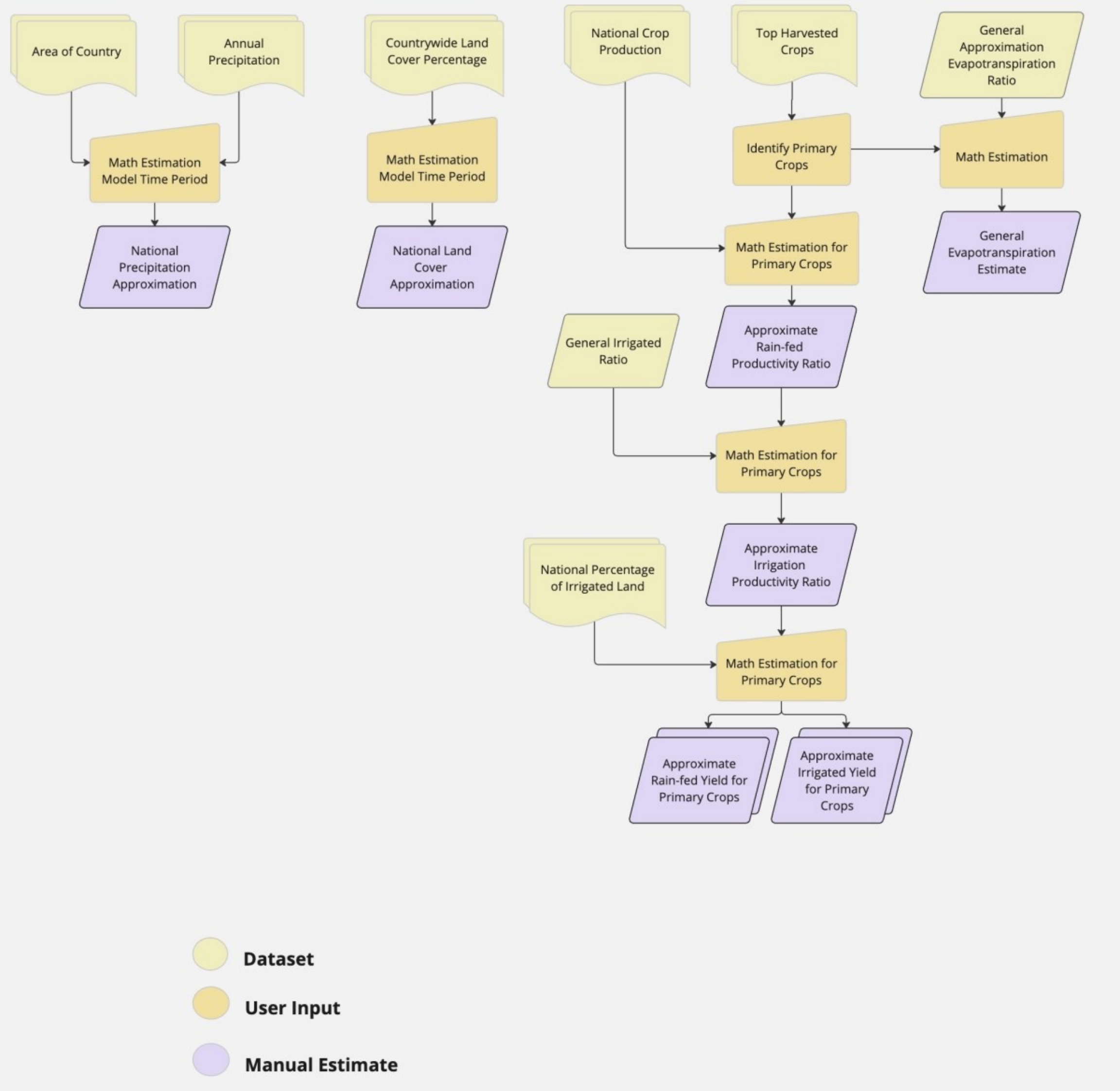

Area of Country
Annual Precipitation
Math Estimation Model Time Period
National Precipitation Approximation
Countrywide Land Cover Percentage
Math Estimation Model Time Period
National Land Cover Approximation
National Crop Production
Top Harvested Crops
Identify Primary Crops
General Approximation Evapotranspiration Ratio
Math Estimation
General Evapotranspiration Estimate
Math Estimation for Primary Crops
Approximate Rain-fed Productivity Ratio
General Irrigated Ratio
Math Estimation for Primary Crops
Approximate Irrigation Productivity Ratio
National Percentage of Irrigated Land
Math Estimation for Primary Crops
Approximate Rain-fed Yield for Primary Crops
Approximate Irrigated Yield for Primary Crops
Dataset
User Input
Manual Estimate

Figure 1: Genral workflow of existing methods to collect land and water data for CLEWs modelling utilizing estimations and coarse spatial open source datasets.

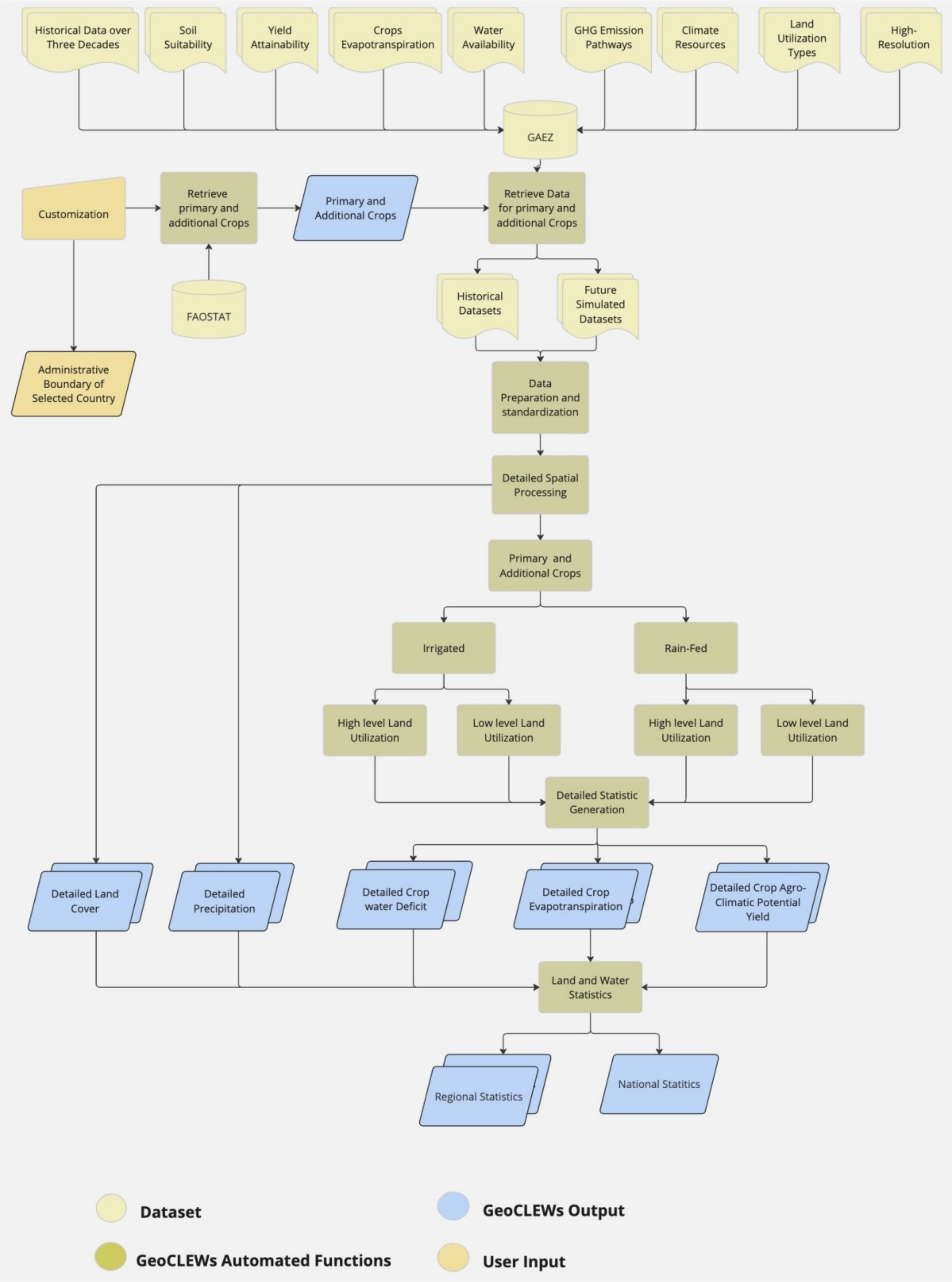

Historical Data over Three Decades
Soil Suitability
Yield Attainability
Crops Evapotranspiration
Water Availability
GHG Emission Pathways
Climate Resources
Land Utilization Types
High-Resolution
GAEZ
Customization
Retrieve primary and additional Crops
Primary and Additional Crops
Retrieve Data for primary and additional Crops
FAOSTAT
Historical Datasets
Future Simulated Datasets
Administrative Boundary of Selected Country
Data Preparation and standardization
Detailed Spatial Processing
Primary and Additional Crops
Irrigated
Rain-Fed
High level Land Utilization
Low level Land Utilization
High level Land Utilization
Low level Land Utilization
Detailed Statistic Generation
Detailed Land Cover
Detailed Precipitation
Detailed Crop water Deficit
Detailed Crop Evapotranspiration
Detailed Crop Agro-Climatic Potential Yield
Land and Water Statistics
Regional Statistics
National Statitics
Dataset
GeoCLEWs Output
GeoCLEWs Automated Functions
User Input

Figure 2: **GeoCLEWs workflow generating detailed land and water statistics automatically utilizing high resolution GAEZ v4 datasets.**

1.1. Outline

This thesis consists of five chapters including this Introduction, an Overview of GeoCLEWs, Methodology, a Case Study, and Conclusions. The following outlines the details covered in each chapter:

1. **Introduction**: This chapter consists of a research overview and the significance of conducting this study including Background, Knowledge Gaps, and Research Objectives. The Background includes a detailed presentation of relevant information and key concepts. The literature review and gaps are discussed within The Knowledge Gaps, followed by the Research Objectives established to address identified gaps.

2. **Overview of GeoCLEWs**: This chapter provides an overview of GeoCLEWs highlighting its advantages and additional features compared to the initial script upon which GeoCLEWs is built.

3. **Methodology**: The Methodology presents an overview of the data preprocessing for the script and an overview of the main methods and details of the script.

4. **Case Study**: This chapter presents a case study conducted in Kenya, including the process of setting up GeoCLEWs to generate detailed land and water statistics. It also covers outcomes from the CLEWs model using produced statistics.

5. **Conclusions**: This chapter outlines a summary and analysis of key achievements presented in this thesis and highlights significant advantages of GeoCLEWs in comparison to existing methods, associated limitations and future work.

1.2. Background

Climate Change: Impacts, Sources

The negative consequences of climate change are intensifying, leading to critical health, economic, and environmental issues all around the world [13]. Climate change describes the significant long-term shift in climate patterns including global temperature increases, wind alterations, changes in precipitation patterns, and any climate-related variables. Climate change is responsible for increase carbon dioxide concentration and temperature of the Earth as well as changing the atmospheric circulation [14]. It can disrupt the biosphere and alter the biodiversity and species distribution as a result of temperature rise [15]. Anthropogenic climate change has detrimental impacts on ecosystems [16] and dramatically affects millions of people manifesting through various challenges such as drought, food and water scarcity, extreme weather, floods, earthquakes, and vector-borne and waterborne diseases [17]–[19].

Climate change is primarily caused by greenhouse gas (GHG) emissions such as carbon dioxide (CO_2), nitrous oxide (N_2O), and methane (CH_4) trapping heat in the Earth's atmosphere [20]. CO_2 constitutes the majority of GHG emissions [21] and there has been a considerable increase in global CO_2 releases according to Emissions Gap Report 2019 [22], revealing an urgent need to identify main contributors and take effective actions. A significant proportion of the global population relies on fossil fuels for domestic, commercial, and agricultural purposes, which releases GHGs contributing to global warming and climate change [23]. There are two main sources of GHG emissions, natural systems and human activities [24]; natural sources include forest fires, volcanoes, earthquakes, wetlands, and oceans, and anthropogenic emissions mostly originate from energy production, forestry, industry, and land-use change. The natural system is capable of self-balancing, while human-caused emissions place additional pressure on the Earth’s system. Fossil fuels are the significant contributor to meeting the energy demand of the growing population, which results in considerable GHG emissions [18]. In addition, unsustainable farming practices, deforestation, and land degradation accelerate climate change, which impacts numerous people and disproportionately affects vulnerable populations [25].

Climate Change: Mitigation Strategies and Challenges

Mitigation and adaptation efforts are being made to avoid climate change effects as well as ensure access to green energy, clean water, and food by establishing national policies and international agreements. In addition to climate change, growing energy demands put extra pressure on resources resulting in widespread energy shortages, and scarcity of food and water. One of the key solutions to address climate change is changing the energy production strategy and the most promising alternatives to fossil fuels are renewable energy resources; however, long-term policy planning and sustainable transition are highly challenging [26], [27]. Well-designed and effective policies facilitate the adaptation of renewable sources of energy and associated advancements [27].

Well-informed sustainable planning necessitates identifying interlinkages between water, land, energy systems, and climate change. The electricity sector involves a broader use of land when utilizing renewable sources of energy in contrast to spatially compact fossil fuel sources [28]. Furthermore, it is highly important to study interactions between climate change, water systems, and agricultural activities to sustainably manage food and water resources. The agriculture sector is one of the substantial contributors to GHG emissions and is one of the biggest consumers of water and land resources as well [29]. Climate change exerts a significant influence on the cultivating systems and producibility of crops by affecting water availability, soil suitability, and temperature [30]. Crop suitability analysis is important to address the challenge of increasing land requirements, water demand and GHG emissions in the agriculture sector [29]. According to the simulation conducted with historical weather data and future climate scenarios, crop productivity rises as a result of climate change, while crop water deficits are negatively impacted [30]. The value chain of the provision of WEF resources is extremely interlinked and contributes to the climate change [10] presenting a significant challenge in achieving sustainable goals.

Policymakers impose regional and global climate actions to achieve sustainable goals while they are facing several challenges due to the complexity of climate, land, energy, and water nexus assessments [3], which highlights the necessity of conducting policy formulation under the cohesive analysis of the cross-sectoral nexus interdependencies [31]. Science-policy collaboration plays an important role in broadening the knowledge of the human-environment system and numerous interrelated concerns with regard to targets embodied in sustainable development goals in the contexts of environmental protection and resource scarcity [32]. Considering the long-term and extensive impacts of strategies, establishing a science-policy relationship is increasingly in demand [33] to assess the intervention of policies and priorities made by decision influencers on interlinked nexus and their involvement in climate change. Identifying and minimizing trade-offs between sustainable management and socio-economic plans is imperative to accommodate the needs of the present and future generations while mitigating the effects of climate change [8].

Nexus Approach: Integrated Assessment Models

The Nexus approach provides scientists and other stakeholders with opportunities for potential policy interventions and sustainable development pathways [1]. The activities that take place within a system are represented by sectors; water supply, electricity, and agriculture sectors are embedded within the water, energy, and land systems respectively [35]. The Water-Energy-Food nexus focuses on interlinkages between these essential sectors and the nexus approach empowers policymakers to minimize trade-offs and maximize synergies utilizing an integrated assessment [6]. The nexus approach appeared in the early 1980s while the complex relationships among three essential elements including water, energy, and food, known as the WEF nexus, were formally recognized in publications as early as 2008 [3].

It is important to incorporate multiple dimensions beyond the WEF sectors to reach nexus thinking and address the complexity of the multi-sectoral resources; however, there are few models that cover all WEF simultaneously and adopt interdisciplinary methods in analyzing the nexus [36]. To promote policy analysis and the evaluation of climate change mitigation strategies, a number of tools and frameworks have been developed. Noteworthy among these tools are MuSIASEM [37], LEAP [38], WEAP [39], and MESSAGE [40]. MuSIASEM, or Multi-Scale Integrated Analysis of Societal and Ecosystem Metabolism, provides an integrated assessment of the WEF nexus in relation to sustainability. Initially designed to examine energy metabolic patterns, MuSIASEM has undergone expansion to encompass considerations of food and water in relation to ecological and socio-economic variables across multiple scales [41]. The Long-range Energy Alternatives Planning system (LEAP), is an integrated, scenario-based modelling tool to track energy resource production and utilization in all economic sectors considering energy and non-energy sector GHG emissions; the upgraded version explores climate, health and crop benefits. WEAP, or Water Evaluation And Planning systems, simulates both the engineered and natural water resources for integrated water resources planning. The Model for Energy Supply Systems and their General Environmental Impact (MESSAGE) is an optimization framework that considers the environmental effects of energy supply strategies when developing scenarios and analyzing long-term energy policies. NExus Solution Tool (NEST) [42] is an open modelling platform for integrating WEF resource optimization, which utilizes MESSAGEix [43], developed based on MESSAGE, to model energy system.

Global Agro-Ecological Zones

State-of-the-art geo-processing technologies and accurate analysis of spatial data carried out over extended periods are increasingly producing high-resolution spatial data beneficial for CLEWs assessment. Agro-Ecological Zone (AEZ) [44] is an open source framework including detailed historical and projected spatial information, commonly used for developing CLEWs model [45]. It was developed by The Food and Agriculture Organization of the United Nations (FAO) in association with the International Institute for Applied Systems Analysis (IIASA), which relies on detailed land assessment principles collected over three decades. The AEZ methodology was initially established to evaluate the demand for food due to an increasing population while considering the limitations of natural resources, particularly in developing countries. The AEZ framework has been completed and refined over the years by analyzing various plant eco-physiological features, edaphic characteristics, and crop suitability to provide agro-climatically attainability.

Since 2000, FAO published an extended global-scale framework named Global Agro-Ecological Zone (GAEZ) utilizing a wide range of attributes and data including climatic parameters, crop attainability, topography, water sources, land-use dynamics, and population distribution. It has been publicly available in digital format and the latest version is GAEZ v4 [11], which contains the 2010 baseline, 1961-2010 historical climatic conditions, and 2011-2100 future agro-ecological simulated data. GAEZ provides information on current and projected ecological conditions, water availability, crop suitability, and land cover supporting local, national, and global sustainable planning. GAEZ framework offers datasets with a fine resolution of 30 arc-seconds and 5 arc-minutes improving detailed integrated assessment modelling, which provides policymakers with a comprehensive analysis of biophysical limitations and opportunities for securing WEF management as well as addressing the consequences of climate change.

GAEZ portal v4 supplies comprehensive spatial datasets using a wide range of data sources and techniques. Detailed agro-climatic information is delivered using the GIS resources database, historical and future climate attributes, soil and terrain resources, land cover protected areas, biodiversity areas, water resources, and Land Utilization Types (LUT). Information is subject to influence from the water source utilized (i.e., rain-fed or irrigated) as well as the assumed degree of inputs and management practices. For clarification, an explanation of the GAEZ specifications is provided below.

LUTs are a concept that empirically describes differences in crop varieties and production systems. Technical requirements for agricultural production in a specific socioeconomic environment correspond to an LUT. The type of primary yield, the sort of water supply, typical cultivation techniques, and the intended consumption of the produce contribute to unique LUT. In the GAEZ v4 framework, more than 1000 crop/LUT and management combinations are distinguished and individually evaluated for both rain-fed and irrigated environments [44]. Low, intermediate, and high input levels are the three generic levels of LUT that are established in GAEZ. Low-level refers to the traditional farming practices, lack of added plant nutrients, and labour-intensive manual techniques. The intermediate input level indicates that the system of agriculture is partially market-oriented under an improved management assumption. The high input level includes a fully mechanized advanced management system with the best use of nutritional and chemical additives at a full commercial production level. Adopting different levels of management results in multiple assessment outputs.

CLEWs Modelling with OSeMOSYS

CLEWs is an open source framework to provide a comprehensive integrated assessment of land, energy, and water systems [6], utilizing the nexus approach to represent synergies, and trade-offs within these systems [35]. In addition, it assists in analyzing interactions and quantifying their contribution to climate change. The CLEWs framework demonstrates effective performance in highlighting interlinkages providing opportunities to increase the synergies and minimize the trade-offs [7]. It has been utilized in numerous research efforts across the world with the modelling adjusted based on specific geographical constraints and national priorities [35], [46]–[48]. CLEWs represents the nexus interdependencies and impacts of various strategies on the natural resources, environment and climate change.

The CLEWs framework is built within the OSeMOSYS, Open Source Energy Modelling System, which has been widely used in several studies due to its flexibility, functionality, and availability to run and interpret modelling scenarios [49]–[51]. OSeMOSYS is a long-term optimization model considering a wide variety of system costs to find the most affordable solution according to the demands and constraints [52], [53]. OSeMOSYS is a bottom-up capacity expansion modelling framework using a linear optimization program to meet energy demand.

A wide range of data is required to create climate, land, water, and energy systems as well as define their interaction for instance, water demand for agricultural activities or energy requirement for water treatment and distribution. In developing CLEWs model, different methods and datasets have been utilized to assess resource availability and demand. Welsch et al. [46] developed CLEWs model for Mauritius using AEZ and WEAP to develop land-use and water systems respectively. Arianpoo et al.[54] estimated crop suitability using GAEZ v3 and collected power system information from various data sources including international and provincial datasets. Shivakumar et al. [55] presented a spatial clustering approach using GAEZ v3 datasets to generate land and water systems based on cross-regional similarities. Similarly, Kuling et al. [56] employed the same clustering approach to develop land and water systems combined with an energy system generated by OSeMOSYS Global [57]. OSeMOSYS Global is an open source energy model generator, simplifying the time-consuming process of data gathering, analyzing and validation using a peer-reviewed open dataset. It offers an automated Snakemake workflow streamlining the energy modelling process.

Compared to the manual entry process, modellers can develop and scale up CLEWs models considerably more quickly and consistently using clewsy [58], which is an effective and freely available Python package leveraging land, water, and electricity information along with an input configuration file to generate data file in CSV format. Clewsy not only simplifies creating CLEWs model in OSeMOSYS but also integrates seamlessly with OSeMOSYS Global. Its command line structure requires an input file with YAML (YAML Ain't Markup Language) format to build the CLEWs model structure starting with creating the energy system, followed by developing the land-use structure [59]. OSeMOSYS Tools for Energy python package (Otoole) [60] is essential in this process to convert the clewsy output into a format that is compatible with the latest and user-friendly version of CLEWs User Interface (UI). The adjusted output generated by Otoole can be imported into UI, allowing users to manually add constraints and design scenarios. After completion of the model design, it is also possible to generate results using the publicly accessible OSeMOSYS Cloud, which employs the OSeMOSYS model to perform the optimization process and identify the least expensive system configuration to address demands and constraints within scenarios.

1.3. The Knowledge Gap

Nexus modelling requires detailed spatial data to achieve in-depth cross-regional interdependencies among WEF systems [61]; this necessity arises from the key role of spatial diversity in CLEWs assessment. The low-resolution spatial data limits understanding of nexus components such as water availability, edaphic factors, land cover, and renewable energy accessibility differ remarkably by geographical location. Agricultural sectors rely on energy availability, water supply, and soil sustainability; subsequently, climate change can significantly affect the distribution of agricultural production [36]. Variable renewable energy technologies depend on weather-driven sources that differ drastically across geographical areas; the incorporation of these technologies into capacity expansion models demands the inclusion of a thorough spatial comprehension [62].

Although information with high spatial resolution will significantly foster the identification of interlinkages resulting in a more reliable and achievable optimum solution, there is an obvious absence of detailed spatial representation within the CLEWs assessment [55]. Numerous research endeavours following the CLEWs framework have employed land and water data featuring low resolution, laborious and manual GIS processing, aggregated measurements, estimated statistics, and open source accumulated data sources [7], [63]–[67]. Detailed data can facilitate the exploitation and management of the resources, and incomplete knowledge hinders identifying interconnections within the nexus [34]. Insufficient understanding pertaining to the spatial-temporal changes in resource availability and accessibility may lead to policies that do not optimize resource utilization; however, a limited number of nexus assessment techniques have integrated multi-sector demands to combine detailed land utilization and spatial crop zoning [36].

Despite the availability of high-resolution land and water data from GAEZ v4, CLEWs models are unable to utilize this valuable database due to a lack of effective, accessible, and CLEWs-compatible tools. Before this thesis and GeoCLEWs there was no available tool proven effective in utilizing GAEZ v4 for geoprocessing the land and water systems within the CLEWs framework. Deprecated tools such as CLEWs GIS Processing tool [9], have been rendered impractical due to incompatibility with the updated GAEZ database and the inaccessibility of former datasets. The previous version of GAEZ datasets being unavailable presents a substantial challenge to the existing tools implementing detailed land and water processing for CLEWs modelling and renders them impractical [56].

The high cost of computational complexity associated with detailed spatial processing presents a significant challenge that needs to be addressed. The high-resolution spatial data integration enhances the energy systems optimization model outcomes while coming at a heavy cost in terms of model dimensions and computing complexity. The finer geographic resolution is emphasized, however, it usually leads to long-running times [8]. Studies covering a large geographic area demand some sort of aggregation in order to deliver In-depth insight at a less detailed level [35]; this fact underscores the necessity of spatial computational support to fully utilize WEF datasets collected from various sources and geographical locations [2] while analyzing various sustainable development scenarios using CLEWs framework. Effective spatial clustering methodologies were previously employed within the CLEWs framework to reduce the computational cost of GAEZ data assessment; however, their practicality has waned due to unavailable GAEZ input data sources and the substantial manual effort [55].

Public availability and transparency are important factors in the climate, land, energy and water nexus modelling [6]. The complexity of energy systems, the uncertainty of the transition to renewable energy, and the rapid development of low-carbon technologies reveal the high demand for public accessibility of nexus modelling tools. In addition, there has been an increasing request for transparent model-informed policy formulation towards the implementation of sustainable development [52].

1.4. Research Objectives

Our research objectives address these gaps by developing an effective and open source tool to incorporate detailed land and water data statistics.

- Develop a reproducible processing tool utilizing high-resolution GAEZ v4 datasets to generate detailed land and water statistics. This tool addresses the lack of functional and replicable methods compatible with the most updated GAEZ datasets to develop the CLEWs model.

- Generate outputs that adhere to compatibility standards with clewsy, which is essential to developing the CLEWs model. clewsy utilizes land and water statistics as input to streamline CLEWs model development. This feature accelerates the process of developing models.

- Implement regional aggregation to reduce number of total regions that addresses the computational complexity within CLEWs model. This strategic approach is beneficial because a larger number of regions in the CLEWs model results in higher computational demands, which could negatively affect the model's efficiency.

- Implement innovative strategies and approaches to minimize manual effort in data collection, preprocessing, analysis, and generating detailed land and water statistics. It enhances efficiency and productivity by automating key functions and taking effective steps toward creating a fully automated CLEWs modelling framework by reducing manual effort. Simplifying and automating the process of land and water data collection, preparation, process, and result generation provides non-technical users with the opportunity to generate detailed spatial statistics without employing complicated and time-consuming geospatial processing.

A case study was developed to evaluate the functionality of GeoCLEWs while exploring real-world scenarios. The case study serves as an effective method to validate GeoCLEWs and improve its functionality. This thesis employs open source tools and open datasets to promote sustainable collaboration; the developed tool is released on the freely accessible GitHub platform including transparent, self-described, and reproducible scripts and essential supplementary documents to foster user contribution. Appendix A details information on GitHub repository of GeoCLEWs.

Chapter 2.

Overview of GeoCLEWs

This research addresses the identified literature gaps and presents a new open source tool named GeoCLEWs to deeply assess high-resolution land and water data without increasing computational complexity. GeoCLEWs utilizes the agro-climatic data from the most updated GAEZ database and generates required land and water statistics for developing the CLEWs model. Furthermore, it involves a number of additional features that streamline and automate all processing steps including FAOSTAT (Food and Agriculture Organization of the United Nations) [69] and GAEZ data collection, preparation, regional aggregation, and processing, which offers modellers a chance to employ this tool without prior knowledge and experience in complicated spatial processing improving decision making.

GeoCLEWs is built upon the foundation concept of the CLEWs GIS Processing tool [9], which was rendered non-functional due to some limitations including the inaccessibility of GAEZ v3 as input data. CLEWs GIS Processing, referred to as the *initial script* throughout this thesis, is designed based on input data from GAEZ v3 and fundamentally depends on a specific data type from this database and cannot function without it. Global AEZ resources were published in 2000 (v2), 2012 (v3) [70] and 2021 (v4) [11] and outdated databases are no longer available. The recently released GAEZ version 4 is substantially different from the prior one and does not provide ASCII data type, which is an essential input for CLEWs GIS Processing tool; subsequently, in the absence of required input data this tool has become inoperable. Currently, there is no available functional tool for processing land and water data sourced from GAEZ v4 and generating statistics to develop CLEWs model.

Compared to the initial script, GeoCLEWs has undergone substantial modifications and includes numerous newly added functions leading to improved performance, accessibility, and flexibility in processing high-resolution land and water data. Significant adjustments and improvements have been made to the newly developed tool compared to CLEWs GIS Processing, which is released under an MIT License permitting modification and distribution without restriction. In addition, the functionality of the initial script has been optimized and revised to achieve more efficiency and less complexity. The initial script, while serving as a valuable starting point, includes considerable manual input. In contrast, GeoCLEWs requires only minor user customization to accommodate various projects' needs and improve its functionality and automatically performs all required steps to generate detailed land and water statistics. In the following, the improvements and distinctions between the initial script and this thesis are described in detail; Table 1 displays a summary of additional functionalities.

