TECHNISCHE UNIVERSITEIT DELFT

APPLYING DEEP LEARNING VS MACHINE LEARNING MODELS TO REPRODUCE DRY SPELLS AT POINT SCALE FROM SATELLITE INFORMATION IN A DATA-SCARCE REGION: THE CASE OF NORTHERN GHANA

M.SC. THESIS

PANAGIOTIS MAVRITSAKIS 5135400

CIVIL ENGINEERING M.SC. | TRACK: WATER MANAGEMENT

TU DELFT

Table of Contents

Table of Figures

Table of Tables

Abstract

Large parts of the world rely on rainfed agriculture for their food security. In Africa, 90% of the agricultural yields rely only on precipitation for irrigation purposes and approximately 80% of the population's livelihood is highly dependent on its food production. Parts of Ghana are prone to droughts and flood events due to increasing variability of precipitation phenomena. Crop growth is sensitive to the wet- and dry spell phenomena during the rainy season. To support rural communities and small farmer in their efforts to adapt to climate change and natural variability, it is crucial to have good predictions of rainfall and related dry/wet spell indices.

This research constitutes an attempt to assess the dry spell patterns in the northern region of Ghana, near Burkina Faso. We aim to develop a model which by exploiting satellite products overcomes the poor temporal and spatial coverage of existing ground precipitation measurements. For this purpose 14 meteorological stations featuring different temporal coverage are used together with satellite-based precipitation products.

Conventional machine-learning and deep-learning algorithms were compared in an attempt to establish a link between satellite products and field rainfall data for dry spell assessment. The deep-learning architecture used should be able to efficiently manipulate spatial data. Hence, Convolutional Neural Networks were used in order to detect spatial patterns in the satellite data.

Using these models we will attempt to exploit the long temporal coverage of the satellite products in order to overcome the poor temporal and spatial coverage of existing ground precipitation measurements. Doing that, our final objective is to enhance our knowledge about the dry spell characteristics and, thus, provide more reliable climatic information to the farmers in the area of Northern Ghana.

1. Introduction

1.1 Relevance of the study

Precipitation is a major component of agriculture. Large parts of the world rely on rainfed agriculture for their food security. In particular, in Africa, 90% of the agricultural yields rely only in precipitation for irrigation purposes (Fischer et al., 2013; Froidurot & Diedhiou, 2017) and approximately 80% of the population's livelihood is highly dependent on this food production (Collier et al., 2008; Gyasi et al., 1997; J. Rockström, 2000). Taking into account that the population of Africa is expected to double in 2050 (Population Reference Bureau, 2019) together with the shift to more water resource-intensive diets in sub-Saharan Africa (Savenije, 2000; WWAP, 2000), pressure for more efficient water management increases. Apart from this, additional pressure is forced to farmers in sub-Saharan Africa, as it is more likely that they are going to be heavily impacted by climate change, enhancing the erratic patterns of rainfall in the region (Sanoussi et al., 2015). In fact, during the last decades the frequency of false start of the rainy seasons and the duration of dry spells during the rainy season in the region have increased remarkably (Salack et al., 2016).

Crop growth is sensitive to the wet- and dry-spell phenomena during the rainy season. In Ghana demands for operational predictions of rainfall and related indices are necessary to support rural communities and especially smallholder farmers (Gbangou et al., 2019; Gbangou et al., 2020; Sultan et al., 2020), as many of them are making efforts to adapt to the climatic variability. For the agricultural productivity to be increased, extensive weather and climatic information related to wet- and dry spell occurrence is needed for smallholder farmers in West Africa (Codjoe et al., 2014; Yaro, 2013) and, in particular, in the coastal savannah of Ghana delta area (Gbangou et al., 2019). Nevertheless, as Masinde et al. (2012) states, the provided services by the public authorities have been of little use for smallholder farmers. Indeed, the drought assessment should always make the stakeholder central and think about which indicator is useful for him/her (Savenije, 1999).

Unfortunately, despite the efforts made to enhance smallholder agricultural production, the agricultural productivity in the region is still low (Rockström et al., 2004), even though approximately 65% of the Ghanaian population owning or working in farms (Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ), 2019).

According to GIZ (2019), the increasing variability of precipitation phenomena leads to approximately 15% of the Ghanaian land to be prone to droughts and floods. Recent studies show a delay in the arrival of the wet season and a reduction in the total amount of rain (Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ), 2019). An adaptive technique used to protect against droughts is to plant crops far apart, in the expense of crop productivity. Ellis (1998) mentions that other adaptive techniques used by Ghanaians include planting different crop types (further expained in Asfaw et al., 2019), differently located farm types, intercropping, animal husbandry, hunting and gathering, trade and migratory wage labour.

Given the lack of irrigation in the area, dry spells during the growing season can be crucial for the production of agricultural yields. In addition to this, coarse time-scale indices like cumulative rainfall can be deceptive, because they do not account for the temporal variation of precipitation, giving the impression that there is no lack of green (or white-to-green) water when it actually is (Savenije, 2000; Usman & Reason, 2004).Therefore, the need for efficient predictions and efficient hydrological modelling arises. Until now assessing the risk of dry spells in the region is limited by the forecast skill of either the global numerical weather prediction models (Vogel et al., 2018) or by the low confidence CMIP5 rainfall projections (Christensen et al., 2013), as these models fail to perform adequately in climate conditions where the nature of precipitation is convective (Maranan et al., 2018). The spatial resolution can greatly influence model outcomes and models using raster based precipitation data outperform models that use precipitation derived from point measurements (Guo et al., 2004; Smith et al., 2004). Hence, finer resolution precipitation data are crucial for improving our understanding of basin-scale hydrology (Badas et al., 2006). Unfortunately, providing localized precipitation data is restricted by the sparse distribution of rain gauges (Wilheit, 1986). The problem arises not only because of the smaller number of in-situ measurements in the region, but also because of reduced representativeness of data-scarce regions in the calibration process of remote-sensing precipitation products.

1.2 Research questions

Overall, the trigger of this research is to provide efficient climatic information to the smallholder farmers of Northern Ghana. Our approach is to use the longer temporal coverage of the remote-sensing products in order to downscale the existing ground observations. The term downscaling refers to a procedure which takes information known at large scales as input to make predictions at higher resolutions. The major problem of ground-truthing is that the satellite-based value is on average different from the point observation. Point precipitation measurements are considered more reliable and representative of the scale of the plots of small-holder farmers than the existing satellite data, as the reanalysis of the satellite products is global leading to under-representation of gauge-scarce regions like West Africa (Beck et al., 2017).

Thus, for the main research question of the research we aim to reproduce the long and short dry spell sequences seen by the rain gauges (point scale) in the region of Northern Ghana based on satellite observations using different Machine Learning (ML) and Deep Learning (DL) tools. In this application dry spell occurrence is defined as five consecutive days with precipitation below 1mm/day. Several ML classification models are compared to different Artificial Neural Network (ANN) models that reproduce dry spell sequences with conventional classification algorithms. A classification model tries to draw some conclusion from the input values given for training. It predicts the class labels for the new data.

The different networks are trained using satellite images with constant spatial extent as input and a Rain/No Rain in daily time-scale as label. This model utilizes the satellite images with temporal coverage longer than the gauge time-series in order to reproduce rainfall intermittency time-series to the temporal extent of the satellite dataset. Moreover, in order to incorporate the climatological information of the area, models featuring each stations latitude were also tested. As a next step, the overall performance of the classification models is assessed, examining the reasons of mismatch between the different data sources.

Deep learning has started to take place in many hydrological applications, especially in studies involving remote-sensing data. Leinonen et al. (2020) used Generative Adversarial Networks to enhance the spatial resolution of remote-sensing precipitation data, managing decent results. Yang et al. (2019) assessed the potential added value of Long Short-Term Memory (LSTM) ANNs (Hochreiter & Schmidhuber, 1997) to improve flood modeling. Kratzert et al. (2018) proved that LSTM trained for rainfall-runoff modeling can outperform state-of-the-art process-based hydrological models. Overall, an extensive review of deep learning applications in hydrology can be found in (Sit et al., 2020).

Apart from the DL approaches, a Multi-linear Regression (MLR) model has already been tested on the data, with the help of the scikit-learn machine learning library (Pedregosa, 2011) and the Keras deep learning library (Chollet, 2015), using as input the 9-nearest-to-the-gauge pixels of the satellite precipitation time series and the mean value of these pixels in the previous time-step. The labels used were the ground observations. Unfortunately, extending the spatial and temporal characteristics of the input data did not add any additional value in the regression process. Hence, this MLR approach will not be elaborated any further in this report.

Having reproduced the dry spell sequences for a longer temporal extent, the second research question of this thesis is to examine the possibility to apply the trained downscaling model over a region wider than Northern Ghana beyond for locations at which we used rain gauge data to train the classification model. Given the data availability in countries near North Ghana, it becomes possible to test the model in certain locations in Benin, Burkina Faso and Niger and even wider. By doing that, we are able to link the dry spell occurrence to latitude and compare against the climatological information of other studies, as explained in Chapter 2.2.

All the data analysis was performed in Python programming language (Van Rossum, G. & Drake, 2009) and visualized by Matplotlib (Hunter, 2007).

The manuscript is organized as follows. In Chapter 2 an overview of the region is presented: political and financial conditions in the region, climatic conditions and precipitation generating mechanisms and mapping of the aforementioned region. The available datasets, both remote-sensing products and gauge measurements, of precipitation in the wider region (northern Ghana, Burkina Faso, Benin, Niger and Mali) are considered in Chapter 3. A deeper analysis of the properties of the existing datasets and an analysis of their consistency is conducted in Chapter 4. In Chapter 5 both the ML and DL methodological approaches that are used for downscaling of the existing datasets are presented. Also, in Chapter 5 the dry spell assessment indices and their relevance to smallholder farmers in the region are examined. In Chapter 6 the results of the classification models are presented and further interpreted. Moreover, some more sophisticated models are introduced and examined as well. Chapter 7 presents the main conclusions of the research and provides recommendations for further research. Chapter 8 contains the bibliography of the study.

2. Case study

2.1 Region of interest

The British colonial rule in Ghana ended in 1957 and the democratization of Ghana occurred as part of the ‘third wave’ of democratization of African countries during the 90’s (Debrah, 2009). North Ghana has been over time the poorest part of the country, being less urbanized than the southern one. As Laube et al. (2012) states: *“the area suffers from difficult climatic conditions, relatively high population density and patterns of underdevelopment, which are the result of discriminatory colonial and post-colonial policies, while the population -consisting of a number of relatively small ethnic groups-is largely dependent on agriculture.”*. Levels of poverty have been increasing, despite poverty decreasing in the other regions of Ghana (Ghana Statistical Service).

The regional area of Southwest Africa, which includes Ghana, contains a broad range of ecosystems. An overview of the physical geography and the political boundaries of the region of interest together with several important urban centers are depicted in Figure 1. The GIS layer was preprocessed in the open-source QGIS software (QGIS, 2009) was acquired by (Tappan et al., 2016).

Ghana’s sovereign financial sector is agriculture, being world's second largest cocoa producer. Except from that, Ghana also produces beans, palm, pineapples, cotton, tomatoes, bananas, coconuts and tobacco. It can be stated that the financial discrepancies between northern and southern Ghana are a subject of the different ecological zones of the two regions. The northern part of the country is dominated by the savannah agro-ecological zone while in the southern part of the country land-use varies between the deciduous forest, evergreen and the coastal savannah agro-ecological zone (Figure 2).

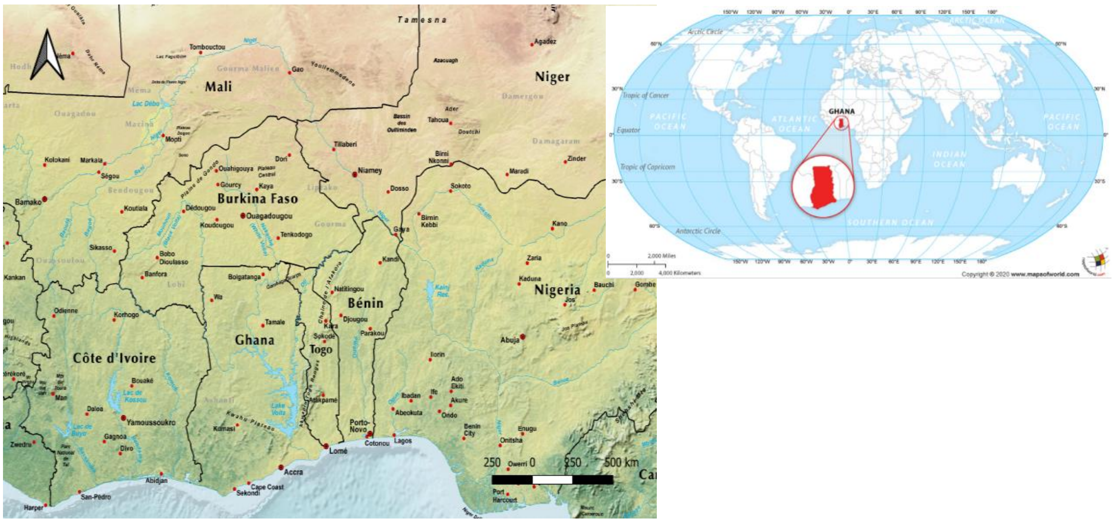


Figure 1: Land cover and political boundaries in the Southwest Africa region. Major cities in the area are depicted with red. World map provided for reference of Ghana by (mapsofworld.com, 2020).

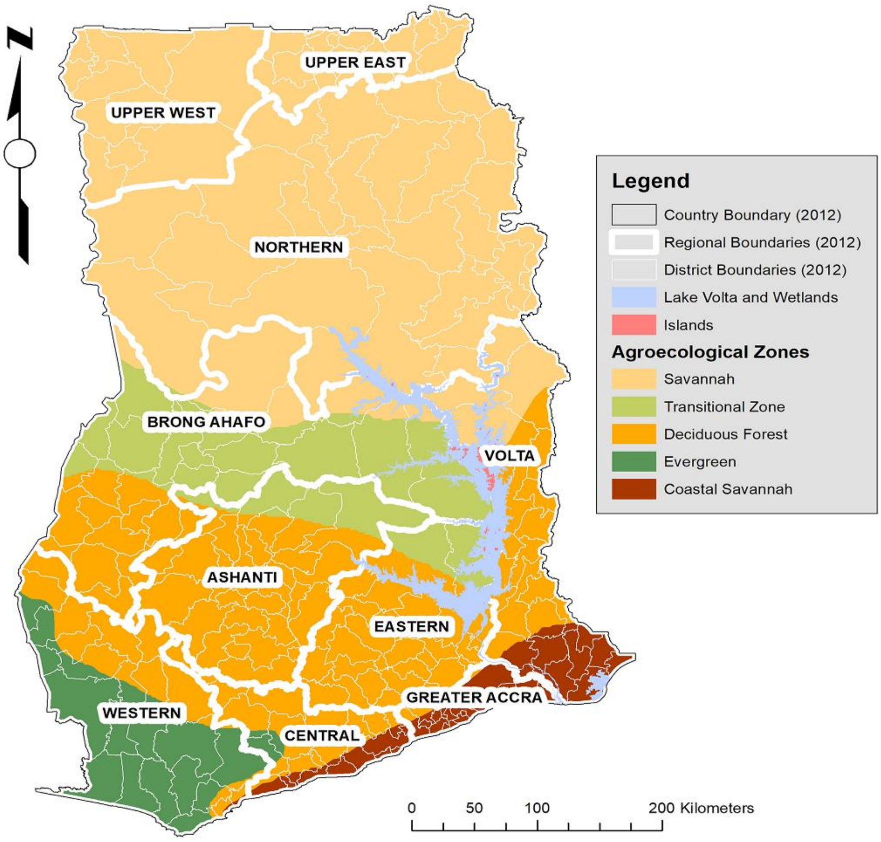


Figure 2: Agro-ecological zones of Ghana, source: (Harvest Choice, 2020).

Apart from northern Ghana, hydrometeorological data in the region are also available in Burkina Faso (3 gauges), Benin (2 gauges), Mali (2 gauges) and Niger (1 gauge). An overview of the available rainfall-measuring stations will be presented in Chapter 3.2.

2.2 Climatic conditions

Most of West Africa regions feature one wet season, usually between April and October, and one dry six-month period where almost no precipitation is observed. The precipitation mechanisms in the region are highly related with the occurrence of two types of air masses: one hot and humid and the other cool and dry. This front moves north and south, following the position of the sun, with a lag of 1 to 2 months (Physical Geography, 2020).

The dry season in the western African region lengthens and annual precipitation decreases at higher latitudes. Southern West Africa is characterized by a wide range of rainfall types and is mainly generated by the West African monsoon (WAM) system, which features an erratic yearly pattern (Diatta & Fink, 2014). Mesoscale Convective Systems (MCSs), large-scale horizontal convective formations, are the sovereign precipitation mechanism in West Africa. However, studies argue that West Africa is characterized by less-organized precipitation mechanisms with varying climatological characteristics (Fink et al., 2006). Nevertheless, not a lot of research has been devoted into this direction.

Several studies have been conducted regarding the rainfall patterns of Southwest Africa. Out of these, three main rainfall patterns emerge. Firstly, a precipitation mechanism dominating Southwest Africa are local convective rainstorms, mostly characterized by erratic patterns. The coastal region of South West Africa is more prone to convective rainstorms, where convective mechanisms generate up to 50% of total annual rainfall (Acheampong, 1982). On the contrary, in the savannah region convective rainfall accounts for approximately 25% of total annual rainfall (Omotosho, 1985). Secondly, monsoon rains dominate the region of Southwest Africa mostly between April and October. Thirdly, squall-line MCSs generate more than 50% of the rainfall in the savannah region (Omotosho, 1985).

A dry squall from east to west is the cause of most of the rainfall showers in the region (Maranan et al., 2018). Dry squalls result to an extremely localized rainfall pattern: while a certain area in Northern Ghana might experience extreme rainfall intensity, a neighbouring area might not experience rainfall at all.

