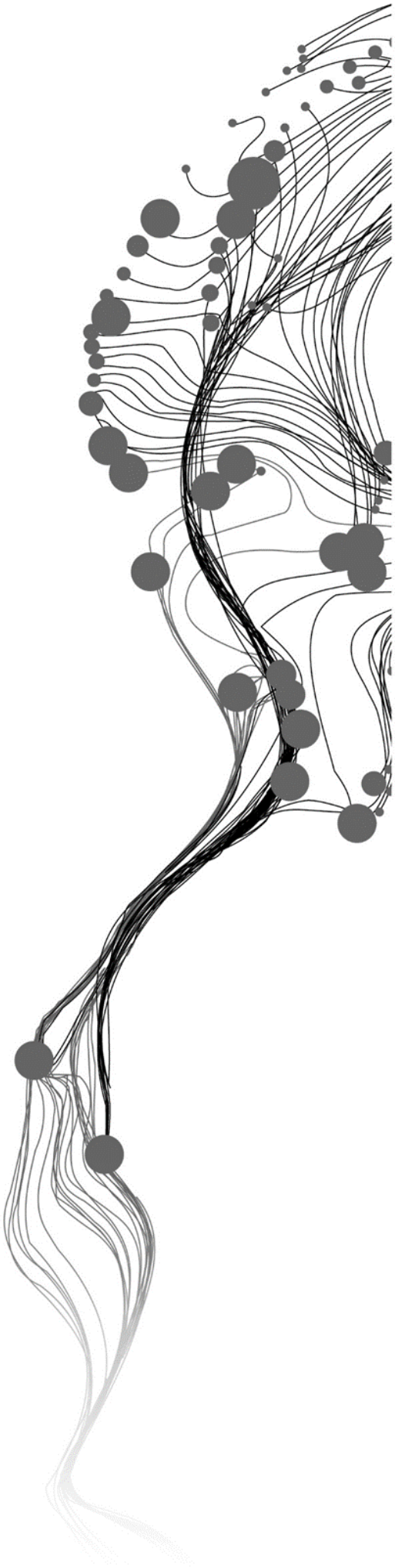
Triple Sensor Approach for Monitoring Water and Climate.

JEAN DE LA CROIX, ISHIMWE
July, 2020

SUPERVISORS:
Dr.ir. C.M.M. Mannaerts (Chris)
Ir. G.N. Parodi (Gabriel)



Triple Sensor Approach for Monitoring Water and Climate.

JEAN DE LA CROIX, ISHIMWE
Enschede, The Netherlands, July, 2020

Thesis submitted to the Faculty of Geo-Information Science and Earth Observation of the University of Twente in partial fulfilment of the requirements for the degree of Master of Science in Geo-information Science and Earth Observation.
Specialization: Water Resources and Environmental Management.

SUPERVISORS:
Dr.ir. C.M.M. Mannaerts (Chris)
Ir. G.N. Parodi (Gabriel)

THESIS ASSESSMENT BOARD:
Dr.ir. C.M.M. Mannaerts (First Supervisor)
Ir. G.N. Parodi (Second Supervisor)
Dr. Ing. T.H.M. Rientjes (Chair)
Dr.ir. R.L.G. Lemmens, (External Examiner, Department of Geo-information Processing)

ABSTRACT

Despite the proliferation of sophisticated climate variables measurement techniques, the resulting datasets remain subject to errors due to the complex nature of the variables being estimated. This raises a formidable challenge in water management domains where proper decisions depend heavily on the reliability of the datasets. In this study, the triple sensor approach for the monitoring of water and climate was investigated by testing climatic variables from mutually independent sources on the extended triple collocation method to establish the conditions for its effective applicability for the identification of reliable climatic datasets.

The factors affecting rainfall measurement systems (as the main research variable) were described and the differences leading to reliability concerns about the resulting datasets were appreciated. Further, datasets of rainfall, short-wave incoming solar radiation, and 2m air temperature were statistically engaged in different set-ups using the extended triple collocation covariance analysis and the results were used to judge the most reliable at several locations in and around the Lake Naivasha basin in Kenya. Some of the used datasets were obtained from independent citizen observers in the study area, to emphasize the potential of their involvement in data collection especially in data scarce areas.

It was concluded that the triple sensor approach allows a more absolute standpoint to assess the reliability of datasets of many climatic variables at a given location compared to the usual dual dataset comparison used for validation except for situations where the triple collocation initial assumptions are violated. Data ranges, variability and size of the time-series were noted as factors contributing to a potential violation of the assumptions and consequently resulting in biased outcomes which cannot be used decisively about the corresponding datasets reliability. The ability of the triple sensor approach to penalize datasets with outliers, as a rather a frequently observed characteristic in climatic datasets, was demonstrated for rainfall datasets. While the triple sensor approach has potential, it cannot always be used as a substitute for other data filtering techniques such as the identification of missing data values before any statistical analysis for dataset validation. It was also noted that the data requirements of the triple sensor approach also limit its application.

Lastly, the triple sensor toolbox implemented in ILWIS, as an important communicative tool to stakeholders in scientific research, was tested and suggestions for the improvement of its usability, presentation, and interpretation of its results were outlined.

KEYWORDS: Extended triple collocation, Triple sensor, Covariance analysis, error variance, correlation coefficient, signal-to-noise-ratio, satellite observations, in-situ measurements, climatic model, citizen observer, reliability.

ACKNOWLEDGMENTS

First and foremost, I would like to extend my gratitude to the Dutch Government which enabled me through the Orange Knowledge Program to pursue a Masters's course at the University of Twente.

I would like to thank also the student affairs and Ir. A.M Arno Van Lieshout for your moral support and always ensuring a good study environment.

To Dr. Ir. Chris Mannaerts and Ir. Gabriel Parodi, thank you for supervising my work with interest and allowing me to benefit from your vast experience.

To Ir. Bas Retsios, thanks for always being available when I needed help.

Special thanks also go to the Trans-African Hydro-Meteorological Observatory (TAHMO) for allowing access to their datasets in the study area. The datasets have been tremendously helpful.

To Sammy, Helen and Dr Joost your support during the fieldwork is appreciated.

To my classmates and friends met at ITC, thanks for the lovely moments we shared. I am very glad I met you.

TABLE OF CONTENTS

1. INTRODUCTION 1
1.1. Background 1
1.2. Problem statement 2
1.3. Research Objectives 2
1.4. Research questions 3
1.5. Scientific and societal justification 3
2. LITERATURE REVIEW 4
2.1. Description of precipitation estimation techniques and associated limitations 4
2.2. Backgrounds of the standard triple collocation and the extended triple collocation methods 6
2.3. The general principals of the triple collocation methods 6
3. METHODOLOGY 10
3.1. Flowchart 10
3.2. An overview of the main activities 10
3.3. An overview of the tools 12
4. STUDY AREA AND DATASETS 13
4.1. Study area 13
4.2. Research datasets 16
5. RESULTS PRESENTATION AND DISCUSSION 39
5.1. Root mean square error and correlation coefficient results 39
5.2. Results interpretation and discussion 44
5.3. The triple sensor ILWIS toolbox 45
6. CONCLUSIONS AND RECOMMENDATIONS 48
6.1. Conclusions 48
6.2. Recommendations 50
7. APPENDIX 55

LIST OF FIGURES

Figure 3.1: Methodology flowchart. 10
Figure 4.1: Study area map. 13
Figure 4.2: Nunjoro farm weather station. 15
Figure 4.3: Spatial visualization of CHIRPS monthly rainfall(January 2018). 17
Figure 4.4: Spatial visualization of ERA5 monthly rainfall(January 2018). 18
Figure 4.5: Daily rainfall at Moi Forces Academy and Ole Tipis GHS. 19
Figure 4.6: Daily rainfall at Delamere farm and Taita Mauche SS. 20
Figure 4.7: Pentad rainfall at Moi Forces Academy and Ole Tipis GHS. 21
Figure 4.8: Pentad rainfall at Delamere farm and Taita Mauche SS. 22
Figure 4.9: Dekad rainfall at Moi Forces Academy and Ole Tipis GHS. 23
Figure 4.10: Dekad rainfall at Delamere farm and Taita Mauche SS. 24
Figure 4.11: Original and Altered TAHMO In-situ rainfall datasets at Ole Tipis GHS. 28
Figure 4.12: Sample ERA5 SW solar radiation spatial visualization. 30
Figure 4.13: Sample CERES SW radiation spatial variability. 31
Figure 4.14: 2017 SW incoming solar radiation and Karima GHS and Moi Forces Academy. 32
Figure 4.15: 2018 SW incoming solar radiation at Karima GHS and Moi Forces Academy. 33
Figure 4.16: 2019 SW incoming solar radiation at Karima GHS and Moi Forces academy. 34
Figure 4.17: Spatial visualization of ERA5 air temperature(January 2018) 36
Figure 4.18: Sample spatial variability. 37
Figure 4.19: Monthly air temperature at Ole Tipis GHS and Moi Forces Academy. 37
Figure 5.1: Triple sensor data input. 45
Figure 5.2: ETC results initial attributes table. 46
Figure 5.3: ETC final attribute table with colour values. 47
Figure 5.4: ETC results visualization. 47

LIST OF TABLES

Table 3.1: Triplets formation example. 11
Table 4.1: Correlation coefficient analysis, rainfall datasets(2018&2019). 25
Table 4.2: Correlation coefficient analysis, PCP datasets for 2018&2019 combined. 26
Table 4.3: Calculation of upper and lower limits. 27
Table 4.4: Correlation analysis-altered dataset. 28
Table 4.5: Correlation analysis in the dataset with missing data. 29
Table 4.6: Correlation analysis, radiation datasets(2017,2018, and 2019 separately.) 35
Table 4.7: Correlation analysis-daily hourly SW incoming radiation(2017,2018,2019 combinedly) 35
Table 4.8: Correlation analysis, air temperature. 38
Table 5.1: ETC results for rainfall(2018&2019). 39
Table 5.2: ETC results for rainfall(2018&2019 combined time-series). 40
Table 5.3: ETC results for the dataset with outliers 41
Table 5.4: ETC results for the dataset with missing values. 41
Table 5.5: ETC results for solar radiation. 42
Table 5.6: ETC results for solar radiation(2017,2018 and 2019 combinedly). 42
Table 5.7: ETC results for air temperature. 43
Table 7.1: Ground weather stations coordinates. 55

LIST OF ACRONYMS

AWS	Automated Weather Stations
CAMS	Climate Anomaly Monitoring System
CCD	Cold Cloud Duration.
CERES	Clouds and Earth's Radiant Energy System
CHIRPS	Climate Hazards Group InfraRed Precipitation with Station Data
CPC	Climate Prediction Center
CSV	Comma Separated Values
ECMWF	European Centre for Medium-Range Weather Forecasts
ETC	Extended Triple Collocation.
GDAL	The Geospatial Data Abstraction Library
GEO	Geostationary Satellites
GHCN	Global Historical Climatology Network
GIS	Geographical Information System.
GRIB	General Regularly-distributed Information in Binary form.
ILWIS	The Integrated Land and Water Information System.
ISOD	The "In Situ and Online Data Toolbox
LEO	Low Earth Orbit
MARS	Meteorological Archival and Retrieval System
MODIS:	Moderate Resolution Imaging Spectroradiometer
NASA	National Aeronautics and Space Administration
NCEP	National Centers for Environmental Prediction
NetCDF	Network Common Data Form
NOAA	National Oceanic and Atmospheric Administration
RMSE	Root Mean Square Error
TAHMO	The Trans-African Hydro-Meteorological Observatory.
TC	Triple Collocation.
TOA	Top Of Atmosphere
UbSNR	Unbiased Signal-To-Noise ratio.

1. INTRODUCTION

1.1. Background

One of the most key aspects of climate monitoring is the necessity for reliable climatic datasets because information about their temporal and spatial distributions have a consequential effect in water resources management decisions (Ebert, Janowiak, & Kidd, 2007). For example, the paramount role of accurate quantification of precipitation as an essential climate variable on local and global scales can never be neglected in domains such as water cycle studies and agriculture. Several climate data sources using sophisticated instruments and methods have existed and improved over time but remain, nonetheless, error-prone in the face of the complex nature of climate evolution and inter-woven physical processes.

For instance, In situ measurements of climatic variables always represent a specific point in space where the sensor is located which may not necessarily be sufficiently representative of the entire area of interest. While satellite and radar-based observations allow larger spatial coverages, they are subject to equipment malfunctions or algorithm failure which results in data gaps. Alternatively, numerical climate models are another important source of climate data. Besides the complex equations used in numerical models to produce climate datasets, observations errors emanating from model initial conditions can undoubtedly propagate into model outputs (Tolstykh & Frolov, 2005).

Most of the available knowledge about climatic variables occurrence and their distribution is based on data disseminated by national and international level meteorological organizations. However, another source of data that is seemingly emerging fast in scientific research is the data obtained from independent citizen observers. Silvertown, (2009) outlines independent observers as independent data collection, analysis, and dissemination of the data for scientific endeavors and public engagement in scientific discussion. In climate sciences, independent observers' data has been acknowledged as an alternative solution for data scarcity because they enable the acquisition of large amounts of data with high spatial-temporal resolution essential for the improvement of our understanding of the environment. (Muller et al., 2015).

On the other hand, skepticism in regards to independent observers datasets lies mainly in data quality, knowledge, the willingness of the independent observers, as well as poorly documented operational processes. Despite the limitations, datasets obtained from trained independent observers, have proven to be important in various climatology studies (Eney & Petzold, 1988).

1.2. Problem statement

With the massive amounts of climatic datasets presently available, users in relevant domains require robust data filtering and quality assessment methods which allow sufficient information about possible errors in datasets obtained from different sources before admissibility in scientific applications. A conscientious selection concerning the exigency of the intended application is therefore required, considering that a method that can account for all characteristics between the measurement system and the target variable does not exist (Entekhabi, Reichle, Koster, & Crow, 2010).

The triple sensor monitoring of water and climate is an approach based on the statistical engagement of 3 climatic datasets obtained using mutually independent measurement systems (sensors) to determine the most reliable at a given location and time. It is rooted in the standard Triple Collocation(TC) and Extended Triple Collocation(ETC) statistical methods introduced for geophysical measurement systems calibration and validation when no reliable datasets are available for normal dual comparison. The methods provide, through a covariance analysis of three climatic datasets, useful information about error distribution in the collocated datasets which can be used to judge the fidelity of the corresponding measurement systems.

While the general rationale behind the triple sensor approach may seem straightforward, several constraints exist and it is, therefore, important to scrutinize it in different set-ups and further characterize the conditions of its appropriateness for climatic datasets reliability evaluation.

1.3. Research Objectives

1.3.1. Main objective

The primary objective of this research is to apply the Extended Triple Collocation method to determine the most reliable climatic data source between In-measurements, Satellite-based observations, and model-based data at a particular location and time.

1.3.2. Specific objectives:

- To describe the nature of rainfall measurement systems(as the main research variable) and identify the factors affecting the reliability of the resulting datasets.
- Test datasets obtained from the previously-mentioned measurement systems on the ETC method and evaluate their performance for rainfall, solar radiation, and surface air temperature.
- Analyze the effects of different time aggregation levels, the presence of outliers, and missing data on the ETC analysis outcomes for the used datasets.
- Contribute further to the improvement of the triple sensor toolbox currently implemented in ILWIS.

1.4. Research questions

- How do the limitations of existing rainfall estimation methods (In-situ measurements, Satellite-based observations, and climatic models) affect the reliability of the resulting datasets?
- How can the triple sensor approach be used for the evaluation of climatic datasets' reliability and what are the associated constraints?
- How can the performance of the triple sensor approach currently implemented in ILWIS be improved?

1.5. Scientific and societal justification

Geophysical measurement verification methods for data quality assessment are meaningful if they can provide decisive information about the level of reliability that can be associated with the datasets being investigated (Stanski, Wilson, & Burrows, 1989). The expected outcome of this study is the identification of the appropriate conditions required for efficient use of the triple sensor approach to determine the most reliable climatic data sources at a given location and time, as novel scientific use of the triple collocation technique.

In a more societal context, the outcomes of this work will further improve the quality of information for public use in related domains and it will raise more awareness to users who would normally rely solely on a single data source. It will also emphasize the efficacy of independent citizen climatic data collectors especially in countries where measurements are scarce.

2. LITERATURE REVIEW

2.1. Description of precipitation estimation techniques and associated limitations

Precipitation is any form of water particles(solid or liquid) that fall from the atmosphere and reach the earth's surface. When masses of warm air and moist air encounter masses of cold air, the formation of droplets, which may become rain(for instance in the current study area), crystals or snow, occurs. When the formed droplets become too heavy, they precipitate towards the earth's surface. Precipitation is one of the most essential climatic variables and it plays a crucial role during the characterization of the climate and its changes. In studies such as Watson & Challinor, (2013), erroneous rainfall datasets have been noted to be detrimental to crop productivity model robustness. In the following sub-sections, methods used by the three main sources of rainfall datasets are described.

2.1.1. In-Situ rainfall measurements

In-situ rainfall measurements are performed by a rain gauge which is usually placed at an open location to collect rainwater, rainfall is expressed as the height of the accumulated water in millimeters. A typical rain gauge is a tipping bucket; it consists of an opening that allows water inside, and the water is collected by two small buckets in an alternating sequence. Every bucket tipping is recorded electronically, the amount of rainfall at the location of the instrument will be the product of the amount of water required for a bucket to tip and the number of recorded tippings.

According to Kim et al., (2014), rain gauge data are the most accurate representation of rainfall at a precise location, but mechanical problems or operational inefficiencies may introduce errors. Various errors in datasets can occur because of the destruction of the gauge by animals or even humans. Kidd & Huffman, (2011) explain that gauge data often underestimate precipitation amounts because of wind effects or rain particles that evaporate before reaching the gauge. Furthermore, the underestimation will intensify if the rain gauge is located under dense vegetation (for example in a forest with dense crowns) and consequently most rainwater will be intercepted and evaporated before reaching the gauge. Another critical aspect is that rain gauges give spot samples of rainfall which cannot be sufficiently representative in areas with heterogeneous rainfall distribution.

2.1.2. Satellite-based rainfall estimates

Satellite-based observations provide datasets with high spatial and temporal resolutions over the span of their orbits depending on the characteristics of rainfall in a given region, and they are the only instruments whose measurements can be used to obtain homogeneous rainfall estimates over a given area (Tapiador et al., 2012).

In the visible region (VIS) of the spectrum, the sensors allow the distinction between clouds and the surface due to the higher brightness associated with clouds. The resulting imagery can be used to distinguish between different types of clouds and their brightness can be related to the occurrence of rainfall. However, the VIS imagery relies on sunlight and can therefore only be obtained during the day. Another important limitation is that the relationship between cloud brightness and precipitation at the surface is often too low. (Kidd & Huffman, 2011).

Thermal InfraRed sensors with high spatial and temporal resolutions measure clouds temperatures and observations can be performed during day and night, and when mounted on Geostationary satellites, they can achieve vast areal coverages (Dembélé & Zwart, 2016). The assumption used by infrared measurements is that low cloud temperatures suggest massive development of clouds in the vertical which can be related to rainfall using the Cold Cloud Duration(CCD) method. Unfortunately, cold clouds do not necessarily precipitate and may often have multiple layers that are imperceptible to the infrared sensors.