Table 1: Functionality comparison between the initial script and this thesis.

Analytical Process	CLEWs GIS Processing (Initial Script)	GeoCLEWs (This Thesis)
Initialization and Configuration	1. Importing necessary modules 2. Manually input crop names 3. Manually choose admin-level 4. Manually choose projection system 5. Manually choose topological classification 6. Directory initialization	1. Importing necessary modules 2. Manually input country name 3. Manually choose admin level 4. Manually define projection system 5. Manually choose topological classification 6. Manually select RCP 7. User-customized aggregation 8. User-customized region extraction 9. Directory initialization
Data Collection and Preparation		10. Finding country code for producing results in a clewsy-compatible format 11. Automated identification of the Top 10 crops from FAOSTAT 12. Categorizing crops information 13. Display results of crop categorization 14. CLEWs-compliant crop naming convention 15. Automatically import GAEZ data according to user input 16. GAEZ and FAO data correction 17. CLEWs-compliant GAEZ data naming convention 18. Filtering GAEZ data according to user configuration 19. Download GAEZ raster files

Analytical Process	CLEWs GIS Processing (Initial Script)	GeoCLEWs (This Thesis)
Generating Land Cells	7. Provide spatial index using ASCII grid 8. Create polygons 9. Fixing missing values 10. Coastal area correction 11. Total area re-estimation & calibration	20. Generating georeferenced point grid from any arbitrary geographical region 21. Create polygons with adjustable size to match the resolution of data and computational time 22. Generate georeferenced point grid 23. Total area re-estimation & calibration
Geospatial Attributes Extraction to Regions	12. Define continues and categorical raster files functions 13. Extract spatial values	24. Crop GAEZ global raster data 25. Define continues and categorical raster files functions 26. Extract spatial features
Calculating Region Summaries	14. National summary stats 15. Calculating region summaries 16. Generate tabular statistics 17. Produce interactive graphs	27. National summary stats 28. CLEWs-compatible unit conversions 29. Averaging additional crops for national summary 30. Calculating region summaries 31. Regional aggregation 32. Perform region extraction 33. Region summary calculations of the aggregated regions 34. Averaging additional crops for region summaries 35. Export tabular clewsy-compatible results 36. Generate interactive graphs

2.1. Initialization and User Configuration

The first part of GeoCLEWs contains the initial preparations including importing necessary modules, user configuration, and directory initialization. GeoCLEWs is created using the Jupyter Notebook environment and requires importing different libraries, packages, and dependencies to run successfully. In configuration part, users can customize the project setup to suit different projects and users' needs. Directory initialization has a well-designed and transparent structure that makes the script easy to reproduce using input data and generates output in the corresponding directories.

GeoCLEWs minimizes manual intervention; users input the name of the country of interest and the script automatically identifies, extracts, and process all required crop codes and country code in the following parts. Furthermore, GeoCLEWs provides an opportunity to retrieve GAEZ data based on the user's preferred Representative Concentration Pathways (RCPs), which represent different levels of greenhouse gas emissions and associated climate changes. A recent extension is the possibility of region aggregation and extraction which is not available in the prior code. GeoCLEWs generates results on the national scale as well as subnational administrative divisions such as provinces and counties. The number of regions escalates the computational problem within CLEWs modelling, and aggregating subnational regions during land and water statistics can speed up CLEWs computational processing. GeoCLEWs offers this option to users to group the subnational divisions based on their preferred number of groups. There is another new function to extract a specific subnational region from aggregation. Users have the capability to customize the aggregation method and remove a particular administrative area according to their preferences. This feature enables an in-depth examination of a particular region while grouping the remaining areas.

2.2. Data Collection and Preparation

This part is an innovation to the prior script including automatic data collection and preparation. It consists of identifying and extracting required data from FAOSTAT and GAEZ datasets and implementing preprocess and modification; nonetheless, the initial code lacked these capabilities. Considering user-defined configuration, GeoCLEWs collects primary types of crops from FAOSTAT and implements adjustments to use them to extract corresponding crop agro-ecological raster files from GAEZ portal. These raster files will be used in the following part to extract required spatial attributes for the region of study.

GeoCLEWs determines the top priority of crop types from the FAOSTAT dataset according to the user-defined country, extracts the top ten crops, and categorizes them into two groups. This feature is highly important to reduce manual effort and enhance the efficiency of the CLEWs analysis since it processes the top ten crops leading to more detailed assessment. The top five are classified as the “main crops” that have the highest harvested area, and remaining crops, rated 6 to 10, are grouped as “additional crops”. New script performs additional functions to calculate an average value of additional crops and add them as one single crop to the assessment. Averaging function reduces complexity issues and preserves further agro-ecological information about commonly grown crops in the region of interest. This supplementary evaluation presents an effective improvement since land and water evaluation are not limited to 5 top-harvested agricultural products.

GeoCLEWs is able to automatically detect, collect, filter, preprocess, and download all required GAEZ land and water raster files based on user configuration and identified primary and additional crops. This tool detects essential raster files and downloads them containing necessary agro-ecological potential yield, crop water deficit and evapotranspiration information for CLEWs processing. Considering the fact that the detailed land and water assessment is significantly data intensive, this advancement considerably reduces time-consuming manual input and human errors, as well as facilitates modification and standardization.

2.3. Land Cells Generation

GeoCLEWs generates land cells from regularly distributed point grid, which serve as the smallest georeferenced parcels of the region. Land cells are created based on geographical administrative boundary of the country or any arbitrary region. The script implements modification and accuracy validation to obtain precise georeferenced land cells across the study area, which will be used in the next part for extracting spatial data from downloaded GAEZ raster files.

The initial script utilized an ASCII grid file from GAEZ v3 datasets to produce land cells that poses several challenges, including outdated information, unavailability, and manual geoprocessing. To address these issues, an efficient approach is adopted to generate land cells based administrative boundary of the total area of study. This method improves the performance of GeoCLEWs to operate without dependency on specific input data as the starting point. It has a general functionality and adaptability to a wide variety of projects with various objectives and configurations. Another advantage is that users have the opportunity to utilize any arbitrary geographical boundary as the input data as well as use the proposed open source administrative boundary available in the GADM portal. In addition, the spacing of point grids is adjustable to optimize high-resolution raster data extraction and processing time. The default setup of point grids concentration results in higher accuracy although it is possible to modify that easily to reduce computational complexity.

2.4. Geospatial Attributes Extraction

This part includes extracting agro-ecological attributes from GAEZ raster files and seamlessly incorporating them into the corresponding land cells. GeoCLEWs retrieves spatial attributes including crop agro-climatic potential yield, water deficit, evapotranspiration, precipitation, and land cover. It employs a new function to modify GAEZ raster files with global coverage and automatically extract the geographical boundary of the region. Automatically clipping global raster files is highly beneficial for reducing the processing time of geospatial feature extraction. This function is not available in the prior code and, consequently, a great deal of time was required to manually clip raster files.

2.5. Calculating Region Summaries

In the last part of the script, GeoCLEWs groups land cells based on administrative boundary and generates outputs for developing CLEWs model. According to the user-defined admin level, land cells located in the similar administrative region are classified in the same region, which is named cluster in this thesis. It collects spatial attributes of land cells within the same cluster to process land and water statistics of each administrative region individually. GeoCLEWs offers the possibility to aggregate admin regions to reduce processing computation of CLEWs modelling. Regional aggregation refers to the process of combining data from different administrative regions into a larger, resulting in the formation of new aggregated region clusters.

In addition, this thesis incorporates the agro-ecological statistics of five additional crops into land and water assessment. GeoCLEWs implement additional crops averaging process to calculate an average value for additional crops, which serves as one single added crop including average statistics of crop potential yield, water deficit, evapotranspiration, named “Other Crop”.

Outputs are provided in CSV formats for developing CLEWs model as well as interactive graphs for comprehensive analysis presenting several advantages.

GeoCLEWs implements CLEWs-compatible unit conversions promoting consistency with the framework and comparability with prior studies. Statistical results of land and water data including land use, potential agronomic yields, crop evapotranspiration, crop water deficit, and precipitation are computed according to the unit standardization within the CLEWs framework. Additionally, final outputs are provided in a tabular format that is designed based on the clewsy structure. GeoCLEWs outcome along with additional information including electricity and configuration file can seamlessly be imported into CLEWs UI using clewsy, boosting the efficiency of the new tool.

Chapter 3.

Methodology

This chapter outlines the process of data preprocessing and the functioning of GeoCLEWs. Data Preparation presents a well-designed data collection and preparation structure that accelerates automated land and water assessment while minimizing manual errors. The required FAOSTAT and GAEZ datasets are retrieved in an efficient storage memory format enabling GeoCLEWs to collect data without user intervention.

The second section of this chapter details operations and outputs of GeoCLEWs, including extracting high-resolution GAEZ land and water datasets, implementing geoprocessing, and generating detailed statistics for CLEWs modelling. GeoCLEWs involves iterative stages consisting of script development, calibration, implementing improvement, and revalidation. The script consists of five main parts: initialization and configuration, data collection and preparation, land cell generation, spatial attributes extraction, regional aggregation, and statistical calculation. It was highly important in this thesis to develop GeoCLEWs utilizing open source tools and data as well as make it freely available to all users to promote contribution to sustainable development.

3.1. Data Preparation

Generating detailed land and water statistics requires different types of data, encompassing primary crops in the region of study, crop agro-ecological information, precipitation, land cover, and administrative boundary of study area. GeoCLEWs utilizes FAOSTAT to identify the most harvested types of crops in each country, then it processes the identified crops' agro-ecological raster files from GAEZ portal along with land cover and precipitation datasets. The administrative boundaries are employed to extract spatial attributes from raster files and generate detailed land and water statistics for specific admin regions. The source and method used to prepare the datasets are described in detail below:

- Preparation crop statistics from FAOSTAT.
- Preprocessing GAEZ datasets.

- Administrative boundary.

Preparation Crop Statistics from FAOSTAT

The explanation of extracting the crop statistics is presented here, which facilitates finding out the most vital types of crops for CLEWs modelling in the selected region without user intervention while running GeoCLEWs. Evaluating cropland areas allocated for harvests can reveal the most significant crops in a region, which are crucial in land management. Subsequently, crops with higher harvested agricultural land are considered more important, which should be involved in climate, land, energy, and water analysis. Open source FAOSTAT database [69] that delivers great sources of data related to agricultural activities globally. GeoCLEWs retrieves crop statistics from the FAOSTAT portal, which produces datasets on a yearly basis including harvested area measurements of a wide variety of types of crops.

This thesis presents a new method that can speed up and simplify primary crop identification compared to the conventional method. Traditionally, modellers manually retrieve the most recent datasets from the FAOSTAT portal and process the results to identify the most important crop kinds for the study area. However, GeoCLEWs eliminates this manual effort by automating this step. At the time of writing this thesis, FAOSTAT 2020 contains the most updated and completed harvested area datasets. Therefore, the statistics of harvested area 2020 from the FAO database are downloaded, preprocessed, and stored as FAOSTAT_2020.csv. This document includes global official measurements of numerous harvested crops. The generated CSV file is utilized by GeoCLEWs for evaluating the priority of crops and selecting them in Part 2: FAOSTAT and GAEZ Data Collection and Preparation of the script. Appendix B illustrates an example of data from the FAOSTAT_2020.csv file, displaying the harvested figures in hectares.

Preprocessing GAEZ Datasets

This thesis presents a new method for precipitation, land cover, and crop agro-ecological data collection and preparation to provide a unique opportunity to obtain all essential high-resolution spatial data without user intervention. Global Agro-Ecological Zoning version 4 database is selected as the source of spatial data offering historical and projected information on land and water. The GAEZ v4 portal is an interactive data access facility that allows users a variety of analytical outputs in addition to providing visualization. Nonetheless, users face several challenges in figuring out voluminous GAEZ documentation as well as finding and retrieving required land and water data from its portal. Modellers currently must spend a great deal of time downloading the necessary historical and projected data in raster format from portals; as an example, the Kenya case study discussed in Chapter 4 involves 110 GAEZ raster files. Furthermore, some of the essential GAEZ raster data require additional spatial modification by adopting GIS techniques, which presents additional difficulties for non-technical users and decision-makers. In this study, high-resolution land and water data processing is streamlined by providing GAEZ data identification and extraction automatically; this allows to perform GeoCLEWs and generate outputs without prior knowledge of GIS. A thorough study and evaluation of extensive documentation of GAEZ v4 is implemented during this research to find out the specifications of the spatial data including data type, resolution, measurements, data sources, processing methodology, accuracy, and limitations. This step also plays a vital role in generating a full automotive workflow to build the CLEWs model. Essential GAEZ data are analyzed, preprocessed, and stored in the same directory as GeoCLEWs, which are utilized in GeoCLEWs Part 2: FAOSTAT and GAEZ Data Collection and Preparation and Part 4: Geospatial Attributes Extraction to . The following details of GAEZ preprocessing are presented. The subsequent content outlines the categories of raster data that must be obtained from the GAEZ v4 portal for the purpose of executing comprehensive land and water data processing for CLEWs modelling. Table 2 provides more details of essential land and water on the GAEZ portal.

Table 2: GAEZ v4 data collection guideline.

Spatial Data	Theme	Sub-theme	Variable	Unit	Type
Agro-climatic potential yield	**3**	**Agro-climatic yield**	**Agro-climatic potential yield**	**kg DW/ha**	**Continuous**
Crop Water Deficit	**3**	**Growth cycle attributes**	**crop water deficit**	**mm**	**Continuous**
Crop Evapotranspiration	**3**	**Growth cycle attribute**	**Crop-specific actual evapotranspiration**	**mm**	**Continuous**
Precipitation	**2**	**Moisture regime**	**Annual precipitation**	**mm**	**Continuous**
Land Cover	**1**	**Land cover**	**Dominant land cover**	**class**	**Categorical**

- **Agro-climatic potential yield:** This dataset is supplied in the format of continuous raster files. This can be downloaded from theme 3, which is Agroclimatic Potential Yield under the sub-theme named Agro-climatic yield. GAEZ yields for crops that are taken into account for CLEWs processing are given in kg dry weight per hectare (kg DW/ha) while there are a few exceptions. The yields are given in 10kg dry weight per hectare for alfalfa, miscanthus, Napier grass, reed canary grass, pasture legumes, and grasses. The yields for sugar beet, sugar cane, and olives are measured in kilograms of sugar per hectare and kilograms of oil per hectare, respectively. Kilogram lint per hectare is how cotton yields are expressed. Considering the majority of yield, this study set the default to kg DW/ha.

- **Crop water deficit**: This dataset is provided in a continuous raster file in millimetres that can be downloaded from theme 3, the sub-theme of Growth cycle attributes.

- **Crop evapotranspiration**: Crop-specific actual evapotranspiration is provided in a continuous raster file under the sub-theme of the Growth cycle attribute in theme 3, which is delivered in millimetres.

- **Precipitation**: The second theme under the sub-theme of Moisture regime, users can find the *annual precipitation* variable, which is given in millimetres as a continuous raster file.

- **Land cover**: This is available for download under theme 1 classified as the Dominant land cover variable, which is categorized and provided in LERC compressed format (Limited Error Raster Compression).

Categorical Data Preprocessing

Categorical raster files divide information into distinct groups and a specific method is employed to prepare the discrete file for CLEWs modelling. Land cover is the only categorical raster file that is essential for CLEWs modelling, classifying landscape into land cover categories or land use such as water, cropland, and built-up area. GAEZ v4 categorized landscape into 11 classes and offered the land cover with a high spatial resolution of 30 arc-seconds, Table 3 represents the description of land cover classification in GAEZ v4. It reduced the size of this land cover dataset using the LERC (Limited Error Raster Compression) technique. LERC data encounter compatibility issues with various versions of Python geoprocessing packages, including Rasterio and GDAL, as well as potential conflicts with other installed libraries. In order to address this problem, open source QGIS is employed to convert LERC to compatible GeoTIFF (Geographic Tagged Image File Format) format leading to more efficient land cover processing. The LERC compressed land cover is exported to GeoTIFF while retaining the palette information by using the Translate tool from the GDAL/OGR toolbox in QGIS. Figure 3 illustrates the GAEZ land cover map including the classification legend generated in QGIS. In this research, a preprocessed and converted format of the land cover data has been generated, named *LCType_ncb.tif* and included in the *global_raster_input* directory. This raster file provides coverage on a global scale and can be utilized in any land and water processing project with various geographical locations. This method minimizes the need for manual data collection and streamlines the GIS manipulation.

Table 3: GAEZ land cover classification description.

Land Cover Type	Description
LCType1	More than 75% Cropland
LCType2	More than 75% Tree-covered land
LCType3	More than 75% Grassland shrub or herbaceous cover
LCType4	More than 75% Sparsely vegetated or bare
LCType5	50 – 75% Cropland
LCType6	50 – 75% Tree-covered land
LCType7	50 – 75% Grassland shrub or herbaceous cover
LCType8	50 – 75% Sparsely vegetated or bare
LCType9	More than 50% Artificial surface
LCType10	Other land cover association
LCType11	Water permanent snow glaciers

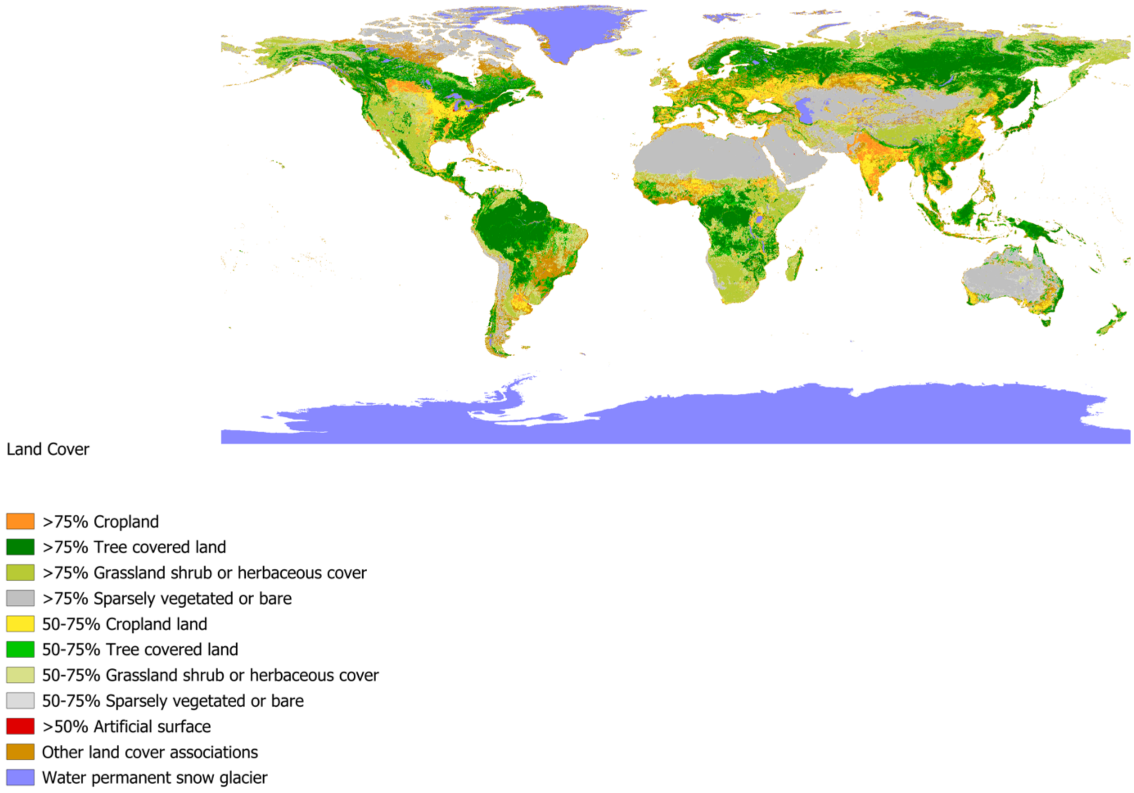

Land Cover
🟧 >75% Cropland
🟩 >75% Tree covered land
🟩 >75% Grassland shrub or herbaceous cover
⬜ >75% Sparsely vegetated or bare
🟨 50-75% Cropland land
🟩 50-75% Tree covered land
🟩 50-75% Grassland shrub or herbaceous cover
⬜ 50-75% Sparsely vegetated or bare
🟥 >50% Artificial surface
🟫 Other land cover associations
🟦 Water permanent snow glacier

Figure 3: GAEZ Land Cover classification with a global converge.

Continuous Data Preprocessing

Continuous geospatial datasets are collected and preprocessed to smooth the procedure of spatial processing. Continuous raster files provide spatial values that fluctuate smoothly over the region. All essential GAEZ raster data for CLEWs modelling, except the land cover, are accessible in the continuous format including crop agro-climatic potential yield, crop water deficit, crop evapotranspiration, and precipitation. To attain a comprehensive and meticulous approach to land and water processing, separate datasets including information on all available types of crops in GAEZ are created in CSV format. These files contain essential information that enables GeoCLEWs without user intervention to filter, process, and download required raster files based on the dominant crop type across the region of interest. That is highly useful to eliminate the time-consuming process of data collection and take steps toward fully automating CLEWs modelling.

The generation and modification of continuous datasets vary depending on the requirements of the CLEWs frameworks and the data specification. Precipitation raster data including projected values for the period of 2011-2040 is processed and stored inside the *global_raster_input* folder to automate the land and water analysis. This is a single, globally-coverage raster file that can be used in all CLEWs projects. Nonetheless, the three remaining datasets highly vary based on the type of the crop and separate raster data is processed for each crop considering water supply, management input, and climate data source. The latest version of GAEZ offers projected spatial datasets on a long-term estimation from 2011 to 2040 based on high input level, which can lead to unrealistic estimations in developing counties or regions that rely on traditional farming practices. In this research, the spatial datasets based on low-level management are provided by integrating historical information collected from 1981 to 2010, which is calculated based on low-level input. Subsequently, two individual CSV datasets, Low Input and High Input, are generated for potential yield (yld), crop water deficit (cwd), and evapotranspiration (evt). The climate data sources of historical and projected datasets are CRUTS32 and HadGM2-ES as suited for agricultural applications leading to more reliable results. It comprises a total of six datasets as follows:

- GAEZ_yld_High_Input
- GAEZ_yld_Low_Input

- GAEZ_cwd_High_Input
- GAEZ_cwd_Low_Input
- GAEZ_evt_High_Input
- GAEZ_evt_Low_Input

Administrative Boundary

Within the framework of this research, national and subnational spatial statistics are extracted from GAEZ raster files using administrative boundaries in shapefile format. GeoCLEWs is flexible and compatible with any arbitrary shapefile, in addition, this study offers an opportunity to use Global Administrative Area (GADM) [71] as an updated and open source database to streamline the process of open source data collection. The GADM is a freely available database that maps the administrative regions of all countries, at a wide variety of levels of administrative divisions. Admin level 0 refers to the national scale, and level 1 indicates primary administrative divisions within a country such as provinces. Subsequently, the detailed levels of admin refer to subdivisions of the corresponding higher admin level. The GADM offers spatial data in Shapefile format by nation, and users can easily use this portal to store the preferred administrative boundaries.

3.2. GeoCLEWs

GeoCLEWs is designed to simplify, automate, and facilitate the high-resolution historical and projected geospatial analysis required for CLEWs modelling. It is a self-documented code that includes explanations within the script as a description and guideline regarding the components to assist users and developers in clearly understanding the script's performance. Therefore, they can effectively utilize GeoCLEWs and collaborate in improving this open source tool. The following provides an explanation of the developing process of GeoCLEWs, including in-depth descriptions of each component, functionality clarification, and results examples. GeoCLEWs notebook is designed based on five main analytical parts and details are shown in Figure 4.

- **Part 1**: Initialization and configuration.
- **Part 2**: FAOSTAT and GAEZ data collection and preparation.

- **Part 3**: Region generation.

- **Part 4**: Geospatial attributes extraction to regions.

- **Part 5**: Key summary statistics calculation and generate outputs for further use in CLEWs modelling.

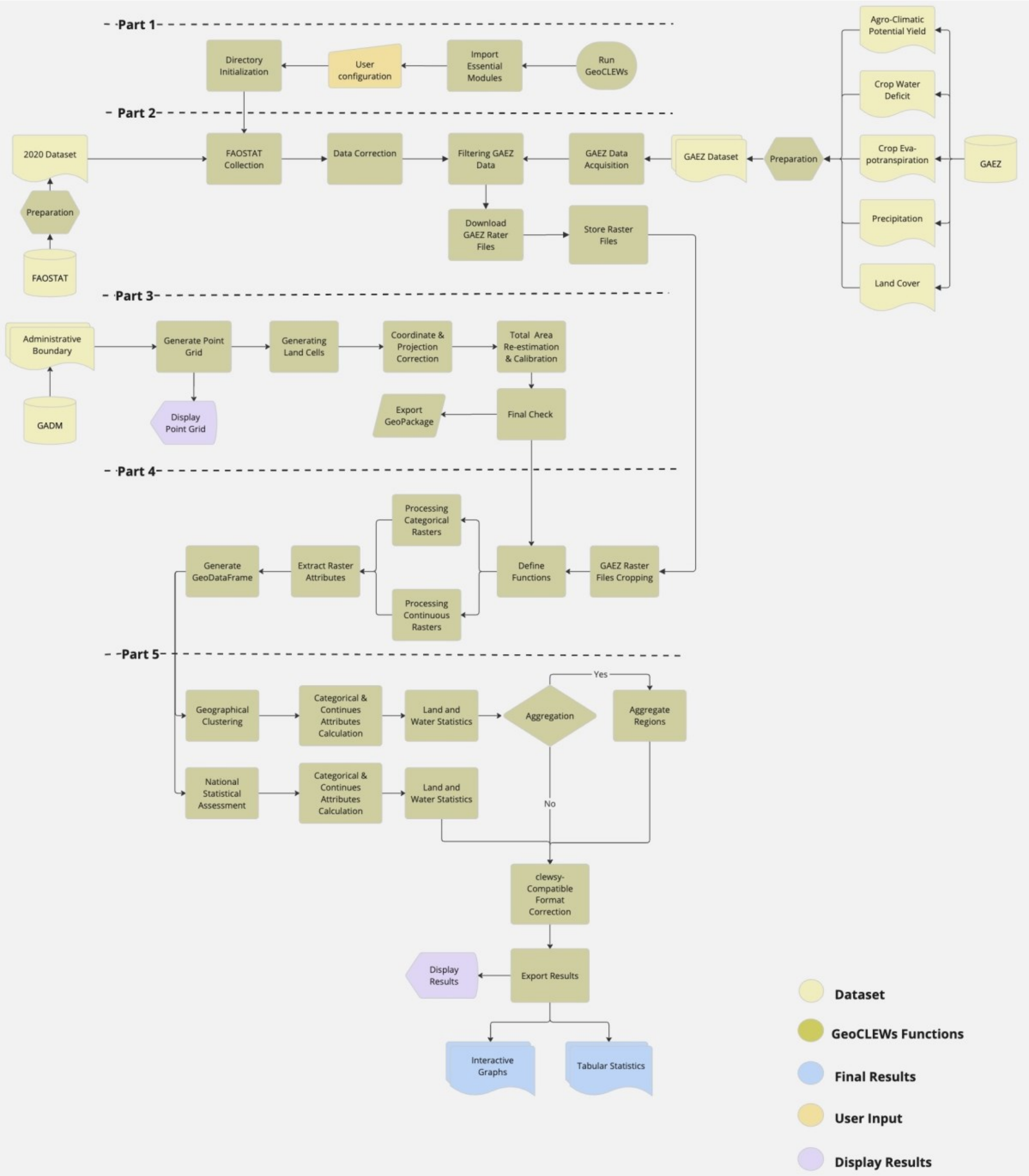


Figure 4: GeoCLEWs Flowchart – illustrating key functions and processes.

Part 1: Initialization and Configuration

GeoCLEWs uses the most accessible and straightforward solution for all users with various levels of programming expertise. This part includes the initial setups of the project including:

- 1.1. Importing essential modules.
- 1.2. User configuration.
- 1.3. Directory initialization and structure.

It starts with importing required Python packages and dependencies based on functions involved in the script to seamlessly run the code. Then, GeoCLEWs requires user input to define configuration such as country name and aggregation setup to customize to outputs according to their project need. Finally, the directory initialization assists to arrange the input folders to retrieve required datasets for script to run and organize output folders to store different type of results in the corresponding directory without manual intervention.

Part 1.1. Importing Essential Modules

GeoCLEWs requires a specific configuration of libraries and associated dependencies installed in a Python environment to avoid any version conflict and run the code smoothly. It was quite challenging to implement all the analysis steps and geospatial process using Python modules which are publicly available and commonly used by a wide range of users. One of the issues was module dependencies conflicted with each other because of various reasons such as version incompatibility or installation order. The optimum setup of the essential packages and modules is identified and provided to perform all steps successfully and accurately. Before running the script, a specific isolated environment is created using Conda, which is an important step to manage dependencies and install the optimum setup with a proper installation order. The new environment includes necessary Python modules and packages with compatible versions to run GeoCLEWs smoothly and avoid conflicts. GeoCLEWs has been successfully tested and verified on Windows machines. However, there may be incompatibility issues with other operating systems due to differences in Python packages or their versions. GeoCLEWs starts the process by importing essential open source Python libraries and dependencies from the initially created Conda environment, which are used in the following parts for various spatial and non-spatial functions.