Overall, the erratic rainfall pattern of the region favors the occurrence of dry spells. The climatology of the Southwest Africa region is summarized in Figure 3, in means of mean annual rainfall and average of wet-season months:

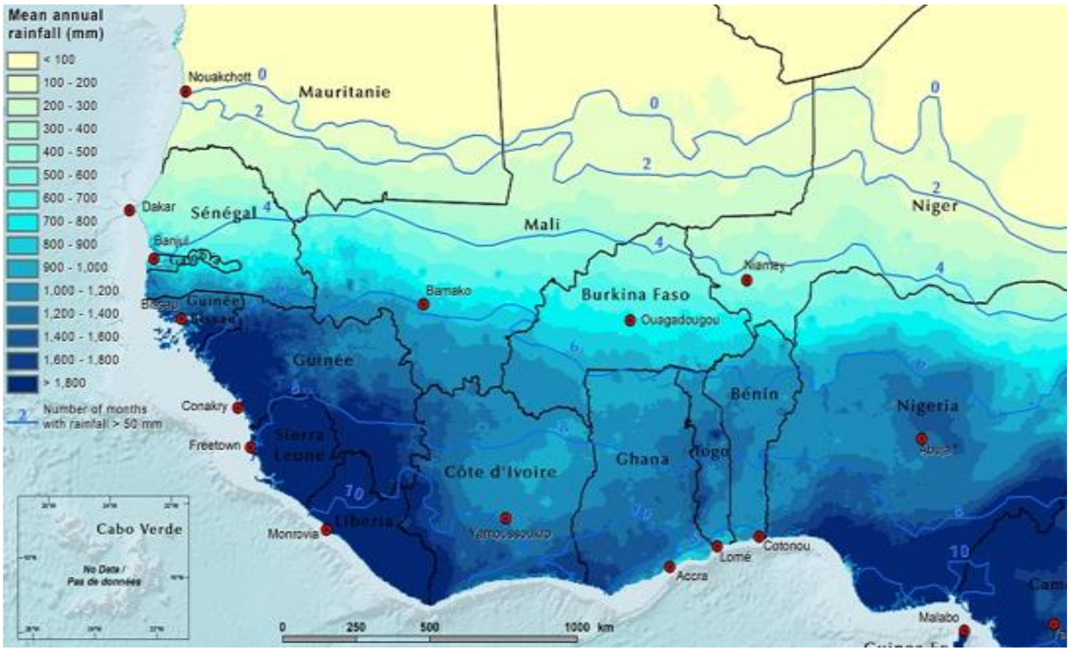


Figure 3: Mean annual rainfall 1981–2014, with number of months of 50 mm or more of rainfall. Map by (Tappan et al., 2016)

The northern part of Ghana is more challenging for agricultural activities than the southern one, as the relative humidity tends to decrease from southern to northern Ghana (Maranan et al., 2018). The absence of the tsetse fly is favorable for drought-resistant crops.

3. Data

3.1 Satellite products

There are six different precipitation satellite products available covering West Africa: CHIRPS, CMORPH, IMERG, MSWEP, RFE and TAMSAT dataset.

The Climate Hazards Infra-Red Precipitation with Stations (CHIRPS) is a satellite product based on climate model outputs, geostationary thermal infrared imagery and in situ precipitation observations (Funk et al., 2014) . The dataset has 0.05° spatial resolution at daily time scale and covers the period from 1981 onwards.

The Climate Prediction Center MORPHing (CMORPH) method dataset (Xie & Xiong, 2011) provides daily rainfall at 0.25° spatial resolution and covers the period from 1998 to the near present.

The Integrated Multi-satellite Retrievals (IMERG) for Global Precipitation Measurement mission (GPM) (Tan & Huffman, 2019) dataset provides a new interesting version of the Tropical Rainfall Measuring Mission (TRMM) Multi-Satellite Precipitation Analysis of the GPM mission. The dataset which is a satellite retrieval product that combines multiple sources of inputs and calibrated with gauge observations, contains daily and 0.1° spatial resolution estimates. The dataset is available from 2000 to the near present.

The Multi-Source Weighted-Ensemble Precipitation (MSWEP) (Beck et al., 2017; Beck et al., 2019) is a High spatial (0.1°) and temporal (3 hourly) resolution product that covers the period from 1979 to the near present. MSWEP optimally merges a wide range of gauge, satellite, and reanalysis data to provide reliable precipitation estimates over the entire globe. Thus, we should probably expect that gauge-scarce regions like West Africa are probably under-represented in this product.

The African Rainfall Estimation (RFE) (Herman et al., 1997) is a product from the NOAA Climate Prediction Center mainly produced for Famine Early Warning Systems Network to assist in disaster-monitoring activities over Africa. The product combines different satellite products with rain gauges to estimate daily rainfall at 0.1° spatial resolution. The database is available from 2001 to the near present.

The Tropical Applications of Meteorology Using Satellite Data (TAMSAT) and Ground-Based Observations that merge MeteoSat IR data and gauge rainfall measurements (Maidment et al., 2017). The database provides daily rainfall estimates at 0.0375° spatial resolution at daily time scale and covers the period from 1983 to 2016.

An overview of the available satellite products is presented in Table 1:

Table 1: Overview of the available satellite products together with their spatio-temporal coverage and spatial and temporal scale.

SATELLITE PRODUCT	SPATIAL RESOLUTION	TEMPORAL RESOLUTION	SPATIAL COVERAGE	FROM
CHIRPS V2.0	0.05° (~5km)	Daily	Africa	1981
CMORPH_BLD V1.0	0.25° (~25km)	Daily	Global	1998
IMERG V6.0	0.1° (~10km)	Daily	Global	2000
MSWEP V2.0	0.1° (~10km)	Daily	Global	1979
RFE V2.0	0.1° (~10km)	Daily	Africa	2001
TAMSAT V3.0	0.0375° (~4km)	Daily	Africa	1983
PGF V3.0	0.25° (~25km)	Daily	Global	1948

3.2 Point measurements

The ground measurement dataset available consists of 14 stations spread across Ghana (northern part), Burkina Faso, Benin, Mali and Niger. The different datasets feature different temporal coverage, but all of them in a daily scale.

In Figure 4 the locations together with the elevation of the 14 stations is depicted. The map was generated in (GIS, 2009) using a Digital Elevation Map layer of Africa generated by (Verdin, 2017). From this map it can be derived that the elevation of all the precipitation measuring stations is between 200 and 500 m. This is important because convective mechanisms are likely to play a role in precipitation generation.

The discrepancies between the different stations' time-series are explained by the fact that the data originate from three different databases:

- The AMMA database which features 4 out of 14 stations (Tillaberi [Niger], Tobre [Benin], Tara [Mali] and Agoufou [Mali]) and its data availability on average extends from 2000 to 2017, with discrepancies in individual datasets.

- The MARLOES database, which is the oldest database available in the area with measurements starting from 1940, overlapping even the early years of the remote sensing products. This database features six stations: Lawra, Navrongo, Ouahigouya, Tamale, Wa and Zuarungu. All of them are located in northern Ghana, except Ouahigouya which lies in Burkina Faso.

- The Wascal database with four stations (Aniabisi [Ghana], Poudri [Benin], Lare [Burkina Faso] and Yabogane [Burkina Faso]) and extremely limited data availability, due to short time-series.

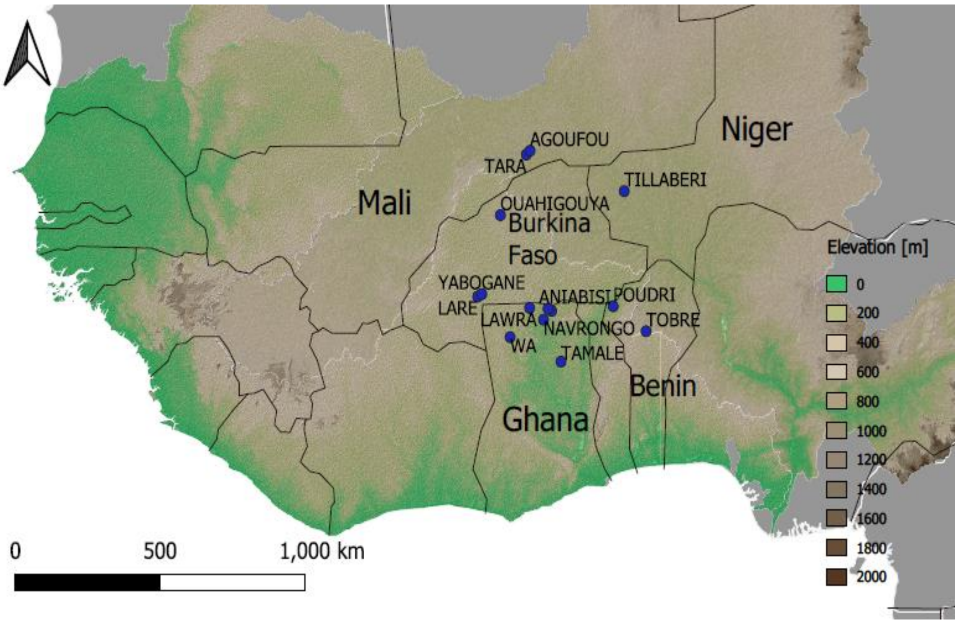


Figure 4: Location and elevation of the 14 stations.

3.3 Satellite-to-ground data match

The first step needed towards assessing the match of the satellite products and the ground observations is selecting the optimal satellite precipitation datasets for our research. The main criteria for selecting the satellite products used in the study are: (a) adequate match of the satellite data and the gauge measurements, usually assessed by the confusion matrix or the performance of the classification model for each individual product; the correlation coefficient of the point observations versus the nearest pixel value of the satellite precipitation product is also presented but it is not the optimal metric for this application, and (b) adequate overlapping time period between the point observations and the satellite data, ensuring enough data to train the models introduced in Chapter 5.

At first, the quality of each satellite product was assessed by determining performance metrics based on:

- the hits or True Positives (TPs), when both the satellite and the point measurement report rainfall for a certain day;
- the false alarms or False Positives (FPs), when the “predicting” satellite product is reporting a rainy day but no rainfall is actually measured on the ground station;
- the misses or False Negatives (FNs), when the “predicting” satellite product reports no rainfall but the gauge does;
- and the True Negatives (TNs) when both sources report no rainfall.

The arrays in which the performance metrics are computed have to be in a binary form. Therefore, a yet crucial but subjective factor is determining the threshold of rainfall which distinguishes wet and dry days. Several studies propose their own threshold values based on each application (more on that on Chapter 5).

Based on the aforementioned elements, the four performance metrics that were used to assess each product's performance are: (a) the Probability Of Detection (POD) (of rainfall), a fraction of hits over the sum of hits and misses, with zero being the worst-performing and one the optimal value, (b) the Success Ratio (SR), a fraction of the sum of hits and misses over the hits and false alarms, (c) the Threat Score (TS) as the ratio of hits over hits, misses and false alarms, with zero being the worst-performing and one the best-performing value, and (d) the Frequency Bias (FBias), a fraction of the sum of hits the false alarms over the sum of hits and misses, with one being its optimal value.

The overall research of defining the best fitting satellite product for West Africa was conducted by (Agoungbome, unpublished) and the reader can refer to this study for the integrated reasoning assessing the satellite precipitation products. As (Agoungbome, unpublished) suggests: *"all the products perform poorly in detecting dry spells as the dry spell length increases, however MSWEP, IMERG, RFE and CMORPH perform better than the others with about 50% Success Ratio of dry spells detected"*. Generally, it is expected that global satellite products will under-perform in gauge-scarce regions like West Africa. Indicatively, in Figure 5 an overview of the performance of the satellite products for a dry spell length of 5-days is depicted in the form of a Roebber performance diagram. A Roebber performance diagram exploits the geometric relationship between four measures of dichotomous forecast performance: probability of detection (POD), false alarm ratio or its opposite, the success ratio (SR), bias and threat score.

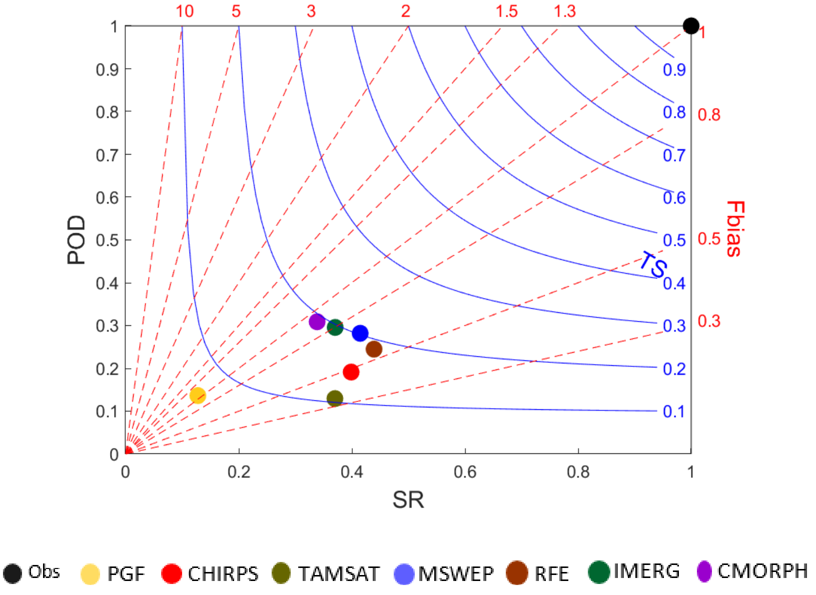


Figure 5: Roebber performance diagram for a dry spell length of 5-days, illustrating SR on the x-axis, POD on y-axis and FBias in dashed-line where POD, SR, and TS measure ratio of accurate detection. The black dot represents the performance of an "ideal" dataset. Source: (Agoungbome, unpublished)

According to (Agoungbome, unpublished), when comparing extensively the MSWEP dataset to the CMORPH dataset: *"MSWEP shows a good (0.73) fit in the semi-arid zone in terms of annual rainfall with 71 % of agreement"* and *"MSWEP shows a better fit in the semi-arid zone (8°N to 12°N)"*. Ultimately, it is derived from the study that the MSWEP dataset is the optimal satellite product to use, not only because of its relatively good fitting-performance, but also because of its high temporal coverage and spatial resolution. However, in Chapter 7 the selection of the optimal satellite datasets to use is going to be made based on their classification performance.

The next step is to compare the product in a day-to-day analysis with the gauge rainfall measurements both in terms of absolute rainfall values and binary values. This was achieved by matching each gauge the pixel of the satellite product containing it.

Supporting the climatological analysis made in Chapter 3.2, the boxplots of monthly accumulated precipitation presented in Figure 6 and Figure 7 for an example station, do prove the notion of a highly-separated wet and dry period in the region with July and August being the most wet months during the rainy season. Besides, it can be noted that the point observations present higher variance than the satellite observations.

The analysis on monthly precipitation values of both satellite and gauge data underlines the discrepancies in terms of variance and absolute rainfall between the two data sources, but comparing the dry spell statistics of both sources is more informative for our application.

When it comes to the dry spell statistics of the point measurements, latitude oriented discrepancies can be distinguished in the dry spell properties of each station. Dry spell occurrence is defined as five consecutive days or more with precipitation below 1 mm/day (more on that at Chapter 5.1). The most northern station available, station Agoufou, reports an average annual frequency of dry spells of between 15 and 16 in every wet season, while average maximum annual dry spell length is 35.50 days. The lower-latitude station Tamale reports on average less than 10 5-day dry spells per rainy season with a mean maximum dry spell length of approximately 14 days. Mid-latitude station's Ouahigouya dry spell characteristics are balanced: 12 5-day dry spells on average per wet season and 32 days mean maximum annual dry spell length. Overall, data availability for each station varies significantly from 2 years (Aniabisi, Poudri, Lare and Yabogane) to 75 years (Lawra, Navrongo, Ouahigouya, Tamale, Wa and Zuarungu).

The discrepancies between the satellite and gauge precipitation data in terms of dry spell statistics do not have an absolute pattern and vary between stations, e.g. while for station Agoufou the satellite data report an increased (relatively to the point measurement statistics) average maximum annual dry spell length of 49.50 days, for station Tamale the average dry spell per rainy season drops to 9 5-day dry spells and the mean maximum dry spell length to 8.50 days. The same holds for Ouahigouya station: dry spell occurrence drops to 10.50 5-day dry spells on average per wet season and the mean maximum annual dry spell length drops to 28 days.

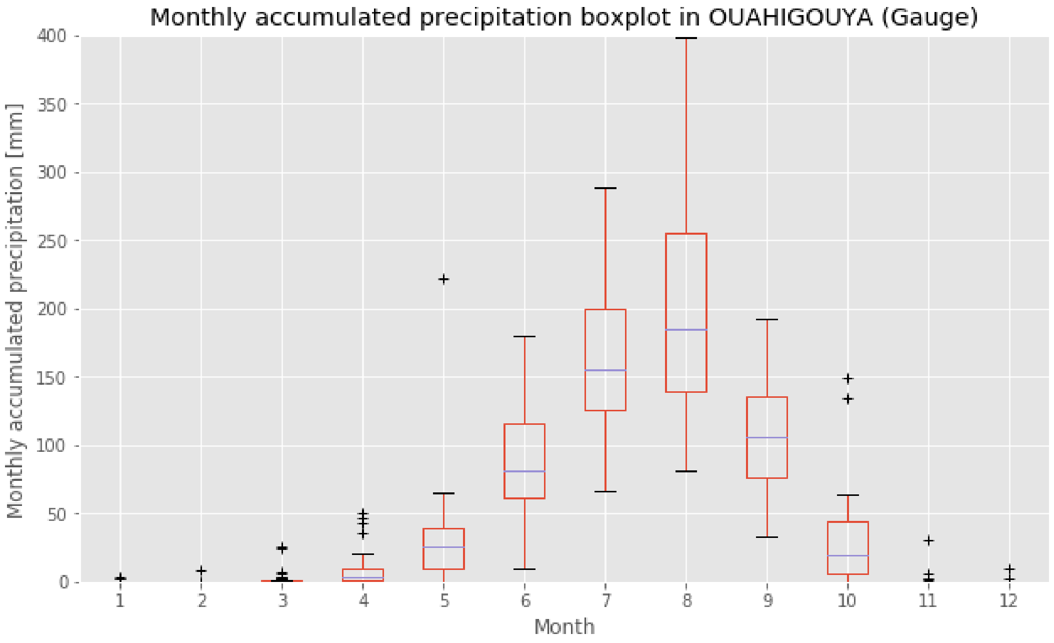


Figure 6: Monthly accumulated precipitation boxplot for Ouahigouya station, ground observations.

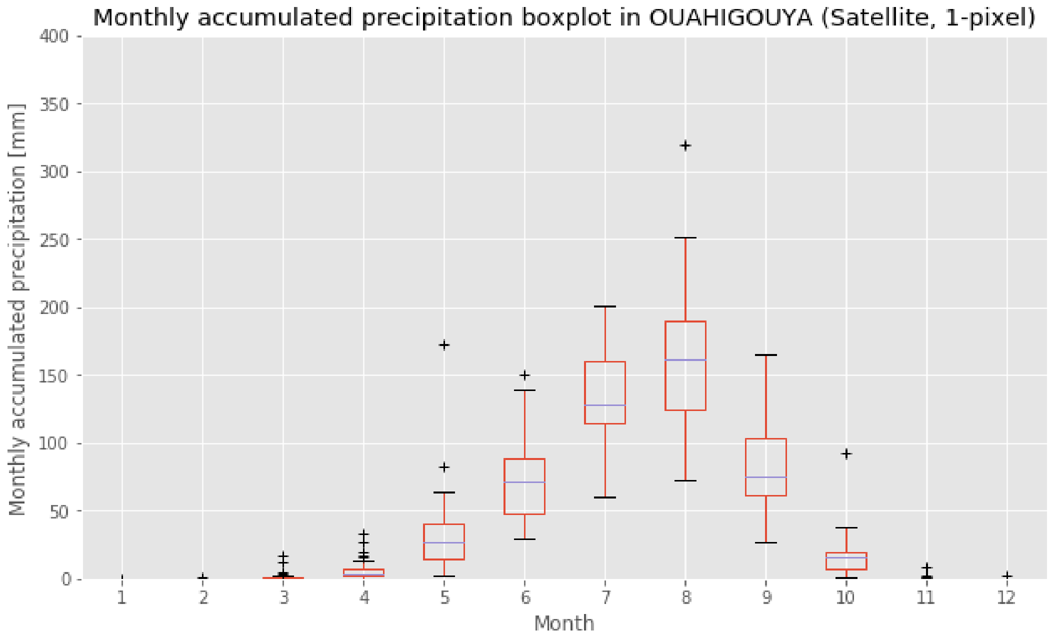


Figure 7: Monthly accumulated precipitation boxplot for Ouahigouya station, remote-sensing observations.