A more direct approach to producing more reliable rainfall estimates from satellites is the microwave region of the electromagnetic. Microwave sensors provide information about the depth of the clouds and layer characteristics, which are then converted to the formation of precipitation (Duan, Liu, Tuo, Chiogna, & Disse, 2016). The principle of Passive Microwave methods is that the surface radiates energy which is affected by water vapor and precipitation particles present in the atmosphere before reaching the sensor. The received signal will be mixed with the radiometric component of the emission and scattering of rain particles in the clouds. For the active microwave sensors, a pulse is emitted by a radar instrument from the satellite towards the clouds and the surface, the returning signal will be attenuated by particles in the atmosphere. Microwave rainfall products are obtained by signal attenuation-correction algorithms(Kidd & Huffman, 2011). The main disadvantage of microwave sensors is the low spatial and temporal resolution because they are onboard Low Earth Orbiting(LEO) satellites.

2.1.3. Model-based rainfall data

Another vital source of rainfall data is numerical climatic models such as the European Centre for Medium-Range Weather Forecasts or the Global Forecast System. They are computer algorithms that attempt to study the transformation of the state of the atmosphere by modeling its interlinked physical processes using complex mathematical equations. Climatic models require datasets of numerous climatic variables for the definition of model initial conditions in a process known as Data Assimilation (Malardel Sylvie, 2019). Different models exist depending on their purpose or coverage (global, national, or local), spatial, and temporal resolutions.

One of the methods used by climatic models is the 4-D variational data assimilation which combines observational data, obtained using techniques such as satellite-based observations and in-situ measurements,

with model-generated forecasts in an attempt to reduce a cost function between the two data sources and reach an optimal fit (Apte, 2015). The imperfections in observational datasets will inevitably be reproduced in the climatic model outputs. As climatic models rely on mathematical equations, another issue is how accurately can science model the intricate laws that govern the evolution of the atmosphere.

2.2. Backgrounds of the standard triple collocation and the extended triple collocation methods

The triple collocation introduced by Stoffelen, (1998), is a powerful statistical method that is used to estimate root mean square errors of at least three independent geophysical measurement systems without considering any of them as a reference. It presupposes a linear error model where errors in measurement systems are unrelated to each other and the target variable, which is valuable for the quantification of error structures in measurements when the true quantity of the target variable is unknown (Vogelzang & Stoffelen, 2012). The triple collocation method has arisen from the need to combine measurements performed from ground stations and those generated by models or satellites. Some of the notable examples of its application are error estimation and calibration of scatterometer winds, soil moisture retrievals, and characterization of spatial and temporal error patterns in precipitation datasets.

While the triple collocation method is useful for quantifying just one system performance metric (the RMSE), other more robust statistical performance metrics are needed for the calibration and validation of various geophysical variables. The extended triple collocation introduced by McColl et al., (2014), uses the same hypothesis and assumptions as the standard triple collocation to produce an unbiased signal-to-noise ratio and correlation coefficients between the collocated measurement systems and the unknown true quantity of the target variable.

2.3. The general principals of the triple collocation methods

2.3.1. The standard triple collocation analysis

The triple collocation of independent datasets is based on the availability of three independent (spatially and temporally collocated) measurement systems which attempt to estimate the same geophysical variable, whereby each of the measurement systems is related to the unknown true quantity by the following error model:

$$\boldsymbol{X_i = \alpha_i + \beta_i * T + \varepsilon_i} (i \in \{1,2,3\}) \quad \textbf{\textit{(Equation 1)}}$$

Where:

- X_i $(i \in \{1,2,3\})$ representing the measurement systems.
- **T** is the part of the signal common to the 3 collocated measurement systems (the unknown truth).
- α and β are least-squares additive and multiplicative bias terms.

- ε is the zero-mean random error term in each measurement system.

In the same manner, the remaining two measurement systems can also be related to the unknown true quantity of the target variable. The 3 measurement systems could be represented by In-situ measurements obtained by ground stations, a remotely sensed dataset such as Satellite-based observations, a model generated datasets and all represent the same system property.

Further, some assumptions concerning the statistical properties of the error terms are necessary for the triple collocation analysis:

- **Error orthogonality:** This assumption signifies that the errors in the measurement systems are uncorrelated to the true quantity of the target variable, such that $(\boldsymbol{Cov}(\boldsymbol{\varepsilon_i}, \boldsymbol{t}) = 0)$
- **Independence between the measurements:** This signifies that the errors in the measurement systems are not correlated with each other $(\boldsymbol{Cov}(\boldsymbol{\varepsilon_i}, \boldsymbol{\varepsilon_j}) = 0, i \neq j)$.
- **Stationarity of error statistics:** The signal and error properties of the available measurement samples are representative and can, therefore, be extrapolated (variance homoscedasticity and zero-mean error).

The covariances between measurement systems can be expressed as:

$$\boldsymbol{Cov}(\boldsymbol{X_i}, \boldsymbol{X_j}) = E(\boldsymbol{X_i}, \boldsymbol{X_j}) + E(\boldsymbol{X_i},)E(\boldsymbol{X_j}) = \boldsymbol{\beta_i \beta_j \sigma_t^2} + \boldsymbol{\beta_i cov}(\boldsymbol{t}, \boldsymbol{\varepsilon_j}) + \boldsymbol{\beta_j cov}(\boldsymbol{t}, \boldsymbol{\varepsilon_i}) + \boldsymbol{cov}(\boldsymbol{\varepsilon_i}, \boldsymbol{\varepsilon_j})$$

Whereby:

$\mathrm{Var(t)} = \boldsymbol{\sigma_t^2}$. Since we assume no error cross-correlation between measurements systems, and again no correlation between measurement systems' errors and the unknown true quantity of the target variable, the two middle terms of the covariance equation will vanish, and the same applies to the last term of the equation when $\boldsymbol{i \neq j}$ $(i, j \in \{1,2,3\})$. From that, the covariance equation will be as follows:

$$Q_{ij} \equiv cov(\boldsymbol{X_i}, \boldsymbol{X_j}) = \begin{cases} \boldsymbol{\beta_i \beta_j \sigma_t^2} & \boldsymbol{for\ i \neq j} \\ \boldsymbol{\beta_i^2 \sigma_t^2 + \sigma_{\varepsilon_i}^2}, & \boldsymbol{for\ i = j} \end{cases}$$

Where:

$\boldsymbol{var}(\boldsymbol{\varepsilon_i}) = \boldsymbol{\sigma_{\varepsilon_i}^2}$

Now we have six terms in the covariance matrix, and six equations with six unknowns and the system cannot be determined.

If we drop solving for $\boldsymbol{\beta_i}$ and $\boldsymbol{\sigma_t^2}$ and introduce a variable $\boldsymbol{\theta_i = \beta_i \sigma_t}$ the covariance will reduce to:

$$Q_{ij} \equiv cov(\boldsymbol{X_i}, \boldsymbol{X_j}) = \begin{cases} \boldsymbol{\theta_i \theta_j} & \boldsymbol{for\ i \neq j} \\ \boldsymbol{\theta_i^2 + \sigma_{\varepsilon_i}^2}, & \boldsymbol{for\ i = j} \end{cases} \quad \textbf{(}\boldsymbol{Equation\ 2}\textbf{)}$$

We reach the triple collocation estimation of the root mean RMSEs for the measurement systems because, with six equations and six unknowns, the system becomes solvable. The root mean square errors are as follows:

$$\boldsymbol{\sigma_{\varepsilon}} = \begin{bmatrix} \sqrt{Q_{11} - \frac{Q_{12}Q_{13}}{Q_{23}}} \\ \sqrt{Q_{22} - \frac{Q_{12}Q_{23}}{Q_{13}}} \\ \sqrt{Q_{33} - \frac{Q_{13}Q_{23}}{Q_{12}}} \end{bmatrix} \textit{(Equation 3)}$$

2.3.2. The extended triple collocation analysis

The extended triple collocation builds on the standard triple collocation assumptions to estimate the correlation coefficients of the measurement systems' estimates with respect to the unknown truth. The following steps for the derivation of the extended triple collocation method were introduced and demonstrated by McColl et al., (2014):

Knowing that the relationship between the slope (β_i), the correlation between a measurement system ($\mathbf{X}_i$) and the truth ($\boldsymbol{\rho_{t,X_i}}$) is expressed as:

$$\boldsymbol{\beta_i = \rho_{t,X_i} \frac{\sqrt{Q_{ii}}}{\sigma_t}} \quad \textit{(Equation 4)}$$

Where:

- $\boldsymbol{\rho_{t,X_i}}$ is the correlation coefficient between the unknown truth($\mathbf{t}$) and a measurement system ($\boldsymbol{X_i}$), and maintaining that that the unknown truth $\mathbf{t}$ has no measurement error.

From equation 4, we can obtain $\boldsymbol{\theta_i = \rho_{t,X_i}\sqrt{Q_{ii}}}$,

Where:

- $\sqrt{\boldsymbol{Q_{ii}}}$ is the variance of any given measurement system and can be estimated from any of the data, and also $\boldsymbol{\theta_i}$ can as well be solved using equation 2, and $\boldsymbol{\rho_{t,X_i}}$ can also be determined.

The extended triple collocation estimation equation can be written as follows:

$$\boldsymbol{\rho_{t,}}\mathrm{x} = \pm \begin{bmatrix} \sqrt{\frac{Q_{12}Q_{13}}{Q_{11}Q_{23}}} \\ \text{sign}\,(Q_{13}Q_{23})\sqrt{\frac{Q_{12}Q_{23}}{Q_{22}Q_{13}}} \\ \text{sign}\,(Q_{12}Q_{23})\sqrt{\frac{Q_{13}Q_{23}}{Q_{33}Q_{12}}} \end{bmatrix} \textit{(Equation 5)}$$

Whereby:

- $\boldsymbol{\rho_{t,X_i}}$ is correct but there is a sign uncertainty. To resolve the sign uncertainty, it is assumed that, in practice, measurement systems are generally positively correlated to the true unknown quantity of the target variable. A pre-test must be conducted on the datasets being collocated before the extended triple collocation analysis to reduce the likelihood of measurement systems that might be negatively correlated to the truth. This test is conducted by ensuring a positive relationship between the datasets being collocated (themselves) and if the positive relationship does not exist, the datasets are judged to be unsuitable for the extended triple collocation analysis (Chen et al., 2017).

The correlation coefficient derived previously allows new information about the performances of the different measurement systems. From the first error model (equation 1) relating measurement systems to the unknown truth, it can be shown that the squared correlation coefficient is the unbiased signal-to-noise ratio as follows:

$$\boldsymbol{\rho^2_{t,X_i}} = \frac{\boldsymbol{\beta_i^2\sigma_t^2}}{\boldsymbol{\beta_i^2\sigma_t^2+\sigma_{\varepsilon_i}^2}} = \frac{\boldsymbol{ubSNR}}{\boldsymbol{ubSNR+1}} \textit{ (Equation 6)}$$

Whereby:

$ubSNR = \frac{\boldsymbol{var(X_i^2)}}{\boldsymbol{var(\varepsilon_i)}} = \boldsymbol{\beta_i^2\sigma_t^2 + \sigma_{\varepsilon_i}^2}$ is defined as the unbiased signal to noise ratio ranging between 0 and 1. It contains information about the sensitivity of the measurement systems β, the variability of the signal as $\boldsymbol{\sigma_t^2}$, and the variability of the measurement error as $\boldsymbol{\sigma_\varepsilon^2}$.

3. METHODOLOGY

3.1. Flowchart

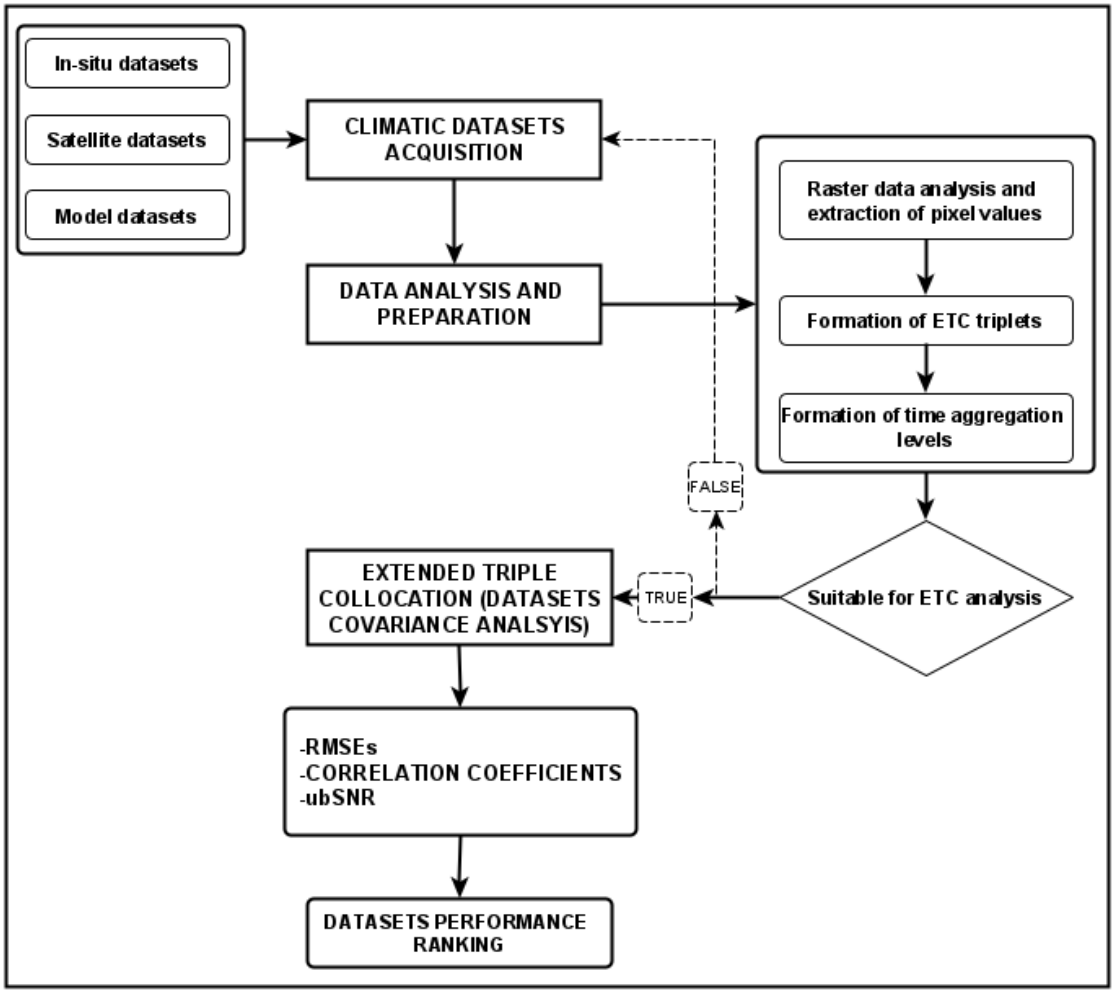


Figure 3.1: Flowchart.

3.2. An overview of the main activities

- **Climatic datasets acquisition and preparation**

Datasets used in the triple sensor approach must originate from 3 independent sources and as can be expected they are in different formats. A data integration step was required to reconcile all datasets into a common interoperable format. The preferred format in this research was the CSV file format. In this sense, the pixels of raster datasets overlapping of ground in-situ weather stations were extracted using the stations' coordinates and they were stored as CSV files for each station. In-situ datasets were acquired already in the CSV format. **(All used datasets are presented separately in the next chapter, section 4.2)**

- **Formation of triplets**

After saving all datasets in a common format (CSV), they must be grouped in triplets for the extended triple collocation analysis. The formation of triplets consists of arranging 3 datasets of the same variable for the same location into a single CSV file, whereby one column contains data values for one dataset and each row contains the date on which the values were recorded. Table 3.2 shows a sample of triplets formed by CHIRPS, ERA5, and In-situ TAHMO rainfall datasets.

Table 3.1: Triplets formation example.

Date	CHIRPS	ERA5	IN-SITU TAHMO
01/01/2018	0	0	0
02/01/2018	0	0	0
03/01/2018	6.94	1	9
04/01/2018	0	0	11
05/01/2018	0	0	0
06/01/2018	8.27	0	0
07/01/2018	0	0	0
08/01/2018	0	0	0
09/01/2018	0	0	5

- **Time aggregation**

The datasets were obtained in daily time aggregation levels. For the analysis of the effects of time aggregation levels, the datasets were aggregated in pentad and dekad. The aggregation for rainfall was done by summing rainfall values of 5 and 10 consecutive days for pentad and dekad time aggregation levels respectively for each dataset. For solar radiation datasets, the aggregation was done by averaging 5 and 10 consecutive days for pentad and dekad aggregation levels.

- **Suitability analysis**

Following the sign ambiguity that was explained for the ETC correlation coefficient in sub-section 2.3.2, a suitability test must be performed to do away with datasets that might be negatively correlated to the unknown true quantity of a given variable.

- **The application of the extended triple collocation**

The application of the extended triple collocation entails performing a covariance analysis of the triplets using the equations presented in section 2.3. The covariance analysis yields root mean square errors and correlation coefficients estimated for each dataset in the triplets table.

- **Ranking**

The estimated correlation coefficients are squared to calculate the unbiased signal to noise ratio for each dataset following equation 6 in sub-section 2.3.2. The unbiased signal to noise ratios of all three datasets are then ranked in a descending order whereby the dataset with the highest value gets rank 1 and is judged as the most reliable between the 3 at that location.

3.3. An overview of the tools

In this research, several software packages were used for various tasks, at different stages of the research and inter-changeably, the main ones are outlined below:

- **GDAL:** The Geospatial Data Abstraction Library was used to translate raster datasets into same, common raster formats for visualization and analysis in different GIS Environments.

- **ILWIS:** The Integrated Land and Water Information System was used for two main tasks: the acquisition of CHIRPS rainfall datasets using the ISOD toolbox and to test sample dataset on the triple sensor toolbox.

- **JavaScript:** A JavaScript code was used to retrieve from the Google Earth Engine model datasets

- **ArcGIS:** The "Extract Multi Values to Points" tool of the Spatial Analyst Toolbox of ArcGIS was used for the extraction of information stored in raster datasets pixels using ground stations point coordinates(for this task Rstudio was also used).

- **RStudio:** R studio was the main workhorse in this research, use was made of it for all the statistical computations involved in the extended triple collocation analysis. It was as well used to make graphs.

- **Microsoft Office:** Microsoft Excel was used for tasks such as triplets formation, time aggregation while Microsoft word was used for general thesis writing.

4. STUDY AREA AND DATASETS

4.1. Study area

4.1.1. Study area location

The study area of this research is located around the Lake Naivasha basin in the Kenyan central and Rift Valley regions at a latitude of 0º 09' to 0º 55'S and a longitude of 36º 09' to 36º 24'E about 75km Northwest of the capital, Nairobi. The highest and lowest altitudes in the basin are 3990m, and 1980m above mean sea level respectively. The total area covered by the basin is 3400km2, whereby lake Naivasha occupies a space of 169km2. (Odongo, Onyando, Mutua, van Oel, & Becht, 2013).