Part 1.2. User Configuration

User configuration provides flexibility to customize the project setup to accommodate users' needs and fit various projects. It is the only part that requires manual input, and all remaining operations are carried out without user intervention; it consists of six main steps:

1. **Country name**: First, users specify the name of the country to perform the analysis. GeoCLEWs utilizes the name of the country to automatically perform a wide range of functions including identifying the 3-letter country code, selecting dominant crops in the region, and generating results with a clewsy-compatible format. The developed script extracts the three-letter ISO code [72] of the selected country, which is a widely recognized code that identifies every country and adheres to CLEW criteria. The source of the ISO country code is provided inside the script for user comfort. It is highly important that users do not need to put any effort into finding out this information and streamline the fully automated process.

2. **RCP**: It supports choosing the Representative Concentration Pathway. RCPs are a collection of scenarios that attempt to quantify various possible future greenhouse gas concentration pathways and corresponding radiative forcing. The GAEZ v4 generates projected information employing four RCPs, with RCP2.6 being the lowest emissions scenario, RCP4.5 and RCP6.0 representing intermediate pathways, and RCP8.5 being the highest emission possibility throughout the 21st century. GeoCLEWs enables users to generate land and water statistics according to specific RCPs assisting in defining more accurate scenarios while developing CLEWs model.

3. **Admin level**: GeoCLEWs is able to process data at different administrative levels according to user preferences. Admin level 0 generates outputs only on a national scale and admin level 1 delivers results at both national and subnational levels, e.g. county.

4. **Aggregation**: It offers region aggregation since the number of geographic areas involved in geo-processing has a negative impact on the complexity of the CLEWs modelling, while GeoCLEWs is capable of producing a comprehensive analysis on as many regions as the user requests. The number of regions to be aggregated into one group can be decided by the user. Moreover, in case of aggregation, there is an additional feature to extract specific areas and avoid combination. Projects that aim to obtain details on a certain region but cumulative information for the remaining area benefit from this new feature. Users can input the first three letters of the selected area for exclusion, and the script extracts the region and generates separate final outputs beside newly created groups of other geographical regions. This new element highly

improves the performance of CLEWs modelling in subsequent steps and optimizes its processing time.

5. **Projection system**: Identifying the proper projection system for the selected country is important since it varies based on the geographical location of the region in study. End-user is provided with a complete explanation, examples, and source of coordinate system worldwide [73] to determine and input required data considering the location of the study area.

Part 1.3. Directory Initialization and Structure

This part contains a well-defined directory structure assisting in an effortless reproducing script. The directory organization is based on a simple design while it can manage multiple projects with different geographic locations in a main directory. GeoCLEWs easily find the relevant input data related to each project and generates results with proper output name. Further modification may be required if the end-user chooses an alternative structure while the proposed structure simplifies the directory design and streamlines perfuming GeoCLEWs. The straightforward structure contains three main directories and three subdirectories:

Main Directories:

- Global_raster_input: This is a directory for downloading GAEZ raster files with global coverage. As it is explained in the section of Preprocessing GAEZ Data the precipitation and land cover raster files have already been preprocessed and moved in this directory as it is a similar input file for all projects. The existing processed data facilitates the collaboration of non-technical end-users by eliminating the step of intricate preprocessing compressed continuous GAEZ raster layers. GeoCLEWs downloads and stores the remaining essential raster files including agro-climatic potential yield, crop water deficit, and crop evapotranspiration based on user-defined configuration. The developed script collects data from GAEZ and FAOSTAT datasets, which are preprocessed and provided along with the Jupyter Notebook code in this research.

- Cropped_raster_input: GeoCLEWs utilizes downloaded raster files saved inside the *global_raster_input* directory and performs cropping functions; the results of this step are stored inside a new directory named *cropped_raster_input*. This function decreases processing time.

- Data: a directory hosted input files and output results.

Subdirectories:

- Data/input: The input subdirectory exists within the Data as the main directory. Users can manually place a Shapefile of the administrative boundary of their preferred geographical region inside this folder. Considering the admin-level customization, proper input data should be saved in this directory. This project provides users with the option of utilizing the open source GADM database including administrative boundaries with different admin-level details on a worldwide coverage as well as employing any arbitrary geographical region in a Shapefile format.

- Data/output: General results in various steps of the process are stored inside the output directory, which is located within the Data folder.
- Data/output/summary_stats: Final land and water analysis for CLEWs modelling including calculated tabular statistics and interactive graphs are exported in this folder as a subdirectory of the output.

Part 2: FAOSTAT and GAEZ Data Collection and Preparation

This part of GeoCLEWs consists of functions to retrieve all required values from FAOSTAT and processed geospatial information from GAEZ datasets. It is entirely carried out without manual intervention, which plays an important role in streamlining the intricate and time-consuming process of spatial data collection and preprocessing.

- 2.1. FAOSTAT collection and preparation.
  - 2.1.1. Retrieve the top 10 harvested crops.
  - 2.1.2. FAOSTAT Standardizing.
- 2.2. GAEZ data collection and preparation.
  - 2.2.1. GAEZ data acquisition.
  - 2.2.2. GAEZ data modification.
  - 2.2.3. GAEZ data standardizing.
  - 2.2.4. GAEZ Data Filtering according to user configuration.
  - 2.2.5. Downloading and storing GAEZ raster files in a clewsy-compatible format.

Considering the user-selected country, it retrieves the agricultural statistics from the Food and Agriculture Organization of the United Nations. The extracted information then is corrected and used to filter the GAEZ datasets. After data cleaning and standardizing, GeoCLEWs downloads the relevant raster files from the GAEZ v4 portal directly and stores in a proper directory for the next operations. As previously stated, the GAEZ presents a wide variety of datasets with various configurations demanding a considerable amount of effort to study full documentation and collect essential data. Automated data collection, correction, and standardization by GeoCLEWs significantly simplify the entire process for end-users and avoid mistakes such as miscalculations, missing data, and potential human errors. It is also a highly important step toward developing a fully automatic CLEWs framework.

Part 2.1. FAOSTAT Collection and Preparation

The script identifies the most important cultivated crops in the user-selected country that play significant roles in the industry, agriculture, and economy of that nation. Since these crops would be affected by any long-term planning regarding managing water, land, and energy sources as well as climate change practices, it is essential to involve historical and projected crop-related agro-climatic factors and analyze crop suitability to achieve informed land use and resource management. First, GeoCLEWs detect all crops cultivated in the chosen country and then select the top 10 harvested ones. Finally, it implements naming standardization to align with the CLEWs standards.

Part 2.1.1. Retrieve Top 10 Crops

An effective method is adopted to identify and choose the most important crops for assessing spatial characteristics without increasing the computational demand. Crops are prioritized and selected based on their coverage of the harvested area. In CLEWs modelling, it is crucial to limit the number of crops for analysis due to computational complexity. Therefore, it is common to utilize information on a maximum of five crops [56] to obtain a reasonable land and water assessment as well as avoid adding extra complexity to the CLEWs modelling. However, there would be a lack of crop attainability examination of the remaining crops. To address this gap, GeoCLEWs consider the top 10 crops in terms of harvested area. For the first top 5 crops, it generates detailed processing for each individually while it produces an average value for additional crop, which are ranked 6 to 10, to optimize processing time in the phase of CLEWs modelling. The estimated average information is presented as a single extra crop, resulting in a total of six crops. This additional feature supports preserving more valuable information and reduces processing time.

GeoCLEWs employs different functions to find out the principal crops. It starts by reading the preprocessed FAOSTAT_2020.csv document, which is extracted from the

FAOSTAT dataset and stored in the same directory. As of writing this thesis, the FAOSTAT version of 2020 is the latest available one. In the future, it can be easily replaced by the most recent FAOSTAT dataset at the moment, with no impact on the scrip operation. Then it sorts the figures in accordance with the proportion of the land allocated to harvesting. After that, the top ten crops are identified and classified into two groups, ‘Main Crops’ and ‘Other Crops’. The Main crops collection represents the top 5 harvested crops and the Other Crops group consists of the second top 5 crops. The script displays two generated groups and their members for clarification, so users are aware of the crops involved in their process and they have the opportunity to manually modify the list to accommodate the specific requirements of their projects.

Part 2.1.2. FAOSTAT Standardizing

The main purpose of this piece of code is to convert the format of crop names used in the FAO dataset to a standard format. This facilitates transferring data among the GAEZ, FAO datasets and CLEWs model. For this purpose, the naming format used in CLEWs modelling is set as the standard framework. Consequently, the retrieved data from FAOSTAT is converted to that framework. CLEWs model utilizes a specific 3-letter naming convention which is recognizable by CLEWs user interface and can easily be processed through OSeMOSYS Cloud for optimization and visualizing final results. After formatting the name of the selected 10 crops, the script displays to users two generated groups of crops and a list of members in each with a new naming format.

Part 2.2. GAEZ Data Collection and Preparation

GeoCLEWs presents a useful method offering an opportunity for GAEZ data collection without user intervention leading to greater simplification and efficiency. After identifying the important crops in the region of study, it is required to collect all GAEZ information related to those crops as well as other essential land and water information. Datasets including required information of raster files are already collected from the GAEZ v4 portal, preprocessed, and stored in the directory where the Jupyter Notebook is located. These datasets comprise specifications of historical and future estimated raster files. GeoCLEWs acquires essential data from those in accordance with user configuration and implements some modification and standardization. Following that, it filters according to identified primary and additional crops names to select relevant files. Finally, it downloads the corresponding raster files from GAEZ portal. This new feature can improve the international collaboration of users without a geospatial analysis background.

Part 2.2.1. GAEZ Data Acquisition

In the beginning, script reads the preprocessed documents in CSV format that are available in the same directory as GeoCLEWs. It uses Pandas library to read the provided CSV files and keep them in properly named files. In total, six CSV files are obtained, encompassing the agro-climatic potential yield, crop water deficit, and crop evapotranspiration with two levels of input management (low and high) [44]. The low input represents traditional management assumptions, and the high level represents advanced farming systems.

Part 2.2.2. GAEZ Data Modification

Some modifications are needed to classify the obtained CSV files. The GAEZ datasets cover different applications of watering agricultural fields, but it is not properly classified in the datasets. GeoCLEWs adds an extra column to these files to represent their water supply. The two sources of water supply are labelled as:

- Rain-fed: relying only on natural rainfall.
- Irrigation: Using artificial irrigation systems to supplement or replace natural rainfall.

Part 2.2.3. GAEZ Data Standardizing

Data collected in previous part is converted to a standard format using the crop names. GAEZ uses different crop naming formats from the FAOSTAT and CLEWs model. It is beneficial to have the 3-letter CLEWs format as a standard framework and convert the GAEZ crops' names to this framework. This task is implemented by defining a new function named *GAEZ_naming*, which adds a new column to CSV files, converts crop name to the 3-letter CLEWs format, and finally stores new crop names inside the newly added column.

Part 2.2.4. GAEZ Data Filtering According to User Configuration

After implementing modification and correction, it is feasible to extract specific information from the stored datasets in accordance with the user configuration. First, the potential yield, water deficit, and evapotranspiration are filtered according to the user-selected RCPs. Following that, the script collects information related to the FAOSTAT top 10 crops and stores them in a new list. In total, 12 separate lists are generated including agro-ecological information with the high and low input management methods and irrigation and rain-fed water supply.

Part 2.2.5 Downloading and Storing GAEZ Raster Files

GeoCLEWs employs generated lists in the previous part to download required raster files from GAEZ portal. Every crop inside listed datasets possesses unique URL links corresponding to a specific agro-ecological feature, level of agriculture management, water supply method, and RCP. These links refer to exclusive raster files stored within the GAEZ v4 portal. GeoCLEWs retrieves matching URL links and afterward downloads the raster files and stores with TIFF format, which will be utilized for spatial attribute extraction.

Part 3: Generating Land Cells

The third part of the script consists of the functions for creating land cells inside the geographical area of research. GeoCLEWs employs an efficient approach to generate land cells using the boundary of the study region. It generates a regular point grid across geographical area and converts georeferenced points to polygons. These polygons are known as land cells in this thesis. Finally, it checks the accuracy and precision of generated land cells and implements calibration to obtain a reliable vector dataset. Land cells are regularly distributed that fully cover the study area; it is helpful to divide a large geographic zone into small segments to collect detailed spatial features. The process of land cell generation is implemented regardless of the admin level factor. One specific country or any region with arbitrary geographical boundaries is processed in depth considering the highest resolution of the input raster files. In the following parts of the script, the corresponding geospatial attributes, such as agro-ecological features, will be assigned to relevant land cells through a geospatial join.

- 3.1. Generating georeferenced point grid from Shapefile.
- 3.2. Converting points to polygons.
    - 3.2.1. Spatial join.
    - 3.2.2. Generating polygon.
- 3.3. Total area re-estimation and calibration.

- 3.3.1. Area calibration.
- 3.3.2. Final check.
- 3.3.3. Export as GeoPackage.

Part 3.1. Generating Georeferenced Point Grid from Shapefile

The script starts with reading the Shapefile of the administrative boundary stored inside the Input directory. GeoPandas library creates a new GeoDataFrame from the attributes and geometry of the Shapefile. GeoDataFrame is used because it can create tabular data structures including a geometry column storing different geometric shapes. Following that, the coordinate system of the new GeoDataFrame named '*shapefile*' is reprojected to WGS84. Changing the coordinate system to WGS84, World Geodetic System 1984, improves the project consistency.

Then, it creates a point grid with regularly distributed points all over the geographic area of interest. Each point represents a specific location on the Earth with a unique latitude and longitude. The point spacing of the distributed points plays a highly important role in the size of land cells and their coverage. The majority of the GAEZ outputs are provided in standard raster format of 5 arc-minute grid cells (nearly 9 x 9 km at the equator). However, there are some exceptions with a higher resolution, such as 30 arc-second (about 0.9 x 0.9 km). As we are using GAEZ results including both resolutions (5 arc-minute and 30 arc-second), the reasonable distance between points of the generated point grid could be around 9 km, which offers a detailed analysis for CLEWs modelling. Therefore, the default spacing of the point grid is set to 0.09 Decimal Degrees (DD) in both latitude and longitude. Since the coordinate system is reprojected to WGS84, the spacing of 0.09 DD is equal to approximately 9.9 kilometres at the equator.

GeoCLEWs possesses a transparent and modifiable structure; users can easily adjust and customize parameters like spacing to suit their needs. GeoCLEWs generates a plot of the produced point grid representing points' location and administrative boundary allowing users to understand its inner operations. Examples of the produced plot at admin level 0 and admin level 1 are shown in Figure 5 and Figure 6 respectively. The former illustrates point grid on a national scale while the latter includes the administrative boundaries of counties. The red lines represent the administrative boundaries on the national and sub-national levels in Kenya. Points that are located inside the national boundary are shown in yellow colour and blue colour points signify points outside the official border of Kenya which will be excluded from point grid at the end of this part.

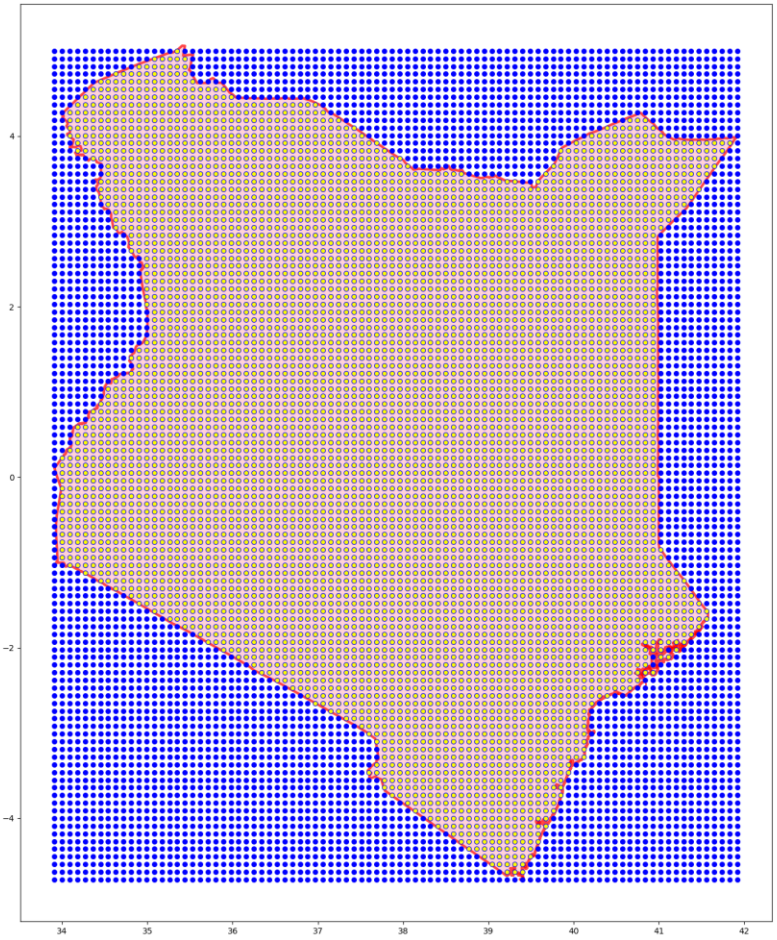


Figure 5: **Point grid at admin level 0, Kenya.**

For detailed analysis, different spacing choices have been examined during this thesis. Figure 6 illustrates a comparison between point grids with 0.09 and 0.18 decimal degrees, which are approximately equal to 9.9 and 19.9 kilometres at the equator respectively. As it is shown, the distance between the points generated in plot (a) is almost half of the spacing in plot (b). The plot (a) has more points closer to each other and subsequently, the land cells generated from this point grid are smaller. Reducing the size of the land cells leads to more in-depth sampling and accordingly improves the precision and efficiency of geospatial attribute analysis in the following parts involved in GeoCLEWs.

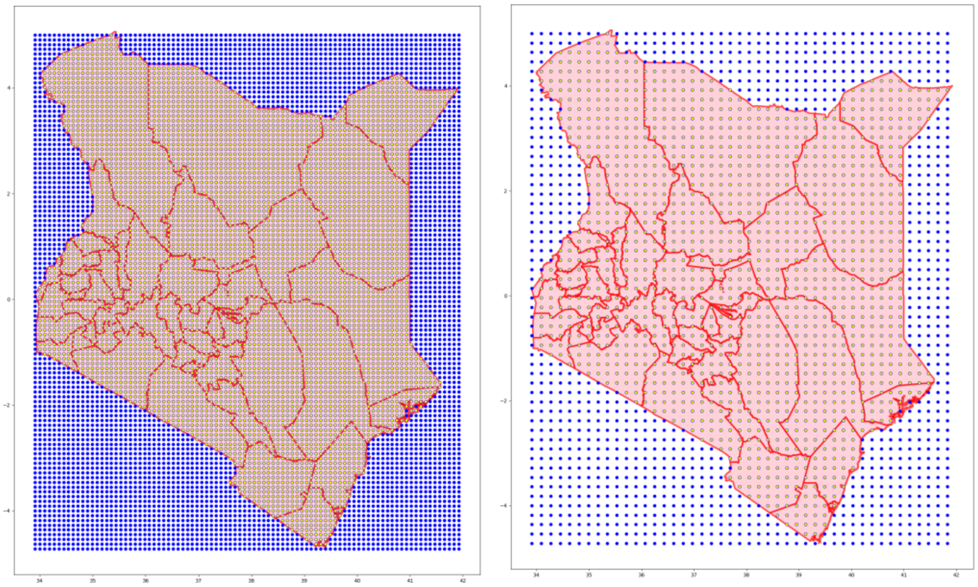


Figure 6: Comparison of generating a point grid at admin level 1 with different spacing from the Kenya case study. Left plot (a) with 0.09 DD and right plot (b) with 0.18 DD.

The final step is data cleaning. The point grid is generated using the highest and lowest longitude and latitude of the region. Some points are created outside of the border, shown in blue colour points in Figure 5. The Shape of the national border is utilized to extract only points inside the main border, which are displayed in yellow. It is worth mentioning, the accuracy of generated points has been validated and each point has a precise georeferenced coordinate in WGS84.

Part 3.2. Converting Points to Land Cells

This part of the script generates land cells by converting inside point grid to polygons. First, the script makes adjustment to reproject the point coordinate system to match the coordinate system of the administrative boundary. This ensures consistency throughout the geoprocessing to achieve accurate spatial assessment. GeoCLEWs also performs some modification prior to spatial join such as renaming column of the GeoDataFrame containing point grid. Spatial join extracts attribute of administrative division from admin Shapefile, which includes the name of admin divisions, and assigns this attribute to corresponding points. Results of this step represent the geographical division of each point. Following that, points are converted to polygons using square buffer method. These polygons that include projected coordinates and attribute of admin division serve as land cells in the following parts.

Part 3.2.1. Spatial Join

This part outlines the necessary adjustments and some implementation methods for running spatial join between administrative boundary and point grid GeoDataFrames (GDF). It is required to make some corrections before implementing the spatial join including:

1. The coordinate system of the extracted points inside the boundary of the region of study is set to the WGS84 to run an accurate spatial join. As mentioned, the administrative boundary has been reprojected to WGS84, and it is required to reproject the generated point grid to the same coordinated system to preserve consistency over the spatial assessment.
2. The GDF of the grid points is renamed to '*clustered_gdf'*, which will be used as the base for the following geoprocessing as well as regional aggregation in Part 5*.*
3. The name and data type of the index column of the clustered file should be changed into the '*cluster*' (referring to administrative region in this thesis) and string respectively. This column represents the administrative division of each point.

After modification, the '*clustered_gdf'* and *admin Shapefile* are ready for spatial join according to the admin level. The result is '*clustered_gdf'* with a geometry column storing the coordinates of points and a new cluster column signifying the administrative division of that point. For transparency, the script prints three rows of the updated GDF. As it is shown in Table 4 the GDF has the information of each point inside the georeferenced point grid including geometry and name of the subnational division at level 1 under the column named 'cluster'; in this example, the first three points are located in Migori County in Kenya. In the case of analysis at admin level 0, the region of interest is processed as a unified unit on a national scale representing the code of country for all points, for example, in Table 5 all points are allocated to the country of Kenya showing with KEN as the country ISO code.

Table 4: Result of spatial join between GDF of point grid and administrative Shapefile at admin level 1, representing first three spatial objects in GDF with point geometry and assigned administrative region cluster at admin level 1.

No	geometry	index_right	cluster
0	POINT (33.99959 -0.94042)	26	Migori
1	POINT (33.99959 -0.85042)	26	Migori
2	POINT (34.08959 -1.03042)	26	Migori

Table 5: Result of spatial join between GDF of point grid and administrative Shapefile at admin level 0, representing first three spatial objects in GDF with point geometry and assigned administrative region cluster at admin level 0.

No	geometry	index_right	cluster
0	POINT (33.99959 -0.94042)	0	KEN
1	POINT (33.99959 -0.85042)	0	KEN
2	POINT (34.08959 -1.03042)	0	KEN

The last modification to the 'clustered_gdf' is converting the admin division names at admin level 1 or higher to the 3-letter to align with CLEWs naming format. The first three letters of subregions are extracted as the name of that subregion. Table 6 illustrates the same points in the GDF with three-letter code as cluster names; MIG represents Migori County in Kenya.

Table 6: Converting admin division names to 3-letter naming format within the GDF of points, representing first three spatial objects in GDF with point geometry and assigned administrative region cluster at admin level 1.

No	geometry	index_right	cluster
0	POINT (33.99959 -0.94042)	26	MIG
1	POINT (33.99959 -0.85042)	26	MIG

Part 3.2.2. Generating Polygons

In this phase, each point inside the ‘clustered_gdf’ is used to create an exclusive polygon. The script creates polygons using Shapely Python package that is useful for geometric operations and supports generating new objects [74]. Shapely creates buffer around points inside GDF and produces polygons. The style of the buffer is set to CapStyle3, which creates a square buffer around each point. The buffer value is set to split the distance between two neighbour points, and subsequently, generated polygons fully cover the total geographic area; the buffer value is calculated based on half the spacing value. The created polygons serve as land cells in the following parts. Table 7 demonstrates that GDF is updated, and the geometry of the same data in Table 4 is transformed from points to polygons.

Table 7: Creating land cells by generating polygons from points inside grid points located in Kenya, representing first three spatial objects in GDF with polygon geometry and assigned administrative region cluster at admin level 1.

No	geometry	index_right	cluster
0	POLYGON ((34.04459 -0.89542, 34.04459 -0.98542...	26	MIG
1	POLYGON ((34.04459 -0.80542, 34.04459 -0.89542...	26	MIG
2	POLYGON ((34.13459 -0.98542, 34.13459 -1.07542...	26	MIG

Part 3.3. Total Area Re-Estimation & Calibration

The last phase of Part 3 executes functions to resolve issues related to area misclassification and miscalculation. These issues include features that are not properly classified as well as inconsistency caused by points around the administrative boundary. Re-estimated and calibrated areas are presented in square kilometres, which is used as the constant unit of area calculation throughout GeoCLEWs process. An in-depth explanation of various steps of the validation process is presented in the following, which has an important role in the accuracy enhancement of the results.

Part 3.3.1. Area Calibration

GeoCLEWs calculate and compare the total area of generated GDF with area of border of region of interest, and if they are not equal it implements calibration function to solve the miscalculation issue. The script adopts the administrative boundary at level 0 as the correct calculation of total area, which is usually the official national border or external boundary of an arbitrary region. In order to achieve an accurate area calculation of both GDF and administrative boundary, GeoCLEWs transform both spatial datasets to proper projection system that can support accurate area measurements according to the geographic location of the region. Then, the script adds a new column to the generated GDF in previous part and calculates the area of each polygon. After that, it calculates the total area in square kilometres (sq km) and implements a double-check of the estimated figure from the reference Shapefile at admin level 0. In case of a discrepancy between the estimated and official area calculation, GeoCLEWs calibrates the generated GDF and displays a message to the user for transparency. The message clearly compares the estimated value and reference figure from the administrative boundary and finally states the calibrated total area of generated polygons, which should be equal to the reference value.

Part 3.3.2. Final Check

This part implements the final modification and reprojection to transform the GDF from the projection system to the base coordinate system. Since the GAEZ produced raster files based on WGS84 coordinate system, the script reprojects the GDF to this system to extract spatial data from GAEZ files and precisely incorporate to GDF. The output of this section is renamed to '*final_clustered_GAEZ_gdf*' and contains land cells' information, including the geometry of polygons, administrative information, and the estimated area of each polygon in sq km.

Part 3.3.3. Export as GeoPackage

Finally, the '*final_clustered_GAEZ_gdf*' is exported as GeoPackege in the Output directory. It offers an opportunity to share generated GeoDataFrame among different GIS software and platforms since GeoPackage supports various spatial layers and attributes and makes it a convenient way for raster attribute extraction. "Part 3 complete" is displayed at the conclusion of the third part.

Part 4: Geospatial Attributes Extraction to Land Cells

The functions employed in the fourth part of GeoCLEWs extract values from downloaded TIFF-formatted GAEZ raster files and assign them as attributes to the land cells based on their spatial locations. The GAEZ raster images contain values indicating various agro-climatic characteristics including precipitation, land cover, agro-climatic potential yield, crop water deficit, and evapotranspiration. This part consists of clipping GAEZ raster files with global coverage to reduce processing time and utilizing different approaches to process continuous and categorical raster files. At the end, GeoCLEWs produces GDF of land cells with geometry, area, and spatial attributes, which will be used in Part 5 to estimate land and water statistics.

- 4.1. Clipping GAEZ raster files.
- 4.2. Collecting raster files.
- 4.3. Extracting raster values.
  - 4.3.1. Processing continuous datasets.
  - 4.3.2. Processing categorical datasets.
  - 4.3.3. Converting GeoJSON file to GeoDataFrame.
- 4.4. Exporting the GeoDataFrame as a vector layer.

Part 4.1. Clipping GAEZ Raster Files

Clipping function is implemented to reduce the size of the raster images to the extent of the study area as it is highly important to minimize the processing time of all steps involved in land and water analysis. Downloaded raster files from the GAEZ v4 portal have global coverage, stored in the '*global_raster_input*' folder in TIFF format. The administrative boundary at level 0 is utilized to trim the GAEZ raster files in order to decrease the computational processing time. The metadata of each raster image after clipping is updated and then stored in a new folder named '*cropped_raster_input*'. This additional function can effectively reduce the time and manual GIS modification required for geoprocessing.

$$Mean = \frac{\sum Values\ of\ the\ selected\ pixels}{Number\ of\ delecetd\ pixels}$$

Part 4.2. Collecting Raster Files

The script collects and classifies raster files based on their specifications to perform proper processing approach in the next part. There are two types of raster files involved in this process for different purposes; continuous raster files with values that vary smoothly across the area, and categorical raster files that classify information into separate categories. The types of geospatial datasets used in this research are displayed in Table 8.

Table 8: Type of GAEZ raster files.

Type	GAEZ Raster File
Categorical	Land cover
Continues	Precipitation Agro-climatic potential yield Crop water deficit Crop evapotranspiration

GeoCLEWs creates two lists hosting continuous and discrete GAEZ raster files separately. Next, it reads all cropped images from '*cropped_raster_input'* using the Rasterio library. Following that, file names are utilized to classify TIFF files into two classes, numerical and categorical lists. In addition, the final lists are checked to remove the duplication. At the end of this step, the script displays the total number and name of the images in both numerical and categorical lists offering users a chance to review input raster files before implementing the spatial attribute extraction.