As already explained in Chapter 1, the research interest focuses on reproducing the dry spell sequences of only the rainy season, since this climatic information will boost the growing season decision making of the smallholder farmers in the region. In addition, as observed in the Figures above, the months of the dry season barely feature any rainy days. Thus, incorporating these months into the input datasets would increase the class imbalance of the dataset severely, degrading the wet day prediction ability of the classifier. Hence, the research uses only wet season data, the temporal coverage of which varies for each of the stations.

4. Exploratory Data Analysis

4.1 Gauge to remote-sensing precipitation values assessment

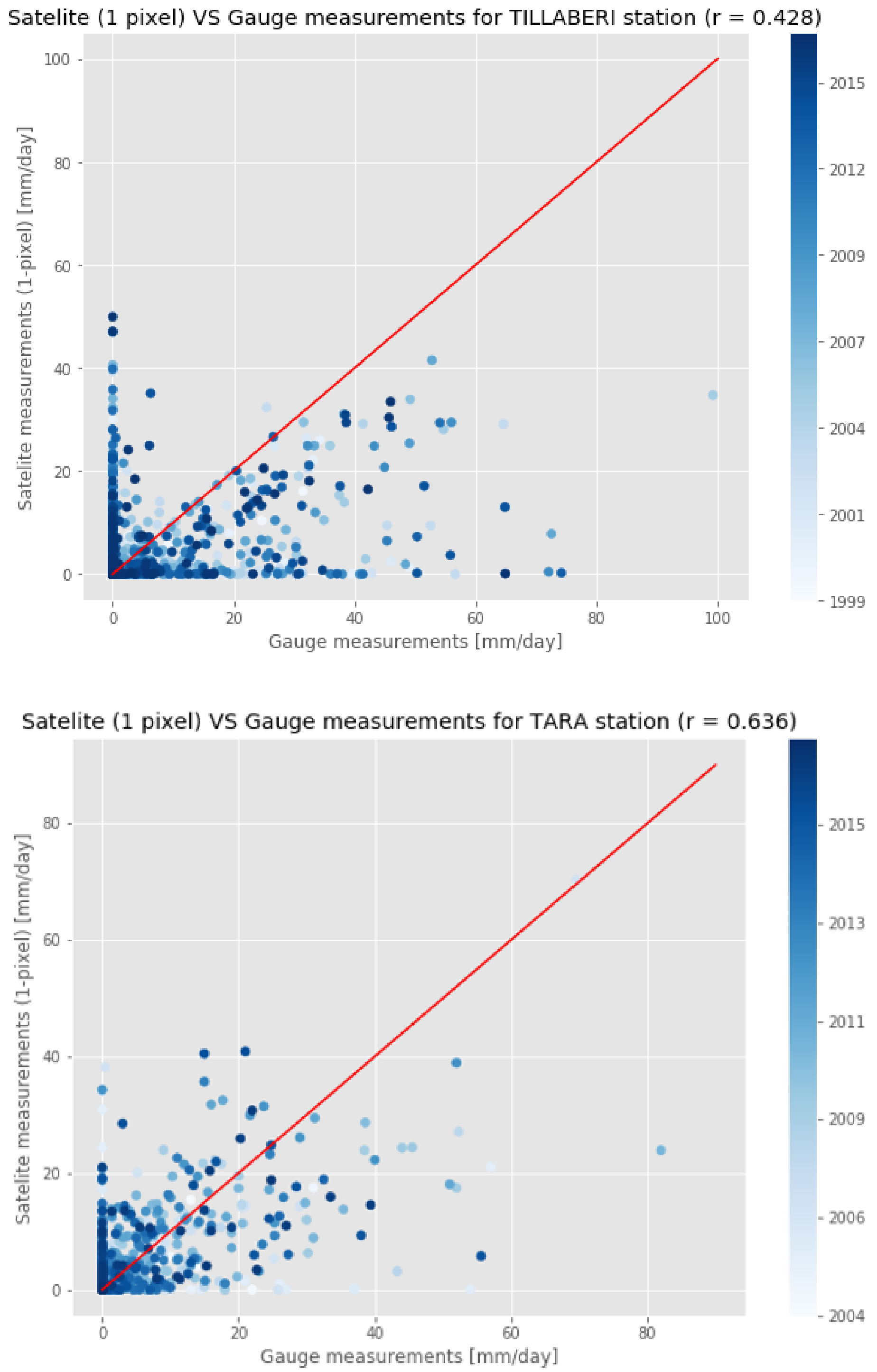
Next, the precipitation values of the satellite products were compared to the ones of the gauge measurements. The comparison was conducted in the overlapping years of the two time-series. Hence, for the AMMA gauges this is done on the years from 2000 until 2016 (the year 2017 which is features in the AMMA database is not included in the MSWEP data), for the MARLOES stations the overlapping starts at the first year of available remote-sensing data (1980) and goes until approximately the first years of 2000. Finally, the WASCAL database is completely overlapped by the remote-sensing data.

The point-to-pixel comparison is primarily assessed through Pearson's product moment correlation coefficient (r). The assessment process reveals that the satellite products perform adequately over some stations (correlation > 0.6), but several stations, in particular 5 out 6 stations from the MARLOES database, perform rather poorly (correlation between 0.3 and 0.4). In Figure 8, the daily values match of the 1-pixel remote-sensing data and the gauge observations for the poor-performing Tillaberi station and the same results are for the overall best-performing station, the Tara station, are depicted.

Overall, the miss-match between the satellite-products and the ground observations is partly dependent on different time-shifts of the data. In particular, we found several cases in which one of the source reports heavy rain over the study area on day x while the second did not report any rain at all (further analysis in Chapter 6.6). However, upon closer inspection, we could see that in the second source, the rain was attributed to day x-1 or x+1. This time-shift leads to time steps with completely miss-matching rainfall values (zero value versus heavy rainfall value), resulting to double penalties: wrong timing leading to two errors, one for each of the time steps.

Taking into account the fact that the satellite data captures the rainfall showers even with a timing discrepancy, a data analysis in coarser time scales could probably provide further insights. It is expected that coarser time scales provide a better match between the data sources, while limiting the type of potential predictions about the dynamics. The aggregation time scale plays a big role in influencing the overall agreement between data sources. For example, the upper Figure 9 image shows that the correlation coefficient for Tillaberi increases from 0.42 to 0.72 when aggregating to 3 days periods, and for Tara, the correlation increases from 0.64 to 0.79 (lower Figure 9 image). The modeling results with coarser time-scale are presented in Chapter 6. Nevertheless, given that the generated dry spell indices of interest have to be in high temporal resolution in order to be relevant to the needs of the smallholder farmers, the modeling process is conducted in the daily scale.

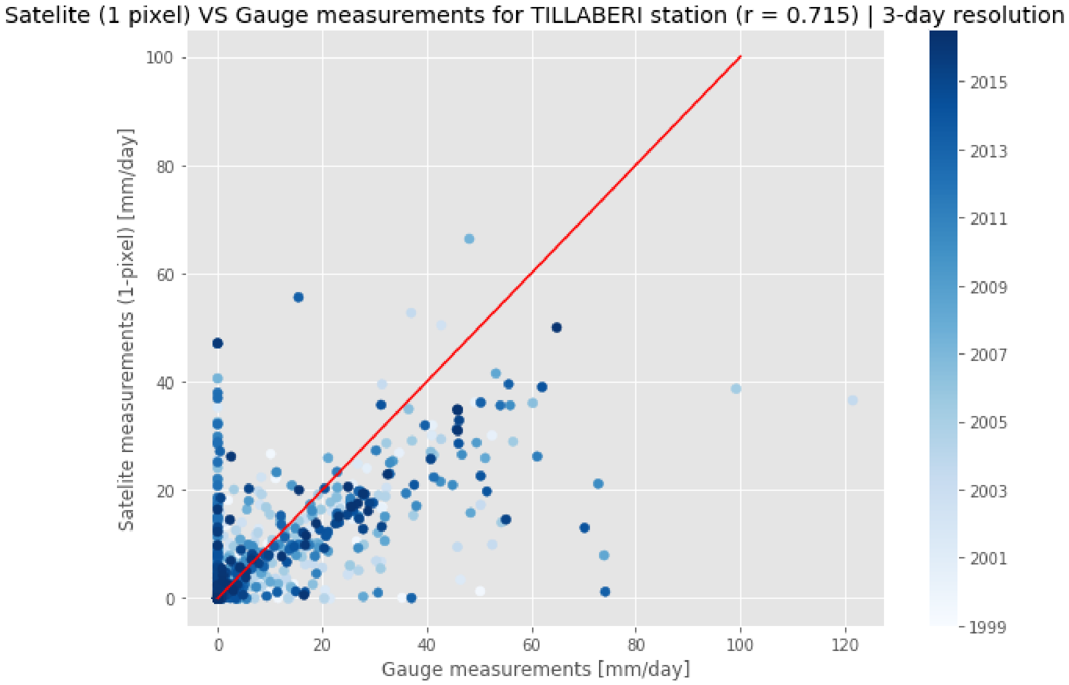
From the Figures below it can be stated that translating the raw data to coarser temporal resolutions mostly enhances the data match in the low-performing stations. However, since the output of interest is to detect the presence of precipitation, discarding the days in which the two sources mismatch in terms of rainfall detection does not add any value to the research.



Satelite (1 pixel) VS Gauge measurements for TILLABERI station (r = 0.428)

Satelite (1 pixel) VS Gauge measurements for TARA station (r = 0.636)

Figure 8: Satellite observation (MSWEP satellite precipitation product) VS ground observations for the overlapping time-series of Tillaberi (upper) and Tara (lower) station. Pearson's product moment correlation included in the title.



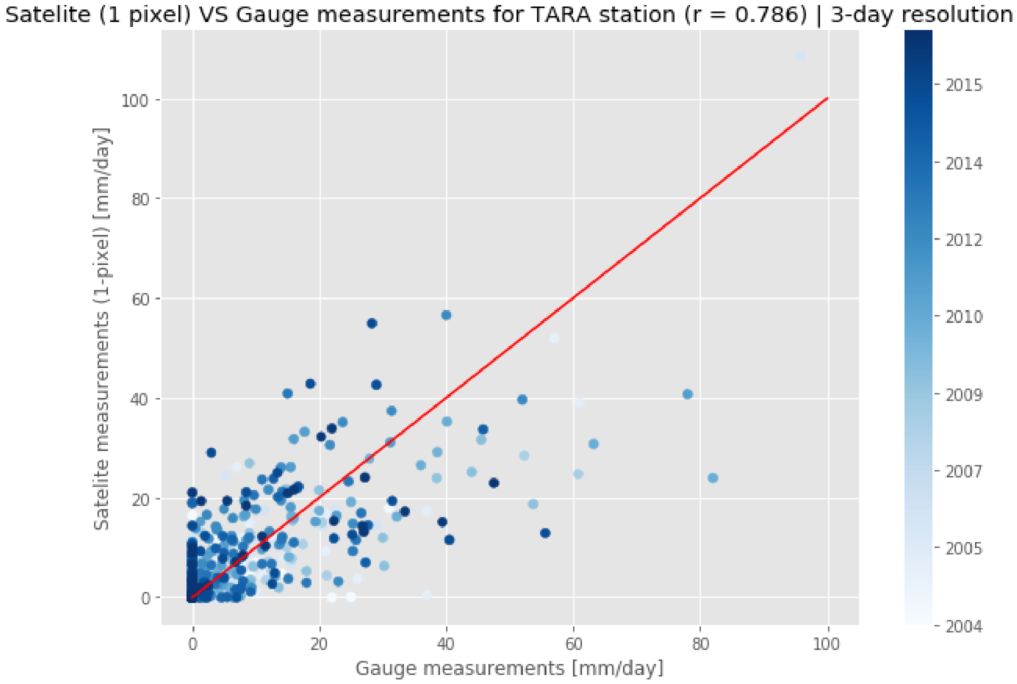
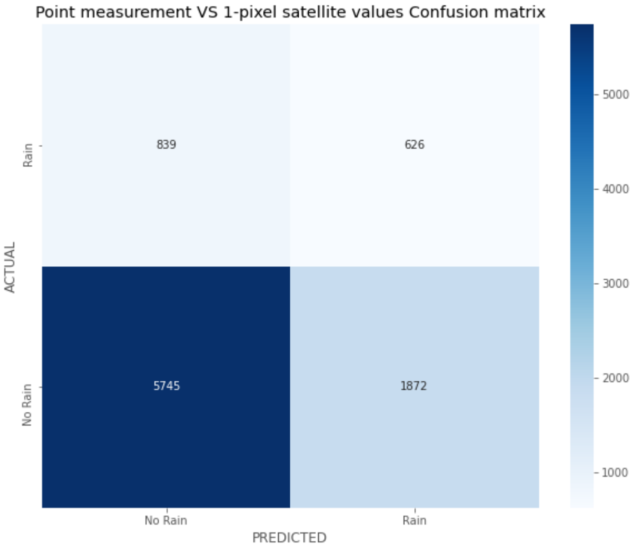


Figure 9: Satellite observation (MSWEP satellite precipitation product) VS ground observations for the overlapping time-series of Tillaberi (upper) and Tara (lower) station on 3-day temporal resolution (raw data). Pearson's product moment correlation included in the title.

Assessing the match between the satellite products and the point measurements using Pearson's correlation provides insights about the match of the raw precipitation values but might not be the optimal way when dealing with a classification problem. Deriving the confusion matrices using a 1mm/day wet/dry day threshold and computing the accuracy and the f1-score is a more informative way of quantifying data quality. Accuracy is computed as a ratio of the correct classification over the total population, while the f1-score is a more sophisticated metric, derived from precision and recall, with more value when facing classification problems where the wet/dry day classes are not equally represented in the data. Precision quantifies the number of positive class predictions that actually belong to the positive class, while recall quantifies the number of positive class predictions made out of all positive examples in the dataset. The f1-score metric provides a single score that balances both the concerns of precision and recall in one number.

In Figure 10 the confusion matrices of the aforementioned stations (Tillaberi and Tara) using all six satellite-based datasets stacked up in one column as predictors, based on the overlapping between each sensor and gauge data, are depicted. The accuracy for Tara station is 0.84 while the one of Tillaberi station is 0.70. However, a direct comparison of the metrics for the two stations would not be fair since the observations in Tara station are more imbalanced than for Tillaberi station. The weighted f1-score for Tara is 0.86 (0.90 for dry days and 0.55 for wet days) while for Tillaberi 0.73 (0.81 for dry days and 0.32 for wet days).



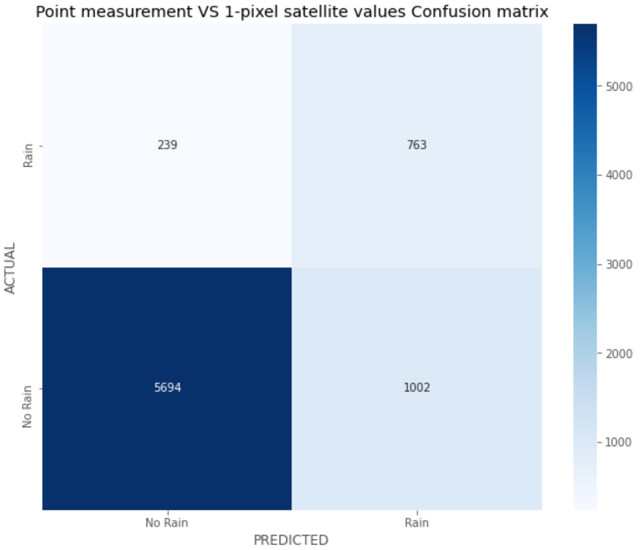


Figure 10: Tillaberi (upper) and Tara (lower) station confusion matrix (1-pixel and point measurements).

In order to investigate the consistency of the data, we computed the double-mass curves (Searcy et al., 1960) between the 14 different gauge time-series and 6 different satellite-based datasets, as shown below. This dataframe is generated by stacking one remote sensing product on top of the other and duplicating days based on the overlap between the gauge and satellite data. Since the graph plots approximately a straight line, the process is homogeneous (Figure 11).

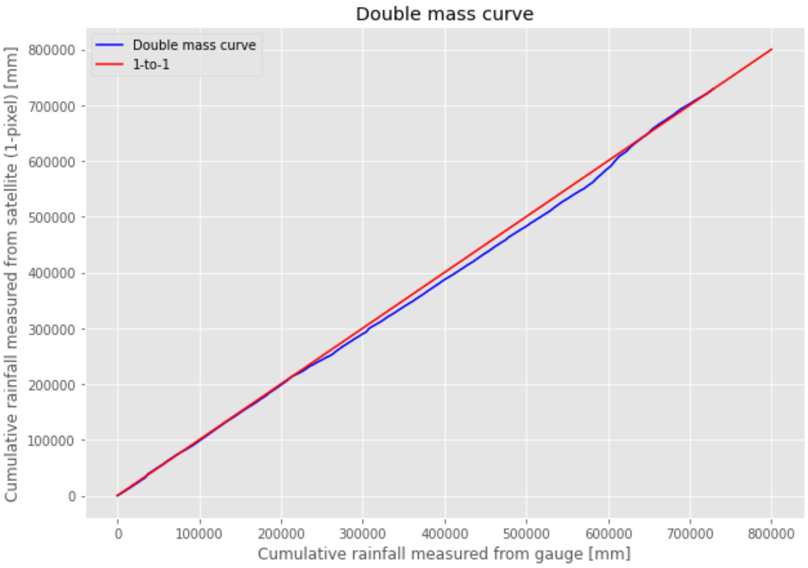


Figure 11: Double mass curve, gauge vs satellite datasets

The total number of days available in this dataframe is 287.333. Distinguishing the wet season for each station is the next step. However, the wet season for each station varies with latitude, as already explained in Figure 3. Approximately, we could define the rainy season from April to September. Discarding all the data of the dry season, the final dataframe contains 145.230 days of 14 different stations matched with 6 different satellite products.

A preliminary analysis of this dataset can be done by transforming the point measurements and 1-pixel (the one containing the gauge) values to binary form by applying the 1 mm/day dry/wet day threshold. The dataframe contains 70,50% dry days. The f1-score of the dry days is 0.74 and the one of the wet days 0.58. This performance should make us suspicious about the mismatch between the satellite-based and gauge data.