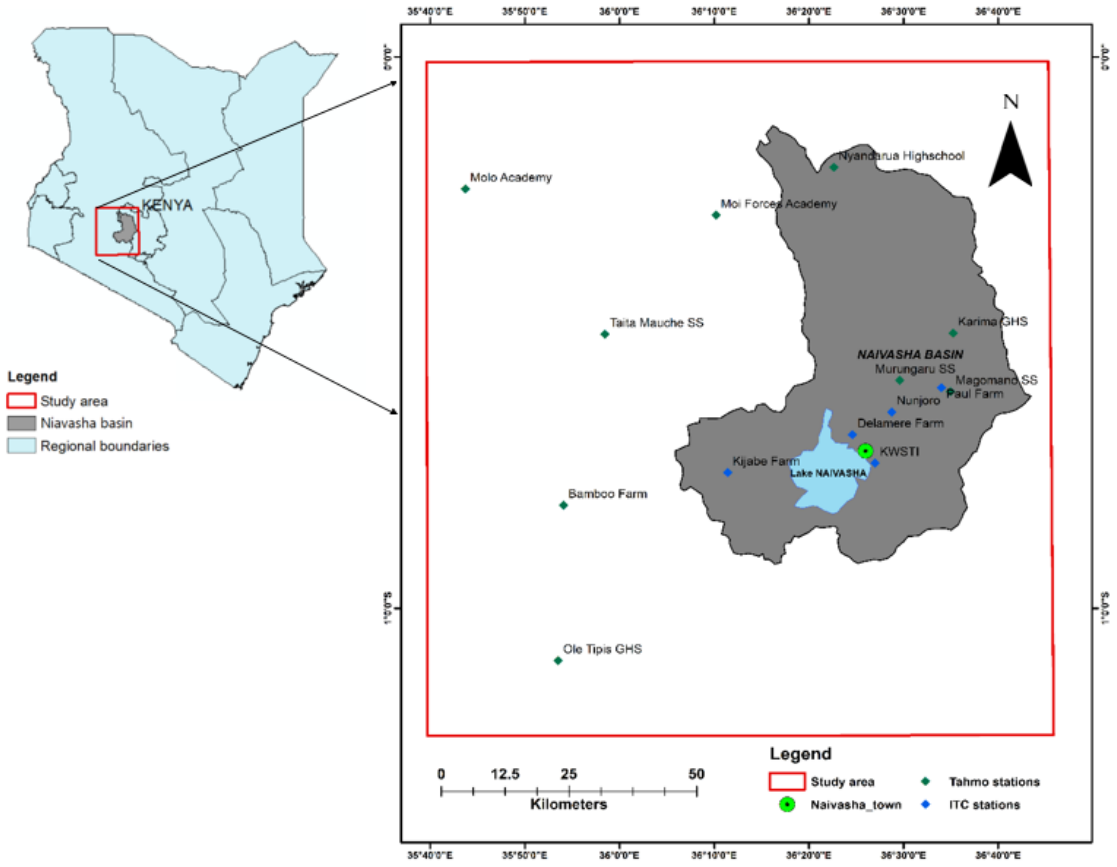


Figure 4.1: Study area map.

The study area map shows a total number of 13 ground meteorological stations scattered in the study area, among which 5 are operated by ITC and the other 8 by the Trans-African Hydro-Meteorological Observatory (TAHMO). The stations were installed to facilitate climatic data availability for scientific research in the area.

4.1.2. Study area description

The basin is characterized prominently by economically significant agricultural activities which have contributed considerably to population growth and consequently water resources management issues in the basin due to increased pressure on the available water resources (Becht, Odada, & Higgins, 2005).

Precipitation regimes vary quite highly because they depend on both local (altitude and relief, The Aberdare Mountains and the Mau Escarpment) and distant processes related to changes in temperature in the oceans. The basin is characterized by long rains from March to June and short rains from October to December. The highly variable climatic conditions in the Lake Naivasha Basin observed by Kuhn, Britz, Willy, & van Oel, (2016) and Odongo et al., (2013), may hamper agriculture practices in the basin as they rely on accurate knowledge of environmental conditions such as rainfall occurrence.

The prediction of environmental conditions, on the other hand, relies on the availability and reliability climatic variables datasets which have been noted to be an issue in the basin by the water management authority in the basin area (van Oel et al., 2013). Careful water management solutions must be the main focus of concerned parties for the benefit of the communities in the Naivasha basin and far beyond.

4.1.3. Fieldwork

For this research, a fieldwork campaign was conducted between the 08-22 Jan 2020 in the study area, and comprised the following activities:

- Familiarizing with the study area and collecting in-situ meteorological datasets that cannot be transferred automatically.
- Visiting all ITC ground meteorological stations and appraising their suitability for in-situ datasets collection for this research and to give recommendations regarding future use.

4.1.3.1. Main fieldwork findings

- During the fieldwork, visits were conducted to all ITC ground stations, and it was found that most data collection instruments were in poor maintenance conditions. For instance, rainfall measurement instruments had been clogged by the accumulation of vegetation debris and restricting rainwater and allowing consequently no rainfall measurements. All ITC stations were deemed unreliable specifically for rainfall analysis in this research. The focus was turned to other in-situ rainfall data sources available in the study area(discussed in later sections). Figure 4.2 illustrates the condition of the ITC rain gauge located at the Nunjoro farm.

Figure 4.2: Nunjoro farm weather station.

- On the other hand, instruments for the measurement of other variables, notably soil moisture, were generally in good condition at most stations, and the datasets were subsequently downloaded during the fieldwork.

- Another important finding of the fieldwork is the rain gauge operated by Delamere farm management as independent citizen observers. This rain gauge had long time-series of rainfall events dating back to 1980, and rainfall records were used as an additional alternative data source.

4.2. Research datasets

The use of the triple sensor approach requires the availability of three mutually independent data sources whereby the datasets represent the same variable, the same spatial and temporal coverages. In the following sub-sections, the used datasets for rainfall, solar radiation, and air temperature are presented.

4.2.1. Rainfall datasets

Rainfall datasets used are from 3 different sources: In-situ rainfall measurements obtained using rain gauges, Satellite-based rainfall products, and Model re-analysis rainfall datasets for the study area.

4.2.1.1. In-situ datasets

In-situ datasets for rainfall were acquired from The trans-African Hydro-Meteorological Observatory (TAHMO) which operates several relatively well-maintained weather stations in the study area. The datasets of daily rainfall estimates were retrieved as CSV files from the Online TAHMO data portal the years 2018 and 2019. After a closer inspection, some of the TAHMO stations were discarded due to large amounts of missing data or because they didn't pass the suitability analysis for the ETC analysis. Three stations were finally used:

- Moi Forces Academy, Ole Tipis GHS, and Taita Mauche for the years 2018 and 2019.
- Delamere farm own rain-gauge rainfall data for the year 2018.

4.2.1.2. Satellite datasets

The Climate Hazards Group Infrared Precipitation with station data is a quasi-global rainfall dataset that combines high resolution (0.05^0, about 5.5km) satellite imagery with ground in-situ stations to form gridded rainfall datasets that can be used by users in several fields1981 (Funk et al., 2015).

The datasets were directly retrieved via the In Situ and Online Data Toolbox (ISOD) in ILWIS for daily time aggregation over the African region. The raster maps of daily rainfall were visualized in ILWIS as maplists to catch a glimpse of the rainfall trend over the continent and the study area in 2018 and 2019. For consistency and integration with other datasets in this research, the pixel values were extracted using TAHMO stations coordinates and saved as CSV files for the ETC analysis. Figure 4.3 is a sample spatial variability visualization of the CHIRPS rainfall over the African continent and in the study area:

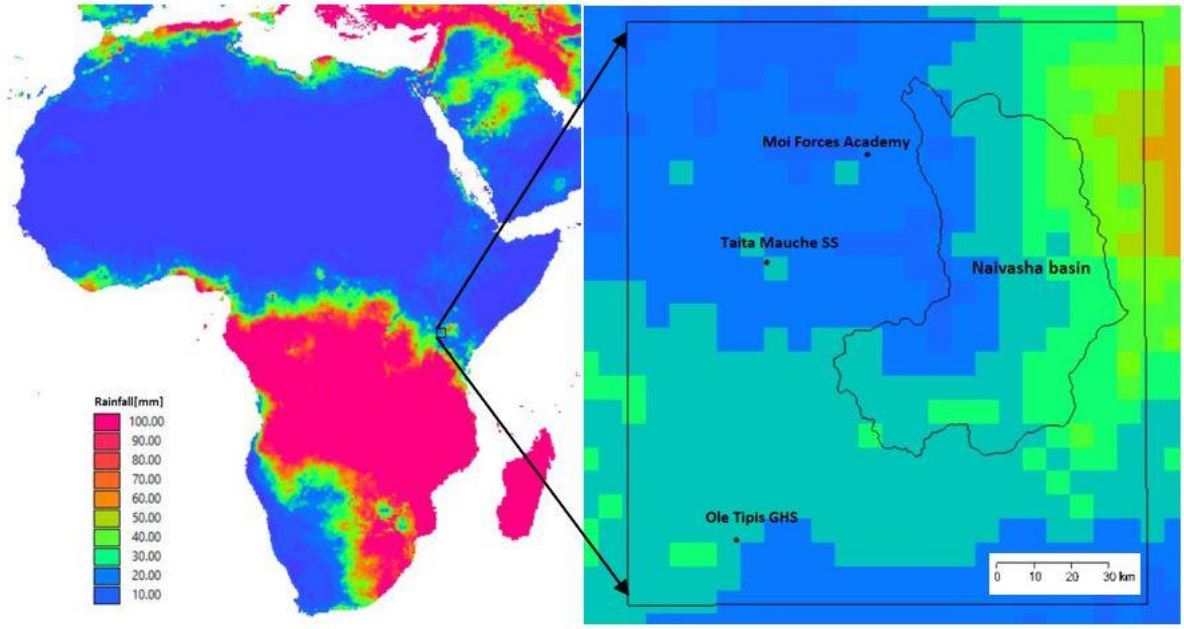


Figure 4.3: Spatial visualization of CHIRPS monthly rainfall(January 2018).

4.2.1.3. Model datasets

For model-generated datasets, use was made of the ECMWF ERA5 which is a climate reanalysis dataset spanning the period from 1950 to present. This model provides high-quality global forecasts and climate reanalyses using data assimilation techniques(4D-VR). The ERA5 datasets are available in both GRIB and NetCDF formats at a regular spatial resolution of 0.25^0 * 0.25^0(roughly 27.75km). For this research, preference was given to ERA5 re-analyses data because it is better compared to its re-analysis predecessors such as the ERA-Interim, in many aspects such as spatial and temporal resolutions (Copernicus Climate Change Services, 2018).

Normally, the datasets can be accessed through MARS (Meteorological Archival and Retrieval System), whereby users can use python scripts to download the data from ECMWF servers directly from the website. The time aggregates which were available in MARS were unsuitable for this research, so daily aggregates of total precipitation were instead retrieved directly from Google Earth Engine to save a lot of energy and computations, by use of a JavaScript code.

The code helps to visualize the preferred variable globally and then the coordinates of the corresponding ground in-situ stations (in this case the TAHMO) were used to extract, plot, and export corresponding pixels values directly in CSV file format. Figure 4.4 is a sample illustration of the ECMWF ERA5 rainfall variability over the study area:

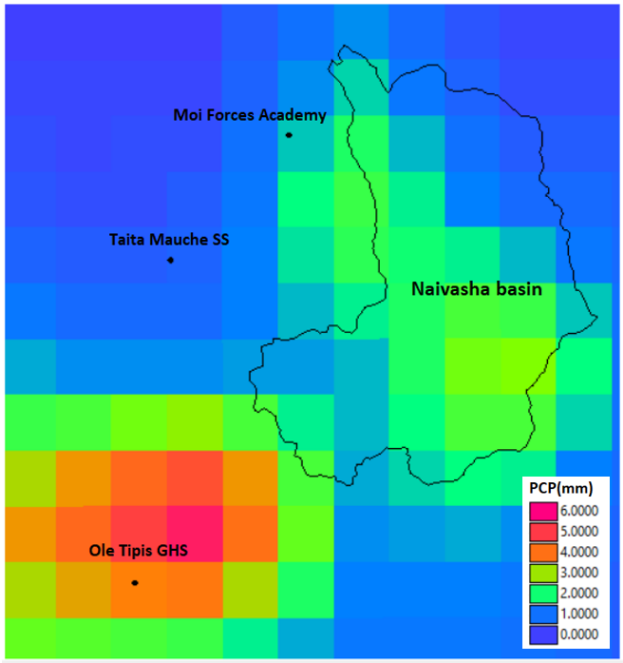


Figure 4.4: Spatial visualization of ERA5 monthly rainfall(January 2018).

4.2.1.4. Datasets temporal visualization

a. Daily rainfall datasets.

In Figure 4.5 we visualize the temporal variability of rainfall at Moi Forces Academy and Ole Tipis GHS for the from 2018 up to 2019(730 days). The first column contains CHIRPS, ERA5, and In-situ TAHMO rainfall estimates at Moi Forces Academy while the second column shows the same datasets at Ole Tipis GHS in the same order.

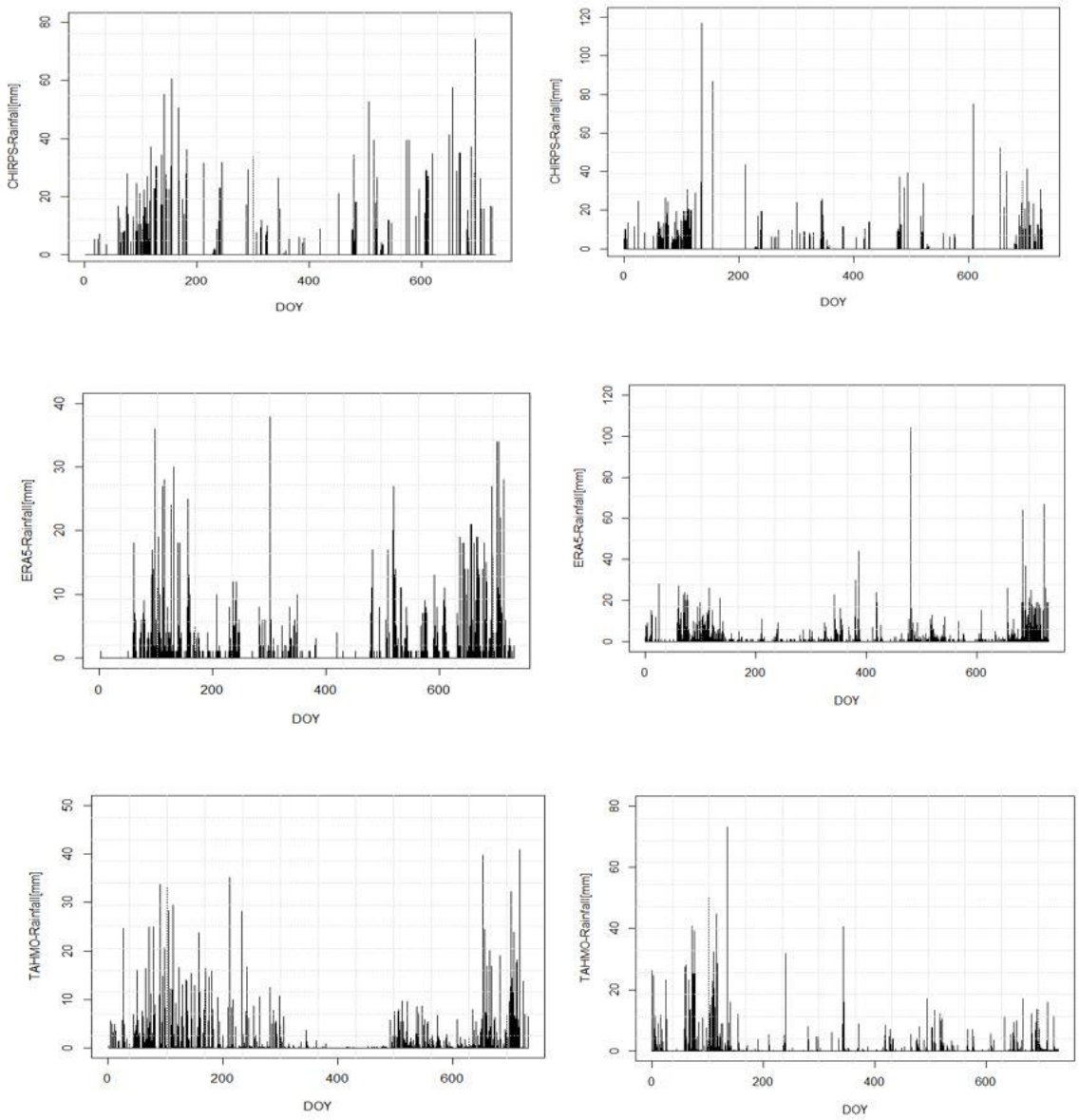


Figure 4.5: Daily rainfall at Moi Forces Academy and Ole Tipis GHS.

In Figure 4.6, rainfall estimates are presented for Delamere Farm and Taita Mauche SS for 2018 and 2019 respectively. The first column contains rainfall at Delamere Farm starting from CHIRPS, ERA5 up to In-situ. The second column presents rainfall in the same order at Taita Mauche SS.