Part 4.3. Extracting Raster Values

At this point, GeoCLEWs adopts modules and functions to extract geospatial attributes from raster datasets and assigns generated values to the GDF. The continuous and categorical are analyzed individually using proper functions. Exclusive functions are defined to extract geographic attributes and calculate statistics since each type of raster has distinctive specifications. The script defines '*processeing_raster_cat*' and '*processing_raster_con*' functions for estimating categorical and continuous values respectively and associates them to land cells. Both algorithms employ the zonal _stats method to determine summary statistics of the pixel values within each land cell vector layer and then assign the result of estimation to the corresponding land cell as a new attribute. The performance of each function is presented below. At the end of this part, the GDF will be updated including land cells with new attributes collected from raster values.

Part 4.3.1. Processing Continuous Datasets

GeoCLEWs employs the '*processing_raster_con*' function to extract pixel values from TIFF images with numerical format and estimate mean values. The '*processing_raster_con'* function analyzes continuous TIFF images and uses spatial relationships to identify the raster cells that intersect with each corresponding land cell in vector layers. The script uses zonal_stats method to analyze and process raster values intersect with vector polygons; it clips the raster images using the polygons and detects pixels inside each polygon's boundary to calculate statistics and assign the new attribute to that specific polygon. GeoCLEWs calculates a summary of statistics including mean, minimum and maximum to provide a comprehensive analysis while the mean value is the only required measurement for CLEWs modelling. For continuous raster files, the mean value is determined by adding up all the pixel values inside a single land cell' polygon and dividing that total by the overall number of pixels in the corresponding polygon, details are presented in Equation 1.

Equation 1: Continuous raster files calculation.

$$Mean = \frac{\sum Pixel\ Values\ of\ selected\ pixels}{Number\ of\ selected\ Pixels}$$

Following that, the calculated mean values are allocated to the corresponding land cells in GDF. Date, time, and completion status are printed after each individual raster file has been processed.

Part 4.3.2. Processing Categorical Datasets

Similar to the prior step, the discrete images are processed by defining and calling the '*processing_raster_cat*' function, which estimates the files inside the categorical list and defines the LCTypes within each land cell. The '*processeing_raster_cat*' function assesses the categorical raster images and defines type of land cover in each cell. It employs the zonal_stats function to mask the raster cells inside each single polygon. Then it assesses the spatial information of cells individually. Land cover is the only discrete raster file required for CLEWs modelling and this function assists in figuring out the LCType of each pixel within an individual polygon. The output of this operation is the total number of pixels belonging to each LCType inside of an individual land cell. At the end, the script displays the name and completion status of processed raster files.

Part 4.3.3. Converting the GeoJSON file to GeoDataFrame

For proceeding to the next part of detailed land and water assessment, outputs from last two parts are converted to GDF due to its flexibility and functionality for geospatial data processing. Hence, a function is defined, named 'geojson_to_gdf', to generate a GDF from an input GeoJSON file. This function is converting the output of '*processing_raster_cat*' and '*processing_raster_con'*, which is GeoJSON file, to a GDF for further processing in next parts.

Part 4.4. Exporting GeoDataFrame as Vector Layer

Part 4.4 exports the generated GDF to CSV and GeoPackage formats with detailed information on all data points in the country. Files are stored in the Output directory and present the name of the administrative boundary of land cells according to the admin level, for example at admin level 1 in the Kenya case study the CSV file shows the name of the county that each cluster is located in. In addition, it contains the total area of each land cell, which is about 100 km^2 because of default spacing of base point grid. Furthermore, processed spatial information such as the mean value of continuous raster data and the area of LCTypes in individual clusters are delivered along with the geometry of the associated polygon. After processing and calculating the necessary data, it is possible to begin the process of statistical land and water computation in Part 5 and generate final outputs for the CLEWs modelling.

Part 5: Statistics Calculation

This part processes the spatial attributes and implements regional aggregation to calculate statistics and generate final results with a clewsy-compatible structure and format. Up to this part, agro-climatic analysis is delivered at the land cell level. GeoCLEWs not only delivers calculated statistics on a national scale but also provides detailed regional analysis based on the user-defined level of administrative division. It calculates continuous and categorical spatial values separately. GeoCLEWs generates tabular outputs in CSV format as well as interactive graphs. Unit of results are adjusted to be compatible with CLEWs modelling, Table 9 represents adjusted units of final results.

GeoCLEWs output can be efficiently combined with additional data for CLEWs modelling, such as electricity information, to create a detailed CLEWs model without implementing complicated and time-consuming spatial processing.

Table 9: GeoCLEWs output units

Agro-Climatic Statistics	**GeoCLEWs Output Unit**
Agro-climatic potential yield	Million tonnes per 1000 km^2
Crop water deficit	Billion cubic meters 1000 km^2.
Crop evapotranspiration	Billion cubic meters 1000 km^2
Precipitation	Billion cubic meters 1000 km^2
Land cover	km^2

In this thesis the regional aggregation method is utilized to preserve essential information and reduce the complexity of the high-resolution spatial data processing. GeoCLEWs groups land cells into clusters, which represent the administrative regions. The script produces statistics on an administrative regional scale using information of land cells inside each cluster. Regional aggregation offers an opportunity to group the region clusters into new aggregated ones and reduce the number of total regions resulting in optimizing computational processing within CLEWs modelling. Below various steps involved in Part 5 are delineated:

- 5.1. National summary statistics.
    - 5.1.1. Collecting names of attributes assigned to land cells.
    - 5.1.2. Land cover and area statistics.
    - 5.1.3. Other variable statistics.
    - 5.1.4. Averaging additional crops.
    - 5.1.5. Exporting national statistics.
- 5.2. Calculating region summaries.
    - 5.2.1. Grouping land cells.
    - 5.2.2. Land cover and area statistics.

    - 5.2.3. Region for land cover GDF.
    - 5.2.4. Other variable statistics.
    - 5.2.5. Averaging additional crops.
    - 5.2.6. Aggregating regions for other variables GDF.
- 5.3. Generate clewsy-compatible statistics.
- 5.4. Generate interactive graphs.

Part 5.1. National Summary Statistics

National statistics are generated by processing continues and categorical spatial attributes of all land cells inside the region of study. First names of spatial attributes are collected to create two separate lists: Land cover values, and remaining agro-ecological values that is named other variables. Then the total area of each LCType is calculated and displayed. After that other variables' statistics are estimated. GeoCLEWs collects other variables of additional crops and implement averaging additional crops process to incorporate the average value of five extra crops into assessment. Finally, national statistics are exported in CSV format.

Part 5.1.1. Collecting Names of Attributes Assigned to Land Cells

Continues and categorical values require exclusive processing approach for statistical calculation. To begin, two separate lists are created to collect names of all attributes extracted from the GAEZ datasets and remove unnecessary features from GDF such as indexes. The generated lists are used to filter GDF in the following parts to calculate land cover and other variables separately.

Part 5.1.2. Land Cover and Area Statistics

The script evaluates the categorical attributes and specifies details on all LCTypes on a national scale. As stated previously, GAEZ classifies images into eleven land cover classes. A new function, named '*cal_LC_sqkm*', is defined to calculate the amount of land covered by each LCType using the number of pixels. The '*cal_LC_sqkm*' determines the total area of every land cover type for the entire region of interest. Equation 2 represents more details on categorical raster data calculation, it gives an example of how to compute the area covered by LCType1 in Region A.

Equation 2: Processing categorical raster files.

$$\rho = \sum Number\ of\ pixels\ in\ Region\ A\ that\ belongs\ to\ LCType1$$

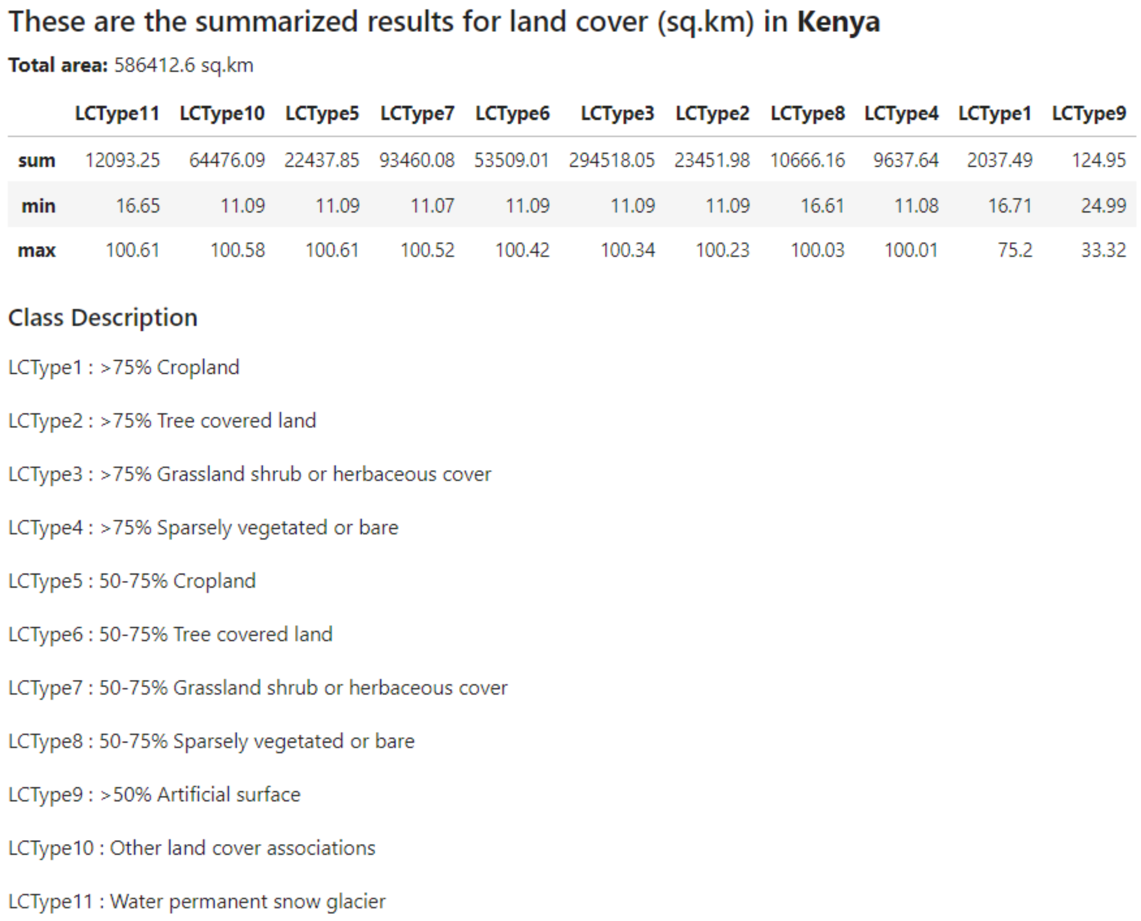
$$\tau = \sum Number\ of\ total\ pixels\ belonging\ to\ Region\ A$$

$$Ratio\ of\ LCType1\ in\ Region\ A:\ \alpha = \frac{\rho}{\tau}$$

$$Total\ area\ of\ Region\ A = \ \beta$$

$$Total\ area\ covered\ with\ LCType1\ in\ Region\ A\ = \alpha \times\ \beta$$

The sum, min and max estimations of the area belonging to each LCType are calculated and presented in square kilometres. The total amount of land in the region that is covered with a certain LCType is shown by the sum figure. The min and max figures represent the minimum and maximum area covered by a particular LCType in a single polygon; the zero value is excluded from the minimum value presentation. The sum of the area is the only figure that used for CLEWs modelling and additional statistics offer users comprehensive analysis. In the end, the script displays the national summary of land cover with an in-depth class description. Figure 7 is an example of the national-scale presentation of land cover statistics from project of Kenya.



These are the summarized results for land cover (sq.km) in **Kenya**

Total area: 586412.6 sq.km

	LCType11	LCType10	LCType5	LCType7	LCType6	LCType3	LCType2	LCType8	LCType4	LCType1	LCType9
sum	12093.25	64476.09	22437.85	93460.08	53509.01	294518.05	23451.98	10666.16	9637.64	2037.49	124.95
min	16.65	11.09	11.09	11.07	11.09	11.09	11.09	16.61	11.08	16.71	24.99
max	100.61	100.58	100.61	100.52	100.42	100.34	100.23	100.03	100.01	75.2	33.32

Class Description

LCType1 : >75% Cropland

LCType2 : >75% Tree covered land

LCType3 : >75% Grassland shrub or herbaceous cover

LCType4 : >75% Sparsely vegetated or bare

LCType5 : 50-75% Cropland

LCType6 : 50-75% Tree covered land

LCType7 : 50-75% Grassland shrub or herbaceous cover

LCType8 : 50-75% Sparsely vegetated or bare

LCType9 : >50% Artificial surface

LCType10 : Other land cover associations

LCType11 : Water permanent snow glacier

Figure 7: National statistics of land cover in Kenya, visually represented by GeoCLEWs in Part 5.1.2.

Part 5.1.3. Other Variable Statistics

In this thesis, other variables refer to spatial attributes extracted from continuous GAEZ raster data, encompassing crop potential yield, crop water deficit, crop evapotranspiration, and precipitation. GeoCLEWs implements unit adjustment and calculates statistics for other variables of top ten crops on a national level. This process considers the spatial measurements of all land cells to estimate the mean, maximum, and minimum agro-climatic parameters for each crop individually along with precipitation estimation. Before calculating these statistics, the script adjusts units of measurements to make them compatible with CLEWS modelling, the unit of output presented in Table 9.

Part 5.1.4. Averaging additional crops

To incorporate the agro-climatic measurements of five additional crops into the land and water assessment, GeoCLEWs an averaging process on these values and adds them to the GDF as one extra crop while statistics of five main crops remain unchanged. One of the value-added features of GeoCLEWs is considering spatial attributes of five extra crops, listed as additional crops in Part 2. Processing and adding averaged statistics as a new extra crop results in preserving valuable data and avoiding unnecessary complexity. The script estimates the mean, maximum, and minimum of other variable statistics of averaged crops and adds as a single extra crop to the national statistics dataset, which contains main crops' parameters and precipitation. The new averaged estimations are considered the sixth crop and named 'Other Crop' which is then converted to the 3-letter 'OTH' to maintain consistency with CLEWs naming standard. At the end of this part, the other variable statistics are displayed with detailed information on the unit of produced results and the reason for unit conversion to improve the transparency of the code.

Table 10 partially illustrates the summarized other variable results in the script, where *OTH yld Rain-fed Low_mean*' refers to statistics for averaged additional crops' agro-climatic potential yield attribute which is rain-fed with low agricultural management level. More information is provided in the following to clarify GeoCLEWs' naming format. In addition, detailed spatial attribute naming guideline is provided in Table 11, and GeoCLEWs crop naming based on CLEWs guideline is presented in Appendix C.

- The first term refers to the crop name, which is converted into 3-letter. For instance, maize and sorghum turned into 'MZE', and 'SOR' respectively.
- The second term demonstrates the agro-climatic attributes such as 'cwd' for crop water deficit.
- The third word provides information about the watering system. It can be 'Rain-fed' and 'Irrigation', using natural rainfall and artificial irrigation systems respectively.
- The last phrase signifies the input level. 'High' is used for advanced agriculture methods and technologies and 'Low' specifies traditional farming practices.

Table 10: **National statistics of other variables in Kenya including additional crops, this is a partial representation of the results due to space limitation.**

Type	BEA cwd Irrigation High_mean	TEA yld Irrigation Low_mean	MZE evt Irrigation Low_mean	MZE cwd Irrigation High_mean	OTH yld Rain-fed Low_mean	OTH yld Rain-fed High_mean
mean	0.4143	0.0079	0.28	0.4374	0.0474	0.1027
min	0	0	0	0	0	0
max	0.7792	0.0522	0.8197	0.907	0.4437	1.6847

Table 11: **GeoCLEWs agro-climatic attribute naming guideline.**

Abbreviation	Agro-climatic Attribute
prc	Precipitation
yld	Agro_climatic Potential Yield
cwd	Crop Water Deficit
evt	Crop Evapotranspiration
Rain-fed	Natural Rainfall
Irrigation	Artificial Irrigation systems
Low	Traditional Agriculture Management
High	Advanced Farming Systems

Part 5.1.5 Exporting National Statistics

The summary of national statistics is produced in CSV format for general analysis. The produced land cover statistics are stored in a file with the format of '*(country code)_LandCover_National_summary.csv*', and other variable estimations are saved as '*(country code)_Parameter_National_summary.csv*'. Using the country code in the process of exporting results offers a chance to easily organize results from different projects and enhance the efficiency of project management using GeoCLEWs.

Part 5.2. Calculating Region Summaries

The area of study is analyzed based on geographical boundaries of administrative regions and generates regional statistical summaries. The region are created based on user-defined admin level and named cluster. For instance, regarding Kenya, GADM admin level 1 provides geographical boundary of 47 counties, and subsequently, 47 clusters are created in total including county's geospatial characteristics. Each land cell belongs to one admin region, which is represented in the cluster column inside GDF. The script rechecks the cluster status of each polygon before implementing aggregation to detect and display land cells that are not assigned to a proper admin division. Any unclassified land cells will be labelled as ‘None’. If it finds any land cells with “None” status, user will receive a warning message to recheck the GDF or admin Shapefile to solve the problem and assign corresponding cluster status to land cells. This step is beneficial for eliminating unexpected errors and improving accuracy while there were no unclassified land cells in any of the projects that have been completed by GeoCLEWs so far thanks to precise classification and accurate data sources.

Following that, all land cells within similar administrative region are grouped in the same cluster. GeoCLEWs adopts almost the same functions and approaches utilized in Part 5.1 to estimate agro-climatic statistics for each administrative region separately. It generates two GDF for processing land cover and other variables using exclusive functions. If user decided to aggregate regions to reduce computational complexity while developing CLEWs model, the script aggregates them into new aggregated clusters. Finally, it produces the statistics of newly created clusters as final output with clewsy compatible format. Detailed technical explanation and examples are provided in the following parts.

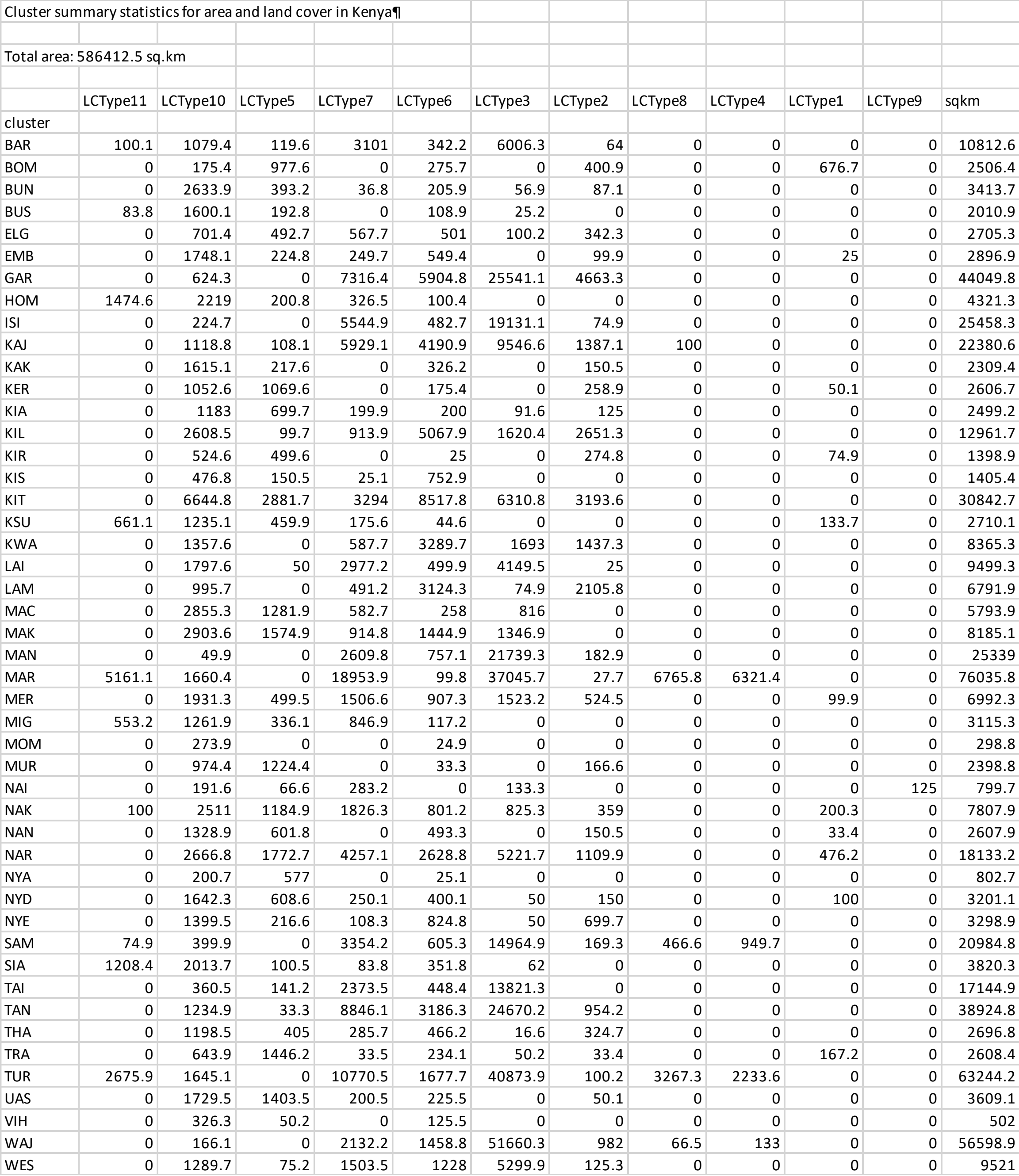
Part 5.2.1. Grouping Land Cells

Land cells are grouped according to the administrative region in which they are located. All land cells belonging to a similar region are assigned to the same cluster. Therefore, land and water statistics of each region are estimated by considering spatial attributes of land cells inside of its boundary.

Part 5.2.2. Land Cover and Area Statistics

The land cover statistics represent the type and area estimation of land-use inside of each individual admin region. The script executes grouped land cells and calculates the cumulative LCTypes area inside each region. As a result, a GDF is created named ‘cluster_lc’, which stores clusters’ land cover statistics. For accuracy validation, GeoCLEWs compares the sum of area of all generated clusters and the area of the entire region of study; two estimated values should be equal to ensure the accuracy of the regional aggregation. The Script displays the land cover summary and area calculation of individual admin regions supplying exhaustive information for user comprehension. An example is displayed in Table 12 which presents admin region statistics of land cover for Kenya. As it is shown, 11 land cover types and associated total area are estimated. The results refer to geoprocessing at admin level 1 in the Kenya case study, and consequently, there are 47 regions linked with the 47 counties in this country. The total area of Kenya is calculated by cumulating the total area of clusters, which is provided in column ‘sqkm’ referring area in square kilometres. By comparing the obtained value and total area of administrative boundary, it is obvious that GeoCLEWs precisely computed clusters’ statistics without missing land cell as well as spatial attributes. Users can easily validate the accuracy and transparency of the developed script.

Table 12: Area and land cover statistics of administrative regions in Kenya, visually represented by GeoCLEWs in Part 5.2.2. of GeoCLEWs.



Cluster summary statistics for area and land cover in Kenya¶

Total area: 586412.5 sq.km

	LCType11	LCType10	LCType5	LCType7	LCType6	LCType3	LCType2	LCType8	LCType4	LCType1	LCType9	sqkm
cluster												
BAR	100.1	1079.4	119.6	3101	342.2	6006.3	64	0	0	0	0	10812.6
BOM	0	175.4	977.6	0	275.7	0	400.9	0	0	676.7	0	2506.4
BUN	0	2633.9	393.2	36.8	205.9	56.9	87.1	0	0	0	0	3413.7
BUS	83.8	1600.1	192.8	0	108.9	25.2	0	0	0	0	0	2010.9
ELG	0	701.4	492.7	567.7	501	100.2	342.3	0	0	0	0	2705.3
EMB	0	1748.1	224.8	249.7	549.4	0	99.9	0	0	25	0	2896.9
GAR	0	624.3	0	7316.4	5904.8	25541.1	4663.3	0	0	0	0	44049.8
HOM	1474.6	2219	200.8	326.5	100.4	0	0	0	0	0	0	4321.3
ISI	0	224.7	0	5544.9	482.7	19131.1	74.9	0	0	0	0	25458.3
KAJ	0	1118.8	108.1	5929.1	4190.9	9546.6	1387.1	100	0	0	0	22380.6
KAK	0	1615.1	217.6	0	326.2	0	150.5	0	0	0	0	2309.4
KER	0	1052.6	1069.6	0	175.4	0	258.9	0	0	50.1	0	2606.7
KIA	0	1183	699.7	199.9	200	91.6	125	0	0	0	0	2499.2
KIL	0	2608.5	99.7	913.9	5067.9	1620.4	2651.3	0	0	0	0	12961.7
KIR	0	524.6	499.6	0	25	0	274.8	0	0	74.9	0	1398.9
KIS	0	476.8	150.5	25.1	752.9	0	0	0	0	0	0	1405.4
KIT	0	6644.8	2881.7	3294	8517.8	6310.8	3193.6	0	0	0	0	30842.7
KSU	661.1	1235.1	459.9	175.6	44.6	0	0	0	0	133.7	0	2710.1
KWA	0	1357.6	0	587.7	3289.7	1693	1437.3	0	0	0	0	8365.3
LAI	0	1797.6	50	2977.2	499.9	4149.5	25	0	0	0	0	9499.3
LAM	0	995.7	0	491.2	3124.3	74.9	2105.8	0	0	0	0	6791.9
MAC	0	2855.3	1281.9	582.7	258	816	0	0	0	0	0	5793.9
MAK	0	2903.6	1574.9	914.8	1444.9	1346.9	0	0	0	0	0	8185.1
MAN	0	49.9	0	2609.8	757.1	21739.3	182.9	0	0	0	0	25339
MAR	5161.1	1660.4	0	18953.9	99.8	37045.7	27.7	6765.8	6321.4	0	0	76035.8
MER	0	1931.3	499.5	1506.6	907.3	1523.2	524.5	0	0	99.9	0	6992.3
MIG	553.2	1261.9	336.1	846.9	117.2	0	0	0	0	0	0	3115.3
MOM	0	273.9	0	0	24.9	0	0	0	0	0	0	298.8
MUR	0	974.4	1224.4	0	33.3	0	166.6	0	0	0	0	2398.8
NAI	0	191.6	66.6	283.2	0	133.3	0	0	0	0	125	799.7
NAK	100	2511	1184.9	1826.3	801.2	825.3	359	0	0	200.3	0	7807.9
NAN	0	1328.9	601.8	0	493.3	0	150.5	0	0	33.4	0	2607.9
NAR	0	2666.8	1772.7	4257.1	2628.8	5221.7	1109.9	0	0	476.2	0	18133.2
NYA	0	200.7	577	0	25.1	0	0	0	0	0	0	802.7
NYD	0	1642.3	608.6	250.1	400.1	50	150	0	0	100	0	3201.1
NYE	0	1399.5	216.6	108.3	824.8	50	699.7	0	0	0	0	3298.9
SAM	74.9	399.9	0	3354.2	605.3	14964.9	169.3	466.6	949.7	0	0	20984.8
SIA	1208.4	2013.7	100.5	83.8	351.8	62	0	0	0	0	0	3820.3
TAI	0	360.5	141.2	2373.5	448.4	13821.3	0	0	0	0	0	17144.9
TAN	0	1234.9	33.3	8846.1	3186.3	24670.2	954.2	0	0	0	0	38924.8
THA	0	1198.5	405	285.7	466.2	16.6	324.7	0	0	0	0	2696.8
TRA	0	643.9	1446.2	33.5	234.1	50.2	33.4	0	0	167.2	0	2608.4
TUR	2675.9	1645.1	0	10770.5	1677.7	40873.9	100.2	3267.3	2233.6	0	0	63244.2
UAS	0	1729.5	1403.5	200.5	225.5	0	50.1	0	0	0	0	3609.1
VIH	0	326.3	50.2	0	125.5	0	0	0	0	0	0	502
WAJ	0	166.1	0	2132.2	1458.8	51660.3	982	66.5	133	0	0	56598.9
WES	0	1289.7	75.2	1503.5	1228	5299.9	125.3	0	0	0	0	9521

Part 5.2.3. Aggregating Regions of Land Cover GDF

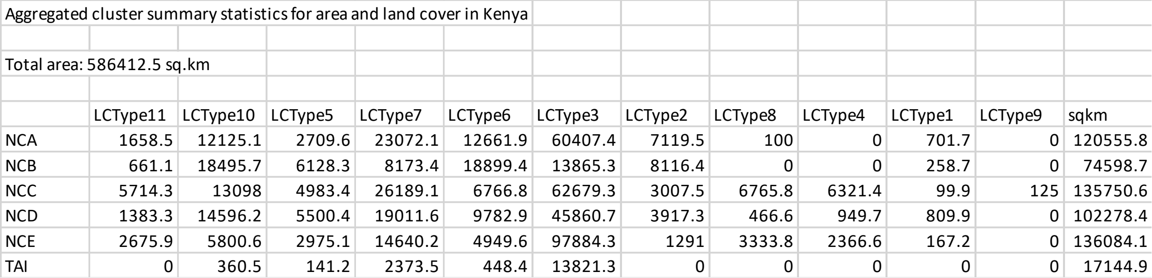
Aggregation is a new enhanced feature of GeoCLEWs that is highly important for reducing processing time during the developing CLEWs model. The computational complexity of the model can be negatively impacted by a large number of regions, reducing the number of clusters improves the efficiency of modelling. GeoCLEWs supports an option for aggregating admin regions; users can decide on performing aggregation during the configuration phase. They also have an opportunity to select the number of final clusters, which are created from grouping admin regions. In addition, excluding a specific region from the aggregation process is feasible, which facilitates a detailed assessment of a particular subnational region.