4.2 Latitude-wise analysis

The climatological information known from previous studies in West Africa is verified: the southern pixels present higher and more varying precipitation values. As shown in the table below, the mean precipitation value of a typical southern pixel in the region is 5-10% higher of the one of a typical northern pixel.

The distribution of satellite precipitation values and the point measurements are both highly imbalanced. The general statistics of the point measurements of all stations and a typical northern and a typical southern pixel are presented in Table 2. The average value of the point measurements is drastically higher than the one of the pixels, even though the satellite data has higher maxima.

In Figure 12 the distributions of a high, mid and low-latitude pixel together with the point measurement distribution are shown. The satellite precipitation values are more imbalanced than the point observations. This characteristic can be attributed to the low Probability of Detection (POD) of precipitation products in gauge-scarce regions, especially for shallow precipitation events (Tan et al., 2016).

Table 2: Descriptive statistics of initial data expressed in mm.

	Gauge	Northern pixel	Southern pixel
mean	15.18	6.69	7.25
std	15.44	9.67	10.38
skewness	2.09	3.44	3.62
kurtosis	6.63	28.47	27.34
min	1.00	0.00	0.00
25%	4.00	0.00	0.00
50%	10.10	3.00	3.85
75%	21.00	10.30	10.70
max	148.20	216.10	194.80

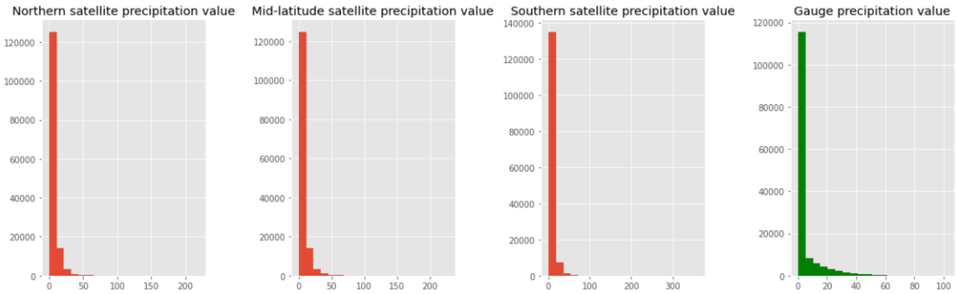


Figure 12: Gauge and high, mid and low-latitude pixel values distributions expressed in mm.

4.3 Assessing satellite precipitation images

In the following, we investigate the spatial patterns of the gridded precipitations data and their match with the point measurements by assessing typical examples of satellite images and groundtruth values. This is important to understand which satellite images are expected to be classified correctly in order to assess the quality of the input data.

Typical examples on the input images of the training data are shown. In the title of each figure the "Label" represents the binary value of the groundtruth, 1 for rainy day and 0 for dry day, and "Rainfall" reports the raw value of precipitation. A maximum value of 40 mm has been set to the colorbar of the satellite images to facilitate the visual comparison of the images. In IMERG and TAMSAT pixel values were aggregated so that the images match the spatial resolution and extent of the CMORPH images, which has a lower resolution. The images have a spatial extent of approximately 800 km x 800 km around each gauge. These three satellite products were selected for the assessment given the individual performance of the classification models in Chapter 6.

Since our goal is to combine the information of multiple different satellite products, we will start by assessing the similarity of the images of those products. The correlation analyses already provided some initial insight into the data. But correlation coefficients are sensitive to spatial mismatches and can not capture the similarity of two images in the presence of shifts or other geometric transformations. Hence, more sophisticated metrics were introduced, like the Co-occurrence matrix (coma) and the Perceptual hash.

The coma representation is calculated by moving through each cell, looking at its value, and counting how many neighbors of each class our central cell has. To implement this in continuous variables like rainfall amount, we first need to transform it to a categorical variable by introducing precipitation classes (e.g. 50 classes from 0 to the maximum value recorded). Then, the co-occurrence matrix will be a 50 by 50 matrix with each cell representing how many times one class is neighbouring with another one.

The co-occurrence matrix representation is two-dimensional, with values of categories in row and columns (Haralick, 1979). It can be converted into a one-dimensional representation called a co-occurrence vector (cove). Using cove of the different images we can assess the similarity of the three images in a certain day for a certain station in pairs of two. This is done by calculating the Jenson-Shannon distance (Fuglede & Topsøe, 2004) of the two normalized vectors. The Jenson-Shannon distance is a value between 0 and 1, where 0 means that two probability functions are identical, and 1 means that they have nothing in common.

The perceptual hash is calculated by averaging out the pixels to 8x8 pixel resolution. Then, the hash of each image is calculated based on whether a pixel has higher or lower value than the average value all 64 pixels. Then, the images are compared by calculating the Hamming distance (Oxford Reference, 2017), with zero indicating identical images and one completely different images. This means counting the number of different individual pixels.

Figure 13 is a typical example of a rainy day that should be classified correctly by the classification model. All three images report heavy rainfall as well as the groundtruth. Through visual inspection, it can be stated that the patterns observed in CMORPH and IMERG look more similar than the one of TAMSAT. Indeed the similarity indices verify this notion: CMORPH and TAMSAT have a Jensen-Shannon distance of 0.49, while CMORPH and IMERG have a Jensen-Shannon distance of 0.29, and TAMSAT and IMERG one of 0.62.

Figure 14 underlines the importance of timing between the gauge and the satellite products. Both CMORPH and IMERG report heavy precipitation around the gauge (they present similar patterns in most of the days) and would probably classify the day as rainy. However, the groundtruth does not report rainfall. CMORPH and TAMSAT have a Jensen-Shannon distance of 0.41 and a Hamming distance of 0.42, CMORPH and IMERG 0.15 and 0.19, and TAMSAT and IMERG 0.40 and 0.39, respectively.

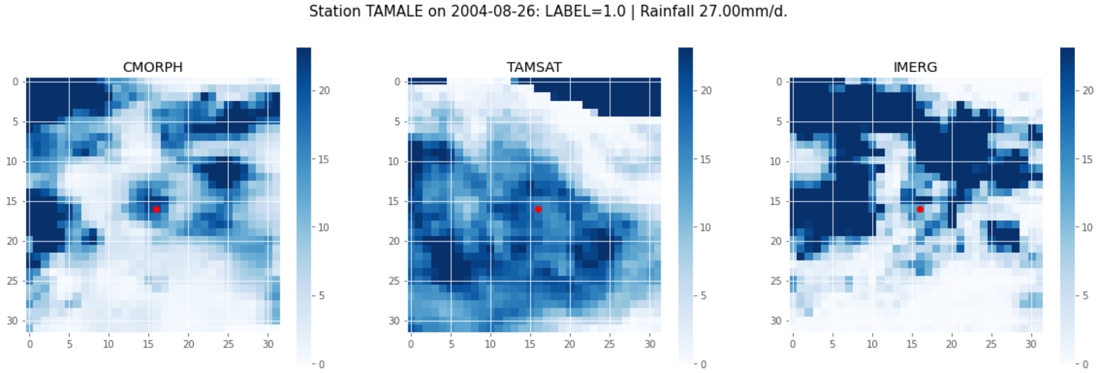


Figure 13: Typical example of rainy day that should be correctly classified by the classifier model

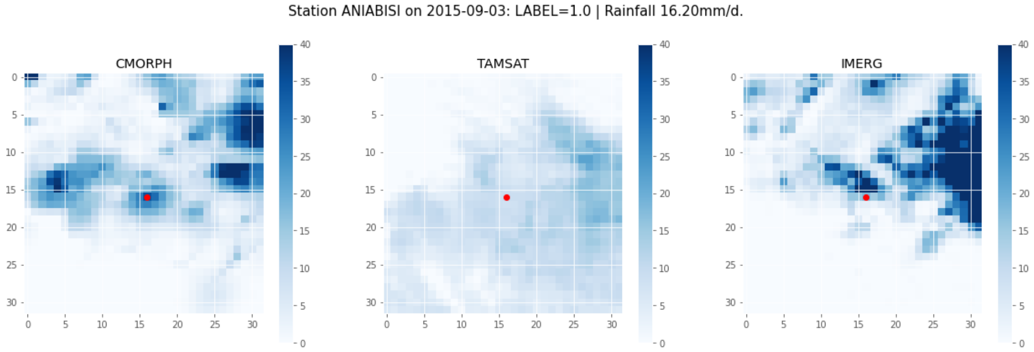


Figure 14: Dry day reporting rainfall in 2 out of 3 satellite sources

Figure 15 is a typical example of a dry day that should be classified correctly. None of the sources report precipitation. CMORPH and TAMSAT have a Jensen-Shannon distance of 0.00, while CMORPH and IMERG have a Jensen-Shannon distance of 0.06, and TAMSAT and IMERG one of 0.06. All three Hamming distances calculated are approximately 0.05.

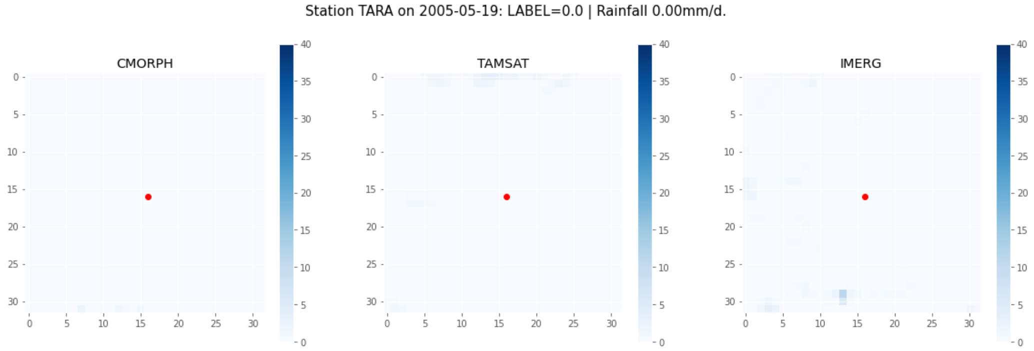


Figure 15: Typical example of dry day that should be correctly classified by the classifier model

Figure 16 is a day that is really likely to be misclassified. The gauge reports no rainfall while all three satellite sources report heavy precipitation close to the station. The Jensen-Shannon distance between CMORPH and IMERG lies around 0.25.

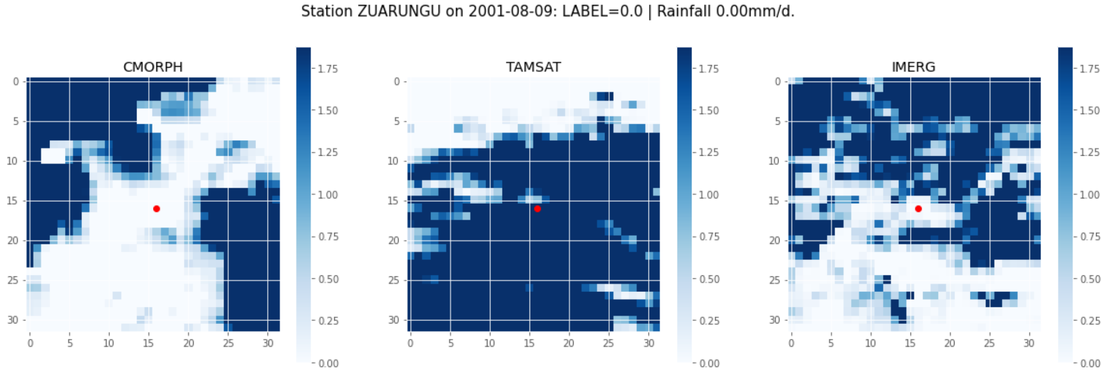


Figure 16: Typical example of a day that will probably be misclassified. All the satellite precipitation products report heavy rainfall whereas the groundtruth reports zero precipitation.

5. Methodology

5.1 Methodological approach

The idea of the present study is to reproduce dry spell sequences by building a model that links the satellite-data to the ground observations. This would be approximately 15 years for the AMMA database stations, almost 20 years for the MARLOES database, most of them during the earlier years that the satellite time-series, and 2-3 years on average for the WASCAL database stations. By training this model, we describe a process in which the remote-sensing data will be used as input, not as a sequence of data, and the process will aim at reproducing a binary Rain/No Rain output. The underlying goal is that this model will gain the ability to detect relevant spatial patterns in satellite data (i.e., features) and learn how to combine these features together into a model for predicting rain-gauge time series in the same region. The data feature a daily time-scale because this temporal scale is relevant to the needs of the smallholder farmers in the region, but models were also tested in different temporal scales.

The aforementioned model can be efficiently represented by an Artificial Neural Network (ANN) which will work as a "black box". Several different ANN architectures are being tested, such as Feed-Forward Neural Networks (FNN) and different Convolutional Neural Networks (CNN).

However, other ML classification models will be used as benchmark models to test the ANN performance to. These will include Logistic Regression, Random Forest, Gaussian Naive Bayes and Support Vector Machine (SVM) and will be explained in detail in Chapter 5.3.

Utilizing the output of the aforementioned ANNs, the final objective is to translate the reproduced dry spell sequences with high temporal extent to dry spell indices. Several drought indices have been used by hydrologists. The selection of the optimal one is always application-specific and stakeholder-centered (Savenije, 1999). In this case the metrics used to assess the dry spell sequences will be the frequency of dry spell occurrence, the maximum dry spell during the growing season and some cumulative rainfall information, as in (Gbangou et al., 2020).

Dry spell is a vague and broad term and should be explicitly defined for each case. In this application, the length of consecutive days that define a dry spell are highly related to crop growth and, thus, to crop types, the drought resistance of these crop types and the water-holding capacity of the soil. All these factors define the critical dry spell, which Gbangou et al. (2020) defines occurrence as five days without rainfall for our region of interest. Furthermore, dry days have to be distinguished from wet days. Setting the threshold between the two is a highly subjective task and usually is application-specific: Froidurot & Diedhiou (2017) sets a threshold of 1 mm/day when analyzing the monsoon system of West Africa, Fischer et al. (2013) uses a threshold value of 0.1 mm/day for the purpose of assessing dry spells in Tanzania, while Enfors & Gordon (2007) set a threshold of 0.8 mm/day for their agricultural-related study. Overall, setting a threshold value is quite subjective. However, most agricultural applications suggest a threshold value of about 1 mm/day. Thus, this value is going to be used in our research.

Overall, there are several factors that dictate the definition of dry spell, which is always application-specific. In this application dry spell occurrence can be defined as five consecutive days with precipitation below 1mm/day.

5.2 On the deep learning approach

Defining the optimal ANN architecture for each application depends on several factors. In our case, we are looking for an ANN architecture that would actually use as input the satellite-based images of precipitation at a fixed extent around any gauge (e.g. 32x32 pixels), while the labels will be the ground observations, constituting a typical supervised-learning modeling case. This process can be efficiently handled by a CNN.

A flow-chart of this process is depicted in Figure 17. The input satellite precipitation images are passed through the feature extraction part and, then, the classifier of the CNN and the output is mapped in [0, 1]. For a more detailed analysis of the technical characteristics of the proposed CNN (convolutional layers, pooling layers, activation functions, padding etc.) the reader can refer to Appendix I.

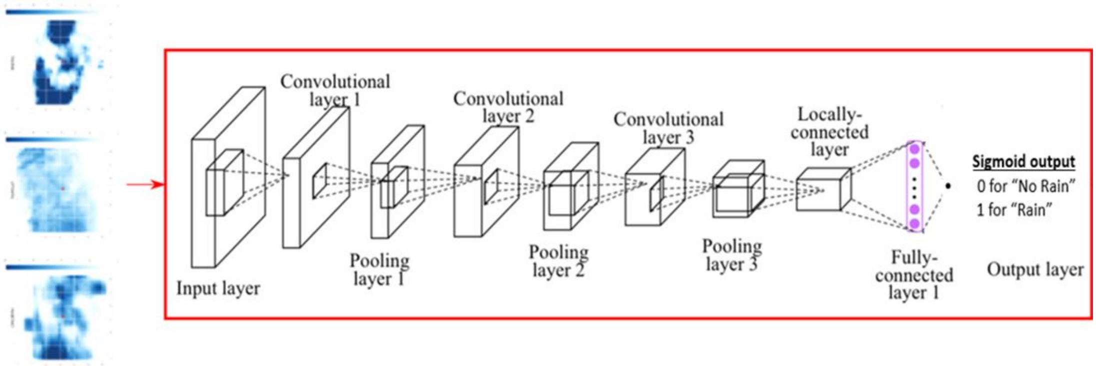


Figure 17: Overview of the CNN modeling process. The input layer features three channels, one for every image, with constant spatial extent. The features extraction section and the classifier section can be clearly distinguished. The output layer features a sigmoid activation function.

The years where the satellite-based data and the point measurements overlap will be used to train and test the model. The number of input nodes depends on the amount of spatial and temporal information we choose to use. The output node will feature a cost function related to the discrepancy between the modeled output of a sigmoid function and observed wet/dry day label. Based on that, the Stochastic Gradient Descent (SGD) will optimize the values of all the trainable parameters of the ANN.

The CNN architecture is chosen to perform this task because of its efficiency in dealing with gridded data and detecting spatial patterns. There are many features that give an advantage to CNNs when handling spatial data (e.g. in image multi-class classification tasks) compared to conventional ANN architectures. One of them is parameter (or weight) sharing. Parameter sharing introduces local connectivity in the convolutional layers of the feature extraction part of the CNN, resulting to each neuron being connected only to a subset of the input image. Thus, the total number of trainable parameters is drastically reduced, minimizing the computational expense and the complexity of the network. Apart from this, with parameter sharing the convolutional layer gains the ability to detect similar spatial patterns in different locations within the input image.

Another concept of CNN that makes it suitable for this particular problem is spatial inductive bias. When referred to inductive bias of an ANN in computer vision, we refer to an ensemble of assumptions that assist the model to generate unknown output. In the case of gridded data handled by a CNN, spatial inductive bias assumes a certain spatial structure of the data.

Two similar but not identical concepts of CNNs are also relevant to our classification task: transitional invariance and transitional equivariance. Translational invariance expresses the ability of the CNN to generate robust output even when the initial input is translated. Translational equivariance expresses the ability of a CNN to detect the location of the spatial pattern of interest in the image without it being in a certain pre-known position.

Even though the aforementioned concepts give a clear advantage to CNNs when dealing with spatial data, the potential of a conventional FNN is also being examined. The number and type of hidden layers, as well as the overall architecture of the FNN, is defined after trial-and-error processes and is a subject closely related to the spatiotemporal characteristics of the data used. The exact architecture of the proposed FNN is also presented in Appendix I.

Testing different ANN architectures, modeling approaches and spatiotemporal information combinations is featured in Chapter 6.

5.3 Benchmark models

In order to objectively compare the models performance simpler ML classification algorithms are used as benchmarks: Logistic Regression, Random Forest, Gaussian Naïve Bayes and Support Vector Machine.

Logistic regression is a regression analysis with the dependent variable being binary and is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal or interval independent variables. Logistic regression is fast and easy to apply but lacks the flexibility that other supervised classification methods provide. It cannot capture complex relationships between variables and is expected to have lower, overall accuracy (Hoffman, 2019).