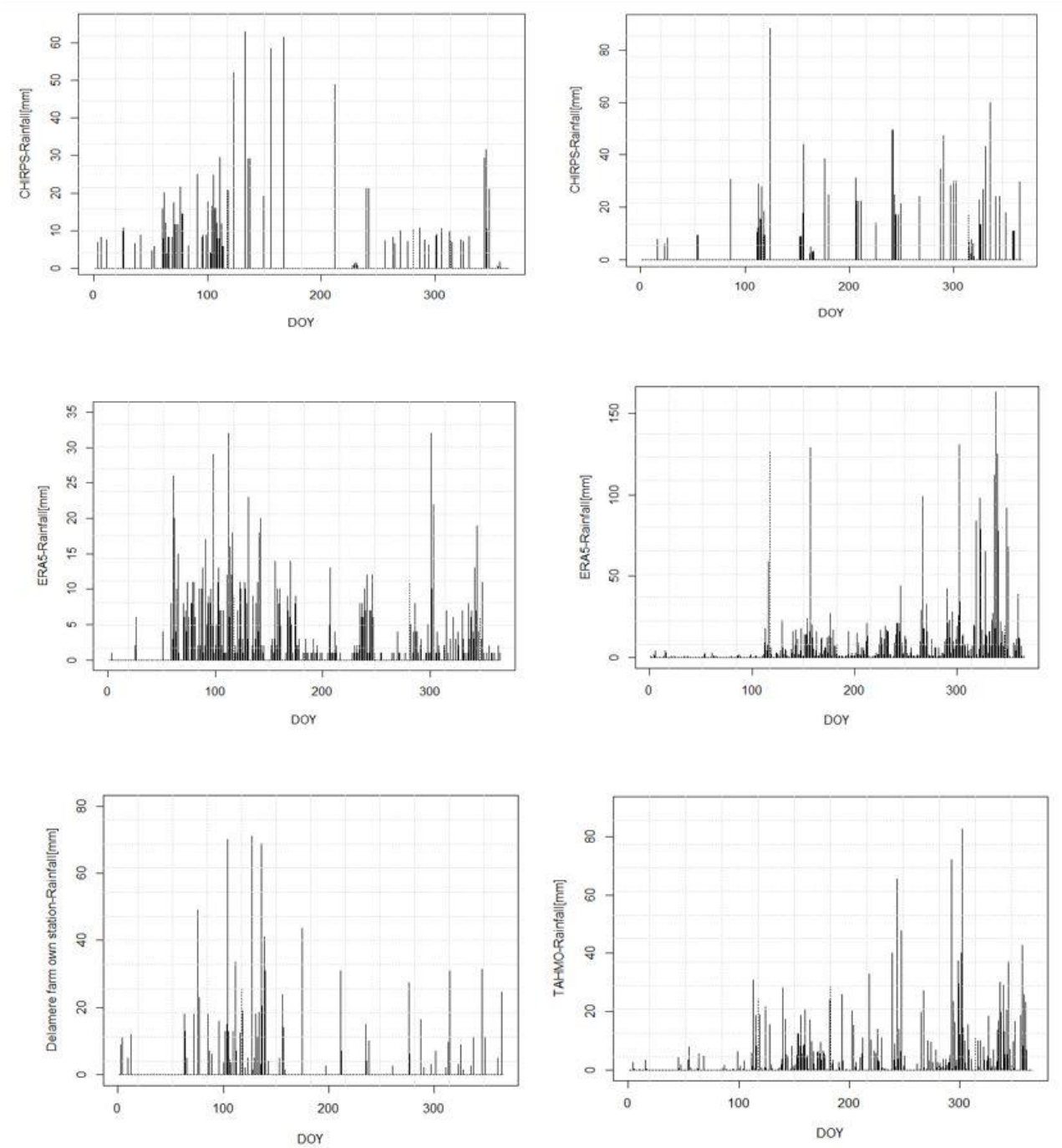


Figure 4.6: Daily rainfall at Delamere farm and Taita Mauche SS.

b. Pentad rainfall aggregates

In Figure 4.7 we present the temporal visualization graphs for pentad rainfall aggregates at Moi Forces Academy and Ole Tipis GHS for 2018 and 2019 combinedly. The first column shows CHIRPS, ERA5, and In-situ TAHMO estimates at Moi Forces Academy while the second column is for Ole Tipis GHS in the same order.

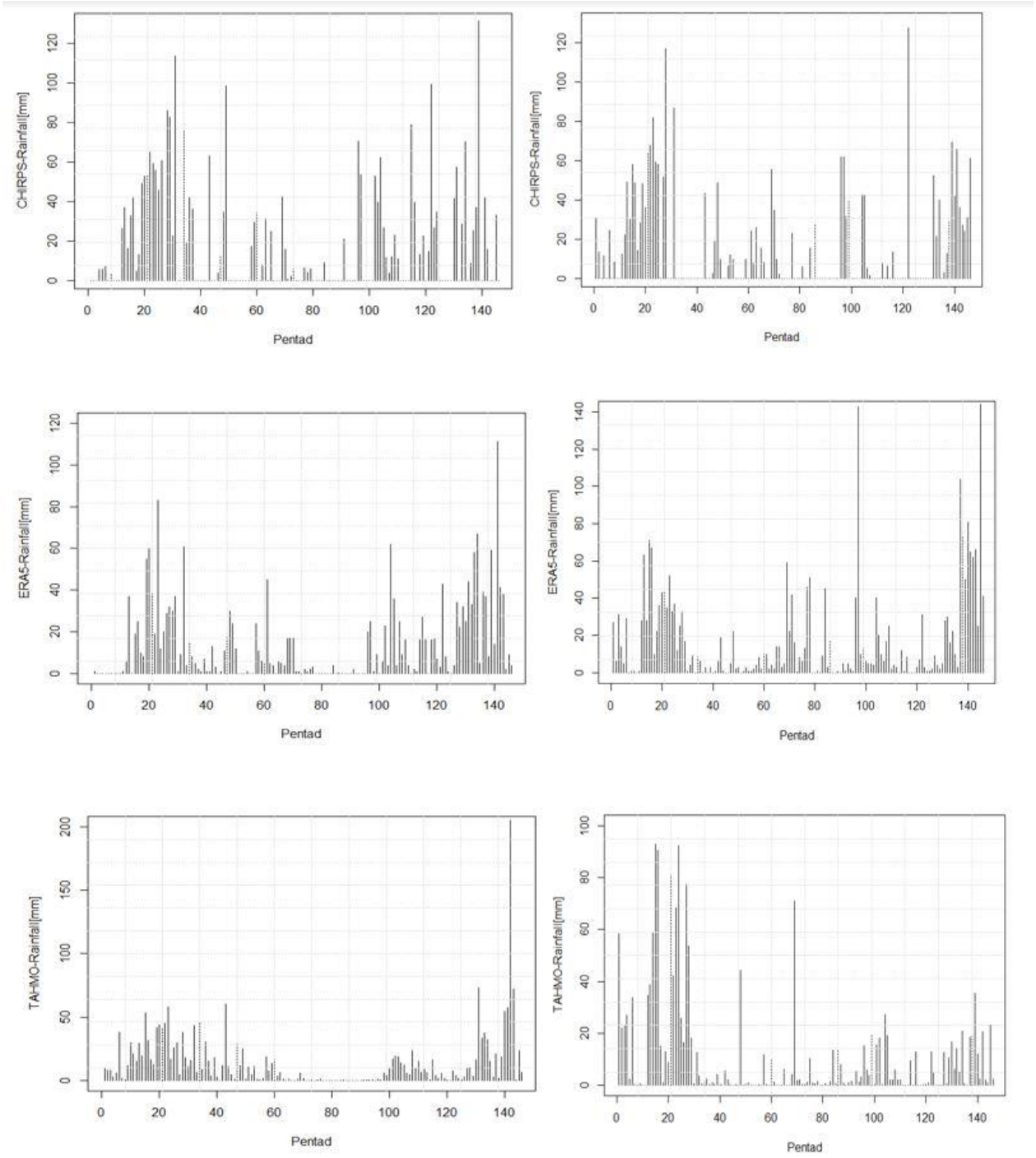


Figure 4.7: Pentad rainfall at Moi Forces Academy and Ole Tipis GHS.

In Figure 4.8, pentad aggregates of rainfall at Delamere farm and Taita Mauche SS are presented in the first column and the second respectively. The aggregates at Delamere farm are for the year 2018 and 2019 at Taita Mauche SS.

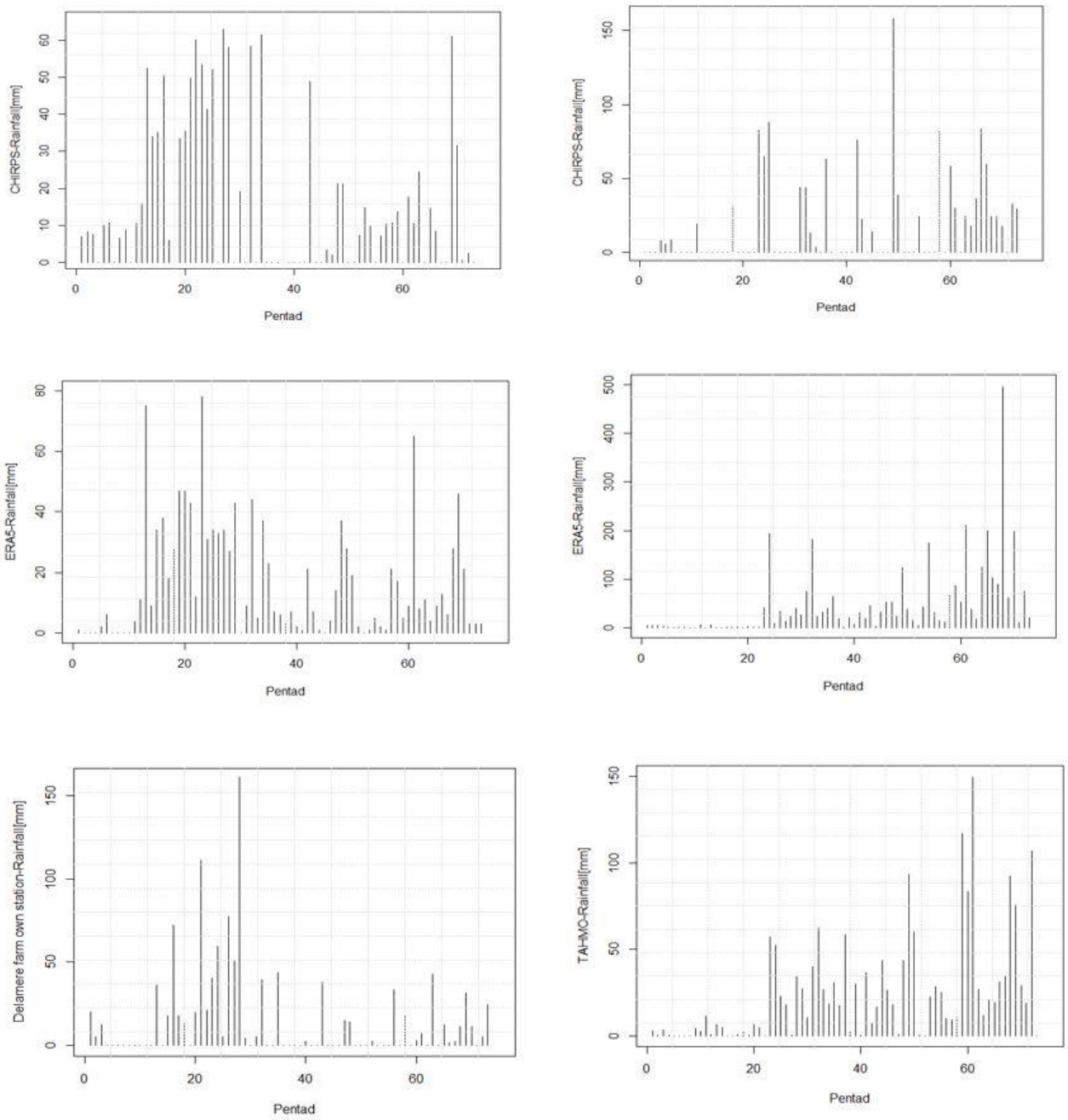


Figure 4.8: Pentad rainfall at Delamere farm and Taita Mauche SS.

c. Dekad rainfall aggregates

In Figure 4.9 we present the temporal visualization graphs for dekad rainfall aggregates at Moi Forces Academy and Ole Tipis GHS for 2018 and 2019 combinedly. The first column shows CHIRPS, ERA5, and In-situ TAHMO estimates at Moi Forces Academy while the second column is for Ole Tipis GHS in the same order.

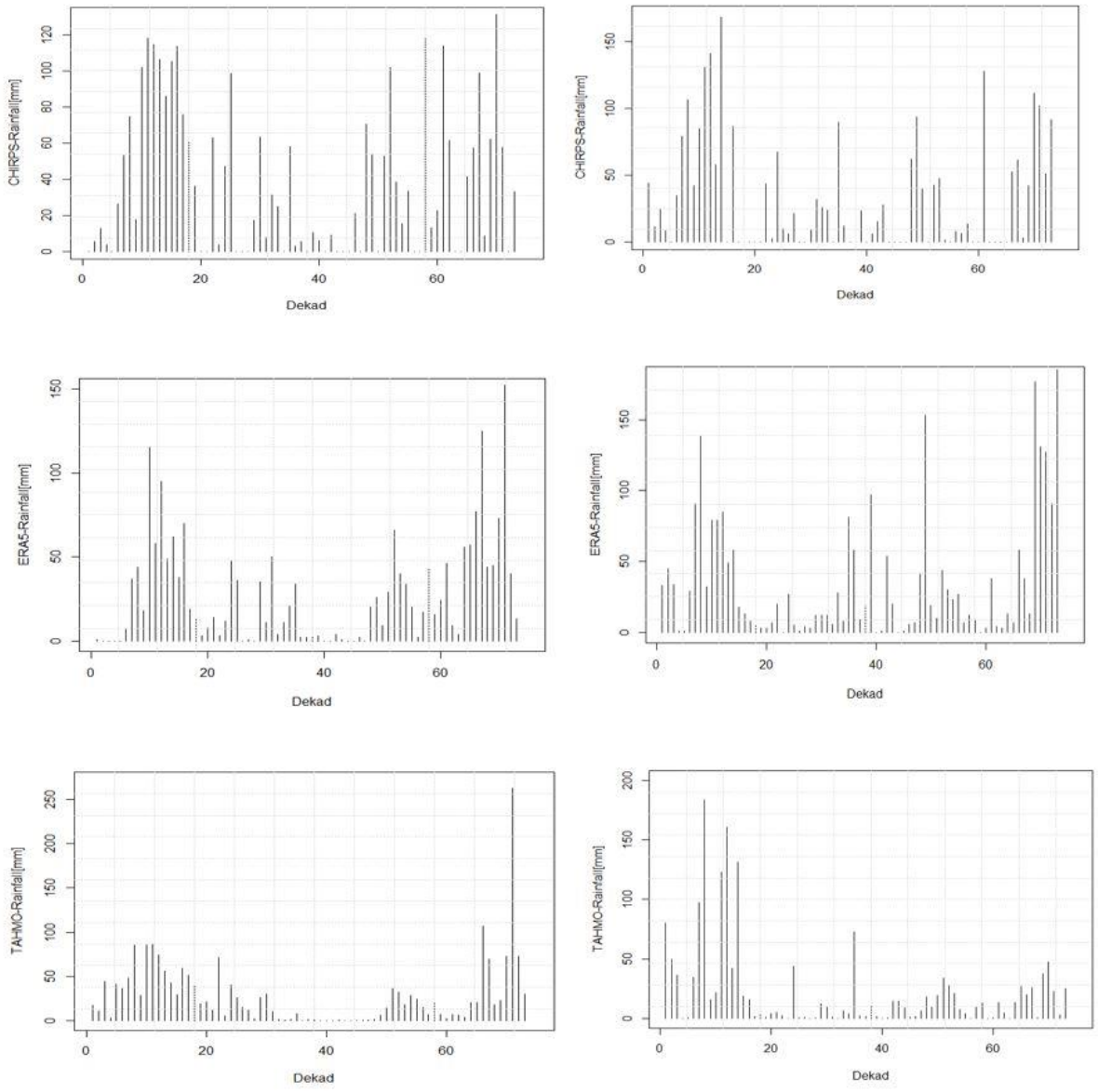


Figure 4.9: Dekad rainfall at Moi Forces Academy and Ole Tipis GHS.

Figure 4.10 contains graphs for the visualization of dekad rainfall aggregates at Delamere Farm for 2018 and Taita Mauche SS for 2019. The first column is for Delamere farm while the second is for Taita Mauche SS.

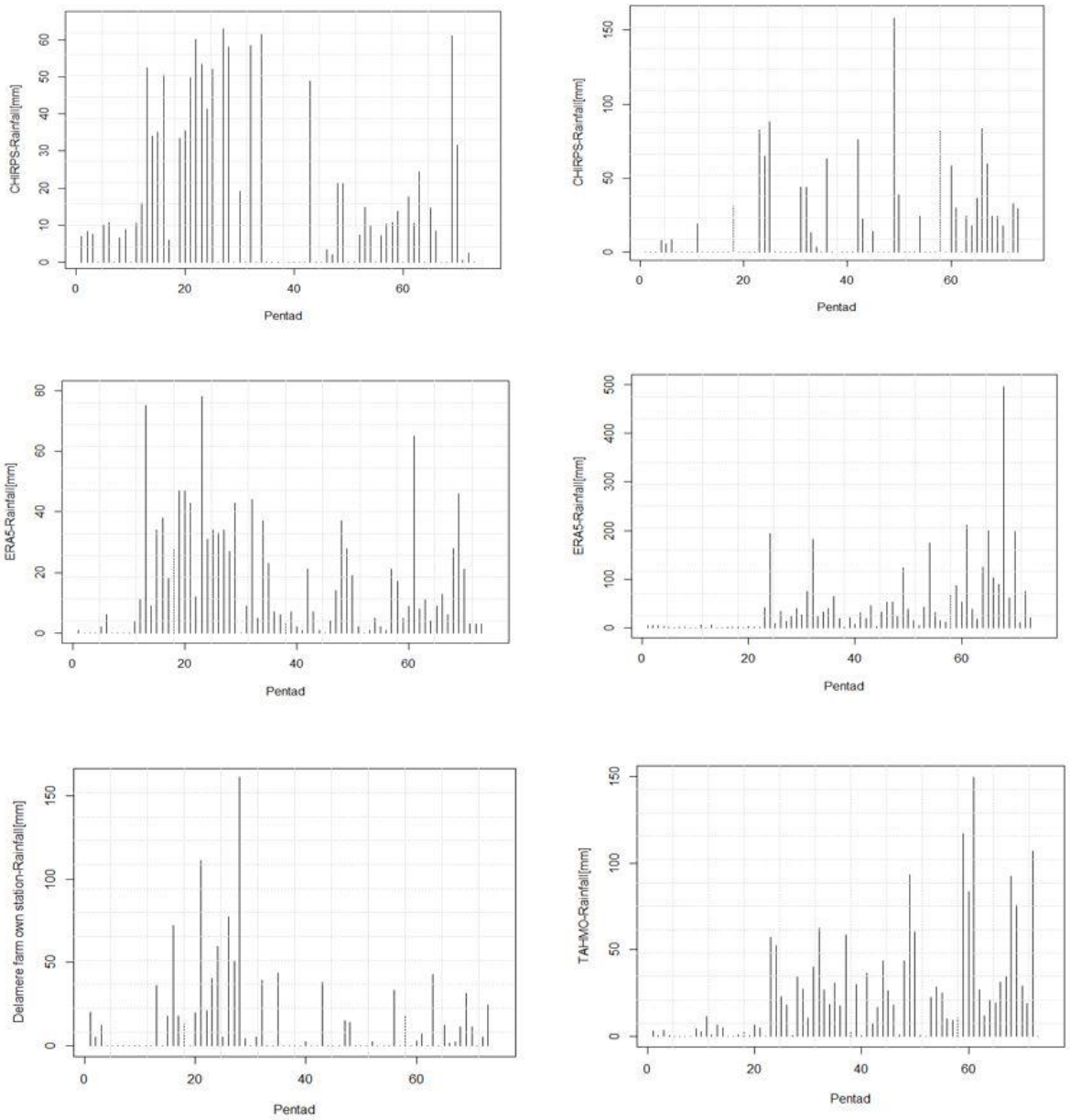


Figure 4.10: Dekad rainfall at Delamere farm and Taita Mauche SS.

4.2.1.5. Evaluation of the rainfall datasets suitability for the extended triple collocation analysis

To resolve the sign ambiguity which arises in the correlation coefficients estimated by the ETC analysis, here a test is conducted to ensure that the chosen datasets are positively correlated to the unknown quantity of the target variable. The test is done by verifying whether the used datasets have a positive relationship with each other.