According to the user-defined number of final clusters, GeoCLEWs implements regional aggregation and group the administrative regions into new aggregated region clusters based on the alphabet order of their name. Since agro-ecological zones vary across the geographical area, there is no assurance that employing a neighbouring method for aggregating regions guarantees the grouping of adjacent regions with similar agro-ecological characteristics. Additionally, these features are not transparent to users, hindering the implementation of supervised aggregation and the informed decision-making on which regions should be grouped into the same new aggregated cluster. Therefore, GeoCLEWs v1.0.0. implements regional aggregation according to the alphabetical order of their names. However, as future work, it is highly recommended to employ spatial clustering based on agro-climatic potential yield criteria to aggregate and classify regions based on their soil suitability similarities.

After implementing regional aggregation to the land cover GDF, the script calculates the land cover statistics of new clusters considering the spatial attributes of associated regions. The aggregated cluster summary for area and land cover is generated and printed for user understanding, which is exported as a CSV file named *'(country code)_LandCover_byCluster_summary.csv'* including detailed LCTypes estimations of the final clusters.

Table 13 illustrates land cover statistics from the Kenya case study after aggregation including statistics of five clusters of aggregated regions and one excluded region named TAI. 46 admin regions turned into 5 new clusters with new 3-letter names: NCA, NCB, NCC, NCD, and NCE. Taite Taveta County, TAI, is excluded from the combination process according to user customization. It is not included in any of the newly created classes, and TAI is added to the new dataset as an individual cluster.

Table 13: Area and land cover statistics of aggregated regions in Kenya, visually represented by GeoCLEWs in Part 5.2.3.



Aggregated cluster summary statistics for area and land cover in Kenya

Total area: 586412.5 sq.km

	LCType11	LCType10	LCType5	LCType7	LCType6	LCType3	LCType2	LCType8	LCType4	LCType1	LCType9	sqkm
NCA	1658.5	12125.1	2709.6	23072.1	12661.9	60407.4	7119.5	100	0	701.7	0	120555.8
NCB	661.1	18495.7	6128.3	8173.4	18899.4	13865.3	8116.4	0	0	258.7	0	74598.7
NCC	5714.3	13098	4983.4	26189.1	6766.8	62679.3	3007.5	6765.8	6321.4	99.9	125	135750.6
NCD	1383.3	14596.2	5500.4	19011.6	9782.9	45860.7	3917.3	466.6	949.7	809.9	0	102278.4
NCE	2675.9	5800.6	2975.1	14640.2	4949.6	97884.3	1291	3333.8	2366.6	167.2	0	136084.1
TAI	0	360.5	141.2	2373.5	448.4	13821.3	0	0	0	0	0	17144.9

Part 5.2.4. Other Variables Statistics

Similar to national scale estimation, the script employs functions to process GDF of continues parameters to compute the mean value of potential yield, water deficit, evapotranspiration, and precipitation with CLEWS compatible unit. It creates a GDF including other variable attributes inside each admin region and utilizes functions to estimate statistics of all admin regions according to user-defined admin level.

Part 5.2.5. Averaging Additional Crops

This part utilizes the similar approach in Part 5.1.4, which collects the spatial values of additional crops and computes the mean value of other variables within each cluster. The attributes of additional crops are removed from cluster statistics and saved in a separate dataset. The average value is generated as Other Crop statistics, which is renamed to 'OTH' as the sixth crop for land and water geoprocessing. Then, this extra information is added to the cluster dataset including the main five crop measurements. The script displays the generated statistics that contain the original region clusters. Table 14 partially represents cluster statistics of other variables after adding the extra crop information as 'OTH'. It shows the mean value of potential yield, crop water deficit, and evapotranspiration classified into rain-fed and irrigation watering systems as well as high and low levels of agriculture practices. Output of this part is a GDF including other variable statistics of all admin regions.

Table 14: Statistics of other variables for administrative regions in Kenya. This is a partial representation due to space limitations.

Region	BEA cwd Irrigation High_mean	TEA yld Irrigation Low_mean	MZE evt Irrigation Low_mean	OTH yld Rain-fed Low_mean
BAR	0.0418	0.0019	0.4136	0.092
BOM	0.0045	0.0114	0.6782	0.1242

Region	BEA cwd Irrigation High_mean	TEA yld Irrigation Low_mean	MZE evt Irrigation Low_mean	OTH yld Rain-fed Low_mean
BUN	0.0342	0.0041	0.4777	0.1066
BUS	0.0099	0.0169	0.4731	0.0695
EMB	0.0565	0.0183	0.4414	0.0871
GAR	0.648	0	0.1912	0.0149
HOM	0.0056	0.027	0.4542	0.0793
ISI	0.6298	0	0.1082	0.0238
KAJ	0.3382	0.0119	0.3547	0.0938
KAK	0.0044	0	0.4899	0.0735
KER	0.0017	0.0127	0.5847	0.1092
KIA	0.073	0.0386	0.5244	0.1403
KIL	0.3286	0.0083	0.4502	0.048
KIR	0.0486	0.0261	0.4573	0.1155
KIS	0.001	0.041	0.5861	0.1011
KIT	0.3189	0.0043	0.3831	0.0464
KSU	0.001	0.0004	0.4731	0.0656
KWA	0.2625	0.0273	0.4596	0.0503
LAI	0.1842	0.0071	0.5097	0.145
LAM	0.3311	0	0.4643	0.0481
MAC	0.1024	0.0188	0.4308	0.1091
MAK	0.2705	0.012	0.3989	0.0811
MAN	0.5835	0.0018	0.1226	0.0076
MAR	0.5595	0.0118	0.1651	0.0219
MER	0.2286	0.0213	0.3944	0.0918
MIG	0.0478	0.0428	0.4558	0.0876
MOM	0.2229	0	0.4539	0.0517
MUR	0.0447	0.0345	0.5111	0.1224
NAI	0.1226	0.0396	0.4814	0.1363
NAK	0.0618	0.0115	0.5323	0.154
NAN	0.0109	0.0014	0.5877	0.1084
NAR	0.1915	0.0164	0.5328	0.1417
NYA	0.001	0.0341	0.6293	0.1017
NYD	0.0505	0.0199	0.3384	0.1507
NYE	0.0199	0.0294	0.4367	0.1274
SAM	0.384	0.0016	0.2227	0.0534
SIA	0.0046	0.0323	0.4553	0.0734
TAI	0.4272	0.033	0.3922	0.0577
TAN	0.5691	0	0.2756	0.0267
THA	0.0972	0.0148	0.4391	0.0744
TRA	0.0703	0.0006	0.5155	0.1437
TUR	0.4293	0.0002	0.1644	0.0152
UAS	0.0143	0.0121	0.5625	0.137
VIH	0.001	0	0.4823	0.0721
WAJ	0.6524	0.0057	0.1331	0.005
WES	0.0796	0.0031	0.3534	0.0981

Part 5.2.6. Aggregating Regions of Other Variable GDF

The methodology of aggregating admin regions inside the other variable GDF is similar to the aggregating land cover GDF. Since land cover and other variable attributes have two distinct GDFs, the aggregation process is implemented in each GDF separately and yields identical new clusters. The script aggregates admin regions into new groups based on user customization. New clusters have the same name and members as clusters created during land cover aggregation. Identical aggregated clusters lead to preserving the consistency and reliability of final results. The aggregated output of other variable statistics from the Kenya case study is demonstrated in Table 15 which includes five new clusters and TAI region as the excluded one from the combination.

Table 15: Other variables statistics of aggregated regions in Kenya. This is a partial representation of agro-climatic characteristics due to space limitations.

Region	BEA cwd Irrigation High_mean	TEA yld Irrigation Low_mean	MZE evt Irrigation Low_mean	MZE cwd Irrigation High_mean
NCA	0.1793	0.0109	0.4049	0.2335
NCB	0.1224	0.0166	0.4918	0.1342
NCC	0.2514	0.0183	0.3878	0.2809
NCD	0.139	0.0161	0.445	0.1708
NCE	0.2078	0.0036	0.3685	0.3119
TAI	0.4272	0.033	0.3922	0.2855

Part 5.3. Generate clewsy-compatible Statistics

GeoCLEWs generates final results of land and water statistics, which serve as the input data for the clewsy tool. GeoCLEWs generates the other variable statistics for each cluster exclusively including potential yield, crop water deficit, and evapotranspiration statistics of five main crops plus one additional crop as well as precipitation information. Exclusive agro-climatic characteristics of each cluster must be generated in a separate CSV file for CLEWs modelling, sample of a tabular results is shown in Table 16. Required modifications to achieve clewsy-compatible outputs consist of the following steps:

1. **Generating exclusive CSV output**: New datasets are created including a specific cluster along with its individual other variable characteristic. As an example, crop water deficit assessment of cluster NCA is extracted and stored in a new dataset. Hence, the number of generated new datasets is equal to the number of final clusters multiplied by 4 because each cluster has a separate

dataset for crop potential yield, water deficit, evapotranspiration, and precipitation. In the Kenya case study, regions are aggregated into 6 clusters and there is a total of 24 final results.

2. **Name correction**: Spatial attributes inside newly generated datasets require name adjustments. As shown in. Table 14 and Table 15, the name of statistics is composed of the abbreviation of agro-climatic attributes including 'yld', 'cwd', 'evt', and 'prc' referring to potential yield, crop water deficit, crop evapotranspiration, and precipitation respectively. These acronyms plus the '_mean' need to be removed from the column heading of every attribute inside each dataset. For example, the name of the water deficit of crop bean with irrigation watering system and high level of land management is renamed from '*BEA cwd Irrigation High_mean*' to '*BEA Irrigation High*'. clewsy is unable to read any data with different naming patterns.

3. **Adding a new column**: A new column is created in each dataset to align with the clewsy framework. The name of the newly added column is changed to 'cluster' with a value of 1.

4. **CSV file rearrangement**: because of special clewsy configuration, it is required to insert 9 empty columns after the first column and then start adding agro-climatic attributes.

5. **Exporting CSV files**: The final corrected datasets are exported in CSV format inside the *summary_stats* folder, Table 16 represents an example of modified CSV result.

Table 16: Final results of crop water deficit of cluster NCA named cluster_result_cwd_NCA.

cluster										BEA Irrigation High	MZE Irrigation High	TEA Irrigation High	MZE Rain-fed Low	BEA Rain-fed High
1										0.1793	0.2335	0.0307	0.0194	0.0305

Part 5.4. Generating Interactive Graphs

Interactive graphs are generated to achieve an in-depth and comprehensive analysis while essential statistics for CLEWs modelling have been generated in the prior step. These graphs provide details statistics for each of the original clusters before aggregation offering an opportunity to improve user comprehension and have an overview of land and water spatial characteristics. The graphs present values in standardized units, which are BCM per 1000 km^2, million tonnes per 1000 km^2, and square kilometre for precipitation, potential yield, and land cover respectively. Figure 8 illustrates examples of interactive graphs from the Kenya case study. The graphs are interactive, and the value of each bar can be visualized on the screen by moving the cursor. The upper image illustrates the value of the potential yield of maize with an artificial irrigation system and high-level agriculture management, figures are presented in million tonnes per 1000 km2. The lower image displays the total area of land cover type 6, which is 50-75% grassland shrub or herbaceous cover, calculated in km2.

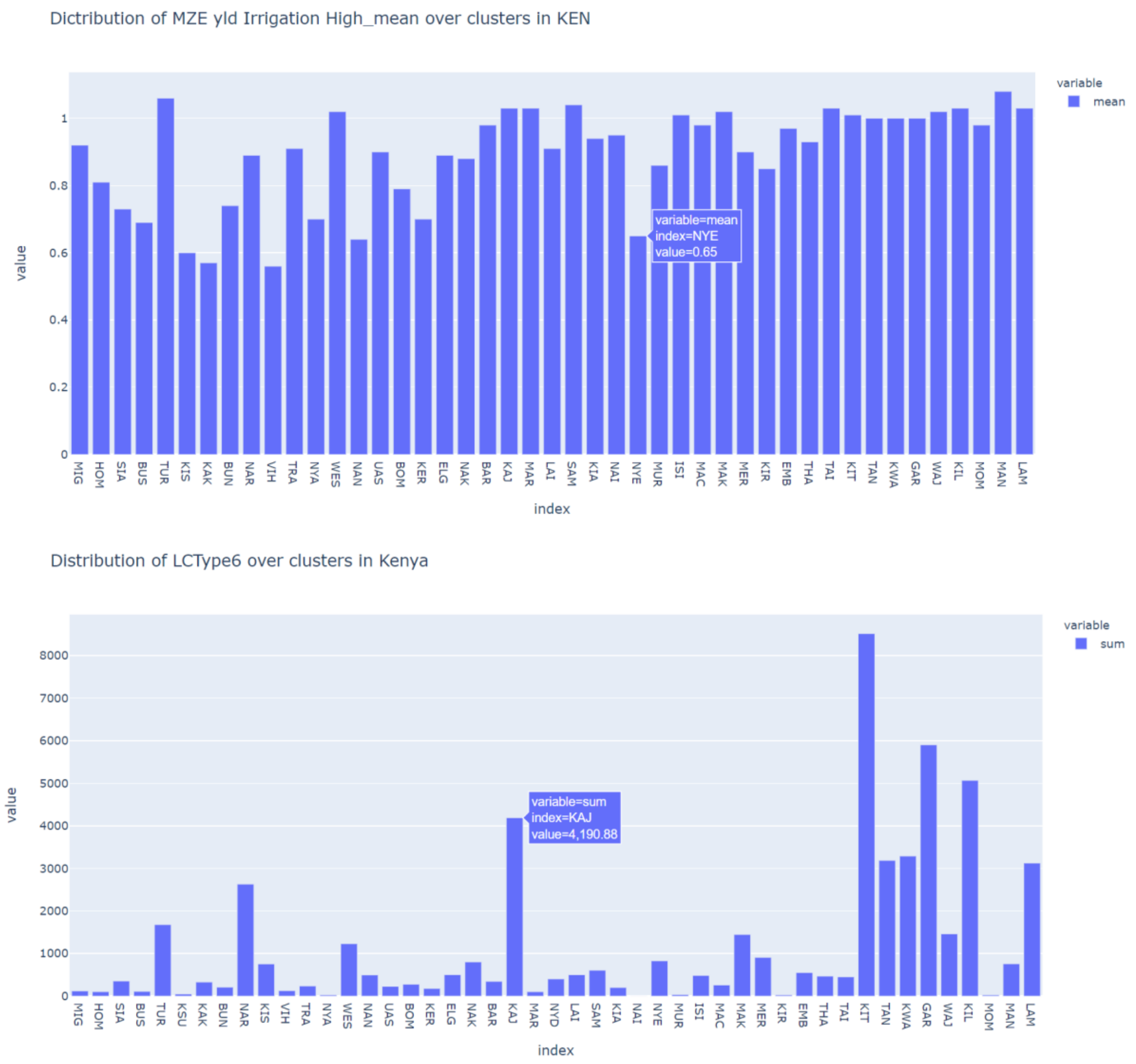


Figure 8: Interactive graphs from the Kenya case study.

Summary of GeoCLEWs Outputs

Figure 9 illustrates detailed land and water statistics for individual region generated by GeoCLEWs that enhances developing land and water systems within CLEWs framework. The following provides a summary of the results produced by GeoCLEWs:

- ***(country code)__LandCover_National_summary***: All LCTypes assessments on a national scale with the sum, minimum, and maximum statistics are provided in a CSV document, Figure 7.

- ***(country code)__LandCover_byCluster_summary***: It is a single CSV file including information coverage of all LCTypes inside of final clusters. It maintains data about aggregated regions in the case of aggregation. Country code refers to the three-letter ISO code of the country e.g. KEN for Kenya, Table 13.

- ***(country code)__Parameter_National_summary***: The mean, minimum, and maximum calculation of precipitation, crop water deficit, crop evapotranspiration, and potential yield at the national level are stored in a single CSV file, Table 10

- ***(country code)_Parameter_byCluster_summary***: This document refers to a single CSV file comprising statistical information on all crops' attributes including water deficit, evapotranspiration, and potential yield values as well as precipitation figures relating to all final clusters, Table 15.

- ***clustering_results_cwd_(cluster code)***: Every final cluster has an individual CSV file including crop water deficit related to five main crops as well as the average value of additional crops. The three-letter code of each cluster is utilized as *cluster code* for naming, Table 16.

- ***clustering_results_evt_(cluster code)***: The crop evapotranspiration statistics of the five main crops and one extra crop are included in a separate CSV file for each final cluster.

- ***clustering_results_(cluster code)***: Results of agro-climatic potential yield statistics for each individual final cluster are provided in separate CSV files. The term 'yld' is removed from the files' name based on clewsy requirements.

- ***clustering_results_prc_(cluster code)****:* Precipitation estimation of every final cluster is stored in CSV documents individually.

- **Interactive graphs**: Multiple interactive graphs representing detailed statistics of all agro-climatic attributes of original clusters before aggregation. Each LCType and every crop have an individual interactive bar chart, Figure 8.

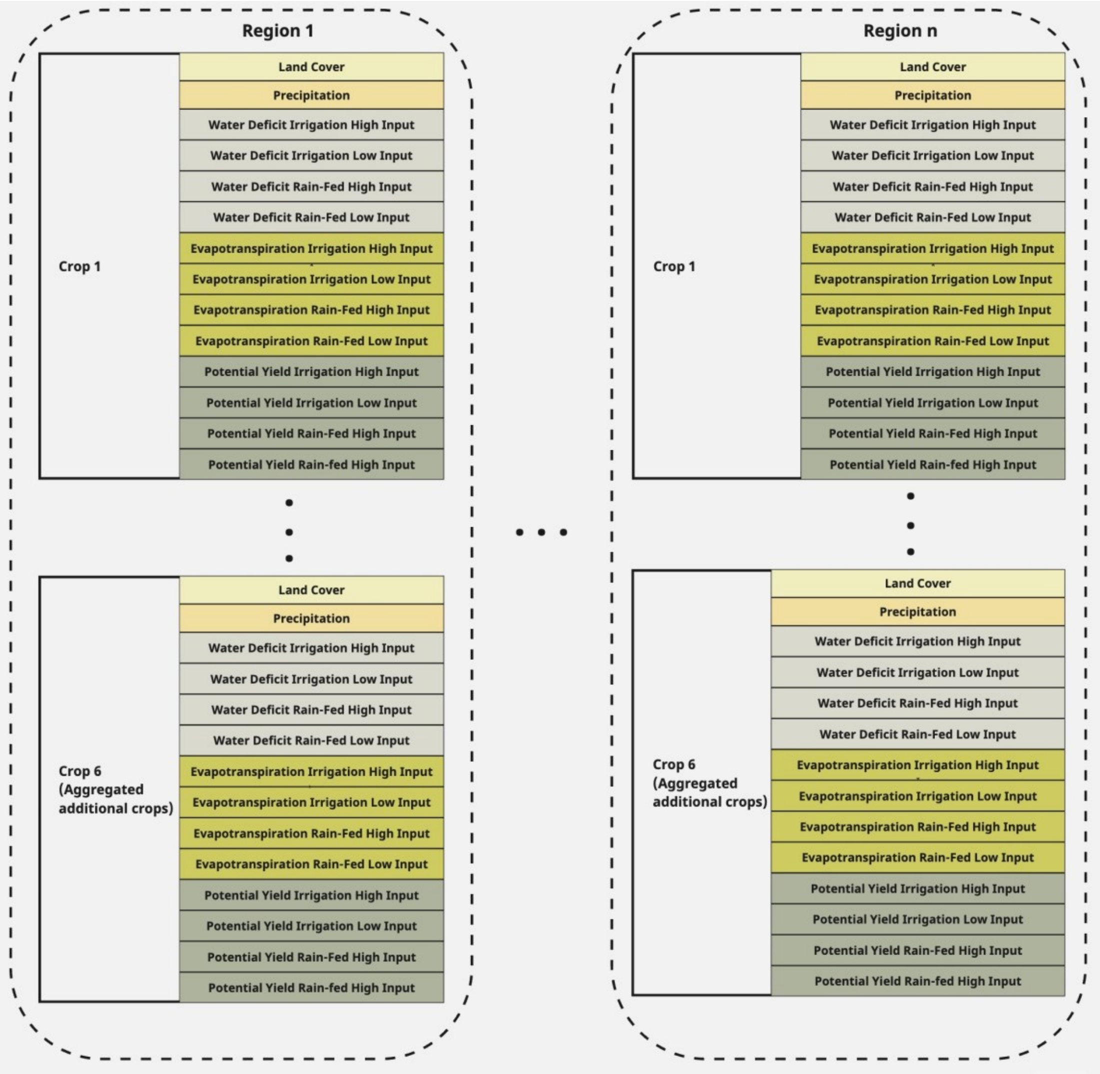
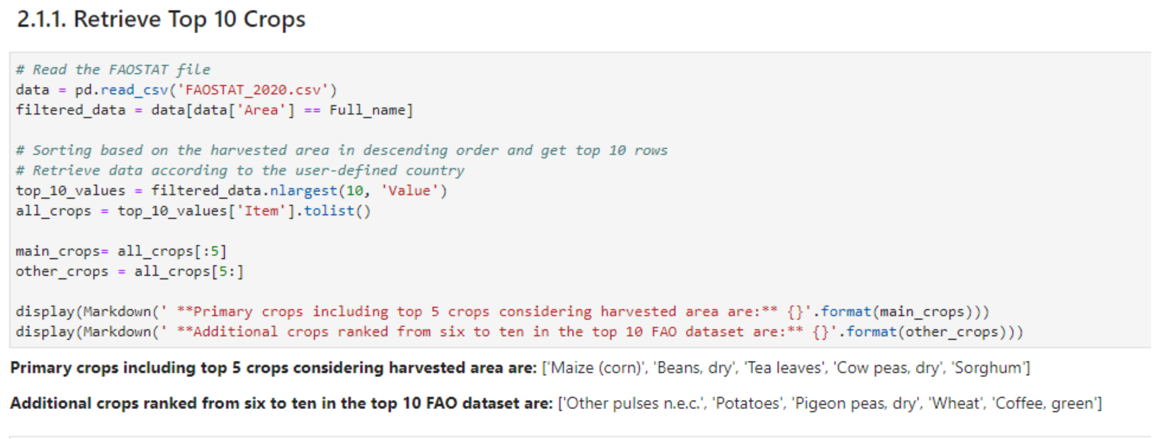

Region 1
Crop 1
Land Cover
Precipitation
Water Deficit Irrigation High Input
Water Deficit Irrigation Low Input
Water Deficit Rain-Fed High Input
Water Deficit Rain-Fed Low Input
Evapotranspiration Irrigation High Input
Evapotranspiration Irrigation Low Input
Evapotranspiration Rain-Fed High Input
Evapotranspiration Rain-Fed Low Input
Potential Yield Irrigation High Input
Potential Yield Irrigation Low Input
Potential Yield Rain-Fed High Input
Potential Yield Rain-fed High Input
Crop 6 (Aggregated additional crops)
Land Cover
Precipitation
Water Deficit Irrigation High Input
Water Deficit Irrigation Low Input
Water Deficit Rain-Fed High Input
Water Deficit Rain-Fed Low Input
Evapotranspiration Irrigation High Input
Evapotranspiration Irrigation Low Input
Evapotranspiration Rain-Fed High Input
Evapotranspiration Rain-Fed Low Input
Potential Yield Irrigation High Input
Potential Yield Irrigation Low Input
Potential Yield Rain-Fed High Input
Potential Yield Rain-fed High Input
Region n
Crop 1
Land Cover
Precipitation
Water Deficit Irrigation High Input
Water Deficit Irrigation Low Input
Water Deficit Rain-Fed High Input
Water Deficit Rain-Fed Low Input
Evapotranspiration Irrigation High Input
Evapotranspiration Irrigation Low Input
Evapotranspiration Rain-Fed High Input
Evapotranspiration Rain-Fed Low Input
Potential Yield Irrigation High Input
Potential Yield Irrigation Low Input
Potential Yield Rain-Fed High Input
Potential Yield Rain-fed High Input
Crop 6 (Aggregated additional crops)
Land Cover
Precipitation
Water Deficit Irrigation High Input
Water Deficit Irrigation Low Input
Water Deficit Rain-Fed High Input
Water Deficit Rain-Fed Low Input
Evapotranspiration Irrigation High Input
Evapotranspiration Irrigation Low Input
Evapotranspiration Rain-Fed High Input
Evapotranspiration Rain-Fed Low Input
Potential Yield Irrigation High Input
Potential Yield Irrigation Low Input
Potential Yield Rain-Fed High Input
Potential Yield Rain-fed High Input

Figure 9: Detailed land and water statistics for each region generated by GeoCLEWs.

GeoCLEWs Result Validation

Results generated by GeoCLEWs undergo validation procedures to ensure their reliability and accuracy. Proper testing approaches are utilized according to the type and processing methodology of each result. In the following detailed validation testing is provided, Kenyan datasets are chosen for the validation process, although GeoCLEWs allows the same procedures to be applied to any chosen county.

Primary crop identification: in Part 2.1.1. Retrieve Top 10 Crops, GeoCLEWs uses the country name that the user specifies in the configuration part to retrieve the top harvested crops from FAOSTAT. Kenya is the test country and GeoCLEWs displayed names of the crops on the screen including maize (corn), beans, dry, tea leaves, cowpeas, dry, sorghum, other pulses n.e.c., potatoes, pigeon peas, dry, wheat, and coffee green. These crops are categorized into two lists: primary and additional, based on their harvested value, a screenshot of the GeoCLEWs results is shown in Figure 10. The FAOSTAT 2020 is manually downloaded and filtered for Kenya; results are provided in Appendix B. The same results are obtained from manual processing and GeoCLEWs demonstrating the accurate performance of this tool in crop identification.



2.1.1. Retrieve Top 10 Crops

```
# Read the FAOSTAT file
data = pd.read_csv('FAOSTAT_2020.csv')
filtered_data = data[data['Area'] == Full_name]

# Sorting based on the harvested area in descending order and get top 10 rows
# Retrieve data according to the user-defined country
top_10_values = filtered_data.nlargest(10, 'Value')
all_crops = top_10_values['Item'].tolist()

main_crops= all_crops[:5]
other_crops = all_crops[5:]

display(Markdown(' **Primary crops including top 5 crops considering harvested area are:** {}'.format(main_crops)))
display(Markdown(' **Additional crops ranked from six to ten in the top 10 FAO dataset are:** {}'.format(other_crops)))
```

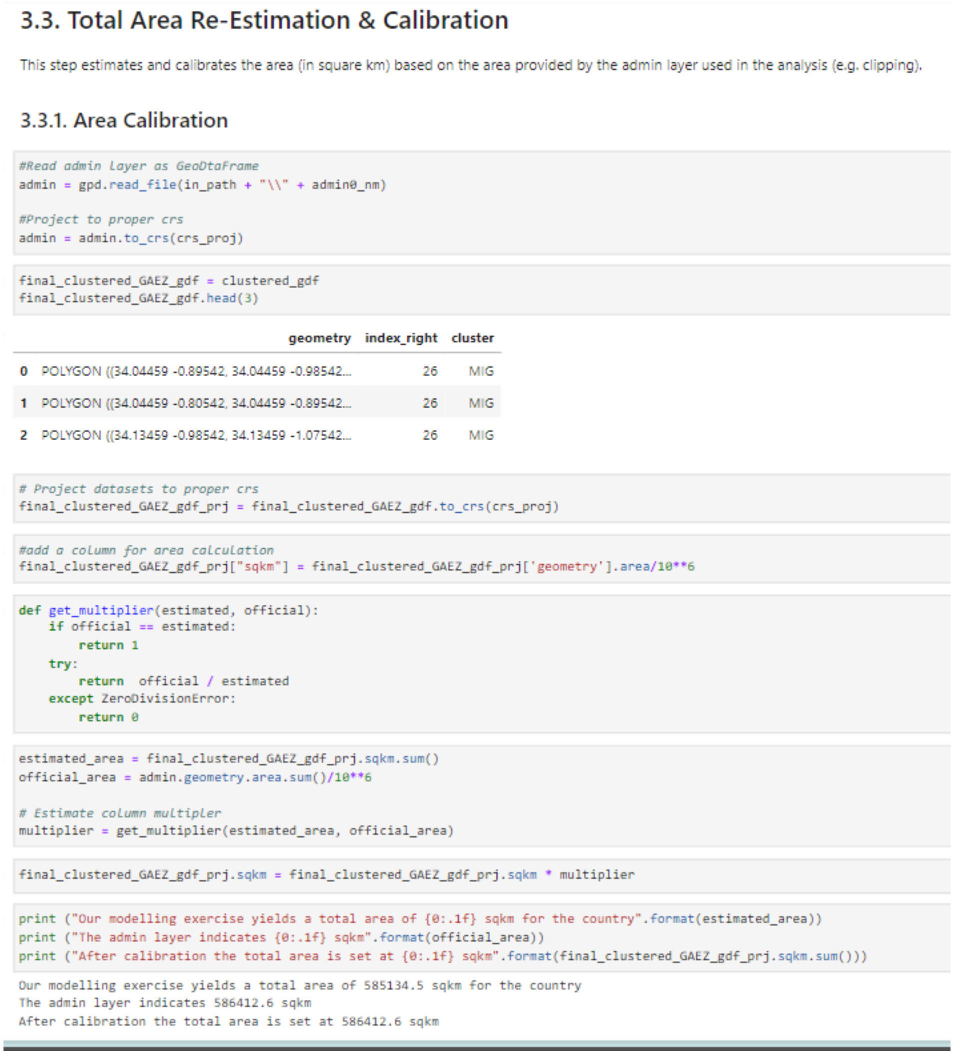
Primary crops including top 5 crops considering harvested area are: ['Maize (corn)', 'Beans, dry', 'Tea leaves', 'Cow peas, dry', 'Sorghum']

Additional crops ranked from six to ten in the top 10 FAO dataset are: ['Other pulses n.e.c.', 'Potatoes', 'Pigeon peas, dry', 'Wheat', 'Coffee, green']

Figure 10: A screenshot of results of Part 2.1.1 for Kenya, showing the top ten crops identified

GAEZ data collection: CLEWs model requires information on potential yield, evapotranspiration, and water deficit of all selected crops and GeoCLEWs retrieves corresponding raster files of these crops from GAZE v4. The GAEZ datasets are classified based on water supply system (rain-fed and irrigation) and input management level (High and Low), which is in total 12 raster files for each crop plus precipitation and land cover processed TIFF files. This tool displays the list of collected raster files in Part 2.2.5 for transparency, the evaluation reveals that GeoCLEWs accurately obtained all required GAEZ files.