As (Siddharth Misra, 2020) states, Random Forest classifier is an ensemble method that trains several decision trees in parallel with bootstrapping followed by aggregation. Bootstrapping indicates that several individual decision trees are trained in parallel on various subsets of the training dataset using different subsets of available features. Random Forest is able to classify large data with accuracy. It acts as a tree predictor where every tree depends on the random vector values. The basic concept behind this is that a group of "weak learners" may come together to build a "strong learner" (Paul & Bhatia, 2020).

Gaussian Naive Bayes is a classification technique where the calculation of the probabilities for each hypothesis are simplified to make their calculation tractable. The values of each attribute value are assumed to be conditionally independent given the target value. Even though this assumption in many cases does not hold, Naive Bayes performs surprisingly on these data too (Brownlee, 2016).

The tasks that Support Vector Machine (SVM) models complete are to find a separating line (or hyperplane) between data of two classes and to learn the classification and regression rules of the input data. SVMs basically descend from the structural risk minimization principle and, thus, confront overfitting (Vahid Mohammadi, 2019).

5.4 Modeling overview

Overall, the process scheme of all the modeling strategies used in this research are:

1. The main CNN classifier, discussed in further detail in Appendix I;
2. Four benchmark ML classifiers (Logistic Regression, Random Forest, Gaussian Naïve Bayes and Support Vector Machine) and one benchmark FNN;
3. An alternative of the CNN classifier which also incorporates the stations latitude as an input variable, further discussed in Chapter 6.2;
4. Different modeling approaches attempting to confront class imbalance, presented in Chapter 6.4;
5. Several "sophisticated" ensembles of models discussed in Chapter 6.5 built in an attempt to combine the information of different satellite precipitation products, featuring majority vote models, models that stack different satellite data in parallel input channels and multi-resolution models.
6. Three state-of-art classifiers used for multi-class classification fine-tuned and tested for this binary classification task, presented in Appendix II.

6. Results

The performances of the models introduced in Chapter 5 are tested and compared. Given the results acquired, decisions are made on differentiating the modeling process in an attempt to force better results.

6.1 Performance on total dataframe

The total dataframe is generated by stacking one remote sensing product on top of the other and duplicating days based on the overlap between the gauge and satellite data for all the 14 different gauge data and 6 different satellite-based datasets available.

The performance of the ANNs introduced in Chapter 5.2 and the benchmark models of Chapter 5.3 were first tested on the total dataframe. Only the results for the input size of 32x32 pixels are shown. Other inputs sizes of 64x64 pixels, 16x16 pixels and 8x8 pixels were also tested, but did not alter drastically the explanatory power of the classification models.

The performance of the FNN was close to the performance of the benchmark models (accuracy of 73,8%) and, thus, will also be considered a benchmark model.

The classifier CNN performed with accuracy of 76.60% in the same test dataset. The accuracy metric using 16x16 pixels around each gauge was slightly reduced (accuracy 75,30%). The precision, recall and f1-score are shown below.

The performance of the four benchmark models and the CNN classifier is presented in the following table:

Table 3: Performance of benchmark models and CNN classifier on total dataframe, with 32x32 pixels input and 1mm/d wet/dry day threshold.

	Logistic Regression	**Gaussian Naive Bayes**	**Random Forrest**	**Support Vector Classification**	**CNN Classifier**
Accuracy:	0.72	0.71	0.75	0.72	0.77
f1-score: **Wet days**	0.42	0.50	0.54	0.41	0.54
f1-score: **Dry days**	0.81	0.81	0.82	0.81	0.83

Below the confusion matrix of the classifier CNN:

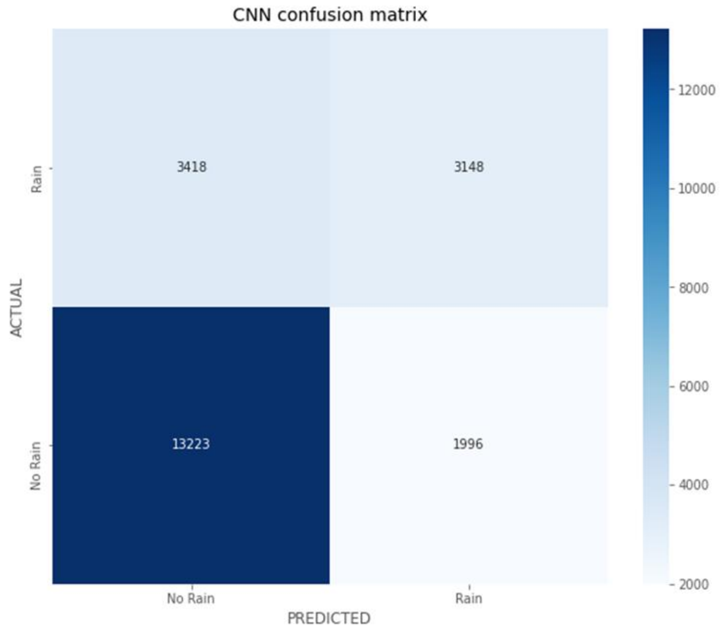


Figure 18: Classifier CNN confusion matrix on total dataframe.

Given that the classes are imbalanced, the same classifier CNN was tested for coarser time-scales. Using as input data 3-day averages at a 32x32 pixels spatial extent the results are improved. The classes are more balanced, the dry days cover 45,72% of the dataset. The overall accuracy is 75%, while the precision, recall and f1-score are presented at the table below for a test dataset of almost 4.500 3-day periods:

Table 4: Classifier CNN performance on 3-day averages of the total dataframe.

	precision	recall	f1-score	support
No Rain	0.81	0.61	0.69	2.070
Rain	0.72	0.87	0.79	2.387

The confusion matrix is depicted in Figure 19.

The higher class balance and the confrontation of the timing mismatch between the data two sources (more on that in Chapter 6.6) leads to improved results for data of coarser temporal resolution. However, since the dry spell of interest for the smallholder farmers is defined as a 5-day long period, the generated dry spell indices have to be in a temporal resolution higher than 3 days. Hence, emphasis is put on reproducing the data in the daily scale.

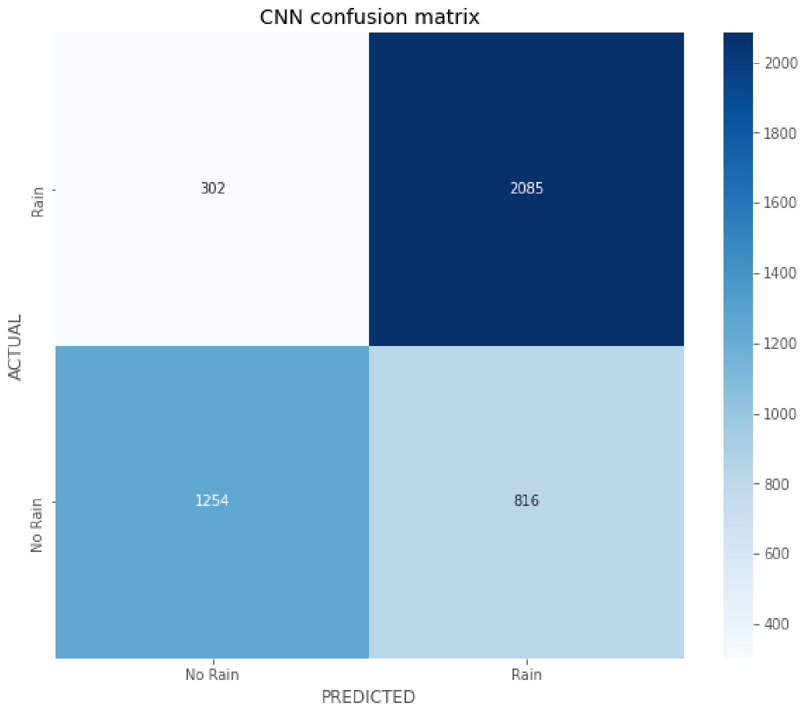


Figure 19: Classifier CNN confusion matrix on 3-day averages of the total dataframe.

6.2 Incorporating latitude

The literature review of Chapter 3.2 and also in Table 2 clearly showed that there is a climatological trend in precipitation in South Africa with latitude. Hence, an attempt was made to incorporate latitude to the regressors of the classifier, so that the total input was 32x32 precipitation pixel-values plus the latitude. The latitude is incorporated as input variable as an additional feature in the classifier part of the CNN. The initial input of the feature extraction part of the CNN is the 32x32 pixels and then the latitude is added during the flattening at the start of the classifier.

Overall, adding the latitude as “static” input did not gave any additional explanatory power to the model. In particular, the overall accuracy stayed constant at 76%, while the dry day and wet day f1-scores were 0.83 and 0.54, respectively, even though there is a “signal” about the climatology in the initial data, as minutely explained in Chapter 4.2.

6.3 Testing individual datasets

The attempt to reproduce the dry spell sequences using all six of the satellite products available did not feature high accuracy. Hence, all the products are going to be tested individually in an attempt to distinguish the best-performing ones. Out of these, the three best performing satellite-based products are used as part of a more complex model that will be introduced later on.

The results of the individual assessment for each satellite product are presented in the following table. From that, it can be distinguished that the three best performing datasets are CMORPH, TAMSAT and IMERG. Except of these, MSWEP also performs adequately given that its classes are relatively more balanced in comparison with other datasets.

Table 5: Individual performance of satellite-based datasets using the CNN classifier.

	Percentage wet days	Accuracy	F1-score dry days	F1-score wet days	F1-score weighted
MSWEP	30,47%	0.75	0.83	0.52	0.73
CMORPH	28,87%	0.77	0.84	0.54	0.75
CHIRPS	30,46%	0.73	0.82	0.46	0.71
TAMSAT	30,62%	0.77	0.84	0.62	0.77
IMERG	28,24%	0.77	0.85	0.56	0.76
RFE	27,89%	0.75	0.83	0.50	0.74

The next step is to distinguish the three best-performing satellite products and create a dataframe based on the overlap between the gauge and all three satellite precipitation products. Given the different temporal coverage of the point measurements and of all three datasets, the dataset is shrinked to 16.026 days where the three satellite products and the gauge data overlap across all 14 stations.

The idea is to use this new dataframe to test more complex models:

- a model that has a (32, 32, 3) input with each input channel being a different satellite product;
- majority vote models where one model is being ran for each satellite product and the final prediction is made based on the level of agreement of the three individual predictions;
- multi-resolution models where each channel has a different temporal resolution (e.g. mean of 2 days, max of 3 days, etc.) of the initial data.

New runs were made to the models testing the new overlapping datasets of each satellite products. The results are shown below, using binary crossentropy as loss function. The last column named "Borderline misclassifications" contains additional info that can be used to better understand the results: it gives the percentage of misclassified days for each dataset that were close to the decision threshold, namely they were between 0.30 and 0.70 with the decision threshold between dry and wet days being 0.50.

Table 6: Individual performance of satellite-based datasets (for the overlapping days of the 3 datasets) using the CNN classifier

	Accuracy	F1-score dry days	F1-score wet days	F1-score weighted	Borderline misclassification
CMORPH	0.76	0.84	0.54	0.75	64,36%
TAMSAT	0.76	0.84	0.58	0.76	65,95%
IMERG	0.78	0.85	0.57	0.77	62,61%

The fact that a high percentage of misclassified days were close to being correctly classified leaves the door open for further investigation.

6.4 Dealing with class imbalance

One challenge that we have to face is that there is an imbalance between the number of dry days (70%) and wet days (30%). The model that tested these datasets ran with binary crossentropy as its loss function, the most common loss function for binary classification. However, under class imbalance the model is seeing much more zeros than ones, as also shown by the recall metrics. Hence, the model was trained to reproduce more dry than wet days because the training loss can be minimized by doing so.

One solution to deal with this class imbalance is to modify the loss function. Loss functions of weighted binary crossentropy and dynamical weighted binary crossentropy are going to be tested in the individual datasets. The results are reported in Table 7 and Table 8. These two functions modify the binary crossentropy function found in Keras by adding a weighting. This weight is determined dynamically for every batch by identifying how many positive and negative classes are present and modifying accordingly.

Overall, the weighted loss functions reproduce more rainy days but do now seem to improve the model in terms of accuracy and f1-score. Except of these, the imbalanced classification oriented loss function “focal loss” was also tested, but did not return improved results either.

Another way of addressing class imbalance is to change the dataset used by under- or over-sampling elements in the training dataset. In cases where instances from the over-represented class are being discarded, the technique is called under-sampling. A run was made for each dataset where days of the over-represented dry day class were discarded from the training dataset. The test dataset was kept with the initial data with ratios of approximately 70:30 dry to wet days. Using under-sampling, the overall accuracy of the model decreased, even though it managed to reproduce more wet days.

Table 7: Individual performance of under-sampled overlapping satellite-based datasets (loss function: binary_crossentropy)

	Accuracy	F1-score dry days	F1-score wet days	F1-score weighted	Borderline misclassification
CMORPH	0.72	0.77	0.62	0.72	57,02%
TAMSAT	0.72	0.77	0.64	0.73	56,62%
IMERG	0.74	0.81	0.61	0.75	58,34%

Table 8: Individual performance of overlapping satellite-based datasets (loss function: dyn_weighted_bincrossentropy)

	Accuracy	F1-score dry days	F1-score wet days	F1-score weighted	Borderline misclassification
CMORPH	0.71	0.78	0.59	0.72	63,93%
TAMSAT	0.73	0.78	0.64	0.74	52,81%
IMERG	0.73	0.80	0.61	0.75	54,50%

In conclusion, none of the methods for dealing with class imbalances that we used in this thesis were able to substantially improve the performance of the classifiers. Thus, the sophisticated models tested featured the initial dataset without under-sampling and binary crossentropy as loss function.

6.5 Trying out more sophisticated models

The dataframe was tested for: (a) a model that stacks-up the satellite products at three different input channels, (b) two different majority vote models where the one will decide the predicted binary output based on the binary outputs of the individual models (one for each satellite product) and one that will decide based on the raw output value of the individual models, and (c) a multi-resolution model where each channel has a different temporal resolution of the initial dataset.

Stacked-up model

This model stacks-up the satellite products and creates a (32, 32, 3) input. The overall accuracy of the model is 0.77, while 60,15% of the misclassified days are between borderline misclassifications.

Its results are presented below:

Table 9: Stacked-up model's performance in terms of precision, recall and f1-score.

	precision	recall	f1-score	support
No Rain	0.81	0.89	0.85	2.292
Rain	0.64	0.47	0.54	914
Weighted	0.76	0.77	0.76	3.206

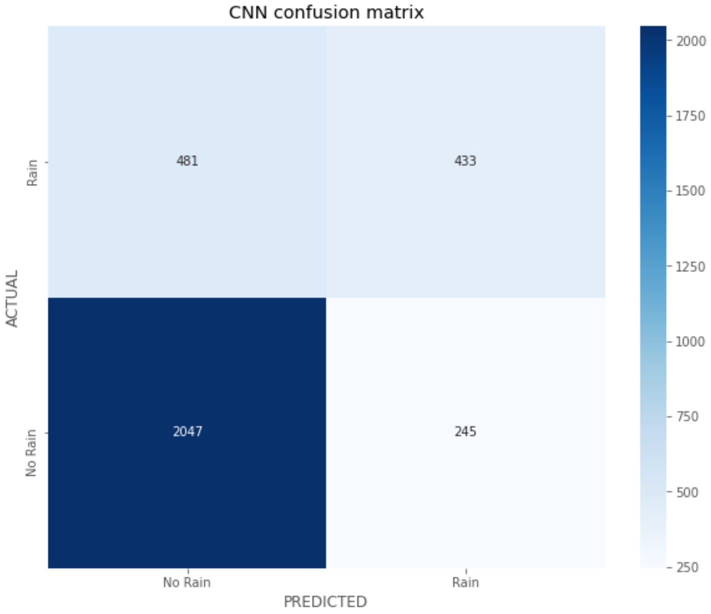


Figure 20: Stacked-up model's confusion matrix

Most of the misclassified days are wet days predicted as dry ones (modeled output below 0.50). A typical example of misclassification is presented in Figure 21 (CMORPH-IMERG Jensen-Shannon distance of 0.23 and Hamming distance of 0.28) where again CMORPH and IMERG are really similar. In this example the mismatch between what the satellite sensors “see” (dry day) and what the gauge reports (heavy precipitation) makes it highly unlikely for any classifier to be correct here.

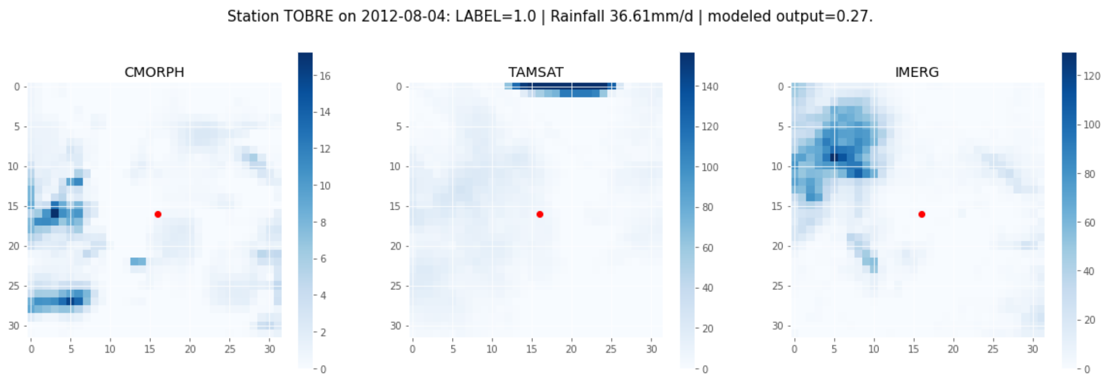


Figure 21: Stacked-up model's misclassified day No.1

Figure 22 presents a case where all three datasets report heavy rainfall and the groundtruth none. Again, this is evidence that a decent amount of the misclassifications (more specific information in Chapter 6.6) are due to the data mismatch and not the modeling approach.

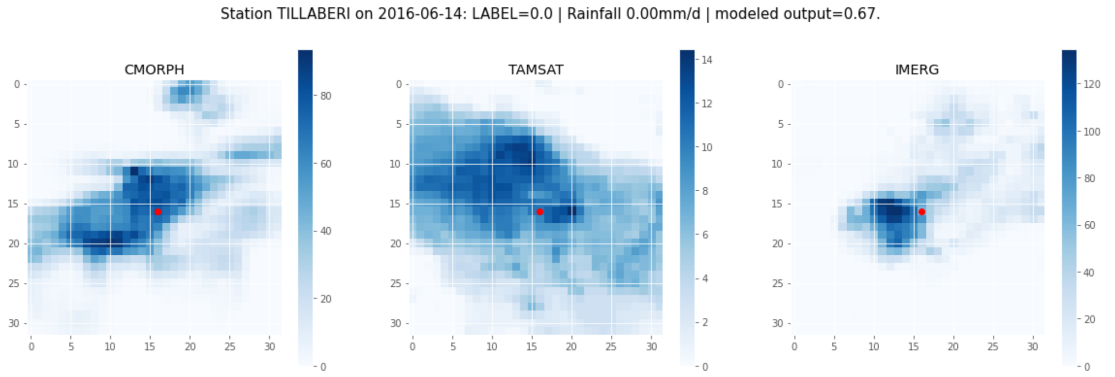


Figure 22: Stacked-up model's misclassified day No.2

Majority vote model 1 – binary outputs

For this model the binary output of the three individual models were used to predict the outcome of a single day. For example, if CMORPH and IMERG predict a wet day (raw out above 0.5 for both) and TAMSAT predicts a dry day (raw output below 0.5), then the day is classified as wet. By examining the performances of the individual models, it can be seen that approximately 90% of the misclassified days are misclassified for all three individual models. Hence, we expect these days to be misclassified by the majority vote models too.