The correlation coefficient between two measurement systems x and y is given by the following equation:

$$\rho_{xy} = \frac{Cov(x,y)}{\sigma_x \sigma_y} \textbf{(}\textit{Equation 7}\textbf{)}$$

Where:

ρ_{xy}= Pearson correlation coefficient.

Cov(x,y)= The covariance of x and y.

σ_x= The standard deviation of x.

σ_y= the standard deviation of y.

4.2.1.6. Correlation analysis

This test was performed on daily, pentad, and dekad time aggregation levels for CHIRPS, ECMWF ERA5, and TAHMO In-situ datasets for the year 2018 and 2019 separately and then combinedly. The results of the relationship analysis are summarized in table 4.1 and table 4.2.

Table 4.1: Correlation coefficient analysis, rainfall datasets(2018&2019).

YEAR	TIME AGGREGATION	DATASET	Moi Forces Academy	Ole Tipis GHS	Delamere farm
2018	DAILY	*CHIRPS-ERA5*	0.217	0.454	0.226
		CHIRPS-IN-SITU	0.312	0.535	0.310
		ERA5-IN-SITU	0.229	0.560	0.192
	PENTAD	*CHIRPS-ERA5*	0.633	0.448	0.704
		CHIRPS-IN-SITU	0.542	0.598	0.713
		ERA5-IN-SITU	0.420	0.580	0.799
	DEKAD	*CHIRPS-ERA5*	0.737	0.747	0.763
		CHIRPS-IN-SITU	0.773	0.742	0.838
		ERA5-IN-SITU	0.512	0.645	0.855
2019		DATASET	Moi Forces Academy	Ole Tipis GHS	Taita Mauche SS
	DAILY	*CHIRPS-ERA5*	0.247	0.416	0.168
		CHIRPS-IN-SITU	0.075	0.462	0.278

		ERA5-IN-SITU	0.139	0.318	0.363
	PENTAD	***CHIRPS-ERA5***	0.344	0.595	0.553
		CHIRPS-IN-SITU	0.393	0.154	0.520
		ERA5-IN-SITU	0.581	0.476	0.460
	DEKAD	***CHIRPS-ERA5***	0.564	0.528	0.701
		CHIRPS-IN-SITU	0.554	0.100	0.646
		ERA5-IN-SITU	0.649	0.585	0.536

Table 4.2: Correlation coefficient analysis, PCP datasets for 2018&2019 combined.

YEAR	TIME AGGREGATION	DATASET	Moi Forces Academy	Ole Tipis GHS
2018&2019 COMBINED	**DAILY**	***CHIRPS-ERA5***	0.234	0.395
		CHIRPS-IN-SITU	0.253	0.485
		ERA5-IN-SITU	0.250	0.328
	PENTAD	***CHIRPS-ERA5***	0.516	0.571
		CHIRPS-IN-SITU	0.309	0.605
		ERA5-IN-SITU	0.495	0.461
	DEKAD	***CHIRPS-ERA5***	0.650	0.681
		CHIRPS-IN-SITU	0.457	0.713
		ERA5-IN-SITU	0.748	0.495

The correlation analysis between the three datasets reveals that in daily time aggregation level the relationship is rather weak (nevertheless positive) and increases slightly with time aggregation levels. The weak relationship between the used measurements gives, on the other hand, more ground to the triple collocation analysis for the identification of the most reliable one. Regardless of the relatively weak relationship, it is positive in all cases, which satisfies fully the requirement for the ETC suitability.

4.2.1.7. Introduction of outliers in daily time aggregation rainfall datasets

In climate sciences, some of the consequences of climate change could be extreme rainfall events which result in outliers. According to Wu, Liu, & Chawla, (2010), outliers in rainfall datasets used in different studies for example for analysis and prediction could skew the results and lead to false conclusions. One of the objectives of this research being the evaluation of the effects of outliers on the results of the ETC analysis, an experimental dataset with outliers must be generated from one of the existing datasets.

Logically we must first establish both the upper and lower limits of the dataset and then introduce random values greater than the upper limit. The Inter-Quartile Range method (IQR) described by Mirzaei et al., (2014) is used to evaluate the distribution of the datasets around the median value and at both ends.
The dataset is divided into quartiles, whereby a quartile is a statistically dividing point that splits the data into quarters which are then used to specify both the upper and the lower limit in the data. The IQR methods steps are shown below:

- Step 1: The data values are arranged from the least to the greatest.
- Step 2: Determine the position of the first quartile by (N+1)/4, whereby N is the number of all data points.
- Step 3: Determine the position of the third quartile by 3*(N+1)/4.
- Step 4: Calculate the interquartile range(IQR) by subtracting the value of the first quartile from the value of the third quartile.
- Step 5: The lower and upper limits of the data are given by (First quartile)-(1.5*IQR), and (Third quartile)+(1.5*IQR) respectively.

For simplicity, daily rainfall estimates at Ole Tipis Girls Highschool (the Year 2018) used for this step. We consider this original dataset as outlier free and then establish its lower and upper limits. The results of the analysis are shown below in table 4.3:

Table 4.3: Calculation of upper and lower limits.

	CHIRPS	ERA5	IN-SITU
Quartile 1	0	0	0
Quartile 3	3.18	4	1.29
IQR	3.18	4	1.29
Upper limit	7.95	10	3.24
Lower limit	-4.77	-6	-1.94

At this stage, we intentionally deteriorate the in-situ dataset by introducing random values greater than the upper limit of the original data (we assume that negative outliers in rainfall datasets should be easily spotted, it is therefore not necessary to introduce any values lower than the lower limit.). Figure 4.11 shows the

original daily rainfall in-situ rainfall at Ole Tipis GHS(first graph) and the generated datasets with outliers(second graph):

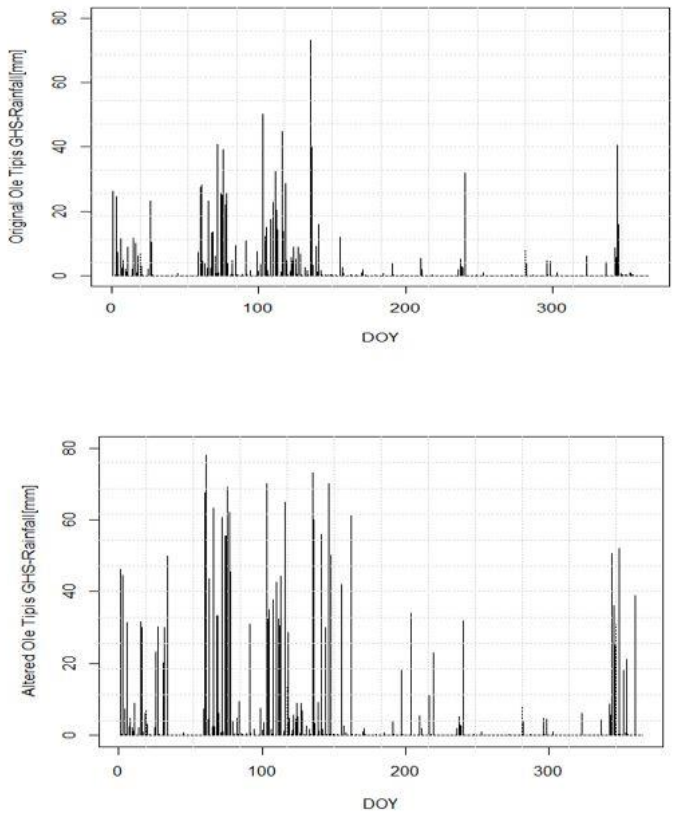


Figure 4.11: Original and Altered TAHMO In-situ rainfall datasets at Ole Tipis GHS.

4.2.1.7.1. Cross-correlation in the dataset with outliers

As done previously, a pre-test must be conducted on the altered dataset to assess its suitability for the triple collocation analysis. It can be noted from table 4.4 that the correlation coefficient between the correlation coefficient reduced from 0.535 to 0.454 between CHIRPS-IN-SITU and from 0.56 to 0.47 between ERA5-IN-SITU. The population variance also increased considerably from 89.02 in the original dataset up to 426.4 in the dataset with outliers.

Table 4.4: Correlation analysis-altered dataset.

LOCATION	CORRELATION COEFFICIENT		
	CHIRPS-ERA5	***CHIRPS-IN-SITU***	***ERA5-IN-SITU***
Ole tipis GHS	0.4542	0.4544	0.47

4.2.1.8. Analysis of the impact of missing data

Another objective of this study is the analysis of the impact of missing data on the results of the triple collocation analysis, for this step, the Taita Mauche SS 2018 daily rainfall dataset was used. This dataset has a large proportion of missing data in the TAHMO in-situ dataset, precisely 36.7%.

The datasets at Taita Mauche SS were first evaluated for their suitability for the triple collocation analysis by analyzing the relationships between them. In table 5 it is shown that a positive relationship exists.

Table 4.5: Correlation analysis in the dataset with missing data.

LOCATION	CORRELATION COEFFICIENT		
	CHIRPS-ERA5	***CHIRPS-IN-SITU***	***ERA5-IN-SITU***
Taita Mauche SS	0.28	0.43	0.5

4.2.2. Solar radiation datasets

Solar radiation is the electromagnetic energy emitted by the sun in all regions of the electromagnetic spectrum, it is the driving factor of several physical processes such as evaporation and precipitation which are at the core of the global climate dynamics. Several measurement systems for this variable exist and it is always necessary to scrutinize the resulting datasets before use as much as for other important climatic variables. For this research, we use the ETC analysis to compare different datasets of the shortwave incoming solar radiation (SW) obtained using the in-situ measurements, satellite-based sensors, and model-based (reanalysis) estimates of the downwelling shortwave solar radiation.

All used datasets were acquired in daily aggregations and different formats for the years 2017, 2018, and 2019. The data preparation such as the extraction of pixel values overlapping locations of in-situ stations and the formation of triplets were performed in the same manner as for rainfall datasets. The major difference in the preparation of the datasets of the rainfall and solar radiation datasets is that for solar radiation the pentad and dekad aggregation levels were performed by averaging 5 and 10 consecutive days respectively after combining the 2017,2018 and 2019 time-series.

The ETC analysis was performed for each year's time-series separately in daily time aggregation and then for daily, pentad, and dekad time aggregation levels for the combined time series.

4.2.2.1. In-situ data

In-situ measurements of the shortwave incoming solar radiation are performed by pyranometers installed in the field and are representative for that specific location, it absorbs the incoming solar radiation in the shortwave solar spectrum. In this study, we use the TAHMO shortwave incoming solar radiation datasets obtained by pyranometers installed at both Karima GHS and Moi Forces Academy in the study area for the year 2017, 2018, and up to November 2019. The datasets were retrieved as CSV files of hourly time aggregation and then daily values were calculated by averaging 24 hours(Daily hourly-based averages).

4.2.2.2. Model data-ECMWF ERA5

The European Centre for Medium-Range Forecast ERA5 is a model-based dataset which provides among others historical estimates of the shortwave solar radiation portion which reaches the Earth Surface (not absorbed or reflected by atmospheric constituents) estimated using data assimilation methods. The ERA5 downwelling shortwave radiation datasets are globally available in 0.1°x0.1° spatial resolution and according to ECMWF, (2019), the datasets are reasonably good compared to those estimated by a pyranometer.

In this study, we make use of the ECMWF ERA5 hourly land datasets retrieved from the Copernicus Data Store in the GRIB raster format for the years 2017, 2018, and up to November 2019. Pixel values overlapping Karima GHS and Moi Forces Academy stations were extracted and the daily hourly-based averages were calculated by averaging 24 hours. Figure 4.13 illustrates the spatial variation of the shortwave incoming solar radiation in the study area for the first day of 2017(sample).

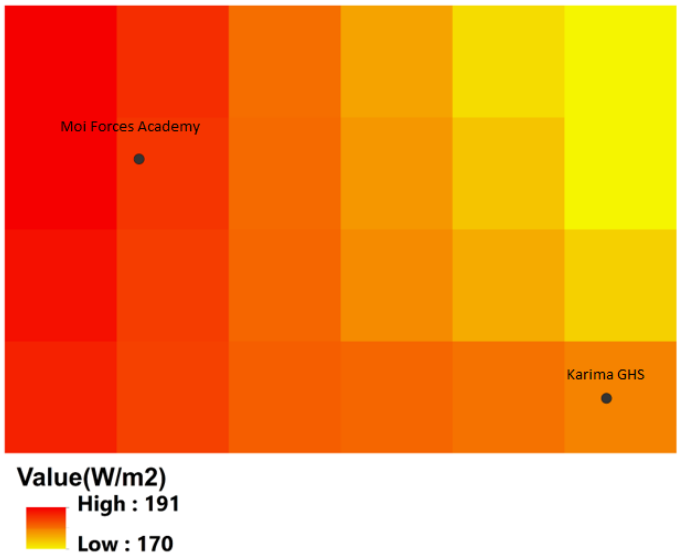


Figure 4.12: Sample ERA5 SW solar radiation spatial visualization.

4.2.2.3. Satellite data

For the satellite-based solar radiation, we use the datasets obtained by the NASA CERES (Clouds and Earth's Radiant Energy System) scanner onboard the Earth Observation System Aqua and Terra satellites. The CERES algorithm assimilates the cloud and aerosol properties (fraction, optical depth, top height, and particle size) retrieved from MODIS and GEO radiances, into radiative transfer models to compute TOA and surface fluxes of solar radiation.

The cloud and aerosol properties are gridded onto 1^{o}x1^{o} and hourly spatial and temporal boxes respectively, and the distribution of the cloud optical depth is estimated for each cloud type. The cloud properties along with other elements such as the temperature and relative humidity profiles are used together in radiative transfer models, as described in detail by Kato, Loeb, Rutan, & Rose, (2015), to compute top of atmosphere and surface fluxes.

The dataset used in this study is Synoptic TOA and surface fluxes and clouds CERES_SYN1deg_Ed4.1 computed daily means (from hourly means) of the shortwave downwelling component of the surface flux in a spatial resolution of 1°x1° for the years 2017, 2018, and up to November 2019 retrieved from the CERES Data Product website. It can be seen in Figure 4.13 that due to the relatively coarse spatial resolution of the CERES solar radiation products, there is not a lot of variability in the study area and the two stations (Karima GHS and Moi Forces Academy) are, in fact, overlapped by a single pixel.

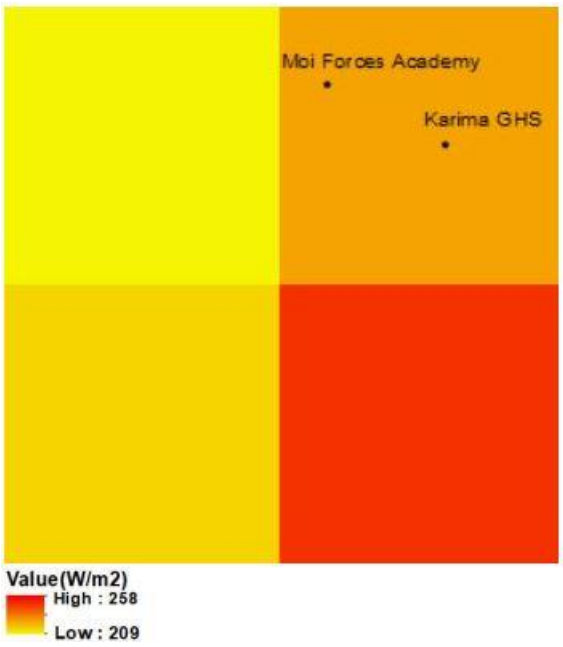


Figure 4.13: Sample CERES SW radiation spatial variability.

4.2.2.4. Solar radiation datasets temporal visualization

In Figures 4.14, 4.15, and 4.16 we visualize the temporal variabilities of daily shortwave incoming solar radiation (W/m²) for the years 2017, 2018, and 2019 respectively, for the three involved data sources. The first column in each figure contains graphs for Karima GHS and the second is for Moi Forces Academy.

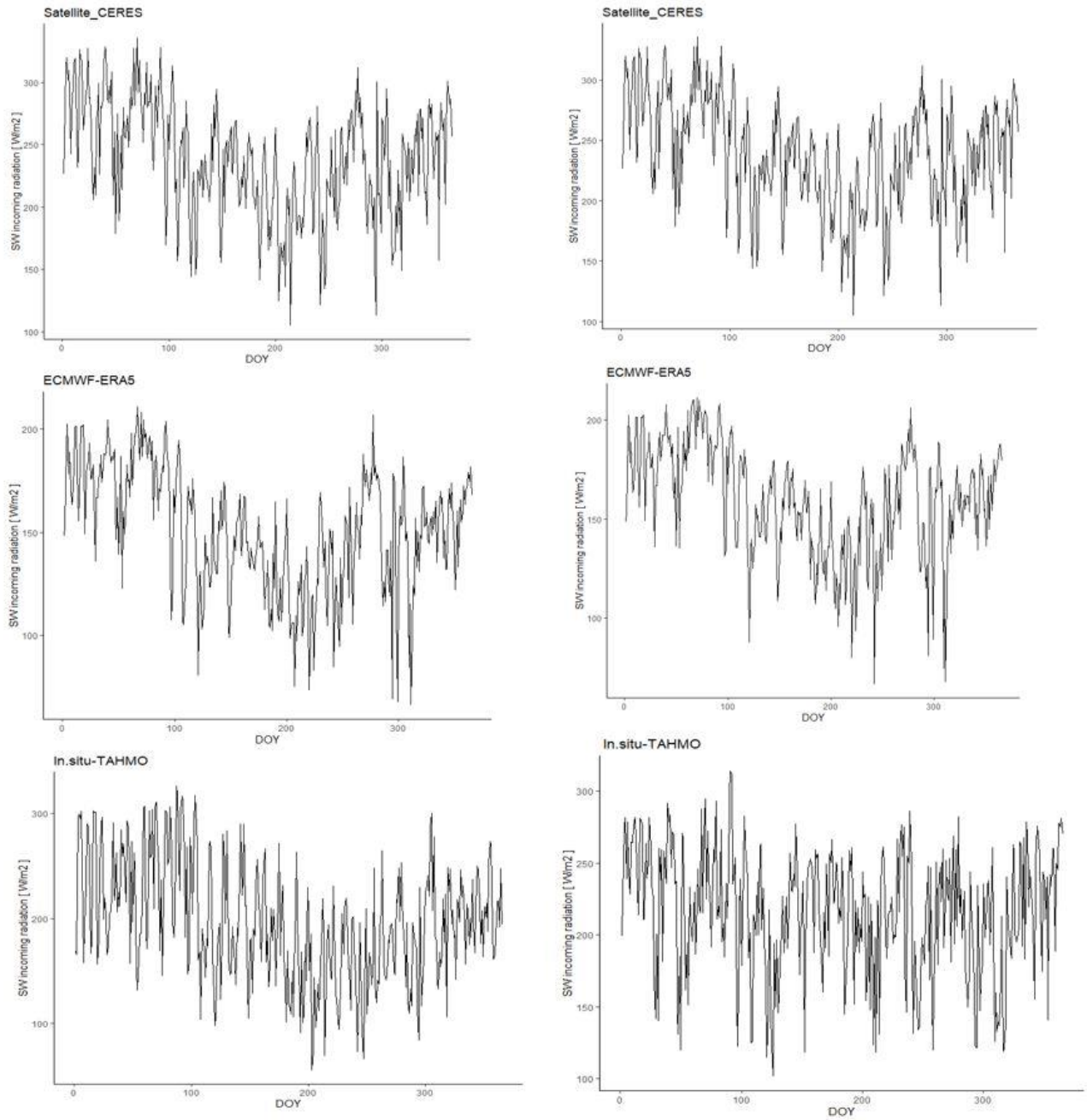


Figure 4.14: 2017 SW incoming solar radiation and Karima GHS and Moi Forces Academy.