Land cell generation: GeoCLEWs utilizes grid points to generate land cells and break down the large geographical region into small segments. These land cells enable the implementation of detailed spatial processing for discrete land segments across the landscape. The generated land cells are evaluated in Part 3.3.1. and 3.3.2. in terms of spatial alignment, coverage, and consistency. The shapefile of the official administrative boundary of the geographical region is downloaded as the reference datasets; GeoCLEWs calculates the total area of the entire region and uses as validation datasets to check the accuracy of the estimated total area of land cells. It implements area calibration to adjust the miscalculated land cells across the region. In Part 3.3.1. Area Calibration of the Kenya assessment, the reference data indicates 586412.6 km^2 and the estimated total area of generated land cells yields 585134.5 km^{2,} which after calibration it demonstrates a total area of 586412.6 km^2 representing the accuracy and reliability of the tool's' performance. The estimated and reference datasets are calculated in the same coordinate system to preserve geoprocessing A screenshot of the area calibration operation for land cells is shown in Figure 11, along with a message that explains how the estimated area and reference data are calculated.



3.3. Total Area Re-Estimation & Calibration

This step estimates and calibrates the area (in square km) based on the area provided by the admin layer used in the analysis (e.g. clipping).

3.3.1. Area Calibration

```
#Read admin Layer as GeoDtaFrame
admin = gpd.read_file(in_path + "\\" + admin0_nm)

#Project to proper crs
admin = admin.to_crs(crs_proj)
```

```
final_clustered_GAEZ_gdf = clustered_gdf
final_clustered_GAEZ_gdf.head(3)
```

	geometry	index_right	cluster
0	POLYGON ((34.04459 -0.89542, 34.04459 -0.98542...	26	MIG
1	POLYGON ((34.04459 -0.80542, 34.04459 -0.89542...	26	MIG
2	POLYGON ((34.13459 -0.98542, 34.13459 -1.07542...	26	MIG

```
# Project datasets to proper crs
final_clustered_GAEZ_gdf_prj = final_clustered_GAEZ_gdf.to_crs(crs_proj)
```

```
#add a column for area calculation
final_clustered_GAEZ_gdf_prj["sqkm"] = final_clustered_GAEZ_gdf_prj['geometry'].area/10**6
```

```
def get_multiplier(estimated, official):
    if official == estimated:
        return 1
    try:
        return  official / estimated
    except ZeroDivisionError:
        return 0
```

```
estimated_area = final_clustered_GAEZ_gdf_prj.sqkm.sum()
official_area = admin.geometry.area.sum()/10**6

# Estimate column multipler
multiplier = get_multiplier(estimated_area, official_area)
```

```
final_clustered_GAEZ_gdf_prj.sqkm = final_clustered_GAEZ_gdf_prj.sqkm * multiplier
```

```
print ("Our modelling exercise yields a total area of {0:.1f} sqkm for the country".format(estimated_area))
print ("The admin layer indicates {0:.1f} sqkm".format(official_area))
print ("After calibration the total area is set at {0:.1f} sqkm".format(final_clustered_GAEZ_gdf_prj.sqkm.sum()))
```

```
Our modelling exercise yields a total area of 585134.5 sqkm for the country
The admin layer indicates 586412.6 sqkm
After calibration the total area is set at 586412.6 sqkm
```

Figure 11: Area calibration of land cells.

Agro-climatic statistic calculation: GeoCLEWs estimates final land and water statistics for CLEWs modelling using categorical and continuous raster files. In terms of validation, two different methods are applied considering the characteristic and methodological approach of the generated outputs.

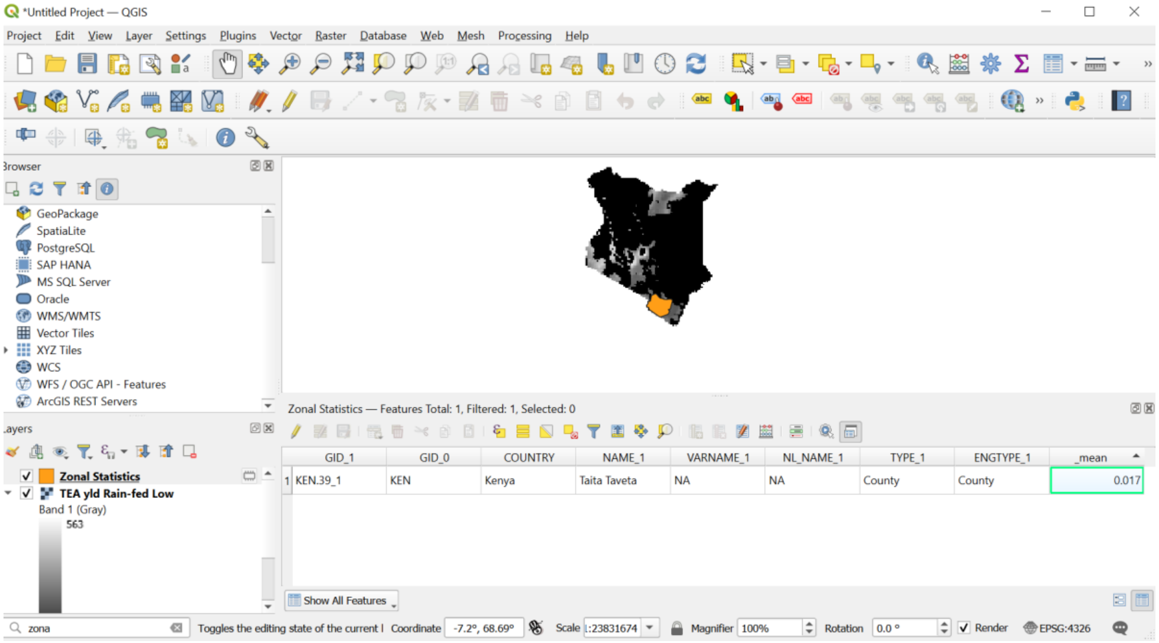
The area and type of land cover for each administrative region are represented by land cover statistics that are generated from categorical sources. As discussed in Sections 5.1.2 and 5.2.2. in order to determine the area of land covered by a certain LCType, GeoCLEWs gathers raster cells from GAEZ raster files and determines how many cells are assigned to each region. For validation, this tool calculates the total area of LCTypes found in each region and shows the results in the "sqkm" column. Additionally, it estimates the sum of areas of all regions by utilizing the value in the "sqkm" column. Table 12 provides the summary statistics of LCTypes in each region and the estimated total area which is equal to the reference data calculation. This demonstrates that throughout the spatial join and extraction process, every land cell was assigned to the appropriate area, and the land cover process was precisely calculated.

Continuous raster datasets are processed to generate crop potential yield, water deficit, evapotranspiration, and precipitation statistics. Randomly chosen raster data are manually processed with QGIS for validation, and the outcomes have been compared with the outputs of GeoCLEWs. Results from Taita Taveta County are shown in Table 17. Both GeoCLEWs and QGIS employs the Zonal Statistics method to calculate the mean vale of raster cells within the administrative boundary of the Taita Taveta County.

Table 17: Randomly selected outputs generated by GeoCLEWs.

Region	MZE cwd Rain-fed High	COW evt Rain-fed High	Precipitation	TEA yld Rain-fed Low
Taita Taveta	0.0708	0.2858	0.6136	0.017

- Potential yield: The raster file of agro_climatic potential yield for tea with the rain-fed and low level of input is loaded into the QGIS shown in Figure 12. This raster is active in the Layer view and provides continuous gray value for the entire Kenya. Zonal Statistics is used to estimate the mean value of the Taita Taveta region, which is displayed on the screen as orange. The computed mean value is taken from the Zonal Statistics table of attributes and represents 0.017 million tonnes per 1000 km2. Both approaches yielded identical values, confirming the accuracy and reliability of this tool's performance.



☑ Zonal Statistics
☑ TEA yld Rain-fed Low
Band 1 (Gray)
563

Zonal Statistics — Features Total: 1, Filtered: 1, Selected: 0

	GID_1	GID_0	COUNTRY	NAME_1	VARNAME_1	NL_NAME_1	TYPE_1	ENGTYPE_1	_mean
1	KEN.39_1	KEN	Kenya	Taita Taveta	NA	NA	County	County	0.017

zona

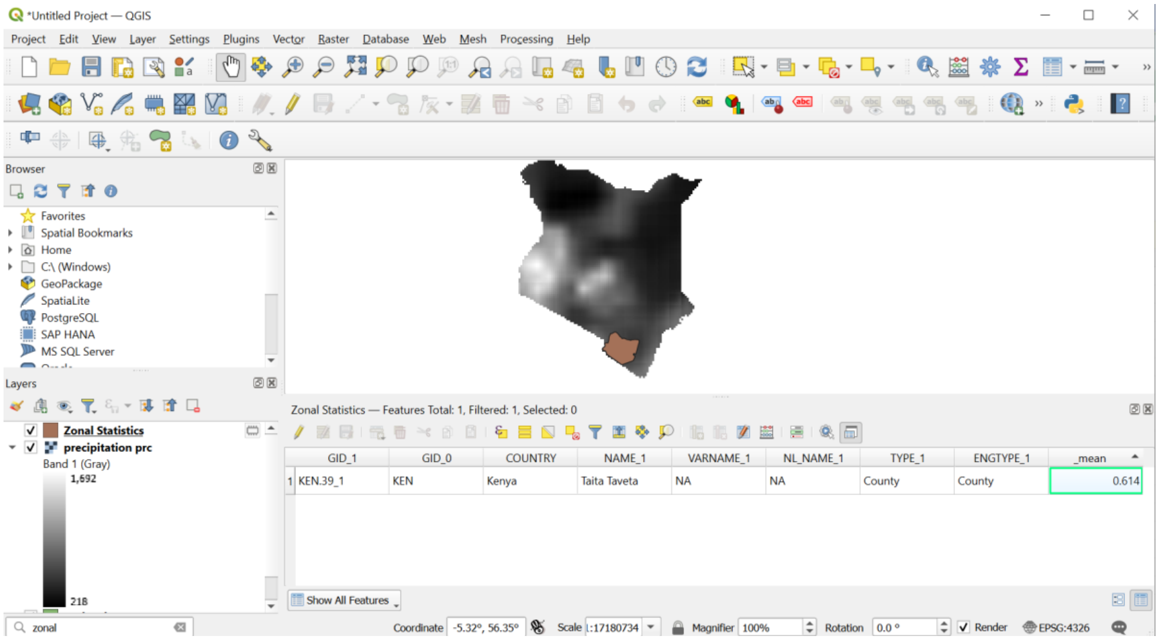
Coordinate -7.2°, 68.69° Scale 1:23831674 Magnifier 100% Rotation 0.0 ° EPSG:4326

Figure 12: Validation process- manual calculation of crop agro-climatic potential yield estimation using QGIS.

- Crop water deficit: The raster file of crop maize with rain-fed watering system and high level of input management is chosen to validate the performance of GeoCLEWs to calculate spatial statistics using continuous crop water deficit GAEZ raster file. GeoCLEWs estimates 0.0708 billion cubic meters 1000 km^2 for Taita Taveta region. There result of implementing Zonal Statistics in QGIS is provided in Appendix D representing maze cwd rain-fed High raster file is active in the Layer view with gray value ranging from 0 to 359. This layer is shown on the screen covering the entire Kenya area. The mean value of the Taita Taveta is calculated with Zonal Statistics shown with purple color. The table of attribute of Zonal Statistic represents the mean value of this county is equal to 0.07 billion cubic meters 1000 km^2.

- Crop evapotranspiration: Randomly raster file of rain-fed cowpea evapotranspiration with high input raster is processed by QGIS and the estimated value as shown in Appendix D is 0.285 billion cubic meters 1000 km^2. The calculated average value of land cells within Taita Taveta is 0.2858 billion cubic meters 1000 km^2 equal to the manual validation process.

- Precipitation: There is a unique raster file providing information on precipitation with global coverage. This raster file is clipped to the administrative boundary of Kenya. Zonal Statistical calculation is implemented in QGIS and the table of attributes represents the mean value of 0.614 billion cubic meters 1000 km^2, Figure 13 illustrates the screenshot of QGIS environment. Table 17 shows that the estimated value from GeoCLEWs is 0.6136 billion cubic meters per 1000 km^2, and any minor differences between the two estimations are attributed to the rounding-up process.



Zonal Statistics — Features Total: 1, Filtered: 1, Selected: 0

	GID_1	GID_0	COUNTRY	NAME_1	VARNAME_1	NL_NAME_1	TYPE_1	ENGTYPE_1	_mean
1	KEN.39_1	KEN	Kenya	Taita Taveta	NA	NA	County	County	0.614

Coordinate -5.32°, 56.35° Scale 1:17180734 Magnifier 100% Rotation 0.0 ° ☑ Render EPSG:4326

Figure 13: Validation process- manual calculation of precipitation for Taita Taveta County using QGIS.

In summary, the outputs from all components involved in GeoCLEWs have been thoroughly assessed, demonstrating the accuracy and reliability of its functionality.

Chapter 4.

Case Study

We applied GeoCLEWs to an energy system model of Kenya and analyzed the implications with a focus on Taita Taveta County (TTC). GeoCLEWs had a significant role in the success of this project as it enables modellers from Taita Tavete University (TTU) to utilize open source high-resolution land and water data to build a more detailed and reliable CLEWs model. It demonstrated effective functionality and applicability by generating detailed statistics at both county and country levels based on stakeholders' preferences providing an opportunity to develop and assess scenarios according to regional and national land and water resource availability and policy planning.

The majority of earlier research in Kenya mainly focused on the energy sector in isolation and failed to delve into the highly interconnected linkages among water, land, and energy for sustainable development. Carvallo et al. [75] examined a low-carbon development pathway for the Kenyan power sector in a series of feasible scenarios for rapidly expanding economies. The expansion of a power system is estimated using SWITCH, a mixed integer linear programme, subject to fulfilling demand forecasts and a variety of operational limitations. Moksnes et al. [76] suggested electrification pathways to guarantee that every Kenyan citizen and community has reliable access to energy by 2030; however, there is less emphasis on mitigating the adverse impacts of GHG emissions. It employed soft linking between the two models, OnSSET and OSeMOSYS to discover the most affordable solution for the entire system on a national scale in Kenya. Akute et al. [77] assessed the impacts of transitioning from biomass stoves to electric stoves on the nation's electrical sector, recommending investigating interactions between different sectors to enable policymakers to make well-informed decisions. Developed CLEWs model in this project using GeoCLEWs empowered local and central governments to make informed decisions and study interlinkages.

Developing the Kenya CLEWs model utilizing GeoCLEWs offered in-depth county-level crop suitability, land-use, and water availability assessments based on historical and simulated future datasets. Various steps of creating the Kenya CLEWs model are presented in Figure 14 which illustrates the requirement of having data on land, water, and electricity, and a configuration file. The Kenya CLEWs model is developed using open source tools and datasets improving transparency and reproducibility of research. In this case study, the land and water systems are built using detailed statistics derived from GeoCLEWs in a clewsy-compatible format. The electricity system is generated by OSeMOSYS Global [57]. The configuration file is in YAML format including basic model specifications such as the period of modelling. clewsy collects generated information on land, water, and electricity systems along with the configuration file and generates the data file. In this case study, Otoole [60], a command-line tool written in Python, is called to modify output CSV files into a format compatible with CLEWs UI. The updated UI enables modellers to change parameters and design scenarios including constraints in a user-friendly environment.

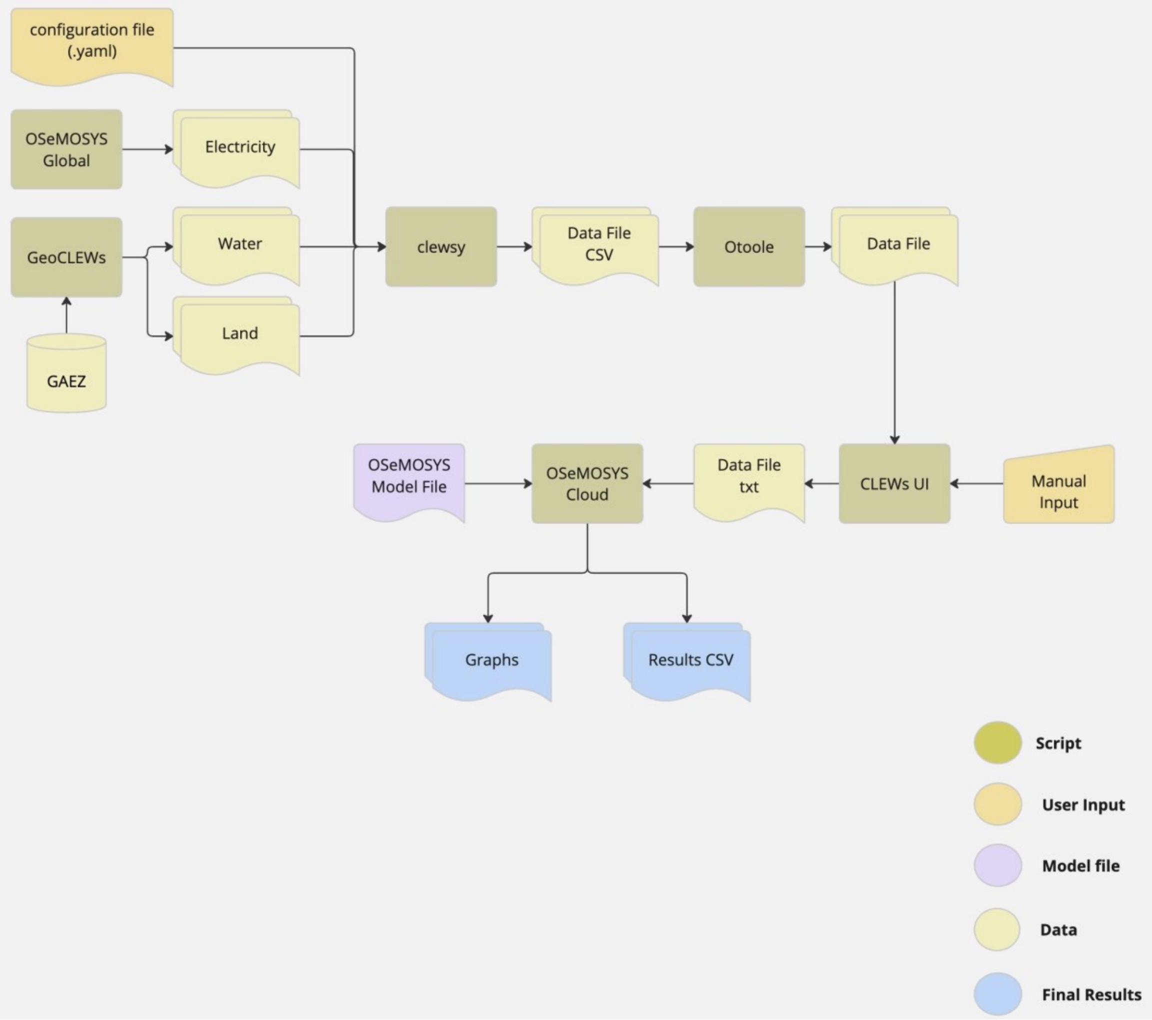


Figure 14: **Flowchart of developing CLEWs model using land and water systems generated by GeoCLEWs.**

4.1. GeoCLEWs: Configurations and Results

This section represents GeoCLEWs setup specification and generated land and water statistics. The configuration of this project is set up based on user preferences and geographical characteristics of the region of study. Additional crops statistics as are not included as this is an additional feature added to GeoCLEWs subsequent to the case study, and crops are selected based on FAOSTAT and user preferences including maize, bean, cowpea, sorghum, sweet potato and tea. The admin level is set to one to generate results at the county administrative level and land cells are generated using point grid with 0.09 DD spacing, which is illustrated in Figure 6. Kenya includes 47 counties in total and stakeholders were interested in having detailed information on this county individually and aggregated statistics on the remaining regions. Therefore, all counties except Taita Taveta were aggregated into groups. Aggregation assists in reducing computational processing while developing the CLEWs model in the following phases. The number of regions to be aggregated into one group is set to 10 in user configuration. As a result, the first 40 counties are classified in the first 4 aggregated region clusters, and the last cluster, NCE, contains the remaining 6 regions. Consequently, GeoCLEWs performed geographical clustering based on user customization, extracted crop attainability, water availability, and land-use information for each of the 47 counties separately; it excluded Taita Taveta County from the cluster dataset. At the conclusion of the aggregating procedure, GeoCLEWs displays a message with the names of the original clusters assigned to each new aggregated. The regional aggregation results are shown in Table 18, county names are changed to the 3-letter format to adhere to CLEWs standards and TAI stands for Taita Taveta County.

Table 18: Geographical cluster aggregation from the Kenya case study.

Aggregated Cluster	Administrative Region
NCA	BAR, BOM, BUN, BUS, ELG, EMB, GAR, HOM, ISI, KAJ.
NCB	KAK, KER, KIA, KIL, KIR, KIS, KIT, KSU, KWA, LAI.
NCC	LAM, MAC, MAK, MAN, MAR, MER, MIG, MOM, MUR, NAI.
NCD	NAK, NAN, NAR, NYA, NYD, NYE, SAM, SIA, TAN, THA.
NCE	TRA, TUR, UAS, VIH, WAJ, WES.
TAI	TAI

GeoCLEWs proceed with the geographical clustering and statistic estimation considering the configuration setup. Land cover statistics and agro-ecological estimations after aggregation in this project are presented in Table 13 and Table 15 respectively. Examples of interactive graphs generated in the Kenya case study are shown in Figure 8 representing the value of the potential yield of maize with an artificial irrigation system and high-level agriculture management, and the total area of land cover type 6, which is 50-75% grassland shrub or herbaceous cover. Table 16 partially provides results of crop water deficit of cluster NCA with clewsy-compatible structure. GeoCLEWs delivered agro-climatic statistics on all primary and additional crops within aggregated clusters and TTC.

4.2. CLEWs Model: Scenario Design and Result Discussion

Kenya CLEWs model employed OSeMOSYS optimization model to find out the optimum solution and system configuration during the period from 2020 to 2035. Required data of land, water and energy systems from GeoCLEWs and OSeMOSYS Global are collected, processed and imported into CLEWs UI. Scenarios are defined inside the user-friendly UI to assess the impacts of different decisions on resource production and consumption. Three scenarios are designed considering the TTC developing plans and national commitments along with a base scenario to compare the impacts, intervention, and feasibility of policies and development practices. Considering input data, parameters, and constraints within each scenario, the Kenya CLEWs model processed variables and generated the optimum configurations to address demands. The following details of scenarios are explained:

- **BASE**: Future outcomes based on current conditions and policies.
- **Green Energy Transition (GET)**: Phase out fossil fuels (oil and gas) and unlock the geothermal potential by investing.
- **Solar and Wind Investment (SWI)**: Increase the capital investment of solar PV and wind, along with phasing out fossil fuels.
- **Increased Forest Cover (IFC):** Increase Taita Taveta Forest to 5% according to county developing commitment.

To give an overview of the functioning of GeoCLEWs, key findings from the Kenya CLEWs model are highlighted and a comparison is made between defined scenarios and BASE conditions.

In comparison with previous method leveraging data with coarse spatial resolution, GeoCLEWs highly increased the level of detail of land and water systems within the Kenya CLEWs model. One of the open-source datasets to obtain land cover estimation OECD [78] containing data up to 2019, but it lacks updates for subsequent years, it provides the percentage of total country area. Table 19 shows land cover statistics for Kenya collected from OECD Stat. CLEWs modellers need to manually calculate the area of land cover on a national scale. Furthermore, these estimations are based on one year and do not support historical and future simulation.

Table 19: Land cover with coarse spatial resolution - percentage of the total country of Kenya collected from OECD Stat.

Tree cover	Grassland	Wetland	Shrubland	Sparse vegetation	Cropland	Artificial surface	Bare area	Inland water
16.7	23.3	1.6	28.1	0.6	24.4	0.1	3.1	2.1

However, GeoCLEWs automatically generates detailed land cover statistic for all regions in the geographical area using high-resolution and comprehensive GAEZ datasets. Figure 15, illustrates land cover types and their area on a national and regional scale from the Base scenario of the Kenya CLEWs model. Thanks to the GeoCLEWs, land cover information of five aggregated regions and Taita Taveta County are calculated in detail offering an opportunity for informed sustainable land management. For example, results it reveals that 13,821.3 km^2 of TTC is covered by Land Cover Type 3 including more than 75% of grassland and woodland shown in orange colour, because of national parks existing in this region that makes land management challenging, which is evaluated using official regional reports. Policymakers in the TTC can greatly benefit from using data from GeoCLEWs to manage their land and handle the issue of inadequate land for sustainable long-term planning.

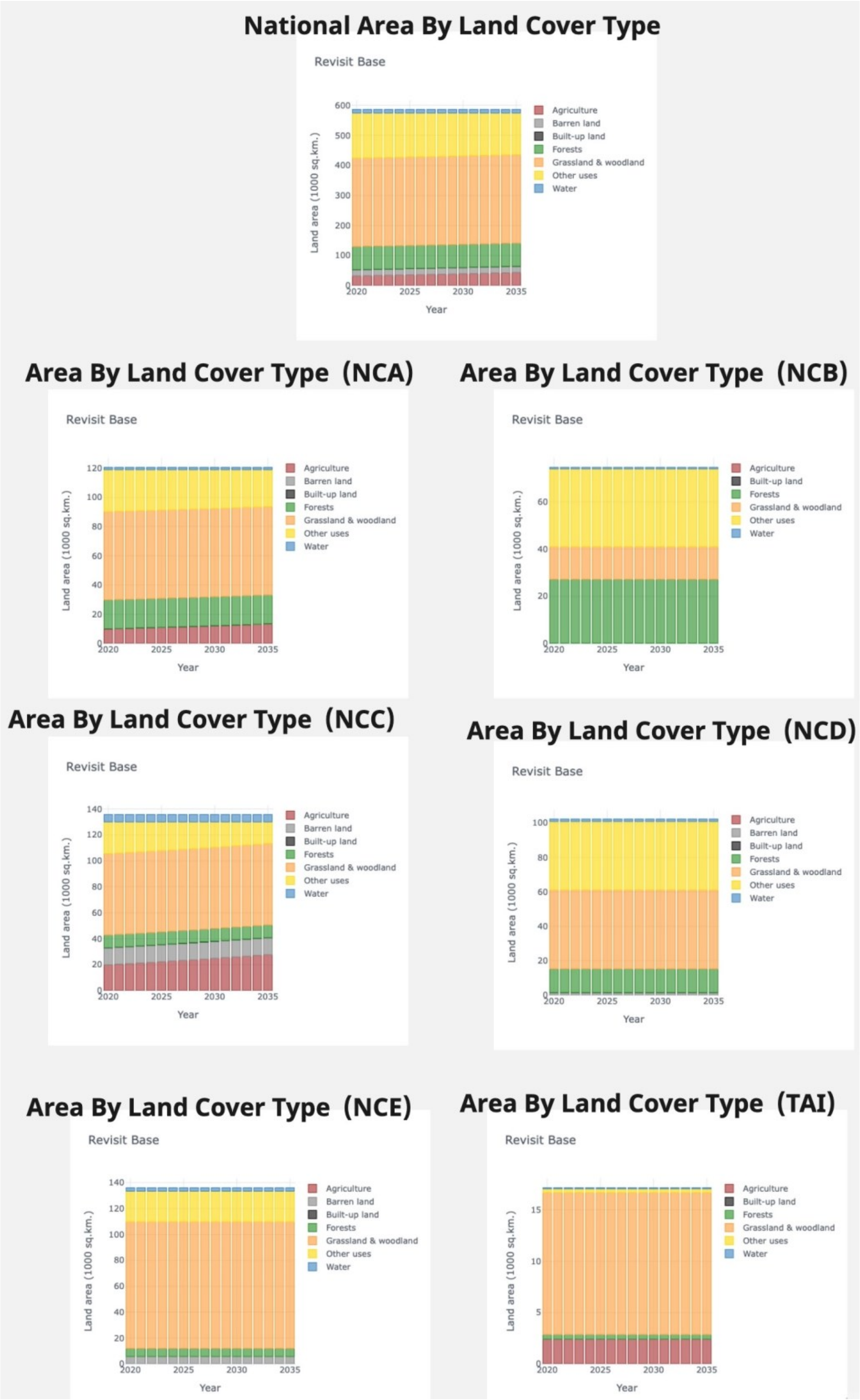


Figure 15: **Land cover results generated by Kenya CLEWs model using detailed GeoCLEWs statistics.**

GeoCLEWs provides the Kenya CLEWs model with the opportunity to utilize statistics on potential yield, water availability, and land cover to precisely describe the kind of crop and the quantity of land that should be allocated to that crop in each region to meet demand during the modelling period. Figure 16 compares the yield and area by crop for the Base and IFC scenarios. It displays the total yield for the country in each year over a 15-year period, as well as the optimum crop combinations to cultivate in each region, taking into account the availability of water, land, and soil suitability. The majority of maize, along with sorghum, sweet potatoes, and cowpea, are produced in the NCC region, whereas the TTC region is a suitable region to cultivate tea and sorghum. GeoCLEWs evaluate the characteristics of each region, and the CLEWs model decides based on these features and interlinkages.