In the following table and figure the performance and the confusion matrix (Figure 23) of this majority vote model are shown. The accuracy is 0.77.

Next, some typical misclassified days are presented. The satellite images of CMORPH and IMERG present on average high similarity, assessed by the similarity indices like the coma or the perceptual hash. In Figure 24 IMERG is the only satellite product that approaches the groundtruth. However, even individual classifier based on IMERG fails.

Table 10: Majority vote model 1 performance in terms of precision, recall and f1-score.

	precision	recall	f1-score	support
No Rain	0.80	0.91	0.85	3.461
Rain	0.63	0.40	0.49	1.347
Weighted	0.75	0.77	0.75	4.808

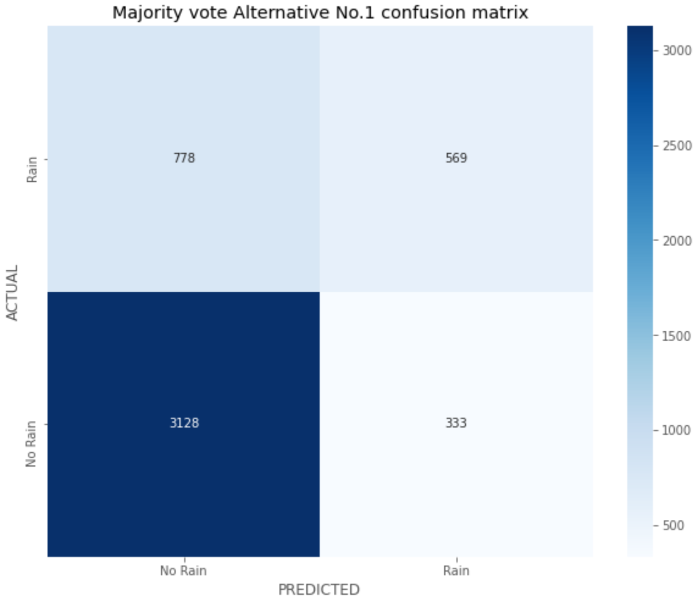


Figure 23: Majority vote model (binary output alternative) confusion matrix

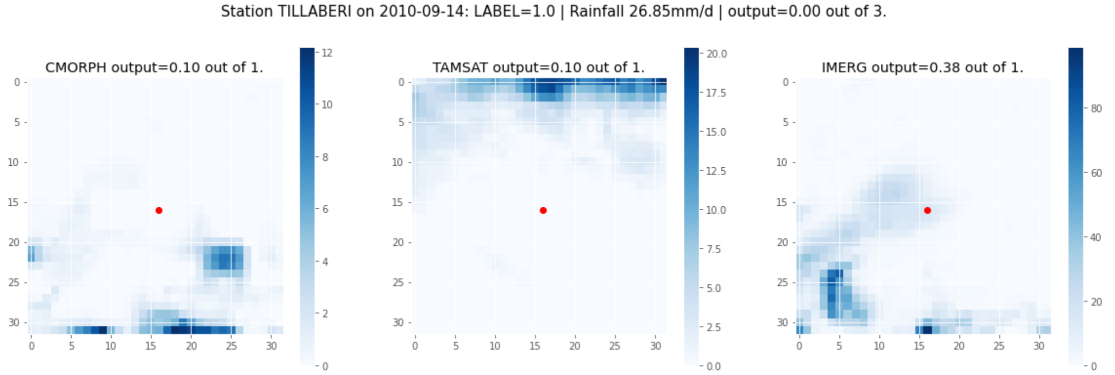


Figure 24: Majority vote model (binary output alternative) misclassification example No.1. The individual outputs of each satellite product can be distinguished above each image.

In Figure 25 all the satellite products report heavy precipitation, with the groundtruth reporting no rainfall at all. Again, one more example of misclassification that should be attributed to the data mismatch and not the modeling approach.

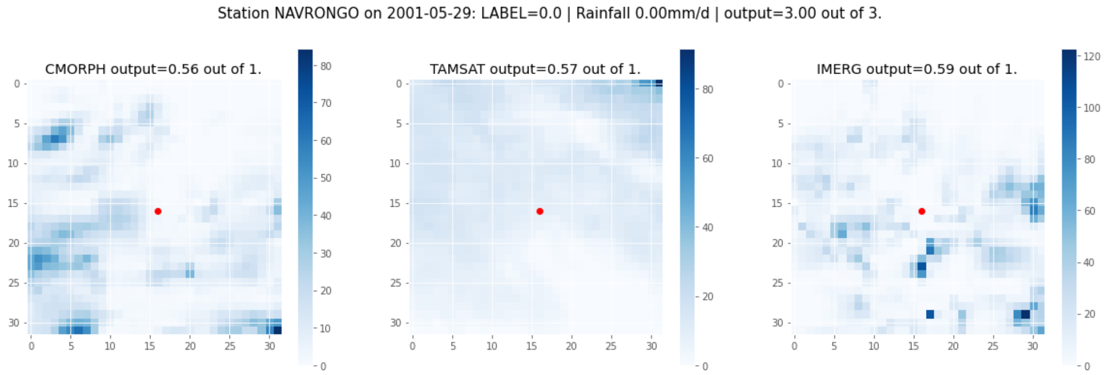


Figure 25: Majority vote model (binary output alternative) misclassification example No.2. The individual outputs of each satellite product can be distinguished above each image.

Majority vote model 2 – raw outputs

In this modeling attempt the raw sigmoid output of the three individual models was summed and used to predict the outcome of a single day. For example, if CMORPH gives an output of 0.55, IMERG 0.33 and TAMSAT 0.67, then the sum is 1.55, above the decision threshold of 1.50 and, thus, the day is classified as wet. The accuracy of the overall model is 0.77. In the table and figure below the performance and the confusion matrix of this majority vote model are displayed.

Table 11: Majority vote model 2 performance in terms of precision, recall and f1-score.

	precision	recall	f1-score	support
No Rain	0.79	0.92	0.85	3.461
Rain	0.65	0.39	0.49	1.347
Weighted	0.76	0.77	0.75	4.808

As the f1-scores and the confusion matrix of Figure 26 suggest, the model has a higher frequency in reproducing dry rather than wet days. In particular, more than 80% of the model output are dry days, whereas they represent circa 70% of the dataset.

Figure 27 is an example of best practice of a majority vote model: majority vote cancels out the misclassification of TAMSAT. Once again the similarity indices (coma and perceptual hash) indicate suchlike patterns between CMORPH and IMERG, which strive the model to a correct classification, despite the misclassification by the TAMSAT classifier.

Figure 28 presents a typical example of misclassification attributed to data mismatch. Specifically, while the gauge report more than 6 mm of precipitation, all three of the satellite products hardly observe any.

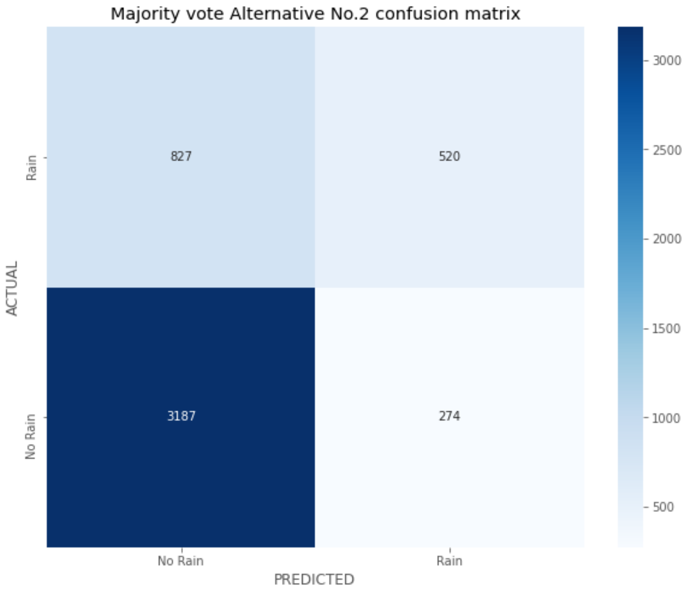


Figure 26: Majority vote model (raw output alternative) confusion matrix

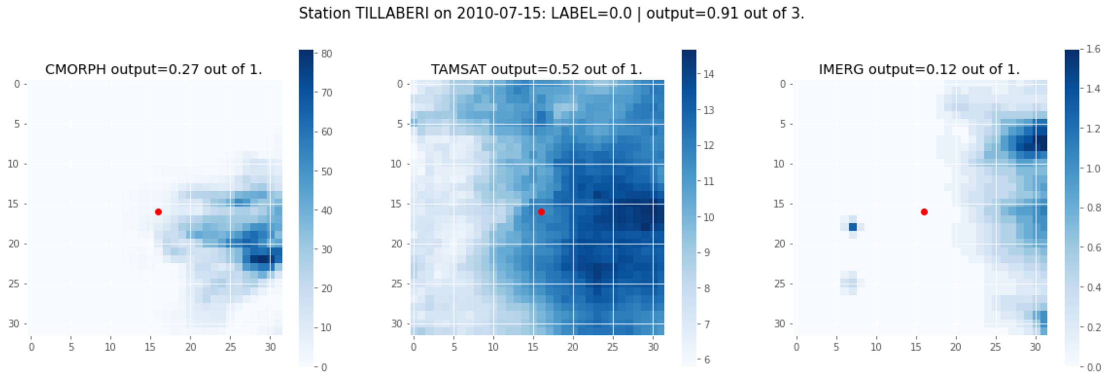


Figure 27: Majority vote model (raw output alternative) correct classification example. The individual outputs of each satellite product can be distinguished above each image.

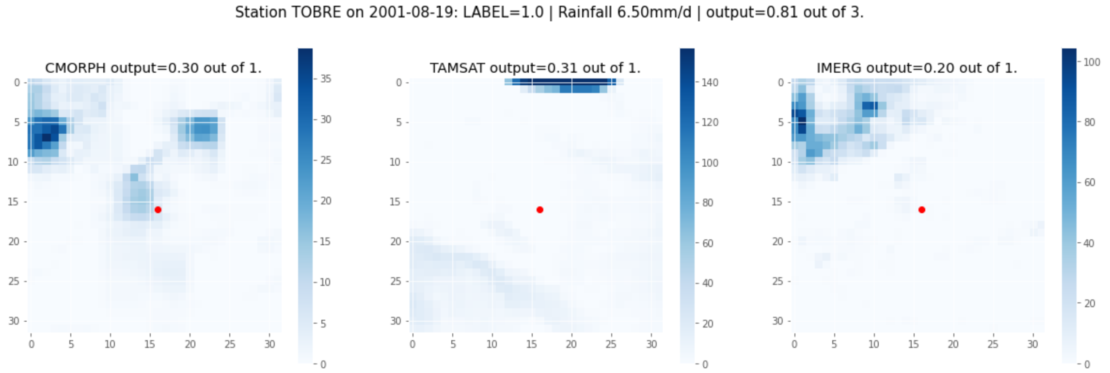


Figure 28: Majority vote model (raw output alternative) misclassification example. The individual outputs of each satellite product can be distinguished above each image.

It is also interesting to examine how close the model comes to correctly classifying the misclassified days. For that reason, the histogram of precipitation values of the wet days that were classified as dry by the model is shown in Figure 29. As can be seen, most misclassifications happened for days with low rainfall. Comparing the histogram of Figure 29 with the histogram representing the entire test data (not presented), the probability of misclassification is heavily weighted towards shallow precipitation events. This can be explained by the fact that the satellite precipitation products are well known to have lower Probability of Detection for low rainfall intensities compared to gauges, as shown by previous studies (Bogerd et al., 2021; Tan et al., 2016). The same pattern can be also observed for the misclassified days of the rest of the models. Overall, in many cases the satellite precipitation products show deficiency in detecting shallow precipitation events, leading to mismatch between the data sources. As (Berg et al., 2010) states he precipitation radar (TRMM) currently providing direct observations of rainfall intensity over the African continent has a sensitivity limited to circa 17 dBZ, leading to low POD of light rainfall events.

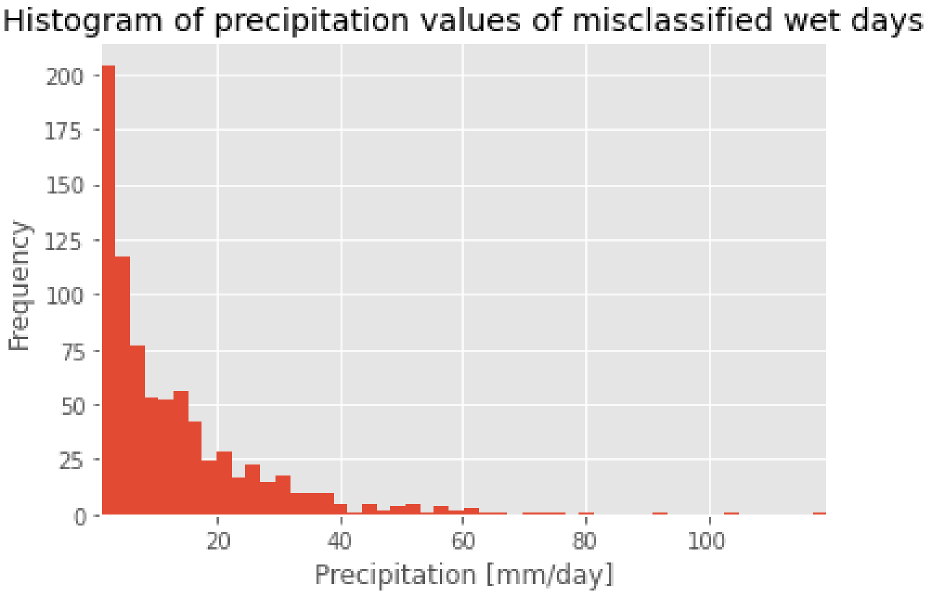


Figure 29: Histogram of precipitation values of misclassified wet days

Overall, setting a lower decision threshold classifies more days as rainy and results in slightly higher performance. Nevertheless, no drastic change is made.

Multi-resolution model

This model takes a (32, 32, 3) input and each channel has a different temporal resolution of the initial data. The data was tested for one channel of the initial data, one channel of 2-day averages and one channel of 2-day maxima. The performance was not altered significantly: the accuracy was 0.77 and the weighted f1-score 0.76. Precision, recall and f1-scores of the multi-resolution model are shown in Table 12.

Except of this attempt, several combinations were tested between channels featuring the initial data, n-day averages and n-day maxima, but the model's performance was robust for all the different combinations.

Table 12: Multi-resolution model performance in terms of precision, recall and f1-score.

	precision	recall	f1-score	support
No Rain	0.80	0.91	0.85	3.461
Rain	0.63	0.40	0.49	1.347
Weighted	0.75	0.77	0.75	4.808

6.6 Studying the properties of the misclassifications

Even in the best-performing model (e.g. Table 10 or Table 11) 23% of the satellite rainfall images were misclassified. The main proven factors contributing to misclassification are timing errors at the daily-scale, positional errors in the sub-daily scale. However, the low performance of the satellite precipitation products over West Africa can also be attributed to satellite sensor deficiencies, data sources mismatch, or random errors.

Moreover, it can be concluded that the spatial extent of the input images has little effect to the performance of the model. Utilizing different spatial extents of the input satellite images, varying from 64x64 pixels to 8x8 pixels around each gauge, the performance of the classification algorithm in terms of accuracy and f1-score was robust.

Besides, the performance of the classification did not alter significantly for different CNN architectures, hyperparameter tuning and model composition. In particular, altering the number of convolutional blocks or layers or trying out different kernel sizes, filters, padding regime, batch normalization, regularization techniques, pooling techniques, dropout layers, and activation functions (tanh and ReLU) resulted in similar performances. Even more sophisticated models featuring multi-resolution channels or majority voting between individual models were overall robust.

During our attempts to reproduce the dry spell sequences as observed by the gauges, the mismatch between the two data sources (point and satellite precipitation values) undermined the performance of all the aforementioned classification models. To verify that the source of error lies in the data itself, a run of the model with all the “faulty” days discarded was made. Days defined as “faulty” are days in which the discrepancy between the precipitation values reported by the gauge and its nearest satellite pixels is above 10 mm. In most of these cases one of the sources does not detect rainfall at all. Running a model having discarded all the “faulty” days (trying out different definitions of what constitutes a “faulty” day) the accuracy exceeds 90%. Given that, it is fair to assume that the source of error lies in the data and not the modeling approach.

Since the data is the main reason of deterioration of the performance of the model, studying the properties of the misclassified frames is an informative step forward.

When it comes to studying the sources of the misclassifications, the daily-scale timing errors can be clearly distinguished in pairs of misclassified frames of the test dataset. In these cases one data source reports (heavy) precipitation in one day while the other (almost) none, while in the following day the opposite happens. This type of misclassification accounts for approximately 15% of the total misclassifications.

Certain examples are presented in the Figures below. In particular, Figure 30 presents two consecutive days in Tillaberi station during the wettest month of the rainy season. During the first day all the satellites report heavy precipitation which probably accounts for the 45 mm of rainfall captured by the gauge during the second day. The opposite holds for Figure 31. During the first day the satellite reports almost zero precipitation while the gauge reports circa 12 mm, whereas during the second day the satellite products report heavy precipitation and the gauge none.

The occurrence of daily-scale timing mismatch is highly correlated with latitude. Specifically, in the cluster of northern gauges, namely Ouahigouya, Tara, Tillaberi and Agoufou, the daily-scale timing errors make up only 0.50-1.00% of the test dataset of each station. This statistic varies between 1.50% and 2.00% for mid-latitude gauges and can reach up to 6.20% for low-latitude stations like Tobre. This characteristic is probably linked to the relatively higher class imbalance of high-latitude stations compared to lower-latitude stations.