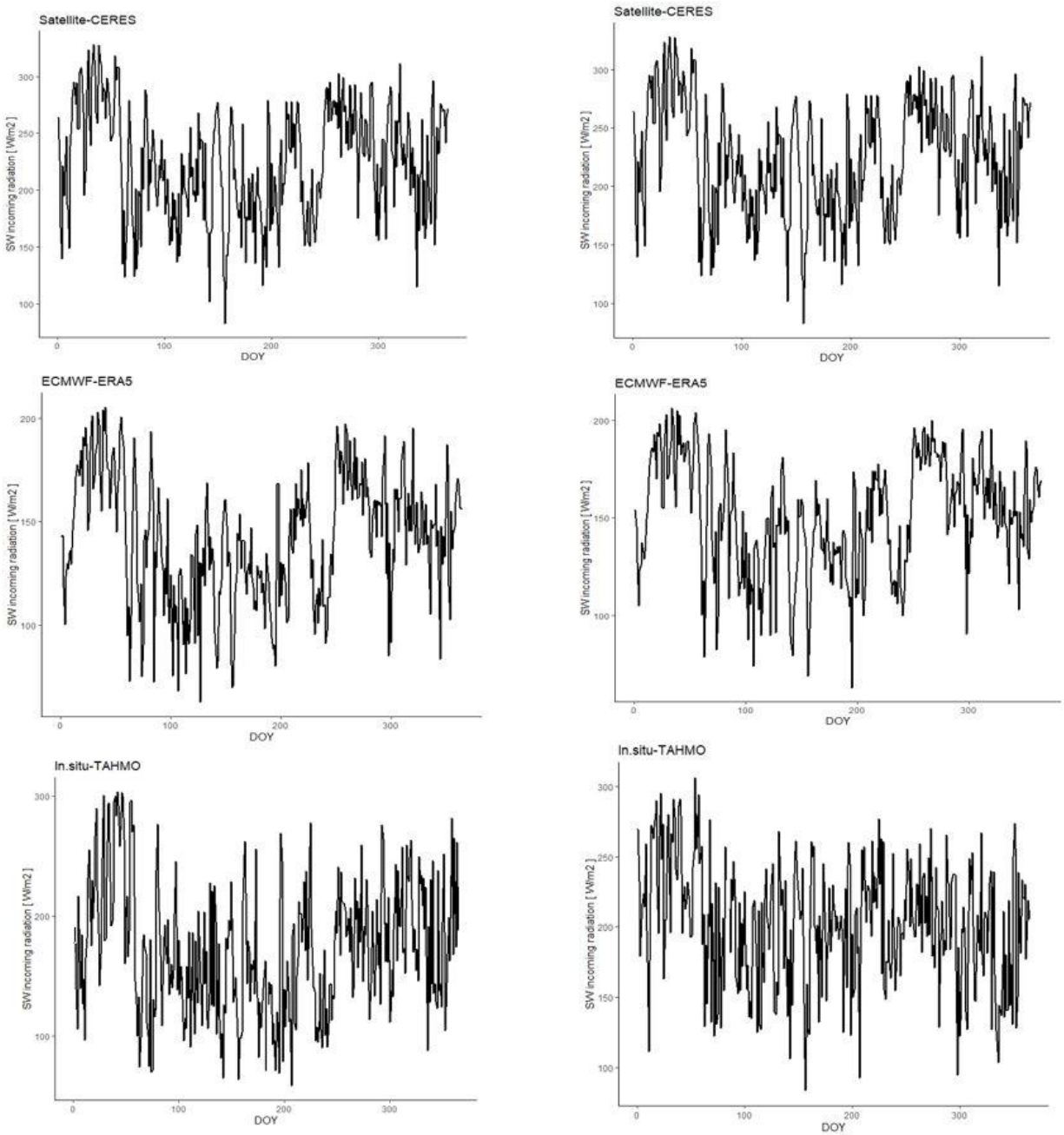


Figure 4.15: 2018 SW incoming solar radiation at Karima GHS and Moi Forces Academy.

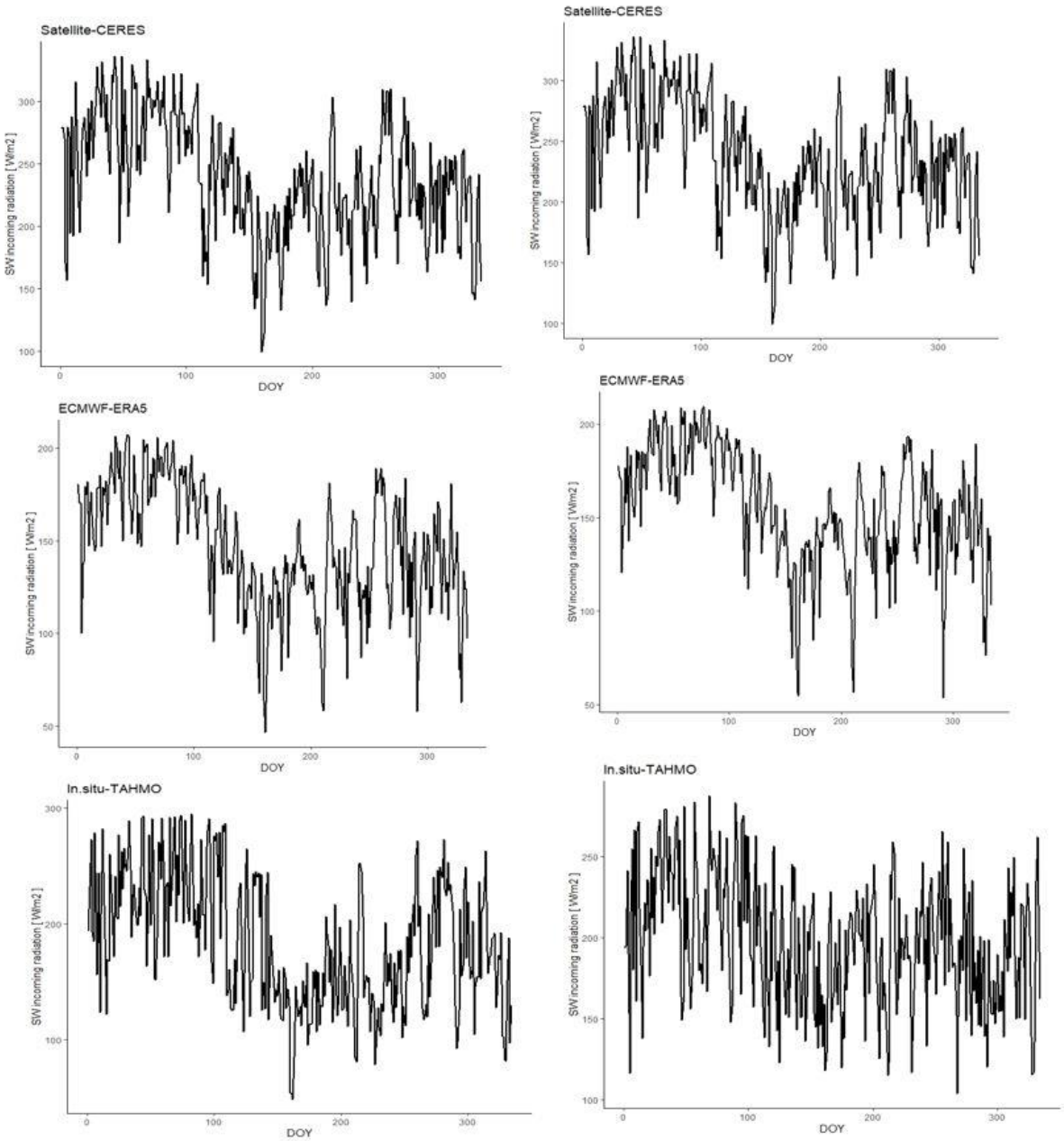


Figure 4.16: 2019 SW incoming solar radiation at Karima GHS and Moi Forces academy.

4.2.2.5. Solar radiation datasets suitability analysis for ETC

In this section, we perform the correlation analysis to ensure the existence of a positive relationship between the used datasets as a pre-requisite for the triple collocation analysis. The results in Tables 4.6 and 4.7 show a positive relationship between the used datasets which satisfies the requirement for the triple collocation analysis.

Table 4.6: Correlation analysis, radiation datasets(2017,2018, and 2019 separately.)

YEAR	DATASET	Moi Forces Academy	Karima GHS
2017	***CERES-ERA5***	0.87	0.88
	CERES-IN-SITU	0.76	0.79
	ERA5-IN-SITU	0.62	0.69
2018	***CERES-ERA5***	0.87	0.88
	CERES-IN-SITU	0.76	0.84
	ERA5-IN-SITU	0.63	0.75
2019	***CERES-ERA5***	0.89	0.88
	CERES-IN-SITU	0.76	0.82
	ERA5-IN-SITU	0.65	0.72

Table 4.7: Correlation analysis-daily hourly SW incoming radiation(2017,2018,2019 combinedly).

YEAR	TIME AGGREGATION		Moi Forces Academy	Karima GHS
2017,2018&2019 COMBINED	DAILY	***CERES-ERA5***	0.88	0.88
		CERES-IN-SITU	0.75	0.81
		ERA5-IN-SITU	0.63	0.72
	PENTAD	***CERES-ERA5***	0.95	0.95
		CERES-IN-SITU	0.79	0.88
		ERA5-IN-SITU	0.72	0.85
	DEKAD	***CERES-ERA5***	0.95	0.96
		CERES-IN-SITU	0.76	0.90
		ERA5-IN-SITU	0.71	0.88

4.2.3. Air temperature datasets

Air temperature (2m) is another crucial variable in the water cycle which plays a key role in water cycle processes such as evaporation and precipitation. In this study, we apply the ETC analysis on monthly means of air temperature from In-situ measurements, ECMWF ERA5 re-analysis, and a gridded surface air temperature product for the years 2018 and 2019:

- For in-situ measurements, monthly means of air temperature (2m) measured at Karima GHS and Moi Forces Academy were retrieved from the TAHMO online data portal in CSV files.

- For the ECMWF ERA5 re-analysis product, use was made of the air temperature means obtained from the Climate Data Service(CDS) in regular grids of 0.1°*0.1°. Figure 4.17 is a sample visualization of the dataset' spatial variability in the study area:

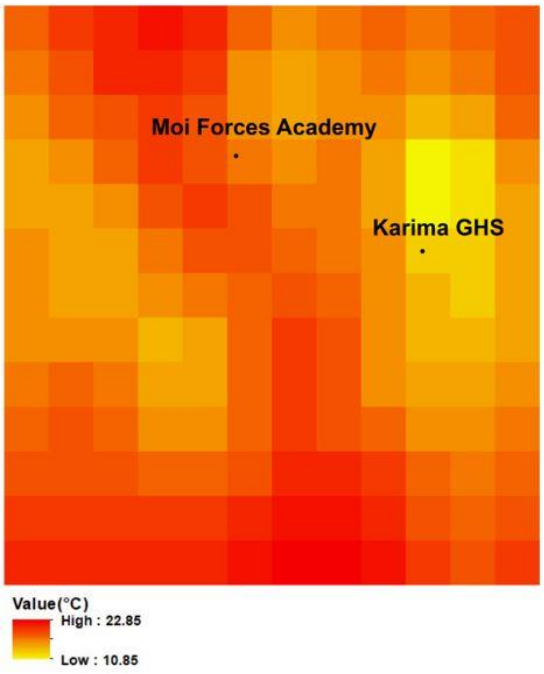


Figure 4.17: Spatial visualization of ERA5 air temperature(January 2018)

- The third dataset used is air temperature NOAA NCEP CPC GHCN_CAMS gridded product which provides global monthly means of air temperature at regular grids of 0.5° *0.5° for the period 1948 up to the present. According to Fan & Dool, (2008), these datasets are obtained by interpolation of in-situ observations. Figure 4.18 is a sample visualization of the GHCN+CAMS air temperature spatial variability in the study area for January 2018:

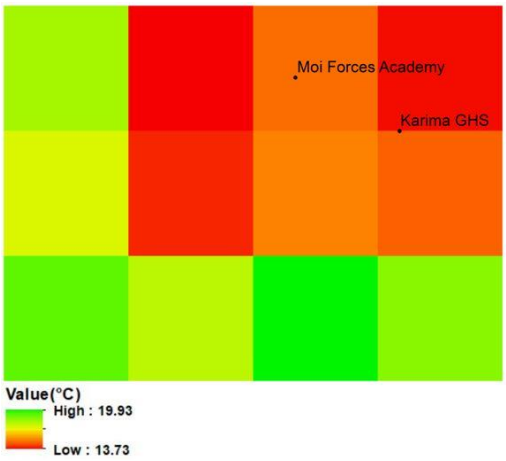


Figure 4.18: Sample spatial variability.

For the two raster-based air temperature products, the coordinates of Moi Forces Academy and Ole Tipis GHS were used to extract pixel values and store them as CSV for the ETC triplets formation(same procedure as rainfall and solar radiation datasets).

4.2.3.1. Visualization of air Temperature datasets temporal variability

In Figure 4.17, the variation of monthly air temperature at Ole Tipis GHS and Moi Forces Academy is presented for the years 2018 and 2019.

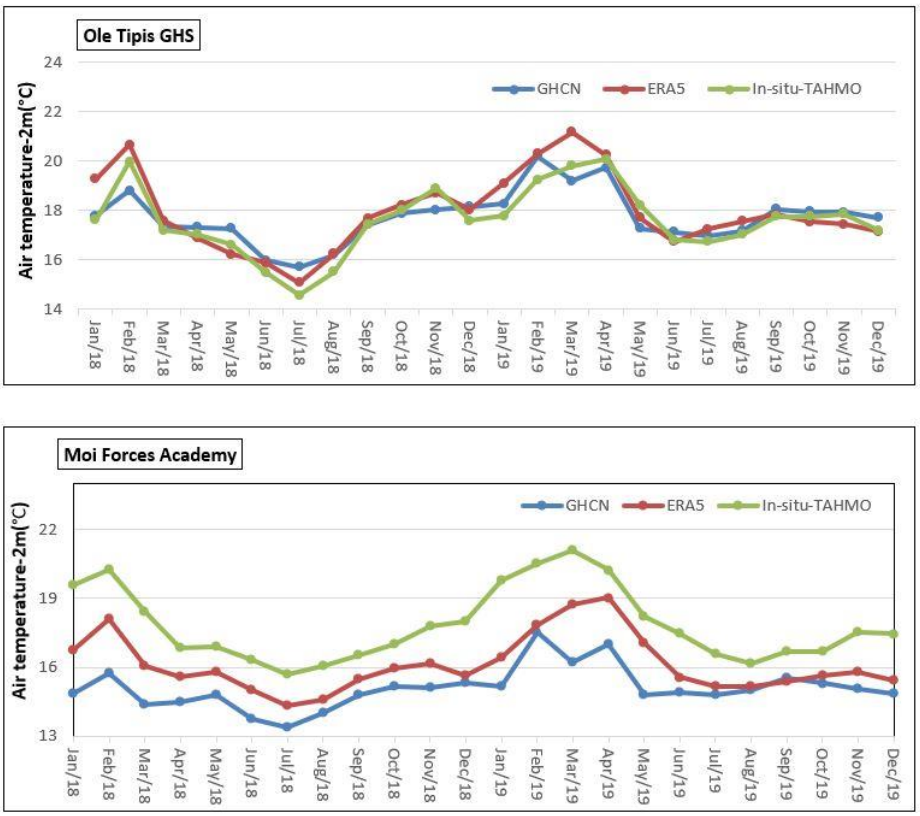


Figure 4.19: Monthly air temperature at Ole Tipis GHS and Moi Forces Academy.

4.2.3.2. Air temperature datasets suitability analysis for ETC

A positive relationship between the used air temperature datasets must be ensured before the ETC analysis as done for the other datasets. The results of the analysis are shown in table 4.8 where at both stations the positive correlation exists between the datasets. This Satisfies the requirement for the datasets' ETC suitability.

Table 4.8: Correlation analysis, air temperature.

LOCATION	CORRELATION COEFFICIENT		
	GHCN + CAMS-ERA5	GHCN + CAMS-TAHMO	*ERA5-TAHMO*
Ole Tipis GHS	0.80	0.73	0.92
Moi Forces Academy	0.90	0.92	0.93

5. RESULTS PRESENTATION AND DISCUSSION

5.1. Root mean square error and correlation coefficient results

5.1.1. Rainfall

5.1.1.1. Results of each year analyzed separately

Table 5.1 contains the ETC estimates of the RMSEs, correlation coefficients (rho), signal to noise ratio (rhosqr) and the corresponding ranks for each rainfall dataset at Delamere farm, Moi Forces Academy and Ole Tipis GHS for the years 2018 and 2019.

The ETC correlation coefficients are relatively low in daily time aggregation level but increase with pentad and dekad aggregations, the same behavior is observed in the normal Pearson correlation coefficients estimated between the datasets before the ETC analysis.

The results are in most cases reasonable except for pentad and dekad time aggregation levels whereby biased RMSEs, correlation coefficients, and signal-to-noise ratio are estimated (negative RMSE, correlation coefficients and signal-to-noise ratio greater than 1).

The most reliable dataset at each station is represented by value 1 in the rank row which corresponds to the highest signal-to-noise ratio. In cases where the signal-to-noise ratio is biased, the corresponding dataset gets a rank 1 because it is always a greater number compared to the other collocated datasets. Its interpretation in terms of reliability is given in later sections.

Table 5.1: ETC results for rainfall(2018&2019).