The IFC scenario considered an increase of 5% of forest cover in TTC by 2035 based on the regional sustainable policy. The TTC land cover graph indicates that there is a limited amount of agricultural land in this county. As a result, CLEWs allocated agricultural land for increasing the county's forest cover, which reduced this region's generation of tea. CLEWs attempted to use other regions, such as the NCC, to compensate for the decreased tea production, but this resulted in a reduction in tea yield from 0.569797 tonnes per hectare in 2035 in the Base scenario to 0.5260727 tonnes per hectare in IFC since TTC is the primary producing region.

GeoCLEWs assist in identifying interlinkages between systems and assess the impact of policies and practices on WEF resources on regional and a national scale. Detailed GeoCLEWs results revealed that increasing forest cover by 5% based on TTC development commitment resulted in a reduction in agricultural land and crop yields in TTC requiring intensified agriculture including advanced farming technologies and high level of land management practices.

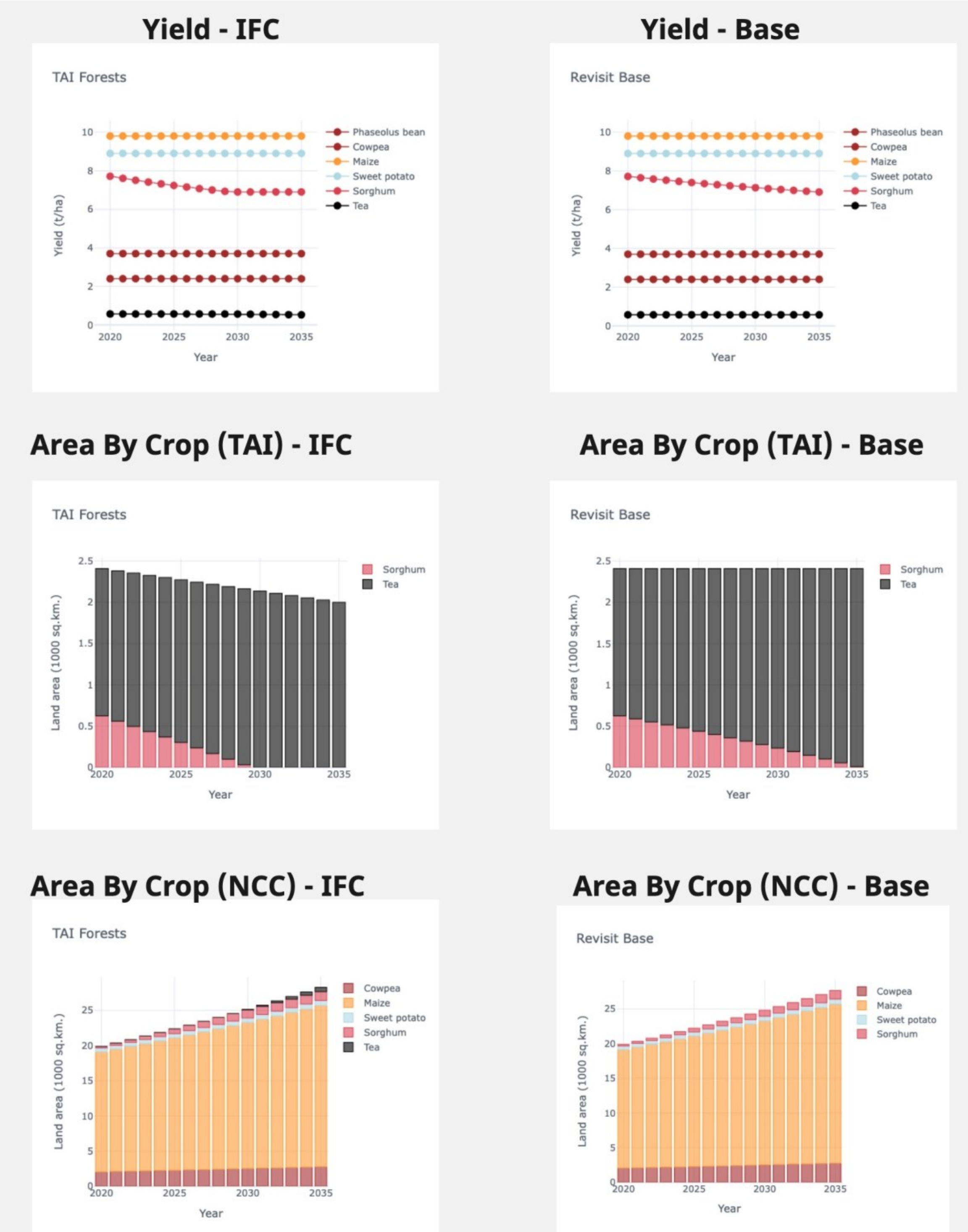


Figure 16: **Comparison of crop yield and area between Base and IFC scenarios generated by the Kenya CLEWs model using GeoCLEWs statistics.**

The power generation results from BASE, GET, and SWI scenarios are visualized in Figure 17 The comparison between developed scenarios and current conditions reveals the overall impact of developing plans. In the Green energy transition, the focus is on increasing capital investment for geothermal energy signifying an essential element of the sustainable energy transition. Eliminating oil and gas resulted in reducing GHG emissions aligning with the Paris Agreement. In addition, GET contributes to less water demand within the power sector, which is extremely advantageous, given the scarcity of water in this area. The SWI scenario's results show that while the generation of electrical power from geothermal has declined, solar and wind electricity production has increased. This scenario requires a significant capital investment to support the development of solar and wind technology. Phasing out fossil fuel-based sources of power caused GHG reduction compared to the BASE case.

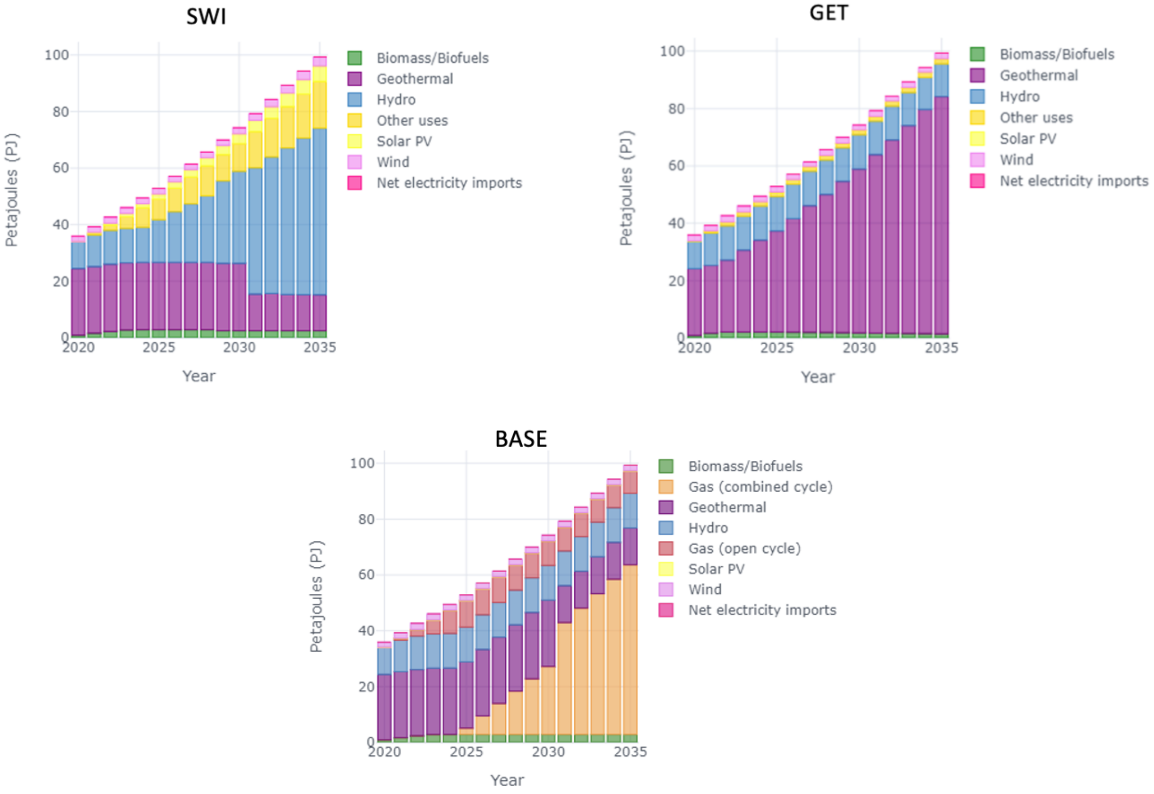


Figure 17: **Graphs of power generation for BASE, GET, and SWI scenarios from the Kenya case study.**

4.3. Benefits and Contributions of the GeoCLEWs in the Kenya Case Study

In summary, comprehensive statistics generated by GeoCLEWs provided the TTU CLEWs modellers to explore the limitations and opportunities of the land and water systems in each region. GeoCLEWs outputs enable stakeholders to explore pressure points and identify trade-offs to reach national and regional development goals, with a focus on TTC. It improved the analysis of policy decisions regarding GHG reduction, water competition, and land-use change. Furthermore, land and water statistics in clewsy-compatible format significantly improved and facilitated the process of developing the CLEWs model assisting TTU modellers to simply adopt data from multi-resources and generate data file for CLEWs UI.

In many cases including the Kenya case study, CLEWs modellers do not have access to reliable detailed land and water datasets and the only available option is using open-source datasets. To generate land and water systems within the CLEWs framework, one option is to manually collect the required information from resources such as FAOSTAT, AQUASTAT [79], and OECD databases and calculate rough estimation on a national scale. Every year, the FAOSTAT dataset provides the area of harvested crops for every country, which can be used to identify the primary crops in the selected country. Then, they need to extract the national production quantity of those crops from the same database to estimate the ratio of production per unit of land. This approach results in obtaining a coarse estimation of crop production across the entire region assuming all regions in the country have the same agro-climatic potential yield ratio for the primary crops, Figure 1 illustrates the general workflow of the existing approach to obtain land and water data for CLEWs modelling utilizing open source datasets.

In addition, CLEWs modellers need to manually do mathematical calculations to estimate the production ratio of irrigated and rainfed crops using general parameters for all crops. They can use the ratio between irrigated and rainfed yields parameter along with the percentage of cultivated land irrigated from the AQUASTAT database to model the following 20 or 30 years. This dataset also contains the annual precipitation information that can be utilized for an approximation of precipitation per unit of land which does not include future climate change as well as specific precipitation data broken down per county or other administrative region.

OECD database provides the percentage of land use on a national scale in previous years, which classifies landscape into nine types of land cover including inland water, bare area, artificial surface, spare vegetation, cropland, shrubland, wetland, grassland, and tree cover.

These roughly estimated statistics do not include comprehensive historical yield production, water availability, future impacts of climate changes, GHG emission pathways, detailed land cover variety, and soil suitability diversity across the country. Moreover, it requires manual unit adjustment, CLEWs naming standardization, and manually entering estimated values for all crops and regions into the CLEWs user interface.

The alternative option for collecting land and water data is utilizing high-resolution GAEZ v4 datasets. Except for GeoCLEWs, there is currently no available and functional tool compatible with GAEZ v4 to generate statistics for CLEWs modelling. Therefore, to correctly collect the necessary raster files and apply the proper approach for modification and statistical calculations, researchers are required to possess an in-depth knowledge of GAEZ documentation in terms of resolution, unit, and parameters of datasets. Furthermore, there is a significant demand for expertise in GIS processing to manually handle raster datasets based on administrative boundaries, ensuring the generation of reliable results.

The replicable workflow of GeoCLEWs significantly accelerates and simplifies the process of producing comprehensive land and water information and collecting spatial attributes per 100 km^2. For instance, CLEWs modellers need to dedicate days or even weeks of work based on their geoprocessing experience and knowledge of GAEZ documentation to generate land and water systems. This is because they must manually filter and collect over 110 raster files, in case of processing 10 crops, from the GAEZ v4 database, employ GIS tools for geoprocessing, and perform calculation processes for all regions and crops individually. However, GeoCLEWs generates detailed results for any country in less than 45 minutes, providing a useful opportunity to customize the project setup and compare outcomes of different geographical areas, admin levels, and GHG emission pathways to obtain a comprehensive understanding of the land and water systems and address different project objectives.

Moreover, GeoCLEWs can improve the precision and accuracy of outcomes by minimizing the possibility of mistakes made by users during manual computation. The manual mathematical estimations and data entry can introduce a level of human error as well as necessitate an extensive amount of time. Considering that the currently available methods for land and water assessment require extensive manual calculation and entry of data, there is a potential for miscalculation and mistakes. Therefore, the automated workflow of GeoCLEWs for data collection, preparation, and standardization as well as computer-based calculation methods can enhance the accuracy and reliability of results by reducing the risk of human errors.

Chapter 5.

Conclusions

This thesis presents GeoCLEWs, an open source tool to process high-resolution land and water data to enhance CLEWs modelling. This research established clear objectives to fill the identified gaps in the existing approaches. The various steps involved in the process of designing GeoCLEWs, addressing challenges, rigorous performance evaluation, and result analysis are presented in detail in the preceding chapters. This conclusion highlights the key elements of the contribution of this study to the climate change, land, water, and energy assessment and pave the way to facilitate the challenge of policy planning within the CLEWs framework.

5.1. Summary of Work and Contributions

This thesis presents a robust reproducible processing tool for generating detailed land and water statistics for CLEWs modelling to address the existing gap in the absence of a functional tool to process the updated GAEZ database. It also adopts effective methods to fulfill the challenge of the high computational complexity of processing finer spatial resolution data. GeoCLEWs is developed by considering the CLEWs framework compatibility and satisfies the clewsy input requirements as a perquisition of CLEWs modelling. GeoCLEWs utilizes open source tools and open datasets to promote global collaboration in sustainable development as well as offer a chance for continuous improvement of this tool through the contribution of developers from around the world. The application of GeoCLEWs is validated and improved within the context of a case study. It enhances CLEWs modelling by producing detailed agro-climatic attainability, water availability, and land-use statistics.

This thesis overcomes several challenges through the successful development of the Python-based land and water processing tool. Adopting only publicly available tools and datasets posed difficulties; the open input data must meet CLEWs requirements with a high level of detail and global coverage. Another difficulty was differences between FAOSTAT, GAEZ, and CLEWs crop naming and classification. Furthermore, it was quite challenging to implement various advanced spatial and non-spatial functions precisely and accurately within an open source, simplified, and automated workflow. The diversity of data types and formats required different Python modules and dependencies, encountered with several incompatibility issues. Preprocessing of datasets, in-depth review of module documentation, and compatibility consideration contributed to creating a compatible virtual environment to smoothly run the script and avoid any conflict. Overcoming these obstacles resulted in the development of a reliable and versatile tool.

GeoCLEWs offers a rich set of key features to enhance high-resolution data processing and generate statistics to improve CLEWs modelling. The primary functionality focuses on promoting accurate land cells and geographical boundary calculation and supports precise spatial attribute extraction, which is extremely important to obtain reliable estimation. Its scalability meets various project requirements since embedded functions enable the process of any arbitrary geographical boundary either at the local level or a group of neighbouring countries. The transparent and self-described tool along with supplementary documentation offers an easy-to-use platform for manipulating and exploring the process, making it accessible to people with different technical backgrounds. Another beneficial feature is that automated FAOSTAT and GAEZ data collection, preprocessing, and assessment streamline intricate geoprocessing of large datasets and reduce human errors.

GeoCLEWs has undergone accuracy assessment and evaluation during the entire development process to ensure the validity, reliability, and accuracy of generated outputs. Input data and processing functions have been validated in various steps to achieve a high level of precision and reliability of geoprocessing, including re-estimation of land cells' area, total area recalculation and output calibration. The generated point grid and raster files are reprojected to the most proper coordinate system according to the geographical longitude and latitude of the selected area. The transparent script displays the results after the completion of each part for user comprehension and accuracy assessment. Furthermore, the results of the case study and making comparison with official local and national statistics confirm the rigorousness of the approach.

5.2. Conclusions and Policy Implications

The application of GeoCLEWs in assessing high-resolution land and water data has yielded compelling results, which represent enhancement in data collection, customization effectiveness, and CLEWs compatibility. The streamlined workflow of data retrieval and preprocessing notably reduced time required for data preparation as well as human error, which enhances the overall efficiency in handling large GAEZ datasets with complex data structures and formats. The customization capabilities empower users to tailor GeoCLEWs to generate results at specific administrative level or aggregated regions. The results offer detailed statistics on crop agro-ecological potential yield, evapotranspiration, water deficit, precipitation, and land cover considering CLEWs consistency and clewsy compatibility, which remarkably facilitate CLEWs model development. Additionally, the interactive graphs and national tabular statistics deliver an intuitive representation and insightful comprehension.

The most significant advantage is that only GeoCLEWs possesses the capability of processing detailed GAEZ v4 datasets; available approaches encounter incompatibility with the updated database and can only process the outdated version released in 2012. Another notable benefit is the substantial reduction in manual effort, in comparison with available methods, GeoCLEWs seamlessly can collect, preprocess, and assess all required information from FAOSTAT and GAEZ datasets without user intervention. In contrast to traditional approaches, the tool developed in this thesis mitigates the requirement of high technical skill in programming and background knowledge of geoprocessing.

Furthermore, GeoCLEWs implements regional aggregation based on administrative boundary that to handle the computational complexity of high-resolution data processing. A considerable number of CLEWs models adopted coarse resolution, aggregated estimation, and approximate calculation due to a lack of detailed historical and projected information or computational challenges. Additionally, GeoCLEWs has no dependency on specific input data suppliers and the base land cells can be generated using any user-defined arbitrary shapefile. It also is capable of processing land and water datasets with TIFF format obtained from sources other than GAEZ; adaptability to custom datasets enhances its functionality.

GeoCLEWs along with supplementary documentation and datasets are publicly available on GitHub enabling users to reproduce the entire process and generate results for CLEWs modelling. The GitHub repository contains the developed tool in the format of a self-documented Jupyter Notebook with interactive computing functionality, which is accompanied by a detailed explanation of the processing steps. The documentation covers detailed instructions for downloading and executing essential Python modules and dependencies on a local system to seamlessly execute the script. The following outlines some notable benefits of publishing GeoCLEWs on the GitHub repository:

- Transparency: GeoCLEWs repository supports source code, required datasets, and documentation that are openly available to everyone promoting reliability and transparency.
- Collaboration: Hosting GeoCLEWs on the main OSeMOSYS GitHub [80], which manages approximately 20 repositories including OSeMOSYS Global, clewsy, and Otool, has greatly increased its visibility and contribution to CLEWs modelling.
- Affordability: Open source GeoCLEWs accessible on GitHub would be highly beneficial to individuals, academic institutions, businesses, and organizations with a limited budget to utilize it without incurring any cost.
- Licensing Flexibility: This tool is developed and published under MIT Licence authorizing it for commercial use, modification, distribution, and private use. The mentioned permissions provide considerable freedom to users to employ GeoCLEWs, modify it to suit their needs, and distribute their results without any restriction.

A case study analysis of Kenya demonstrates capabilities of GeoCLEWs in enhancing CLEWs modelling to capture important interactions among WEF nexus. It enables TTU modellers to process high-resolution land and water data at local and national levels and analyze various CLEWs scenarios based on regional and national sustainable planning. Results at the county level generated by GeoCLEWs revealed the limitations of land availability and detailed potential yield analysis in this region. In addition, detailed land and water statistics improved the Kenya CLEWs model to assess national long-term planning; it outlined that the policy of a 5% increase in forest cover will restrict the amount of land available for agriculture and crop yields, necessitating intensified farming with high-tech farming techniques.

5.3. Limitations and Uncertainties

This thesis employed freely available datasets, despite the reliability of data sources, they can incorporate some degree of uncertainties due to assumptions and methodologies. GAEZ dominant land cover classification carries a sort of generalization that may not fit all projects. The general classification criteria to define GAEZ land cover types may underestimate the region’s specifications and lead to missing some information. For example, LCType9 was unable to detect rural houses in Taita Taveta County. Another limitation posed by data suppliers is that the projected datasets of agro-climatic potential yield from the GAEZ v4 considered only high input levels referring to a fully automated land management system allowing the most efficient use of chemical and nutritional pesticides for commercial production; the updated portal does not cover low and intermediate levels of agricultural activity leading to missing valuable information within CLEWs assessment. However, this thesis incorporates historical agro-climatic potential yield datasets with low input levels to address this limitation which pertains to GAEZ and is not related to GeoCLEWs.

Aggregation is an essential component of large-scale geographic area assessment that may pose uncertainty. Regional aggregation combines the detailed value of all land cells across aggregated administrative regions into a summary. The size of land cells and extent of the region play an important role in representing accurate central tendency and minimizing outliers.

5.4. Future Research

Although GeoCLEWs presents significant advancement in processing high-resolution land and water, it is essential to acknowledge limitations and outline directions for future enhancements. Considering the complex procedure of generating, combining, and transforming required data from various sources to develop CLEWs models as represented in Figure 14, exploring a standard workflow to collect all pieces and create a base CLEWs model can streamline this process and promote contribution to CLEWs assessment. In response, an ongoing study is actively working to address this challenge, we are trying to develop an open source automated workflow to collect, process, and modify required multi-resource data to create a base CLEWs model of any given country or a group of nations [59]. The proposed workflow, CLEWs Global, utilizes land and water statistics supplied by GeoCLEWs along with electricity data generated by OSeMOSYS Global and configuration file to develop CLEWs-compatible data file using clewsy and Otoole. Open source workflow and embedded components will be publicly available and reusable to increase the accessibility of CLEWs modelling and promote informed policy planning.

Enhancing GeoCLEWs to include spatial clustering capabilities would be highly beneficial to detect yield similarity within administrative regions. Agro-climatic potential yield characteristics can vary significantly across admin regions, and spatial clustering preserves vital cross-regional similarities and can be an effective alternative to regional aggregation to reduce computational complexity. We are currently enhancing GeoCLEWs functionality by incorporating spatial clustering, and the updated version of this tool will be released soon.

GeoCLEWs cluster naming convention does not differentiate the different subregions with common first three letters such as Kisumu County and Kisii County in Kenya consequently manual correction is required. Exploring a systematic modification to ensure the uniqueness of generated codes of sub-national regions regarding CLEWs name standardization may prevent the misaggregation of distinct regions.

FAOSTAT delivers crop statistics in generalized categories including less specification compared to the GAEZ database; similarly, yield classification within the CLEWs framework does not support comprehensive GAEZ agro-ecological yield information. Currently, users can utilize the transparent and self-described script to include specific crop names, and GeoCLEWs has the flexibility to collect related information from the GAEZ portal and proceed following steps as usual. However, defining a standard and systematic approach would improve crop analysis.

Integrating the automated calibration between FAOSTAT and GAEZ yields can enhance GeoCLEWs capability. It also can eliminate the need for manual calibration using clewsy and contribute to the development of automated workflow of developing CLEWs model.

References

[1] R. C. Estoque, “Complexity and diversity of nexuses: A review of the nexus approach in the sustainability context,” *Sci. Total Environ.*, vol. 854, p. 158612, Jan. 2023, doi: 10.1016/j.scitotenv.2022.158612.

[2] E. Eftelioglu, Z. Jiang, X. Tang, and S. Shekhar, “The Nexus of Food, Energy, and Water Resources: Visions and Challenges in Spatial Computing,” in *Advances in Geocomputation*, D. A. Griffith, Y. Chun, and D. J. Dean, Eds., in Advances in Geographic Information Science. Cham: Springer International Publishing, 2017, pp. 5–20. doi: 10.1007/978-3-319-22786-3_2.

[3] C. A. Scott, M. Kurian, and J. L. Wescoat, “The Water-Energy-Food Nexus: Enhancing Adaptive Capacity to Complex Global Challenges,” in *Governing the Nexus: Water, Soil and Waste Resources Considering Global Change*, M. Kurian and R. Ardakanian, Eds., Cham: Springer International Publishing, 2015, pp. 15–38. doi: 10.1007/978-3-319-05747-7_2.

[4] C. Zhang, X. Chen, Y. Li, W. Ding, and G. Fu, “Water-energy-food nexus: Concepts, questions and methodologies,” *J. Clean. Prod.*, vol. 195, pp. 625–639, Sep. 2018, doi: 10.1016/j.jclepro.2018.05.194.

[5] M. Howells *et al.*, “Integrated analysis of climate change, land-use, energy and water strategies,” *Nat. Clim. Change*, vol. 3, no. 7, pp. 621–626, Jul. 2013, doi: 10.1038/nclimate1789.

[6] S. Kaddoura and S. El Khatib, “Review of water-energy-food Nexus tools to improve the Nexus modelling approach for integrated policy making,” *Environ. Sci. Policy*, vol. 77, pp. 114–121, Nov. 2017, doi: 10.1016/j.envsci.2017.07.007.

[7] A. Beltramo, E. P. Ramos, C. Taliotis, M. Howells, and W. Usher, “The Global Least-cost user-friendly CLEWs Open-Source Exploratory model,” *Environ. Model. Softw.*, vol. 143, p. 105091, Sep. 2021, doi: 10.1016/j.envsoft.2021.105091.

[8] V. Aryanpur, B. O’Gallachoir, H. Dai, W. Chen, and J. Glynn, “A review of spatial resolution and regionalisation in national-scale energy systems optimisation models,” *Energy Strategy Rev.*, vol. 37, p. 100702, Sep. 2021, doi: 10.1016/j.esr.2021.100702.

[9] A. Korkovelos, A. Shivakumar, and T. Alfstad, “CLEWs GIS processing script.” [Online]. Available: https://github.com/akorkovelos/un-clews-gis-work/blob/51168785883a73eab2e2cf6a5df4c5872c73edc8/CLEWs%20GIS%20Processing.ipynb

[10] D. Gielen, F. Boshell, D. Saygin, M. D. Bazilian, N. Wagner, and R. Gorini, “The role of renewable energy in the global energy transformation,” *Energy Strategy Rev.*, vol. 24, pp. 38–50, Apr. 2019, doi: 10.1016/j.esr.2019.01.006.

[11] FAO and IIASA, “Global Agro-Ecological Zones (GAEZ v4) – Data Portal user’s guide. Rome. https://doi.org/10.4060/cb5167en.”

[12] Y. Saedi and T. Niet, “GeoCLEWs v1.0.0.” [Online]. Available: https://github.com/OSeMOSYS/CLEWs_GAEZ

[13] R. Gasper, A. Blohm, and M. Ruth, “Social and economic impacts of climate change on the urban environment,” *Curr. Opin. Environ. Sustain.*, vol. 3, no. 3, pp. 150–157, May 2011, doi: 10.1016/j.cosust.2010.12.009.

[14] S. C. Pryor, “Climate Change Impacts, Risks, Vulnerability, and Adaptation: An Introduction,” in *Climate Change in the Midwest : Impacts, Risks, Vulnerability, and Adaptation*, Indiana University Press, 2013.

[15] J. C. Bergengren, D. E. Waliser, and Y. L. Yung, “Ecological sensitivity: a biospheric view of climate change,” *Clim. Change*, vol. 107, no. 3, pp. 433–457, Aug. 2011, doi: 10.1007/s10584-011-0065-1.

[16] Y. Malhi *et al.*, “Climate change and ecosystems: threats, opportunities and solutions,” *Philos. Trans. R. Soc. B Biol. Sci.*, vol. 375, no. 1794, p. 20190104, Jan. 2020, doi: 10.1098/rstb.2019.0104.

[17] A. Hurlimann, S. Moosavi, and G. R. Browne, “Urban planning policy must do more to integrate climate change adaptation and mitigation actions,” *Land Use Policy*, vol. 101, p. 105188, Feb. 2021, doi: 10.1016/j.landusepol.2020.105188.

[18] X. Zheng, D. Streimikiene, T. Balezentis, A. Mardani, F. Cavallaro, and H. Liao, “A review of greenhouse gas emission profiles, dynamics, and climate change mitigation efforts across the key climate change players,” *J. Clean. Prod.*, vol. 234, pp. 1113–1133, Oct. 2019, doi: 10.1016/j.jclepro.2019.06.140.

[19] S. Clayton, “Climate anxiety: Psychological responses to climate change,” *J. Anxiety Disord.*, vol. 74, p. 102263, Aug. 2020, doi: 10.1016/j.janxdis.2020.102263.