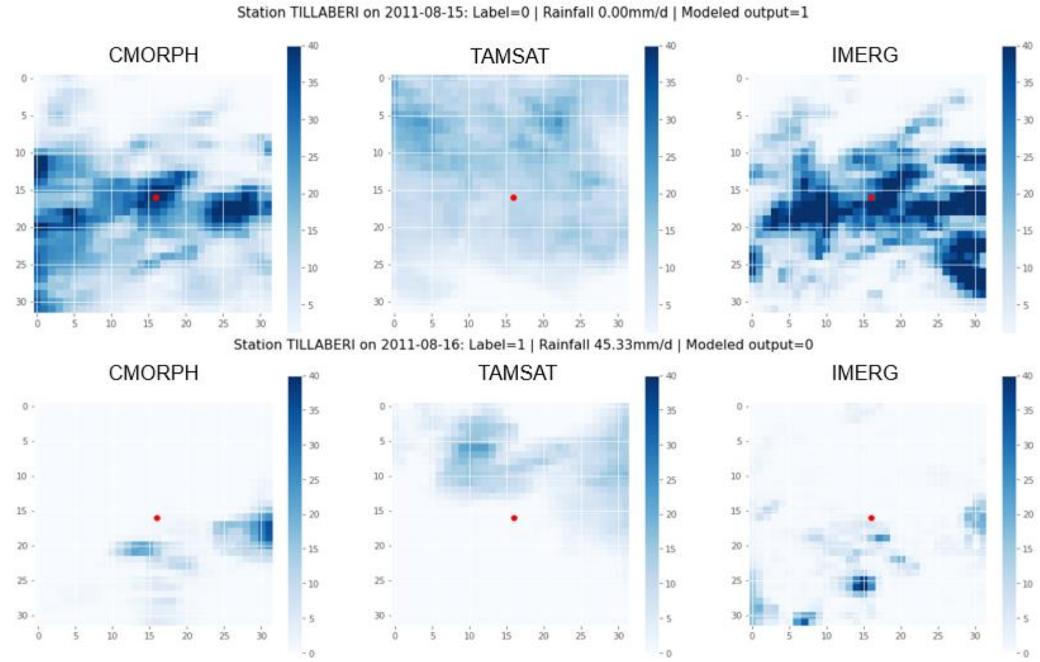

Station TILLABERI on 2011-08-15: Label=0 | Rainfall 0.00mm/d | Modeled output=1
CMORPH
TAMSAT
IMERG
Station TILLABERI on 2011-08-16: Label=1 | Rainfall 45.33mm/d | Modeled output=0
CMORPH
TAMSAT
IMERG
0
5
10
15
20
25
30
35
40

Figure 30: Typical example of daily-scale timing error misclassification. During the first day of the two, the modeled output reports a wet day when the gauge reports a dry day. The opposite happens during the second day for rainfall of approximately 45 mm.

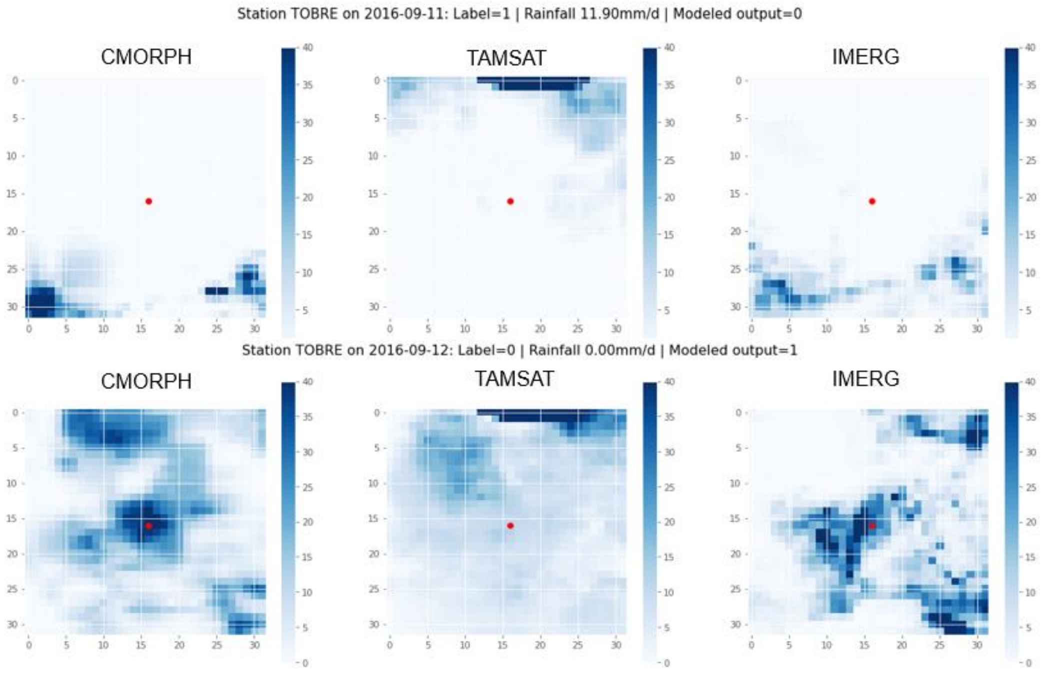

Station TOBRE on 2016-09-11: Label=1 | Rainfall 11.90mm/d | Modeled output=0
CMORPH
TAMSAT
IMERG
Station TOBRE on 2016-09-12: Label=0 | Rainfall 0.00mm/d | Modeled output=1
CMORPH
TAMSAT
IMERG

Figure 31:Typical example of daily-scale timing error misclassification. During the first day 12 mm of rainfall are observed by the gauge while the satellites barely report precipitation. The opposite holds for the next day.

Except for the daily-scale timing errors there is also a timing mismatch observed in the sub-daily scale. Since the timing shift between the different data sources is systematic, whether this will result in daily-scale timing errors or will cause a mismatch in the sub-daily scale is depended on the timing of each precipitation event. The latter errors are defined as positional errors, because they present the same spatial pattern: while the groundtruth is reporting precipitation, the pixels near the gauge report no rainfall and a small-scale precipitation window lies in some distance, usually greater than 50 km, away from the gauge (Figure 32). Even though it is fair to assume that positional errors constitute a repercussion of the timing shift of the two sources, satellite deficiencies detecting the precipitation windows ineffectively are also likely to play a role.

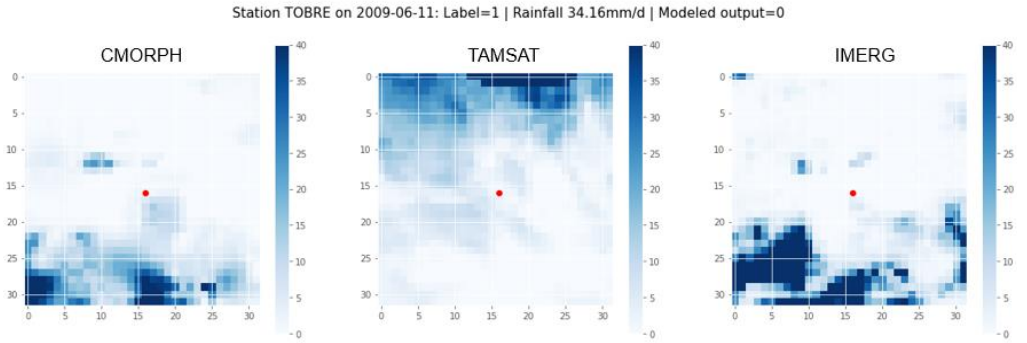

Station TOBRE on 2009-06-11: Label=1 | Rainfall 34.16mm/d | Modeled output=0
CMORPH
TAMSAT
IMERG

Figure 32: Typical example of positional errors leading to misclassification: the satellites are able to capture the precipitation observed by the gauge, but the timing mismatch of the two sources results in the precipitation windows lying in a distance greater than 100 km from the station.

Positional errors, i.e., when the rain cells are at the wrong location in the satellite products, are significantly harder to assess than time-shift errors. The spatial pattern of the positional errors reflects the local and convective nature of precipitation in the region of Western Africa and can be directly linked to the timing mismatch between the two data sources. Circa 20-22% of the misclassifications of the test dataset present this pattern, depended on how the term precipitation window is defined in each case. Besides, the satellite images of the positional errors present on average high similarity (assessed by the co-occurrence matrix similarity index) with frames of the same precipitation products from different samples. Overall, the timing mismatch between the satellite images and the groundtruth is a systematic source of error leading to misclassifications uncontestably attributed to it (daily-scale timing errors) or misclassifications in which the timing error is likely to be the cause of failure (positional errors). When it comes to the latter, there is high but not absolute confidence that these misclassifications can be charged to timing mismatch since the initial data are in the daily-scale.

Studying the rest of the misclassifications, there is no uncontested information about the source of misclassification since there is no specific pattern in the rest of the misclassified samples. On the one hand, since we know that the timing mismatch between the data sources is systematic, it is likely that it will affect the quality of the match between the groundtruth and the satellite precipitation data. On the other hand, it is already known from previous studies that the satellite products tend to underperform in gauge-scarce regions like Western Africa (Beck et al., 2017; Le Coz & Van De Giesen, 2020). Hence, the reasons of misclassification of the remaining frames are a combination between satellite product deficiencies, random errors and timing errors, the individual contribution of which cannot be clearly determined.

A deeper understanding of the properties of the different satellite precipitation products can explain the occurrence of misclassifications. In particular, e.g. IMERG and CMORPH present a relatively low POD of the rainy pixels in many regions of the world, but tends to overestimate rainfall values when rainfall is detected (Bogerd et al., 2021; Tan et al., 2016). In our case, the model has far more wet days misclassified (as presented in the confusion matrices throughout the report) because of the satellite products not detecting rainfall at all. On average 60-70% higher precipitation values are reported in the correctly classified wet days. To be precise, our results indicate that the satellite precipitation products face difficulties identifying shallow precipitation events, as confirmed in previous studies (e.g. Tan et al., 2016).

The spatial rainfall intensity patterns observed in CMORPH and IMERG present high similarity, as indicated by the similarity indices presented in Chapter 4.3 and tend to generate similar outputs during the classification process of the ML and DL models. This is a consequence of both products including passive microwave (PMW) sensors as a primary component of rainfall intensity estimation. The satellite product TAMSAT using only infrared measurements shows much lower similarity. In particular, IMERG unifies a 13-channel PMW imager and displacement vectors derived from infrared (IR) measurements on geosynchronous satellites to produce gridded estimates of rainfall at fine resolution quasi globally, while CMORPH merges geostationary IR images and PMW data from low-orbiting satellites.

6.7 Random forest vs CNN classifier

Approximately 79.50% of the frames misclassified by the Random Forest model were also misclassified by the CNN classifier. In this portion all the daily-scale timing errors and the vast majority of the sub-daily timing errors were included. The frames in which the Random Forest classifier underperformed relatively to the CNN model mostly featured wide spatial distribution of relatively low precipitation values; wet days which were misclassified by the conventional

ML algorithms (Figure 33). The fraction of wet pixels of the satellite images in which the CNN performs better than the Random Forest classifier is on average 28% greater than the one of the frames in which the Random Forest overperforms.

However, in approximately 8 out of 10 of these frames the sigmoid output of the CNN classifier was close to the decision threshold (from 0.30 to 0.70), making them "borderline" correctly classified frames. Given the intrinsic uncertainty attributed to the CNN output, either due to the dropout layers of the model or the stochastic nature of the SGD, there is high chance that the aforementioned frames could be misclassified also by the CNN classifier.

The samples in which the Random Forest algorithm outperforms the CNN model present on average concentrated precipitation windows and minimum distribution of precipitation pixels, as shown in Figure 34.

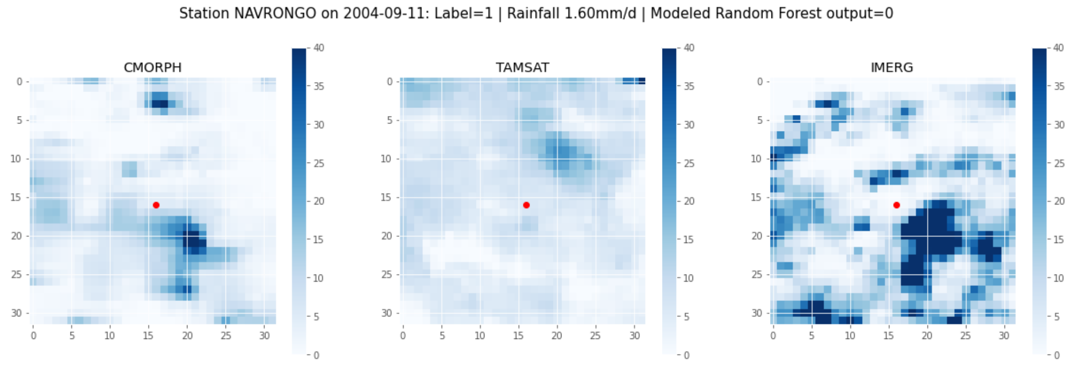


Figure 33: Typical example of a frame in which CNN overperforms in comparison to the Random Forest classifier. Its main characteristic is the wide spatial distribution of rainy pixels.

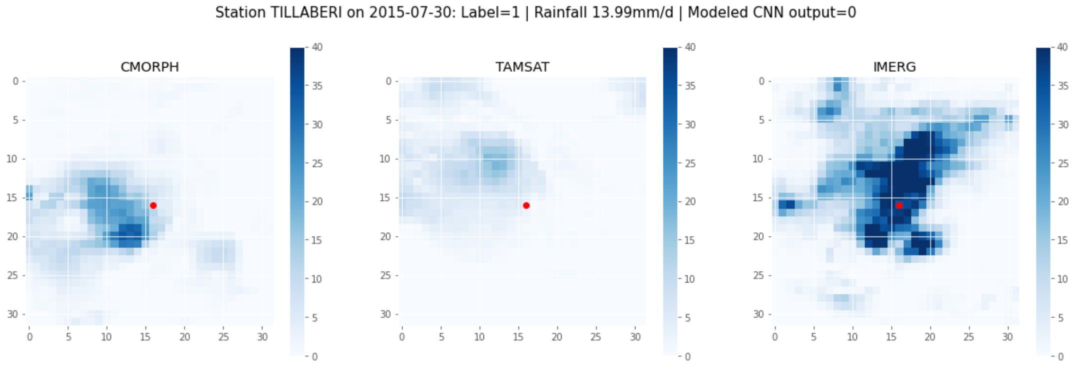


Figure 34: Typical example of a frame in which CNN underperforms in comparison to the Random Forest classifier. This sample features small-scale precipitation windows and narrow spatial distribution of rainy pixels.

It can be assumed that the ability of the feature extraction part of the CNN model to identify and manipulate the spatial properties of the input images is the main reason for the relatively better performance of the CNN algorithm, compared to the conventional ML algorithms. The CNN classifier was able to identify and utilize the spatial properties of precipitation to identify the label of each day, whereas the ML algorithms weigh the contribution of each pixel value individually.

6.8 Global properties

The ability of the CNN classifier and the best-performing ML model (Random Forest) to generalize in climatic conditions different than the ones it was trained on was tested in three different ways:

1. The classifier was trained with the data of the 10 mid- and low-latitude stations, which feature more wet days, and tested for the cluster of the 4 northern stations;
1. The classifier was trained with the data of the 8 mid- and high-latitude stations, which feature more dry days, and tested for the cluster of the 6 southern stations;
2. The classifier was trained for all the available data of this study (14 stations) and tested for 5 stations in neighbouring countries, such as Burkina Faso, Benin and Niger.

The performance of these alternatives in terms of accuracy are presented below:

Table 13: Testing the global properties of the CNN and the Random Forest classifiers using 3 different alternatives.

	Clustered training/testing No.1	**Clustered training/testing No.2**	**Testing on near stations**
Accuracy CNN classifier	77.10%	76.20%	76.80%
Accuracy Random Forest	74.20%	73.60%	72.70%
Training stations	10 mid and low-latitude stations	8 mid and high-latitude stations	14 stations in the region of Northern Ghana
Testing stations	4 high-latitude stations	6 low-latitude stations	5 stations in neighbouring countries

Overall, the performance of the CNN model is robust. However, for the global properties of the classifier to be verified, data from regions with different climatic conditions should be tested. The performance of the Random Forest classifier decreased drastically during the testing of the third alternative. Hence, it is fair to assume that the Random Forest classifier features less generalization ability than the CNN model.

7. Conclusions and recommendations

The main objective of this study was to reproduce the dry spell sequences as seen by the rain gauges in the region of Northern Ghana based on satellite precipitation products using ML and DL algorithms. Using satellite precipitation images as input and binary wet/dry day labels different classification models were trained and tested. The ability of ML and DL models to reproduce dry spell sequences observed by ground sensors (rain gauges) using satellite data was assessed in terms of accuracy and f1-score. Accuracy is computed as a ratio of the correct classification over the total population, while the f1-score is a more sophisticated metric with more value when facing classification problems where the classes (wet or dry day) are not equally represented in the data (class imbalance). The best-case accuracy and f1-score achieved were 77% for both. Station-wise, there is variance with several high latitude stations reporting accuracy above 80%, mostly due to higher class imbalance (i.e., more dry days).

Overall, the performance of the different tested CNN classifiers attempting to reproduce the dry spell sequences was robust with 77% of the days classified correctly. Testing different model architectures and ensembles did not manage to improve the accuracy of the classification model above that, clarifying the robustness of the DL classifier. In addition, the model performance is not linked to the spatial extent of the input satellite images. Trying out different spatial extents staring from 8x8 pixels around the station up to 64x64 pixels, the performance of the CNN classifier in terms of accuracy and f1-score did not alter significantly.

The overall robustness of the different DL models and model ensembles underlines the notion that it is probably the data mismatch the degrades the performance of the model(s). As often stated in ML and DL applications: “Garbage in – Garbage out”.

An important conclusion derived from the research is that the timing mismatch between rain gauge and satellite observations limits the performance of the proposed algorithm. It was shown that approximately 15% of the misclassifications can be attributed to timing errors expressed at the daily scale. These are highly linked to the latitude and, thus, the climatology of each gauge with northern stations reporting only 0.50-1.00% timing errors in the test dataset while in the southern gauges this percentage increases up to circa 6.00%. In addition to this, the timing mismatch of the two data sources (satellite and gauge) is also expressed in the sub-daily-scale through positional errors. Even though the satellite sensors are able to capture the precipitation observed by the gauge during days featuring positional errors, ultimately, they are misclassified because the precipitation windows are captured distant from the station, reflecting timing errors in the sub-daily-scale or other satellite deficiencies. This type of error accounts for approximately 20-22% of the misclassifications by the DL classifier. When it comes to the rest of the misclassifications, there is no clear and uncontested pattern of failure. Knowing already that the satellite precipitation products perform poorly in gauge-scarce regions, it can not be absolutely defined whether the misclassifications should be attributed to conditions that the satellite does not observe, timing errors, random errors, or a combination of these.

Furthermore, the classifier performance was remarkably stable (at approximately 76-77%) trying out different model characteristics, including model architecture, composition or hyperparameters. Differentiating the model composition from single CNN classifiers to more sophisticated ensembles such as majority vote models or multi-resolution input models resulted in similar performance. In addition, adding more convolutional blocks or running the classifier with different kernel sizes, filters, padding regime, pooling techniques, and activation functions did not alter the overall performance.