YEAR	T.AGG	*Location*	*Delamere farm*			*Moi Forces Academy*			*Ole Tipis GHS*		
		Dataset	*CHIRPS*	*ERA5*	*IN-SITU*	*CHIRPS*	*ERA5*	*IN-SITU*	*CHIRPS*	*ERA5*	*IN-SITU*
2018	Daily	RMSE	47	24	64	62	23	23	57	14	25
		rho	0.60	0.37	0.51	0.54	0.40	0.57	0.66	0.69	0.81
		rhosqr	0.36	0.14	0.26	0.29	0.16	0.33	0.44	0.48	0.66
		rank	1	3	2	2	3	1	3	2	1
	Pentad	RMSE	76	178	487	416	167	57	247	70	132
		rho	0.90	0.70	0.60	0.68	0.66	0.88	0.79	0.89	0.90
		rhosqr	0.82	0.49	0.36	0.46	0.43	0.77	0.63	0.79	0.81
		rank	1	2	3	2	3	1	3	2	1
	Dekad	RMSE	-136	471	923	244	286	245	525	247	146
		rho	1.05	0.70	0.73	0.93	0.81	0.80	0.87	0.88	0.97
		rhosqr	1.11	0.49	0.54	0.86	0.65	0.64	0.75	0.78	0.94
		rank	1	3	2	1	2	3	3	2	1
2019		*Location*	*Taita Mauche SS*			*Moi Forces Academy*			*Ole Tipis GHS*		
		Dataset	*CHIRPS*	*ERA5*	*IN-SITU*	*CHIRPS*	*ERA5*	*IN-SITU*	*CHIRPS*	*ERA5*	*IN-SITU*
	Daily	RMSE	57	14	25	47	24	64	62	23	23
		rho	0.36	0.47	0.78	0.36	0.68	0.21	0.78	0.54	0.60

		rhosqr	0.13	0.22	0.60	0.13	0.46	0.04	0.60	0.29	0.35
		rank	3	2	1	2	1	3	1	3	2
	Pentad	RMSE	692	2861	326	574	-364	680	216	480	37
		rho	0.48	0.71	0.82	0.44	1.36	0.35	0.79	0.70	0.66
		rhosqr	0.23	0.51	0.66	0.19	1.84	0.12	0.63	0.49	0.43
		rank	3	2	1	2	1	3	1	2	3
	Dekad	RMSE	1171	5098	1014	1727	-2471	2249	247	1080	78
		rho	0.69	0.81	0.80	0.30	1.76	0.33	0.92	0.76	0.70
		rhosqr	0.48	0.66	0.64	0.09	3.09	0.11	0.84	0.58	0.49
		rank	3	1	2	3	1	2	1	2	3

5.1.1.2. Combined time-series

In table 5.2, the ETC results for the combined time-series of the years 2018 and 2019 for rainfall datasets are shown for Moi Forces Academy and Ole Tipis GHS. The correlation coefficients estimated by the ETC are also relatively low in daily time aggregation levels but augment once more in pentad and dekad aggregation levels.

Biased RMSEs, correlation coefficients, and signal-to-noise ratios observed in the pentad aggregation for 2018 and 2019 rainfall separately no longer exist after combining the two years' time-series. However, such values are still observed in one case at Moi Forces Academy in the dekad time aggregation level.

Biased ETC estimates still get a rank with respect to the other collocated dataset results.

Table 5.2: ETC results for rainfall(2018&2019 combined time-series).

YEAR	T.AGG	*Location*	*Moi Forces Academy*			*Ole Tipis GHS*		
		Dataset	*CHIRPS*	*ERA5*	*IN-SITU*	*CHIRPS*	*ERA5*	*IN-SITU*
2018&2019 COMBINED	Daily	RMSE	71.74	22.81	63.43	35.27	42.15	25.03
		rho	0.47	0.49	0.33	0.77	0.52	0.64
		rhosqr	0.22	0.24	0.11	0.59	0.27	0.40
		rank	2	1	3	1	3	2
	Pentad	RMSE	502.49	63.81	359.63	155.64	359.72	204.82
		rho	0.57	0.91	0.55	0.87	0.66	0.70
		rhosqr	0.32	0.82	0.30	0.75	0.44	0.49
		rank	2	1	3	1	3	2
	Dekad	RMSE	981.84	-66.71	690.28	34.34	1016.03	651.35
		rho	0.63	1.03	0.73	0.99	0.69	0.72
		rhosqr	0.40	1.07	0.53	0.98	0.47	0.52
		rank	3	1	2	1	3	2

5.1.1.3. Root mean square error and correlation coefficient analysis in the daily dataset with outliers

Table 5.3 contains the results of the ETC analysis after introducing artificial outliers into the daily In-situ rainfall dataset and collocating it with the original CHIRPS and ERA5 datasets at Ole Tipis GHS in 2018. An increase of the ETC RMSE is observed in the in-situ dataset (from 25 to 137), which corresponds also to a decrease in the ETC correlation coefficient for this dataset (from 0.81 to 0.69) and the signal-to-noise ratio decreased (from 0.66 to 0.476). In comparison with the original CHIRPS and ERA5 datasets, the original in-situ dataset had been ranked as the most reliable (rank 1) but after the introduction of outliers, it has shifted to the second position.

Table 5.3: ETC results for the dataset with outliers.

	DATASET	*CHIRPS*	*ERA5*	*IN-SITU*
Ole Tipis GHS	**RMSE**	56.86	13.790	137.01
	rho	0.658	0.692	0.69
	rhosqr	0.433	0.479	0.476
	rank	3	1	2

5.1.1.4. Root mean square error and correlation coefficient analysis in the datasets with missing values

Table 5.4 contains the ETC analysis results for datasets with missing values. Here, an In-situ rainfall dataset at Taita Mauche SS with a lot of missing data points was collocated with CHIRPS and ERA5 datasets with no missing values. The estimates of the ETC show that the in-situ dataset has the lowest RMSE, the highest correlation coefficient, the highest signal-to-noise ratio, and gets a rank 1 despite having a large part of missing data.

Table 5.4: ETC results for the dataset with missing values.

	DATASET	*CHIRPS*	*ERA5*	*IN-SITU*
Taita Mauche SS	**RMSE**	76.14	48.57	9.23
	rho	0.49	0.57	0.88
	rhosqr	0.24	0.32	0.77
	rank	3	2	1

5.1.2. Solar radiation

5.1.2.1. Results of each year analyzed separately

Table 5.5 contains the ETC estimates of the RMSEs, correlation coefficients, signal to noise ration, and ranking of the different datasets that were collocated. The results shown are for daily shortwave incoming radiation at Karima GHS and Moi Forces Academy for the years 2017, 2018, and 2019 each separately. The ETC correlation coefficients are all high. In some cases, notably at Moi forces Academy, biased RMSEs, correlation coefficients, and signal-to-noise ratio are observed in each year at Moi Forces Academy and 2019 at Karima GHS. The order of ranks given to the datasets remains uniform in all years and at all stations.

Table 5.5: ETC results for solar radiation.

YEAR	Location	Karima GHS			Moi Forces Academy		
	Dataset	SATELLITE	ERA5	IN-SITU	SATELLITE	ERA5	IN-SITU
2017	RMSE	6.686	198.022	1242.319	-154.99	226.74	862.22
	rho	0.998	0.879	0.787	1.04	0.84	0.74
	rhosqr	0.996	0.773	0.619	1.08	0.71	0.54
	rank	1	2	3	1	2	3
2018	RMSE	51.774	200.28	859.451	-154.151	246.629	849.03
	rho	0.989	0.887	0.847	1.032	0.845	0.741
	rhosqr	0.978	0.787	0.717	1.065	0.714	0.549
	rank	1	2	3	1	2	3
2019	RMSE	-8.821	250.875	1072.454	-69.964	227.209	710.386
	rho	1.002	0.879	0.816	1.015	0.875	0.748
	rhosqr	1.004	0.773	0.666	1.03	0.766	0.56
	rank	1	2	3	1	2	3

5.1.2.2. Results for combined time-series

Table 5.6 contains the ETC results which were estimated for the combined time series of short-wave incoming solar radiation of 2017, 2018, and 2019 at Karima GHS and Moi Forces Academy. The results are for daily, pentad, and dekad aggregation levels. At Karima GHS, the correlation coefficients are all high and within normal range and do not vary considerably with time aggregation levels. At Moi Forces Academy, biased ETC results are observed in all aggregation levels in the satellite-based datasets. The order of ranking of the correlation coefficient remains invariant.

Table 5.6: ETC results for solar radiation(2017,2018 and 2019 combinedly).

T.AGG	Location	Karima GHS			Moi Forces Academy		
	Dataset	SATELLITE	ERA5	IN-SITU	SATELLITE	ERA5	IN-SITU
Daily	RMSE	29.472	214.931	1071.346	-92.23	228.92	851.52

	rho	0.994	0.884	0.820	1.02	0.86	0.74
	rhosqr	0.988	0.781	0.672	1.04	0.74	0.54
	rank	1	2	3	1	2	3
Pentad	RMSE	33.557	58.476	390.85	-42.79	81.08	340.87
	rho	0.989	0.958	0.889	1.01	0.934	0.78
	rhosqr	0.978	0.918	0.790	1.03	0.87	0.6
	rank	1	2	3	1	2	3
Dekad	RMSE	22.325	34.932	258.373	-30.352	52.561	250.883
	rho	0.99	0.969	0.909	1.013	0.948	0.752
	rhosqr	0.980	0.939	0.826	1.026	0.899	0.566
	rank	1	2	3	1	2	3

5.1.3. Air temperature

In table 5.7, the ETC analysis results are shown for the monthly air temperature dataset. At Ole Tipis GHS, the correlation coefficients are all high and the TAHMO dataset has the highest compared to the other datasets. At Moi Forces Academy, biased estimates are once again observed.

Table 5.7: ETC results for air temperature.

Location	*Ole Tipis GHS*			*Moi Forces Academy*		
Dataset	GHCN + CAMS	*ERA5*	*TAHMO*	GHCN + CAMS	*ERA5*	*TAHMO*
RMSE	0.123	0.219	0.098	0.295	-0.010	0.391
rho	0.944	0.953	0.974	0.798	1.003	0.920
rhosqr	0.891	0.908	0.949	0.637	1.006	0.846
rank	3	2	1	3	1	2

5.2. Results interpretation and discussion

5.2.1. Evaluation of the ETC results

The ETC estimates for all variables used in this study (rainfall, solar radiation, and air temperature) are in reasonable ranges, and in most cases, the lowest root mean square errors corresponded to the highest correlation coefficients, signal-to-noise ratio and ranks. To evaluate the magnitudes of the ETC estimates let's follow, let's follow McColl et al., (2014) and consider the unbiased root mean square error $\boldsymbol{\sigma_{\varepsilon}} = 47.356$ and correlation coefficient $\boldsymbol{\rho_{t,X}} = 0.604$ calculated for the CHIRPS satellite daily rainfall product at Delamere farm in the year 2018. By replacing these values in equation 6 (section 2.3.1 of the thesis) and also assuming the sensitivity of the measurement system $\boldsymbol{\beta} \approx \mathbf{1}$, we obtain a realistic estimate $\boldsymbol{\sigma_{t}} \approx 58.485$.

In cases where the biased ETC estimates were observed, it is due to violation of the implicit initial triple collocation assumptions of error stationarity(zero-mean error and variance homoscedasticity) and error orthogonality, the effects of which, are discussed in detail by Tugrul Yilmaz & Crow, (2014). In such cases, the verification cannot be performed because the calculation done in the previous paragraph result in the square root of a negative number.

5.2.2. Effects of time aggregation levels

Biased ETC results were, were noted in many cases for all variables used in this study but they were more prominently noticeable in pentad and dekad datasets of rainfall where the datasets' ranges and variability increased significantly as a result of the aggregation done by summation of consecutive days. It suggests that, for a geographical area like the Lake Naivasha basin with highly varying climatic conditions, increased variability because of the aggregations (which is also partly due to the presence of many zero values in rainfall datasets), the potential violation of the triple collocation assumptions is to be expected. In solar radiation datasets, biased ETC outcomes were noted more remarkably at Karima GHS in 2019 and Moi Forces Academy in all set-ups. The time aggregations affected only marginally the ETC outcomes because the aggregations were made by averaging consecutive days which doesn't change significantly the data ranges.

Regarding the results obtained for combined time-series(2 years for rainfall and 3 years for solar radiation), the ETC estimates improved in pentad rainfall at Moi Forces Academy and dekad solar radiation at Karima GHS. It implies that depending on the magnitudes and distribution of errors in the datasets, the effects of the violation of error stationarity and orthogonality assumptions may be suppressed when the used time-series are sufficiently long.

5.2.3. Effects of outliers

Consistent with the introduction of outliers, the correlation coefficient in the daily rainfall estimates of the in-situ datasets at Ole Tipis GHS in 2018 has decreased from 0.81(highest compared to the CHIRPS and ERA5) in the original dataset to 0.69 (second highest after ERA5). This is directly related to the increment of the root mean square error for the in-situ dataset after the introduction of outliers (from 25 up to 130). It implies that outliers may degrade the outcomes of the extended triple collocation depending on the distribution and amount of outliers present and how they affect the dataset's variability.

5.2.4. Effects of data gaps

The correlation coefficient estimated for the in-situ rainfall at Taita Mauche SS in 2018 has the highest correlation coefficient despite a large amount of missing data compared to CHIRPS and ERA5 products. It indicates that ETC results could be misleading as far as data gaps are concerned due to the simple reason that statistics will be calculated for the available data values only. Depending on the underlying data distribution properties, their variance may be the lowest compared to the other datasets and will result consequently in a higher correlation coefficient.

5.3. The triple sensor ILWIS toolbox

An ILWIS geospatial procedure to perform the ETC analysis and visualize cartographically the results has been created. Here we use it to perform the ETC on selected rainfall datasets to assess its performance and propose improvements. The datasets used for this are the CHIRPS, ECMWF ERA5, and in-situ TAHMO monthly rainfall at Moi Forces Academy, Ole Tipis GHS, and Taita Mauche SS for the year 2019. The toolbox allows 3 data inputs: two raster time-series referred to as maplists and one point map as shown in figure 5.1. In this case, the raster time-series are the CHIRPS and ECMWF ERA5 rainfall products while the point map is produced using in-situ coordinates and corresponding rainfall values.

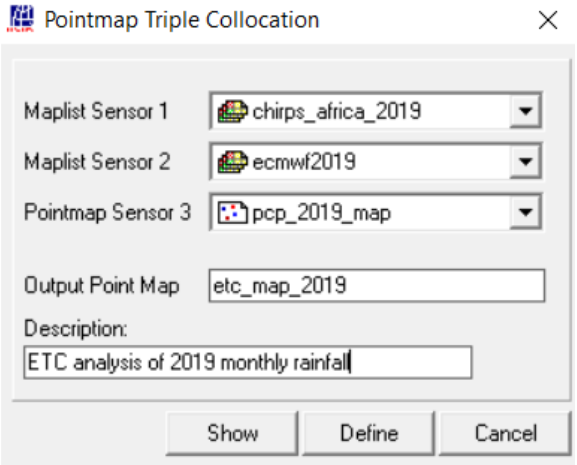


Figure 5.1: Triple sensor data input.

After clicking 'show' the toolbox runs the ETC covariance analysis and yields a point map showing the stations involved and an associated attribute table as follows:

Table "etc_2019_pcp" - ILWIS

File Edit Columns Records View Help

	w1	w2	w3	r1	r2	r3	rhosq1	rhosq2	rhosq3	errvar1	errvar2	errvar3
pnt 1	0.608	1.285	0.516	2	1	3	0.370	1.651	0.266	2615.9	-24.231	9866.066
pnt 2	0.948	0.954	0.680	2	1	3	0.898	0.909	0.463	451.50	0.785	334.751
pnt 3	0.956	0.926	0.880	1	2	3	0.914	0.857	0.774	320.35	1.115	3219.778

Figure 5.2: ETC results initial attributes table.

The attributes w1,w2,w3 are the correlation coefficients calculated for each input(Maplist sensor 1, Maplist sensor 2, and Point map sensor 3). A higher value indicates a higher fidelity in the corresponding dataset with respect to the unknown quantity of the target variable.

The attributes r1,r2, and r3 and the ranking given to the correlation coefficient in descending order. For example at pnt1(which corresponds to the first station Moi Forces Academy), the Maplist sensor 2 corresponding to the ERA5 dataset has the highest correlation coefficient.

The attributes rhosqur1, rhosqr2, and rhosqr3 are the squares of the correlation coefficient which are the unbiased signal to noise ratio shown in equation 4 in the methodology section.

Lastly, the errvar1, errvar2, and errvar3 are the RMSEs calculated using the ETC for each dataset.

The next step is the visualization of the results in a map using RGB values for each location using a Red-Green-Blues triangular legend. The RGB values are calculated as follows:

Blue = 255*rhosq1/max(rhosq1,rhosq2,rhosq3), High correlation coefficient for Satellite data (w1) contribute to the color Blue.

Red = 255*rhosq2/max(rhosq1,rhosq2,rhosq3) High correlation coefficient for ERA5 data (w2) contribute to the colour Red.

Green = 255*rhosq3/max(rhosq1,rhosq2,rhosq3) High correlation coefficient for in_situ data (w3) contribute to the colour Green.

An attribute filed called COLOR_TC is added to the attribute table and will then be used to color the different points depending on which datasets has a higher correlation coefficient:

color_tc:=COLOR(255*rhosq2/max(rhosq1,rhosq2,rhosq3),255*rhosq3/max(rhosq1,rhosq2,rhosq3), 255*rhosq1/max(rhosq1,rhosq2,rhosq3)).

The last column in Figure 5.3 contains the color contributions for each dataset:

Table "etc_2019_pcp" - ILWIS

File Edit Columns Records View Help

	w1	w2	w3	r1	r2	r3	rhosq1	rhosq2	rhosq3	errvar1	errvar2	errvar3	color_tc
pnt 1	0.608	1.285	0.516	2	1	3	0.370	1.651	0.266	2615.9	-24.231	9866.066	(255, 41, 57)
pnt 2	0.948	0.954	0.680	2	1	3	0.898	0.909	0.463	451.50	0.785	334.751	(255,130,252)
pnt 3	0.956	0.926	0.880	1	2	3	0.914	0.857	0.774	320.35	1.115	3219.778	(239,216,255)

Figure 5.3: ETC final attribute table with colour values.

Figure 5.4 is a map visualizing the final output of the ETC analysis in the study area using the triple sensor ILWIS toolbox. It can be seen that the three stations namely; Moi Forces Academy, Taita Mauche SS, and Ole Tipis GHS are represented by different colors, depending on which dataset had the highest correlation coefficient and signal-to-noise ratio than others.