[20] S. A. Montzka, E. J. Dlugokencky, and J. H. Butler, “Non-CO2 greenhouse gases and climate change,” *Nature*, vol. 476, no. 7358, Art. no. 7358, Aug. 2011, doi: 10.1038/nature10322.

[21] Z Pan, D Andrade, and N Goseelin, “Vulnerability of soil carbon reservoirs in the midwest to climate change,” in *Climate Change in the Midwest: Impacts, Risks, Vulnerability, and Adaptation*, 2013. Accessed: Nov. 09, 2023. [Online]. Available: https://scholar.google.com/scholar_lookup?title=Vulnerability%20of%20soil%20carbon%20reservoirs%20in%20the%20midwest%20to%20climate%20change&author=Z.%20Pan&publication_year=2013&pages=92-103

[22] United Nations Environment Programme, “Emissions Gap Report 2019,” Nairobi, 2019. [Online]. Available: https://wedocs.unep.org/bitstream/handle/20.500.11822/30797/EGR2019.pdf?sequence=1&isAllowed=y

[23] F. Martins, C. Felgueiras, M. Smitkova, and N. Caetano, “Analysis of Fossil Fuel Energy Consumption and Environmental Impacts in European Countries,” *Energies*, vol. 12, no. 6, Art. no. 6, Jan. 2019, doi: 10.3390/en12060964.

[24] S. Fawzy, A. I. Osman, J. Doran, and D. W. Rooney, “Strategies for mitigation of climate change: a review,” *Environ. Chem. Lett.*, vol. 18, no. 6, pp. 2069–2094, Nov. 2020, doi: 10.1007/s10311-020-01059-w.

[25] M. A. Benevolenza and L. DeRigne, “The impact of climate change and natural disasters on vulnerable populations: A systematic review of literature,” *J. Hum. Behav. Soc. Environ.*, vol. 29, no. 2, pp. 266–281, Feb. 2019, doi: 10.1080/10911359.2018.1527739.

[26] A. G. Olabi and M. A. Abdelkareem, “Renewable energy and climate change,” *Renew. Sustain. Energy Rev.*, vol. 158, p. 112111, Apr. 2022, doi: 10.1016/j.rser.2022.112111.

[27] P. A. Owusu and S. Asumadu-Sarkodie, “A review of renewable energy sources, sustainability issues and climate change mitigation,” *Cogent Eng.*, vol. 3, no. 1, p. 1167990, Dec. 2016, doi: 10.1080/23311916.2016.1167990.

[28] K. Palmer-Wilson *et al.*, “Impact of land requirements on electricity system decarbonisation pathways,” *Energy Policy*, vol. 129, pp. 193–205, Jun. 2019, doi: 10.1016/j.enpol.2019.01.071.

[29] Y. Tang *et al.*, “Impact assessment of climate change and human activities on GHG emissions and agricultural water use,” *Agric. For. Meteorol.*, vol. 296, p. 108218, Jan. 2021, doi: 10.1016/j.agrformet.2020.108218.

[30] W. N. Smith *et al.*, “Assessing the effects of climate change on crop production and GHG emissions in Canada,” *Agric. Ecosyst. Environ.*, vol. 179, pp. 139–150, Oct. 2013, doi: 10.1016/j.agee.2013.08.015.

[31] M. Bazilian *et al.*, “Considering the energy, water and food nexus: Towards an integrated modelling approach,” *Energy Policy*, vol. 39, no. 12, pp. 7896–7906, Dec. 2011, doi: 10.1016/j.enpol.2011.09.039.

[32] Independent group of scientists, “Global Sustainable Development Report 2019: The Future is Now – Science for Achieving Sustainable Development,” United Nations, New York, first quadrennial Global Sustainable Development Report, 2019. [Online]. Available: https://sustainabledevelopment.un.org/gsdr2019

[33] Marga Gual Soler, Tom C. Wang, and Cynthia R. Robinson, “International Collaboration in Connecting Scientists to Policy | Science & Diplomacy,” *Sci. Dipl.*, 2017, [Online]. Available: https://www.sciencediplomacy.org/article/2017/international-collaboration-in-connecting-scientists-policy

[34] P. Zhang *et al.*, “Food-energy-water (FEW) nexus for urban sustainability: A comprehensive review,” *Resour. Conserv. Recycl.*, vol. 142, pp. 215–224, Mar. 2019, doi: 10.1016/j.resconrec.2018.11.018.

[35] E. P. Ramos *et al.*, “The Climate, Land, Energy, and Water systems (CLEWs) framework: a retrospective of activities and advances to 2019,” *Environ. Res. Lett.*, Dec. 2020, doi: 10.1088/1748-9326/abd34f.

[36] S. Shannak, D. Mabrey, and M. Vittorio, “Moving from theory to practice in the water–energy–food nexus: An evaluation of existing models and frameworks,” *Water-Energy Nexus*, vol. 1, no. 1, pp. 17–25, Jun. 2018, doi: 10.1016/j.wen.2018.04.001.

[37] M. Giampietro and K. Mayumi, “Multiple-Scale Integrated Assessment of Societal Metabolism: Introducing the Approach,” *Popul. Environ.*, vol. 22, pp. 109–153, Nov. 2000, doi: 10.1023/A:1026691623300.

[38] Heaps, C.G, “LEAP: The Low Emissions Analysis Platform.” Stockholm Environment Institute, Somerville, MA, USA., 2022. Accessed: Nov. 10, 2023. [Online]. Available: https://leap.sei.org/

[39] J. Sieber, “WEAP (Water Evaluation And Planning).” Stockholm Environment Institute, Somerville, MA, USA. Accessed: Nov. 10, 2023. [Online]. Available: https://www.weap21.org

[40] International Institute for Applied Systems Analysis (IIASA), “Model for Energy Supply Systems and their General Environmental impact-MESSAGE.” Accessed: Nov. 10, 2023. [Online]. Available: https://previous.iiasa.ac.at/web/home/research/researchPrograms/Energy/MESSAGE.en.html

[41] The Institute of Environmental Science and Technology (ICTA-UAB), “Integrated assessment: sociology, technology and the environment (IASTE),” UAB-Universitat Autònoma de. Accessed: Nov. 10, 2023. [Online]. Available: https://www.uab.cat/web/el-centre-icta-uab/grups-de-recerca-icta-uab/arees-de-recerca-icta-uab-1345819914243.html

[42] A. Vinca *et al.*, “The NExus Solutions Tool (NEST) v1.0: an open platform for optimizing multi-scale energy–water–land system transformations,” *Geosci. Model Dev.*, vol. 13, no. 3, pp. 1095–1121, Mar. 2020, doi: 10.5194/gmd-13-1095-2020.

[43] D. Huppmann *et al.*, “The MESSAGEix Integrated Assessment Model and the ix modeling platform (ixmp): An open framework for integrated and cross-cutting analysis of energy, climate, the environment, and sustainable development,” *Environ. Model. Softw.*, vol. 112, pp. 143–156, Feb. 2019, doi: 10.1016/j.envsoft.2018.11.012.

[44] Fischer, G., Nachtergaele, F.O., van Velthuizen, H.T., Chiozza, F., Franceschini, G., Henry, M., Muchoney, D. and Tramberend, S, *Global agro-ecological zone V4 – Model documentation*. Rome: FAO, 2021. doi: 10.4060/cb4744en.

[45] D. L. Keairns, R. C. Darton, and A. Irabien, “The Energy-Water-Food Nexus,” *Annu. Rev. Chem. Biomol. Eng.*, vol. 7, no. 1, pp. 239–262, 2016, doi: 10.1146/annurev-chembioeng-080615-033539.

[46] M. Welsch *et al.*, “Adding value with CLEWS – Modelling the energy system and its interdependencies for Mauritius,” *Appl. Energy*, vol. 113, pp. 1434–1445, Jan. 2014, doi: 10.1016/j.apenergy.2013.08.083.

[47] V. Sridharan, E. P. Ramos, C. Taliotis, M. Howells, P. Basudde, and I. V. Kinhonhi, “Vulnerability of Uganda’s Electricity Sector to Climate Change: An Integrated Systems Analysis,” in *Handbook of Climate Change Resilience*, W. Leal Filho, Ed., Cham: Springer International Publishing, 2018, pp. 1–30. doi: 10.1007/978-3-319-71025-9_45-1.

[48] J. Dargin, B. Daher, and R. H. Mohtar, “Complexity versus simplicity in water energy food nexus (WEF) assessment tools,” *Sci. Total Environ.*, vol. 650, pp. 1566–1575, Feb. 2019, doi: 10.1016/j.scitotenv.2018.09.080.

[49] M. Welsch, M. Howells, M. Bazilian, J. DeCarolis, S. Hermann, and H. Rogner, “Modelling elements of Smart Grids - Enhancing the OSeMOSYS (Open Source Energy Modelling System) code,” *Energy*, vol. 46, Oct. 2012, doi: 10.1016/j.energy.2012.08.017.

[50] N. V. Emodi, T. Chaiechi, and A. B. M. R. Alam Beg, “Are emission reduction policies effective under climate change conditions? A backcasting and exploratory scenario approach using the LEAP-OSeMOSYS Model,” *Appl. Energy*, vol. 236, pp. 1183–1217, Feb. 2019, doi: 10.1016/j.apenergy.2018.12.045.

[51] M. Howells *et al.*, “OSeMOSYS: The Open Source Energy Modeling System: An introduction to its ethos, structure and development,” *Energy Policy*, vol. 39, no. 10, pp. 5850–5870, Oct. 2011, doi: 10.1016/j.enpol.2011.06.033.

[52] T. Niet, A. Shivakumar, F. Gardumi, W. Usher, E. Williams, and M. Howells, “Developing a community of practice around an open source energy modelling tool,” *Energy Strategy Rev.*, vol. 35, p. 100650, May 2021, doi: 10.1016/j.esr.2021.100650.

[53] F. Gardumi, M. Welsch, M. Howells, and E. Colombo, “Representation of Balancing Options for Variable Renewables in Long-Term Energy System Models: An Application to OSeMOSYS,” *Energies*, vol. 12, no. 12, Art. no. 12, Jan. 2019, doi: 10.3390/en12122366.

[54] N. Arianpoo, F. SINGH, A. S. WRIGHT, and T. Niet, “BC Nexus Model - Impacts of electrification on land and water resources through 2050,” Simon Fraser University, 2021.

[55] A. Shivakumar, T. Alfstad, and T. Niet, “A clustering approach to improve spatial representation in water-energy-food models,” *Environ. Res. Lett.*, vol. 16, no. 11, p. 114027, Oct. 2021, doi: 10.1088/1748-9326/ac2ce9.

[56] Kamaria Kuling, Trevor Barnes,AbhishekShivakumar, MaartenBrinkerink , Taco Niet, “Applying the open-source climate, land, energy, and water systems (CLEWs) model to Canada,” *Energy Strategy Rev.*, vol. 44, p. 100929, Nov. 2022, doi: 10.1016/j.esr.2022.100929.

[57] T. Barnes, A. Shivakumar, M. Brinkerink, and T. Niet, “OSeMOSYS Global, an open-source, open data global electricity system model generator,” *Sci. Data*, vol. 9, no. 1, Art. no. 1, Oct. 2022, doi: 10.1038/s41597-022-01737-0.

[58] T. Niet and A. Shivakumar, “clewsy: Script for building CLEWs models.” OSeMOSYS, 2020. Accessed: Dec. 04, 2022. [Online]. Available: https://github.com/OSeMOSYS/clewsy

[59] K. Kuling, Y. Saedi, T. Barnes, A. Sunder Rajan, T. Niet, “CLEWs Global: An open source, open data Climate, Land, Energy, and Water systems model generator,” Accessed: Oct. 30, 2023. [Online]. Available: https://summit.sfu.ca/libraries/pdf.js/web/viewer.html?file=%2F%2Fsummit.sfu.ca%2F_flysystem%2Ffedora%2F2023-10%2FCLEWs-Global-IEWConferencePaper_0.pdf

[60] W. Usher, H. Henke, and C. Muschner, “OSeMOSYS/otoole: otoole: OSeMOSYS tools for energy work.” Accessed: Nov. 13, 2023. [Online]. Available: https://zenodo.org/records/4730003

[61] M. Al-Saidi and H. Hussein, “The water-energy-food nexus and COVID-19: Towards a systematization of impacts and responses,” *Sci. Total Environ.*, vol. 779, p. 146529, Jul. 2021, doi: 10.1016/j.scitotenv.2021.146529.

[62] W. Cole *et al.*, “Variable Renewable Energy in Long-Term Planning Models: A Multi-Model Perspective,” *Renew. Energy*, 2017.

[63] Y. Saif and A. Almansoori, “An Optimization Framework for the Climate, Land, Energy, and Water (CLEWS) Nexus by a Discrete Optimization Model,” presented at the The 8th International Conference on Applied Energy, Elsevier, May 2017, pp. 3232–3238. doi: 10.1016/j.egypro.2017.03.714.

[64] M. Weirich, *GLOBAL RESOURCE MODELLING OF THE CLIMATE, LAND, ENERGY AND WATER (CLEWS) NEXUS USING THE OPEN SOURCE ENERGY MODELLING SYSTEM (OSEMOSYS)*. 2013. Accessed: Dec. 25, 2022. [Online]. Available: http://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-131696

[65] E. Fejzic, “Renewable energy outlook for the Drina River Basin countries,” KTH Royal Institute of Technology School of Industrial Engineering and Management, Sweden, 2020.

[66] C. Arderne, “A climate, land-use, energy and water nexus assessment of Bolivia,” KTH School of Industrial Engineering and Management, Sweden, 2016.

[67] Daniel Pastor Pascual, “Development of a GIS Model for Water Accounting in Jordan: Focus on Irrigation and Energy Usage in the Water Sector,” KTH School of Industrial Engineering and Management, 2019.

[68] A. Vinca, K. Riahi, A. Rowe, and N. Djilali, “Climate-Land-Energy-Water Nexus Models Across Scales: Progress, Gaps and Best Accessibility Practices,” *Front. Environ. Sci.*, vol. 9, 2021, Accessed: Jan. 21, 2022. [Online]. Available: https://www.frontiersin.org/article/10.3389/fenvs.2021.691523

[69] Food and Agriculture Organization of the United Nations, “FAOSTAT - Crops and Livestock Products.” Accessed: Oct. 08, 2023. [Online]. Available: https://www.fao.org/faostat/en/#data/QCL

[70] IIASA/FAO, “Global Agro-ecological Zones (GAEZ v3.0).” 2010. Accessed: Nov. 28, 2022. [Online]. Available: https://www.gaez.iiasa.ac.at/

[71] “Global Administrative Areas - GADM.” Ausria, Jul. 2022. Accessed: Oct. 05, 2023. [Online]. Available: https://gadm.org/

[72] One World - Nations Online, OWNO, “nationsonline.org Editor Klaus Kästle.” Accessed: Sep. 18, 2023. [Online]. Available: https://www.nationsonline.org/oneworld/disclaimer.htm

[73] K. T. GmbH (https://www.klokantech.com/), “EPSG.io: Coordinate Systems Worldwide.” Accessed: Sep. 18, 2023. [Online]. Available: https://epsg.io

[74] Sean Gillies, “The Shapely User Manual — Shapely 2.0.1 documentation.” Apr. 07, 2023. Accessed: Sep. 22, 2023. [Online]. Available: https://shapely.readthedocs.io/en/stable/manual.html

[75] J.-P. Carvallo, B. J. Shaw, N. I. Avila, and D. M. Kammen, “Sustainable Low-Carbon Expansion for the Power Sector of an Emerging Economy: The Case of Kenya,” *Environ. Sci. Technol.*, vol. 51, no. 17, pp. 10232–10242, Sep. 2017, doi: 10.1021/acs.est.7b00345.

[76] N. Moksnes, A. Korkovelos, D. Mentis, and M. Howells, “Electrification pathways for Kenya–linking spatial electrification analysis and medium to long term energy planning,” *Environ. Res. Lett.*, vol. 12, no. 9, p. 095008, 2017, doi: 10.1088/1748-9326/aa7e18.

[77] M. Akute and C. Cannone, “Effects of switching from biomass stoves to electric stoves and subsequent reduction in resultant emissions in the Kenyan energy sector,” Dec. 2022, doi: https://doi.org/10.21203/rs.3.rs-2353038/v1.

[78] “The Organisation for Economic Co-operation and Development (OECD).” [Online]. Available: https://stats.oecd.org/Index.aspx?DataSetCode=LAND_COVER#

[79] Food and Agriculture Organization of the United Nations, “AQUASTAT - FAO’s Global Information System on Water and Agriculture.” Accessed: Dec. 16, 2023. [Online]. Available: https://www.fao.org/aquastat/en/

[80] “OSeMOSYS - Open-source Energy Modelling System,” GitHub. Accessed: Dec. 04, 2023. [Online]. Available: https://github.com/OSeMOSYS

[81] Will Usher, “OSeMOSYS · GitHub.” Accessed: Oct. 01, 2023. [Online]. Available: https://github.com/OSeMOSYS

Appendix A.

GeoCLEWs Release

The following delineates the details of releasing GeoCLEWs on GitHub to make it freely accessible to everyone and contribute to its development. GeoCLEWs is published on a public platform allowing all users regardless of their affiliation or geographic location to benefit from this tool for a wide variety of projects and collaborate on sustainable developments. GitHub is chosen being a proper platform to distribute GeoCLEWs as hosting a repository is an effective method to share scripts and vital datasets. GitHub is commonly used for distributing open source software and tools. Successful experience of user contribution and developer collaboration played an important role in choosing GitHub as the main platform to publish the results of this research. Furthermore, the OSeMOSYS is available on GitHub [81] with more than 20 repositories. The fact that GeoCLEWs has been granted permission to be published on the main GitHub is greatly valuable and brings two major advantages. First, it contains several open source tools and datasets that are essential or related to GeoCLEWs such as clewsy and Otoole. Second, there is an active community on this GitHub that effectively supports developers and users to reproduce, share, track changes, and contribute to development. GitHub, a well-known platform for hosting source code, makes version control and revision tracking easier, which encourages collaboration.

A new repository is created inside the OSeMOSYS GitHub page for storing GeoCLEWs and supplementary documents. It is named CLEWs_GAEZ, which refers to the idea of processing GAEZ v4 land and water data to create detailed CLEWs models. Then the configuration setup is defined, and the new repository is publicly available. The CLEWs-GAEZ encompasses the developed Jupyter Notebook and complementary documents and datasets.

- LICENSE: GeoCLEWs is published under the MIT license providing permissions for commercial and private use, modification, and distribution aligned with the original code license. The licence is provided on the main page specifying details of permissions.
- README.md: It serves as a brief documentation providing contributors with information on project functionality and setup instructions simplifying comprehending and applying GeoCLEWs.

- Environment.yml: An environment file in YAML format is generated to simplify and speed up the process of building the necessary environment for running GeoCLEWs. Contributors can use the YAML file to create a proper Conda environment and install all required Python packages to run the Jupyter Notebook smoothly.

- GEAZ_Processing: This folder includes GeoCLEWs script and all essential complementary datasets to extract FAOSTAT and GAEZ raster files automatically. Forking the GitHub repository creates a copy of the entire CLEWs_GAEZ inside the users' local system. The structure of GAEZ_Processing folder should remain unchanged to avoid any issues with running the script. The Data folder hosts input and output files. The administrative boundary Shapefiles should be stored inside the Input folder. For clarification, two examples of shapefiles for admin level 0 and user-defined admin level are provided in order to assist users in following the naming format. In addition, the global_raster_input hosts the GAEZ raster files that will be downloaded during GeoCLEWs process. However, the land cover and precipitation raster files in TIFF format are extracted from the GAEZ v4 portal, preprocessed, and stored inside this folder, which is useful to fully automate the land and water processing for CLEWs modelling.

Appendix B.

FAOSTAT Statistics – Partially Presented

Area	Element	Item	Year	Unit	Value	Flag Description	Area
Kenya	Area harvested	Maize (corn)	2020	ha	2135741	Official figure	Kenya
Kenya	Area harvested	Beans, dry	2020	ha	1147709	Official figure	Kenya
Kenya	Area harvested	Tea leaves	2020	ha	269400	Official figure	Kenya
Kenya	Area harvested	Cow peas, dry	2020	ha	239131	Official figure	Kenya
Kenya	Area harvested	Sorghum	2020	ha	219657	Official figure	Kenya
Kenya	Area harvested	Other pulses n.e.c.	2020	ha	198972	Estimated value	Kenya
Kenya	Area harvested	Potatoes	2020	ha	176252	Official figure	Kenya
Kenya	Area harvested	Pigeon peas, dry	2020	ha	133525	Official figure	Kenya
Kenya	Area harvested	Wheat	2020	ha	132231	Official figure	Kenya
Kenya	Area harvested	Coffee, green	2020	ha	119700	Official figure	Kenya
Kenya	Area harvested	Millet	2020	ha	118411	Official figure	Kenya

Appendix C.

GeoCLEWs Crop Naming

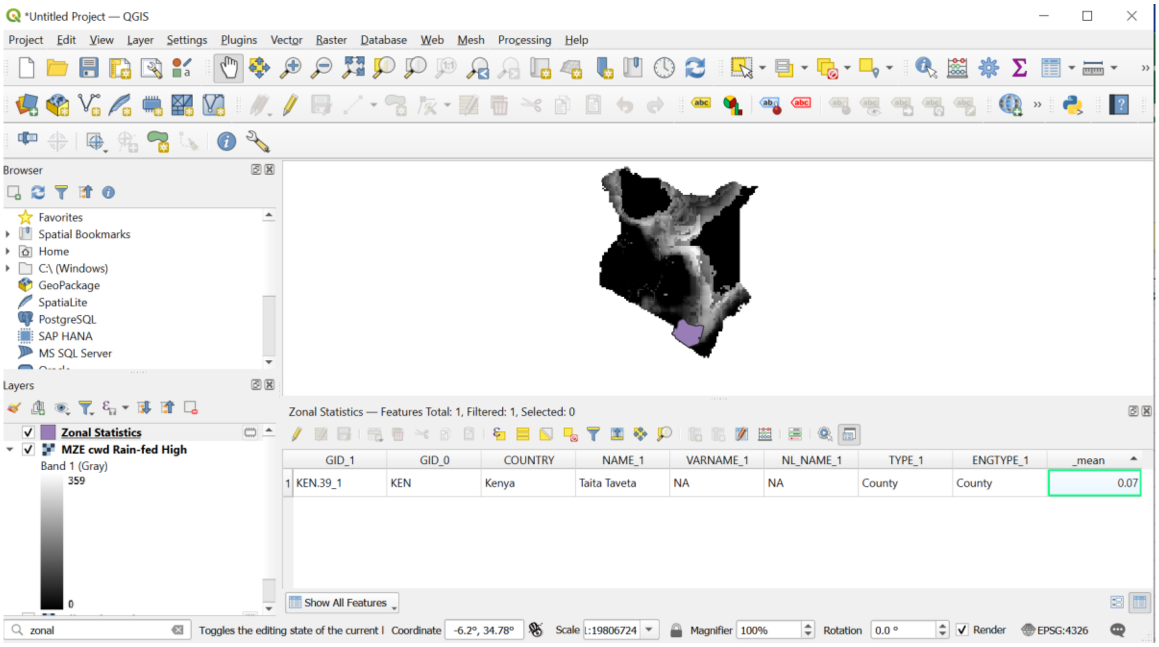
GeoCLEWs Crop Name	FAOSTAT Crop Name	GAEZ Crop Name
ALF	Alfalfa	Alfalfa
ARE	Arecanut	Arecanut
BAN	Bananas	Banana
BRL	Barley	Barley
BEA	Beans, dry	Phaseolus bean
BMX	BMX	BMX
BUC	Buckwheat	Buckwheat
CAB	Cabbages	Cabbage
CRD	Cardamom	Cardamom
CAR	Carrots and turnips	Carrot
CAS	Cassava, fresh	Cassava
CER	Cereals n.e.c.	Cereals
CHI	Chick peas, dry	Chickpea
COC	Cocoa beans	Cocoa
CON	Coconuts, in shell	Coconut
COF	Coffee, green	Coffee
COW	Cow peas, dry	Cowpea
RCD	Dryland rice	Dryland rice
FLA	Flax, processed but not spun	Flax
MTF	Foxtail millet	Foxtail millet
FRU	Fruits	Fruits
GRM	Gram	Gram
GRO	Groundnuts, excluding shelled	Groundnut
JAT	Jatropha	Jatropha
MZE	Maize (corn)	Maize
MIS	Miscanthus	Miscanthus
OAT	Oats	Oat
OIL	Oil palm fruit	Oil palm
OLI	Olives	Olive
ONI	Onions and shallots	Onion
CIT	Other citrus fruit, n.e.c.	Citrus
SGB	Other sugar crops n.e.c.	Sugarbeet

GeoCLEWs Crop Name	FAOSTAT Crop Name	GAEZ Crop Name
MTP	Pears	Pearl millet
PEA	Peas, dry	Dry pea
PIG	Pigeon peas, dry	Pigeonpea
PTW	Potatoes	White potato
RAP	Rape or colza seed	Rapeseed
REE	Reed	Reed canary grass
RUB	Rubber	Rubber
RYE	Rye	Rye
COT	Seed cotton, unginned	Cotton
SOR	Sorghum	Sorghum
SOY	Soya beans	Soybean
SGC	Sugar cane	Sugarcane
SUN	Sunflower seed	Sunflower
PTS	Sweet potatoes	Sweet potato
SWI	Switchgrass	Switchgrass
TEA	Tea leaves	Tea
TOM	Tomatoes	Tomato
TOB	Unmanufactured tobacco	Tobacco
VEG	Vegetables	Vegetables
TEF	Warm C4	Warm C4
RCP	Wetland rice	Wetland rice
WHE	Wheat	Wheat
YAM	Yams	Yam
MLT	Millet	Millet

Appendix D.

GeoCLEWs Validation

Calculating mean values of maize crop water deficit raster in TTC using QGIS.



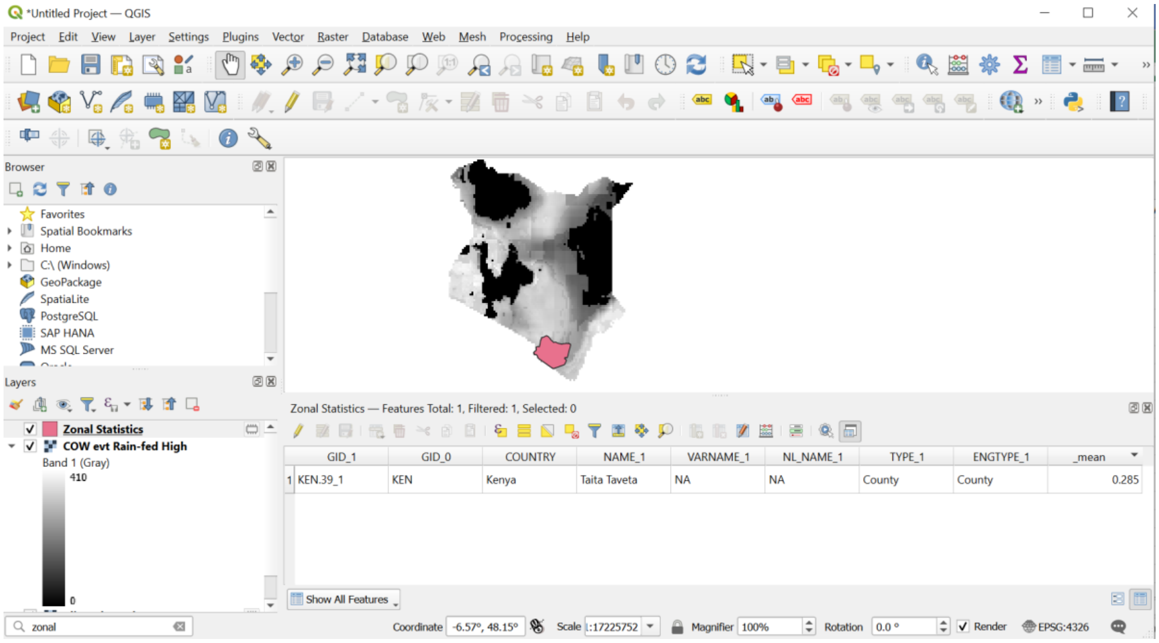
Layers: ☑ Zonal Statistics; ☑ MZE cwd Rain-fed High — Band 1 (Gray) — 359 / 0

Zonal Statistics — Features Total: 1, Filtered: 1, Selected: 0

	GID_1	GID_0	COUNTRY	NAME_1	VARNAME_1	NL_NAME_1	TYPE_1	ENGTYPE_1	_mean
1	KEN.39_1	KEN	Kenya	Taita Taveta	NA	NA	County	County	0.07

Toggles the editing state of the current l Coordinate -6.2°, 34.78° Scale 1:19806724 Magnifier 100% Rotation 0.0 ° ☑ Render EPSG:4326

Processing cowpea evapotranspiration with rain-fed and high input using QGIS.



Layers: ☑ Zonal Statistics; ☑ COW evt Rain-fed High — Band 1 (Gray) — 410 / 0

Zonal Statistics — Features Total: 1, Filtered: 1, Selected: 0

	GID_1	GID_0	COUNTRY	NAME_1	VARNAME_1	NL_NAME_1	TYPE_1	ENGTYPE_1	_mean
1	KEN.39_1	KEN	Kenya	Taita Taveta	NA	NA	County	County	0.285

Coordinate -6.57°, 48.15° Scale 1:17225752 Magnifier 100% Rotation 0.0 ° ☑ Render EPSG:4326