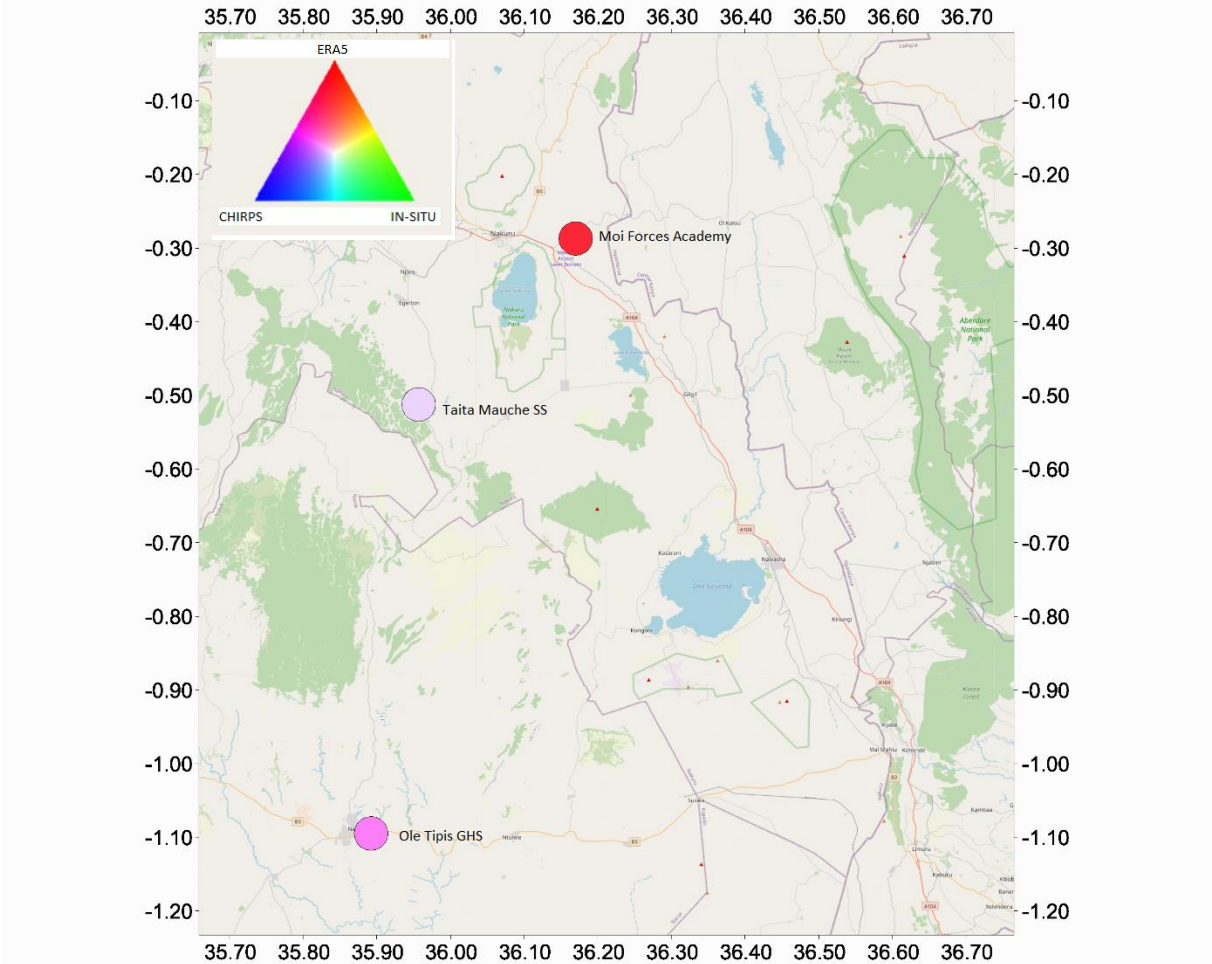


Figure 5.4: ETC results visualization.

6. CONCLUSIONS AND RECOMMENDATIONS

6.1. Conclusions

In this research, we investigated the Triple Sensor Approach for Monitoring water and Climate using the extended triple collocation covariances analysis to identify the most reliable climatic data source between In-situ measurements, Satellite-based observations and model-generated datasets. The following conclusions have been drawn:

- Measurement systems employing different methods will lead to different data estimates. While this seems obvious enough, the differences evidenced by relatively low Pearson correlation coefficients calculated between the datasets in daily rainfall were very remarkable. It emphasizes once more the necessity of a robust approach like the triple sensor method because it allows a more absolute way to assess dataset reliability. The ETC analysis results provide more information about the dataset's measurement system sensitivity, the variability of the signal, and measurement error as noted by McColl et al., (2014). This information is a compelling basis to judge more objectively which dataset can be trusted most at a given location, compared to the normal dual comparison data validation approaches.

- Time aggregation levels can greatly affect the estimates of the ETC analysis results depending on how the datasets are aggregated. For instance, the aggregates of rainfall which were obtained by summation resulted in biased ETC results because of the increased data ranges and variability. For variables where the aggregation is calculated by averaging consecutive data points(e.g. solar radiation), the ETC results are only slightly affected.

- The triple sensor approach has the capability to penalize datasets characterized by the presence of outliers considering the high sensitivity of the ETC covariance analysis to outliers. The ETC results will be skewed depending on the number of outliers and their effect on the data distribution properties. In this study, the introduced outliers increased the corresponding dataset's root mean square error and its correlation coefficient, signal-to-noise ratio, and consequently leading to a lower reliability rank.

- The triple sensor approach is on the other hand is unable to expose datasets with gaps because the ETC statistics are only performed for the part of the datasets with available values. If the datasets with gaps have the lowest variance (observed case in this study) and the highest ETC correlation coefficient and signal-to-noise ratio, it may be misleadingly interpreted as the most reliable.

- Limitations to the applicability of the triple sensor approach have also been noted. A typical case is when the distribution properties(ranges, variability) of the used dataset violate the initial triple collocation assumptions of zero-mean error, variance homoscedasticity, and no-error cross-correlation between the used datasets. This limits the situations in which the triple sensor approach can be employed.

- Another important limitation of the triple sensor approach that was observed is related to its datasets requirement. The ETC analysis requires at least three datasets that must be temporally and spatially collocated. It is not easily achievable because many datasets have differences, especially in temporal resolutions, length of time-series.

- Regarding the data quality of different data sources and types, we observed that especially the in-situ rainfall station data present a high potential source of error. Normally, the people's perception of data quality inter-comparison is that in-situ measurements represent the ground truth, and cannot be biased. This was obviously not the case, as we investigated in this study. Proper maintenance of local automated weather stations (AWS) is of prime importance, to obtain reliable data. This holds especially for rainfall, which usually is still a kind of mechanical measurement (tipping bucket or other principles). For solar radiation, air temperature, and soil moisture sensors other sensors, this is less the case as we observed during fieldwork, although regular control is required for example to change power supply batteries. Too many people think that in-situ water and climate observation ends when a device (station) has been installed, and the data are transmitted automatically (e.g. GPRS, etc.), in weather and other environmental observation is more than that.

- The results of this research emphasize, further, the potential of independent citizen observer climatic datasets in scientific research. It is exemplified by cases where citizen datasets (in-situ rainfall datasets estimated by Delamere farm own rain gauge) performed relatively well and in some scenarios, better than the most commonly used satellite and model datasets. data sources.

6.2. Recommendations

6.2.1. General recommendations

- For the triple sensor approach to allow useful information regarding climatic datasets' reliability, the used datasets should be spatially and temporally collocated and should be representing the same climatic variable. The datasets can be from the same source (for example satellite) as long as the retrieval processes are mutually independent and the distribution properties do not violate any of the triple collocation assumptions. In cases where the assumptions are violated the results will be biased and can therefore not be reliable. It is advised to use different datasets or to attempt manipulating the time-series where possible.

- A suitability analysis of the datasets at hand for the ETC analysis should also be prioritized as well as using other data filtering techniques such as the identification of data before the application of ETC analysis to filter out datasets that can potentially add a margin for mistakes in the ETC results' interpretation.

- In terms of data collection and quality, it was realized that the ITC low-cost Sodaq (Kukua) weather stations in the Naivasha area are of poor manufacturing and electronic component quality, but also poorly maintained especially rain gauges and therefore the datasets cannot be easily used for any scientific research. Another important element is the solar radiation datasets which are available but cannot be used because the conversion factor of the dataset was not provided by the instrument manufacturer. The AWS systems of the TAHMO project and African network showed a variable behavior in data quality. At some locations, rather reliable data time series are collected, where other locations show similar problems as the ITC stations. The time interval between manual check-ups and maintenance activities such as the cleaning of rain gauges is too long. We recommend more frequent check-ups and maintenance to facilitate future research in the area.

- The use of independent citizen observer datasets in climate science has also been emphasized. When available, these datasets can be very useful, but also on the sole condition, that the in-situ citizen observatories and stations undergo regular control and maintenance and proper operational procedures are documented.

6.2.2. Improvement of the triple sensor toolbox in ILWIS

- One of the main setbacks of the triple sensor approach and also generally in scientific research is data availability and accessibility in the required spatial and temporal resolutions. To the average user, the task of data identification and retrieval may be daunting. A data retrieval method through the triple sensor toolbox for the most commonly used and available datasets should be enabled in ILWIS as it has been done for the ISOD toolbox. Also regarding the dataset, the current triple sensor toolbox only allows the entry of two raster map lists and one point map. A lot more flexibility is required because some datasets are only available for example as rasters only or point maps.

- Warning routines can be added in cases where the condition for a proper application of the triple sensor approach are not respected to facilitate the ETC results' interpretation for the datasets' reliability: Spatial and temporal inconsistencies in the datasets, datasets with negative cross-relationships(relating the datasets suitability analysis performed earlier) and biased ETC estimates.

- A simple reliability score can be calculated and added to the ETC results table to ease the interpretation of the results. The squared correlation coefficient of a given dataset would be divided by the maximum correlation observed in all the collocated datasets at that particular location. This means that for the maximum correlation coefficient itself, the reliability score would be 1. As an example, let's take daily rainfall at Delamere farm in 2018 whereby the squared correlation coefficients of the involved datasets are as follows:
  CHIRPS=0.36. ERA5=0.14 and In-situ=0.26.
  The reliability scores would be:
  CHIRPS=0.36/0.36=1, ERA5=0.13/0.36=0.44 and In-situ=0.26/0.36=0.72.

- An automatic legend function for the cartographic visualization of the triple sensor outcomes should be created to allow a complete and smooth visual interpretation. Currently, it is only available in the triple sensor web demo and it is a triangular type of legend based on a mixture of RGB colors to represent the results of the triple sensor approach as shown in figure 24. Each collocated dataset contributes to the mixture of colors and the most dominant color corresponds to the dataset with the highest correlation coefficient(the most reliable). This does not enable a straight forward visual identification of the best performing dataset at that location when the correlation coefficients of the used datasets are close. The legend should rather be calculated using the reliability score proposed previously, in such fashion that a data source would be represented by a unique color every time it achieves the highest reliability score in comparison with the other collocated data sets.

List of references

Apte, A. (2015). An introduction to data assimilation. https://doi.org/10.1007/978-81-322-2547-8_4

Atmospheric Science Data Center. (2017). CERES SYN1deg Ed4A Data Quality Summary, 41. https://doi.org/https://ceres.larc.nasa.gov/dqs.php#level3table

Becht, R., Odada, E., & Higgins, S. (2005). Lake Naivasha: experience and lessons learned brief (Lake basin management initiative): Experience and lessons learned briefs, *5*, 277–298. Retrieved from http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Lake+Naivasha+Experience+and+Lessons+Learned+Brief#0

CERES Team. (n.d.). CERES Data Product Information. Retrieved June 18, 2020, from https://ceres.larc.nasa.gov/data/#synoptic-toa-and-surface-fluxes-and-clouds-syn

Chen, F., Crow, W. T., Colliander, A., Cosh, M. H., Jackson, T. J., Bindlish, R., … Seyfried, M. S. (2017). Application of Triple Collocation in Ground-Based Validation of Soil Moisture Active/Passive (SMAP) Level 2 Data Products. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *10*(2), 489–502. https://doi.org/10.1109/JSTARS.2016.2569998

Copernicus Climate Change Service (C3S). (2019). ERA5-Land hourly data from 1981 to the present. https://doi.org/10.24381/cds.e2161bac

Copernicus Climate Change Services. (2018). ERA5 | ECMWF. Retrieved March 23, 2020, from https://www.ecmwf.int/en/forecasts/datasets/reanalysis-datasets/era5

Dembélé, M., & Zwart, S. J. (2016). Evaluation and comparison of satellite-based rainfall products in Burkina Faso, West Africa. *International Journal of Remote Sensing*, *37*(17), 3995–4014. https://doi.org/10.1080/01431161.2016.1207258

Duan, Z., Liu, J., Tuo, Y., Chiogna, G., & Disse, M. (2016). Evaluation of eight high spatial resolution gridded precipitation products in Adige Basin (Italy) at multiple temporal and spatial scales. *Science of the Total Environment*, *573*, 1536–1553. https://doi.org/10.1016/j.scitotenv.2016.08.213

Ebert, E. E., Janowiak, J. E., & Kidd, C. (2007). Comparison of near-real-time precipitation estimates from satellite observations and numerical models. *Bulletin of the American Meteorological Society*, *88*(1), 47–64. https://doi.org/10.1175/BAMS-88-1-47

ECMWF. (2019). ERA5-Land monthly averaged data from 1981 to the present. https://doi.org/10.24381/cds.68d2bb30

Eney, A. B., & Petzold, D. E. (1988). The spatial variability of rainfall pH in Washington, D.C. Metropolitan area. *Professional Geographer*, *40*(3), 315–326. https://doi.org/10.1111/j.0033-0124.1988.00315.x

Entekhabi, D., Reichle, R. H., Koster, R. D., & Crow, W. T. (2010). Performance Metrics for Soil Moisture Retrievals and Application Requirements. *Journal of Hydrometeorology*, *11*(3), 832–840. https://doi.org/10.1175/2010jhm1223.1

Fan, Y., & Dool, H. Van Den. (2008). A global monthly land surface air temperature analysis for 1948 – present, *113*, 1–18. https://doi.org/10.1029/2007JD008470

Funk, C., Peterson, P., Landsfeld, M., Pedreros, D., Verdin, J., Shukla, S., … Michaelsen, J. (2015). The climate hazards infrared precipitation with stations - A new environmental record for monitoring extremes. *Scientific Data*, *2*. https://doi.org/10.1038/sdata.2015.66

Kato, S., Loeb, N. G., Rutan, D. A., & Rose, F. G. (2015). Clouds and the earth's radiant energy system (CERES) data products for climate research. *Journal of the Meteorological Society of Japan*, *93*(6), 597–612. https://doi.org/10.2151/jmsj.2015-048

Kidd, C., & Huffman, G. (2011). Global precipitation measurement. *Precipitation: Advances in Measurement, Estimation and Prediction*, *353*, 131–169. https://doi.org/10.1007/978-3-540-77655-0_6

Kim, K., Price, K., Whelan, G., Galvin, M., Wolfe, K., Duda, P., … Pachepsky, Y. (2014). Using remote sensing and radar MET data to support watershed assessments comprising IEM. *Proceedings - 7th International Congress on Environmental Modelling and Software: Bold Visions for Environmental Modeling, IEMSs 2014*, *2*(September 2015).

Kuhn, A., Britz, W., Willy, D. K., & van Oel, P. (2016). Simulating the viability of water institutions under volatile rainfall conditions - The case of the Lake Naivasha Basin. *Environmental Modelling and Software*, *75*, 373–387. https://doi.org/10.1016/j.envsoft.2014.08.021

Malardel Sylvie, C. E. de P. M. à M. T. (CEPMMT). (2019). Assimilation of meteorological data. Retrieved October 7, 2019, from www.encyclopedie-environnement.org/en/air-en/weather-forecasting-

models/
McColl, K. A., Vogelzang, J., Konings, A. G., Entekhabi, D., Piles, M., & Stoffelen, A. (2014). Extended triple collocation: Estimating errors and correlation coefficients with respect to an unknown target. *Geophysical Research Letters*, *41*(17), 6229–6236. https://doi.org/10.1002/2014GL061322
Mirzaei, S., Raoof, M., Ghasemi, A., Etaati, H., Moradnezhadi, M., & Mirzaei, Y. (2014). *Bulletin of Environment, Pharmacology and Life Sciences O OR RI IG GI IN NA AL L A AR RT TI IC CL LE E Determination of a Some Simple Methods for Outlier Detection in Maximum Daily Rainfall (Case Study: Baliglichay Watershed Basin-Ardebil Province-Iran). Env. Pharmacol. Life Sci* (Vol. 3).
Muller, C. L., Chapman, L., Johnston, S., Kidd, C., Illingworth, S., Foody, G., … Leigh, R. R. (2015). Crowdsourcing for climate and atmospheric sciences: Current status and future potential. *International Journal of Climatology*, *35*(11), 3185–3203. https://doi.org/10.1002/joc.4210
Odongo, V. O., Onyando, J. O., Mutua, B. M., van Oel, P. R., & Becht, R. (2013). Sensitivity analysis and calibration of the Modified Universal Soil Loss Equation (MUSLE) for the upper Malewa Catchment, Kenya. *International Journal of Sediment Research*, *28*(3), 368–383. https://doi.org/10.1016/S1001-6279(13)60047-5
Silvertown, J. (2009). A new dawn for citizen science. *Trends in Ecology and Evolution*. https://doi.org/10.1016/j.tree.2009.03.017
Stanski, H. R., Wilson, L. J., & Burrows, W. R. (1989). Survey of Common Verification Methods in Meteorology - 2. *Atmospheric Research*, 9–42.
Stoffelen, A. (1998). Toward the true near-surface wind speed: Error modeling and calibration using triple collocation. *Journal of Geophysical Research: Oceans*, *103*(C4), 7755–7766. https://doi.org/10.1029/97jc03180
Tapiador, F. J., Turk, F. J., Petersen, W., Hou, A. Y., García-Ortega, E., Machado, L. A. T., … de Castro, M. (2012). Global precipitation measurement: Methods, datasets and applications. *Atmospheric Research*. https://doi.org/10.1016/j.atmosres.2011.10.021
Tolstykh, M. A., & Frolov, A. V. (2005). Some current problems in numerical weather prediction. *Izvestiya - Atmospheric and Ocean Physics*, *41*(3), 285–295.
Tugrul Yilmaz, M., & Crow, W. T. (2014). Evaluation of assumptions in soil moisture triple collocation analysis. *Journal of Hydrometeorology*, *15*(3), 1293–1302. https://doi.org/10.1175/JHM-D-13-0158.1
van Oel, P. R., Mulatu, D. W., Odongo, V. O., Meins, F. M., Hogeboom, R. J., Becht, R., … van der Veen, A. (2013). The Effects of Groundwater and Surface Water Use on Total Water Availability and Implications for Water Management: The Case of Lake Naivasha, Kenya. *Water Resources Management*, *27*(9), 3477–3492. https://doi.org/10.1007/s11269-013-0359-3
Vogelzang, J., & Stoffelen, A. (2012). Triple collocation Triple collocation.
Watson, J., & Challinor, A. (2013). The relative importance of rainfall, temperature, and yield data for a regional-scale crop model. *Agricultural and Forest Meteorology*, *170*, 47–57. https://doi.org/10.1016/j.agrformet.2012.08.001
Wu, E., Liu, W., & Chawla, S. (2010). Spatio-temporal outlier detection in precipitation data. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Vol. 5840 LNCS, pp. 115–133). https://doi.org/10.1007/978-3-642-12519-5_7

7. APPENDIX

Table 7.1: Ground weather stations coordinates.

Name	X	Y	Z
Moi Forces Academy	36.16998	-0.28712	1936
Karima Girls	36.5875	-0.50084	2523
Ole Tipis Girls	35.89234	-1.09468	1922
Molo Academy	35.7289	-0.23934	2508
Nyandarua Highschool	36.37694	-0.20018	2421
Murungaru Sec	36.49309	-0.58664	2423
Magomano Sec	36.58292	-0.60715	2482
Taita Mauche	35.97424	-0.50307	2441
Kijabe Farm	36.19085	-0.75382	2074
Delamere Farm	36.41029	-0.68526	1910
Paul Farm	36.56659	-0.59977	2468
KWSTI	36.44983	-0.73678	1998
Nunjoro farm	36.47917	-0.64381	2237