Moreover, the CNN classification algorithm managed to score higher accuracy than other conventional ML classification models. The ability of the feature extraction part of the CNN model to efficiently handle spatial information through convolution and pooling leads to the CNN model efficiently recognizing wet days with relatively low values and wide spatial distribution of rainfall.

An additional objective of the research was to test assess global properties of the classifier. The ability of the classification model to generalize in the wider region with similar climatological conditions was tested in three different ways:

1. By splitting the available stations to a training mid- and low-latitude cluster of 10 stations and a testing high-latitude cluster of 4 stations, in which the presence of rainy days is relatively lower;
2. By splitting the available stations to a training mid- and high-latitude cluster of 8 stations and a testing low-latitude cluster of 6 stations;
3. By training the classifier for the 14 available stations and testing its performance for data in neighbouring gauges (Benin, Niger).

Overall, using all three methodologies the performance was robust and, thus, the CNN classifier can efficiently generalize in the region of Western Africa, presenting a more efficient generalization ability than the best-performing ML model (Random Forest).

Several benchmark ML models achieved accuracy close to the one of the best-performing CNN classifiers. In particular, a Random Forest model performed the classification with circa 75% accuracy. Nevertheless, there was a significant percentage of the test dataset (about 2%) that was correctly classified by the CNN model but not the Random Forest one. The CNN classifier and the Random Forest classifier perform well in different patterns of rainfall. The frames in which the CNN classifier overcomes the Random Forest classifier featured wide spatial distribution of relatively low precipitation values, while the samples in which the CNN classifier underperforms relatively to the Random Forest model present on average concentrated precipitation windows and minimum distribution of precipitation pixels. The discrepancies of the spatial patterns are also reflected in the proportion of wet pixels of each pattern: the CNN best-performing pattern features on average approximately 28% more wet pixels than the one of Random Forest.

When it comes to further research recommendations the most informative step forward would be getting more insight in the contributions of various sources (timing/satellite sensor limitations or a combination of both) to the misclassifications. Working with data that break down the contribution of each source (PMW or IR) to the final result can help define the contribution of each type of sensor. Since existing studies suggest that PMW sensor identify the presence of precipitation more accurately than IR sensors, linking the days with or without PMW overpasses to the overall performance of the model could provide insights.

In addition to this, working with data of higher temporal resolution can expand the research possibilities. Given the addressed timing mismatch between the data sources, aggregating the available half-hourly IMERG data to daily scale for different time shifts can determine the daily resolution IMERG product that features the lowest timing discrepancies with the groundtruth. Besides, in regions where point measurements of higher temporal resolution are available, performing the classification with algorithms that feature the ability to account for the sequence of data (ConvLSTM) can limit the timing mismatch and accurately reproduce the temporal precipitation patterns. Nonetheless, working with ConvLSTM models in the daily resolution would not bear results due to the convective and, thus, highly varying nature of rainfall in the region of Northern Ghana.

In addition to better understanding the sources of misclassification, one could also work on improving the usefulness of the downscaled data. One way to do this would be to provide users with more information about the uncertainty affecting individual predictions. This can be achieved either by bootstrapping the input data or by running a Monte Carlo simulation. In the second case, the uncertainty of each output is directly linked to the stochastic nature of the dropout layers and the backpropagation of the ANN. As a next step linking the width of the prediction intervals of each satellite product to its properties could have interest in quantifying the uncertainty of predictions for IR and PMW sensors. Besides, knowing the uncertainty attributed to each frame can improve the overall performance by operating a meta-model which decides in which frames will go forward utilizing the CNN classifier and in which frames the CNN classifier is uncertain about the classification. In the latter case, other classification algorithms can be used that perform better in the uncertain frames. For example, since the Random Forest classifier performs better in frames featuring different characteristics, this added explanatory power can be utilized by the meta-model. Nevertheless, more research has to be put on clarifying the optimal alternative model(s) of the meta-model ensemble.

8. References

Acheampong, P. K. (1982). *Rainfall Anomaly along the Coast of Ghana . Its Nature and Causes Author (s): Peter Kwabenah Acheampong Published by : Taylor & Francis , Ltd . on behalf of the Swedish Society for Anthropology and Geography Stable URL : https://www.jstor.org/stable/520*. *64*(3), 199–211.

Agoungbome, S. M. D. (n.d.). *How well do rainfall products reproduce locally observed rainfall patterns in West Africa? (Powerpoint presentation)*.

Asfaw, S., Scognamillo, A., Caprera, G. Di, Sitko, N., & Ignaciuk, A. (2019). Heterogeneous impact of livelihood diversification on household welfare: Cross-country evidence from Sub-Saharan Africa. *World Development*, *117*, 278–295. https://doi.org/10.1016/j.worlddev.2019.01.017

Badas, M. G., Deidda, R., & Piga, E. (2006). Modulation of homogeneous space-time rainfall cascades to account for orographic influences. *Natural Hazards and Earth System Science*, *6*(3), 427–437. https://doi.org/10.5194/nhess-6-427-2006

Barron, J., Rockström, J., Gichuki, F., & Hatibu, N. (2003). Dry spell analysis and maize yields for two semi-arid locations in east Africa. *Agricultural and Forest Meteorology*, *117*(1–2), 23–37. https://doi.org/10.1016/S0168-1923(03)00037-6

Beck, H. E., van Dijk, A. I. J. M., Levizzani, V., Schellekens, J., Miralles, D. G., Martens, B., & de Roo, A. (2017). MSWEP: 3-hourly 0.25° global gridded precipitation (1979–2015) by merging gauge, satellite, and reanalysis data. *Hydrology and Earth System Sciences*, *21*(1), 589–615. https://doi.org/10.5194/hess-21-589-2017

Beck, H. E., Vergopolan, N., Pan, M., Levizzani, V., Van Dijk, A. I. J. M., Weedon, G. P., Brocca, L., Pappenberger, F., Huffman, G. J., & Wood, E. F. (2017). Global-scale evaluation of 22 precipitation datasets using gauge observations and hydrological modeling. *Hydrology and Earth System Sciences*, *21*(12), 6201–6217. https://doi.org/10.5194/hess-21-6201-2017

Beck, H. E., Wood, E. F., Pan, M., Fisher, C. K., Miralles, D. G., Van Dijk, A. I. J. M., McVicar, T. R., & Adler, R. F. (2019). MSWep v2 Global 3-hourly 0.1° precipitation: Methodology and quantitative assessment. *Bulletin of the American Meteorological Society*, *100*(3), 473–500. https://doi.org/10.1175/BAMS-D-17-0138.1

Berg, W., L'ecuyer, T., & Haynes, J. M. (2010). The distribution of rainfall over oceans from spaceborne radars. *Journal of Applied Meteorology and Climatology*, *49*(3), 535–543. https://doi.org/10.1175/2009JAMC2330.1

Bogerd, L., Overeem, A., Leijnse, H., & Uijlenhoet, R. (2021). A comprehensive five-year evaluation of IMERG Late Run precipitation estimates over the Netherlands A comprehensive five-year evaluation of IMERG Late Run precipitation estimates over the Netherlands. *Journal of Hydrometeorology*.

Brownlee, J. (2016). *Naive Bayes for Machine Learning*. https://machinelearningmastery.com/naive-bayes-for-machine-learning/

Chollet, F. & others. (2015). *Keras.*

Chouinard, O., Plante, S., Weissenberger, S., Noblet, M., & Guillemot, J. (2017). The participative action research approach to climate change adaptation in Atlantic Canadian coastal communities. In *Climate Change Management* (pp. 67–87). Springer. https://doi.org/10.1007/978-3-319-53742-9_5

Christensen, J. H., Kanikicharla, K. K., Aldrian, E., An, S. Il, Albuquerque Cavalcanti, I. F., de Castro, M., Dong, W., Goswami, P., Hall, A., Kanyanga, J. K., Kitoh, A., Kossin, J., Lau, N. C., Renwick, J., Stephenson, D. B., Xie, S. P., Zhou, T., Abraham, L., Ambrizzi, T., … Zou, L. (2013). Climate phenomena and their relevance for future regional climate change. *Climate Change 2013 the Physical Science Basis: Working Group I Contribution to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*, *9781107057*, 1217–1308. https://doi.org/10.1017/CBO9781107415324.028

Codjoe, S. N. A., Owusu, G., & Burkett, V. (2014). Perception, experience, and indigenous knowledge of climate change and variability: The case of Accra, a sub-Saharan African city. *Regional Environmental Change*, *14*(1), 369–383. https://doi.org/10.1007/s10113-013-0500-0

Collier, P., Conway, G., & Venables, T. (2008). Climate change and Africa. *Oxford Review of Economic Policy*, *24*(2), 337–353. https://doi.org/10.1093/oxrep/grn019

Debrah, E. (2009). *The Economy and Regime Change in Ghana, 1992-2004*. 84–113.

Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ). (2019). *An analysis of the agricultural insurance market development in Ghana*.

Diatta, S., & Fink, A. H. (2014). Statistical relationship between remote climate indices and West African monsoon variability. *International Journal of Climatology*, *34*(12), 3348–3367. https://doi.org/10.1002/joc.3912

Ellis, F. (1998). Household strategies and rural livelihood diversification. *Journal of Development Studies*, *35*(1), 1–38. https://doi.org/10.1080/00220389808422553

Enfors, E. I., & Gordon, L. J. (2007). Analysing resilience in dryland agro-ecosystems: a case study of the Makanya catchment in Tanzania over the past 50 years. *Land Degradation & Development*, *18*(6), 680–696. https://doi.org/10.1002/ldr.807

Fink, A. H., Vincent, D. G., & Ermert, V. (2006). Rainfall types in the West African Sudanian zone during the summer monsoon 2002. *Monthly Weather Review*, *134*(8), 2143–2164. https://doi.org/10.1175/MWR3182.1

Fischer, B. M. C., Mul, M. L., & Savenije, H. H. G. (2013). Determining spatial variability of dry spells: A Markov-based method, applied to the Makanya catchment, Tanzania. *Hydrology and Earth System Sciences*, *17*(6), 2161–2170. https://doi.org/10.5194/hess-17-2161-2013

Froidurot, S., & Diedhiou, A. (2017). Characteristics of wet and dry spells in the West African monsoon system. *Atmospheric Science Letters*, *18*(3), 125–131. https://doi.org/10.1002/asl.734

Fuglede, B., & Topsøe, F. (2004). *Jensen-Shannon Divergence and Hilbert space embedding*.

Funk, C., Hoell, A., Shukla, S., Bladé, I., Liebmann, B., Roberts, J. B., Robertson, F. R., & Husak, G. (2014). Predicting East African spring droughts Predicting East African spring droughts using Pacific and Indian Ocean sea surface temperature indices Predicting East African spring droughts. *Hydrol. Earth Syst. Sci. Discuss*, *11*, 3111–3136. https://doi.org/10.5194/hessd-11-3111-2014

Gbangou, T., Ludwig, F., van Slobbe, E., Greuell, W., & Kranjac-Berisavljevic, G. (2020). Rainfall and dry spell occurrence in Ghana: trends and seasonal predictions with a dynamical and a statistical model. *Theoretical and Applied Climatology*, *141*(1–2), 371–387. https://doi.org/10.1007/s00704-020-03212-5

Gbangou, T., Ludwig, F., van Slobbe, E., Hoang, L., & Gordana Kranjac-Berisavljevic. (2019). Seasonal variability and predictability of agro-meteorological indices: Tailoring onset of rainy season estimation to meet farmers' needs in Ghana. *Climate Services*, *14*(January), 19–30. https://doi.org/10.1016/j.cliser.2019.04.002

Gbangou, T., Sarku, R., Van Slobbe, E., Ludwig, F., Kranjac-Berisavljevic, G., & Paparrizos, S. (2020). Coproducing weather forecast information with and for smallholder farmers in Ghana: Evaluation and design principles. *Atmosphere*, *11*(9). https://doi.org/10.3390/atmos11090902

Ghana Statistical Service. (2000). *Poverty Trends in Ghana in the 1990s*. *October*, 1–78.

Guo, J., Liang, X., & Ruby Leung, L. (2004). Impacts of different precipitation data sources on water budgets. *Journal of Hydrology*, *298*(1–4), 311–334. https://doi.org/10.1016/j.jhydrol.2003.08.020

Gyasi, E. A. (Edwin A., 1943-, & Uitto, J. I. (1997). *Environment, biodiversity and agricultural change in West Africa*. United Nations University Press. https://agris.fao.org/agris-search/search.do?recordID=US201300036480

Haralick, R. M. (1979). Statistical and structural approaches to texture. *Proceedings of the IEEE*, *67*(5), 786–804. https://doi.org/10.1109/PROC.1979.11328

HarvestChoice. (2020). *HarvestChoice | IFPRI : International Food Policy Research Institute*. 2020. https://www.ifpri.org/project/harvestchoice

Herman, A., Kumar, V. B., Arkin, P. A., & Kousky, J. V. (1997). Objectively determined 10-day African rainfall estimates created for famine early warning systems. *International Journal of Remote Sensing*, *18*(10), 2147–2159. https://doi.org/10.1080/014311697217800

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, *9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Hoffman, J. I. E. (2019). *Logistic Regression Analysis - an overview | ScienceDirect Topics*. https://www-sciencedirect-com.tudelft.idm.oclc.org/topics/medicine-and-dentistry/logistic-regression-analysis

Hunter, J. D. (2007). "Matplotlib: A 2D Graphics Environment." *Computing in Science & Engineering*, *9*(3), 90–95.

Kaiming, H., Xiangyu, Z., Shaoqing, R., & Jian, S. (2015). Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification Kaiming. *Biochemical and Biophysical Research Communications*, *498*(1), 254–261.

Kratzert, F., Klotz, D., Brenner, C., Schulz, K., & Herrnegger, M. (2018). Rainfall–runoff modelling using Long Short-Term Memory (LSTM) networks. *Hydrology and Earth System Sciences*, *22*(11), 6005–6022. https://doi.org/10.5194/hess-22-6005-2018

Laube, W., Schraven, B., & Awo, M. (2012). Smallholder adaptation to climate change: Dynamics and limits in Northern Ghana. *Climatic Change*, *111*(3), 753–774. https://doi.org/10.1007/s10584-011-0199-1

Le Coz, C., & Van De Giesen, N. (2020). Comparison of rainfall products over sub-saharan africa. *Journal of Hydrometeorology*, *21*(4), 553–596. https://doi.org/10.1175/JHM-D-18-0256.1

Leinonen, J., Nerini, D., & Berne, A. (2020). Stochastic Super-Resolution for Downscaling Time-Evolving Atmospheric Fields with a Generative Adversarial Network. *ArXiv*, 1–14. https://doi.org/10.1109/tgrs.2020.3032790

mapsofworld.com. (2020). *Where is Ghana Located? Location map of Ghana*. https://www.mapsofworld.com/ghana/ghana-location-map.html

Maranan, M., Fink, A. H., & Knippertz, P. (2018). Rainfall types over southern West Africa: Objective identification, climatology and synoptic environment. *Quarterly Journal of the Royal Meteorological Society*, *144*(714), 1628–1648. https://doi.org/10.1002/qj.3345

Masinde, M., Bagula, A., & Muthama, N. J. (2012). The role of ICTs in downscaling and up-scaling integrated weather forecasts for farmers in sub-Saharan Africa. *Proceedings of the Fifth International Conference on Information and Communication Technologies and Development - ICTD '12*, 122. https://doi.org/10.1145/2160673.2160690

Ochola, W. O., & Kerkides, P. (2003). A Markov chain simulation model for predicting critical wet and dry spells in Kenya: Analysing rainfall events in the kano plains. *Irrigation and Drainage*, *52*(4), 327–342. https://doi.org/10.1002/ird.94

Omotosho, B. J. (1985). *THE SEPARATE CONTRIBUTIONS OF LINE SQUALLS , THUNDERSTORMS AND THE MONSOON TO THE TOTAL RAINFALL IN NIGERIA*. *5*, 543–552.

Oxford Reference. (2017). *Hamming distance*. https://www.oxfordreference.com/view/10.1093/oi/authority.20110803095918607

Paul, S., & Bhatia, D. (2020). Smart healthcare for disease diagnosis and prevention. In *Smart Healthcare for Disease Diagnosis and Prevention*. Elsevier. https://doi.org/10.1016/C2018-0-03178-3

Pedregosa, F. et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12(Oct)*, pp.2825–2830.

Physical Geography | West Africa. (2020). 2020. https://eros.usgs.gov/westafrica/physical-geography

Pollack, J. B. (1990). *Recursive Distributed Representations*.

Population Reference Bureau. (2019). 2019 World Population Data Sheet. *2019 World Population Data Sheet, population*. https://www.prb.org/datasheets/

QGIS, D. T. (2009). *QGIS Geographic Information System*. Open Source Geospatial Foundation. http://qgis.org

Rockström, J. (2000). Water resources management in smallholder farms in Eastern and Southern Africa: An overview. *Physics and Chemistry of the Earth, Part B: Hydrology, Oceans and Atmosphere*, *25*(3), 275–283. https://doi.org/10.1016/S1464-1909(00)00015-0

Rockström, Johan, Folke, C., Gordon, L., Hatibu, N., Jewitt, G., Penning de Vries, F., Rwehumbiza, F., Sally, H., Savenije, H., & Schulze, R. (2004). A watershed approach to upgrade rainfed agriculture in water scarce regions through Water System Innovations: An integrated research initiative on water for food and rural livelihoods in balance with ecosystem functions. *Physics and Chemistry of the Earth*, *29*(15-18 SPEC.ISS.), 1109–1118. https://doi.org/10.1016/j.pce.2004.09.016

Salack, S., Klein, C., Giannini, A., Sarr, B., Worou, O. N., Belko, N., Bliefernicht, J., & Kunstman, H. (2016). Global warming induced hybrid rainy seasons in the Sahel. *Environmental Research Letters*, *11*(10), 104008. https://doi.org/10.1088/1748-9326/11/10/104008

Sanoussi, A., Mohamed, L., Seyni, S., & David, A. G. (2015). Adapting to climate variability and change in smallholder farming communities: A case study from Burkina Faso, Chad and Niger. *Journal of Agricultural Extension and Rural Development*, *7*(1), 16–27. https://doi.org/10.5897/jaerd14.0595

Savenije, H. H. G. (1999). *Dealing with Water Scarcity*. 1999. https://www.infona.pl/resource/bwmeta1.element.elsevier-a0d14a42-25d9-3dfe-96e6-176090676e38

Savenije, H. H. G. (2000). Water scarcity indicators; the deception of the numbers. *Physics and Chemistry of the Earth, Part B: Hydrology, Oceans and Atmosphere*, *25*(3), 199–204. https://doi.org/10.1016/S1464-1909(00)00004-6

Searcy, J. K., Hardison, C. H., & Langbein, W. B. (1960). Double-mass curves, with a section fitting curves to cyclic data. In *Water Supply Paper*. https://doi.org/10.3133/wsp1541B

Siddharth Misra, H. L. (2020). Machine Learning for Subsurface Characterization. In *Machine Learning for Subsurface Characterization*. Elsevier. https://doi.org/10.1016/c2018-0-01926-x

Sit, M., Demiray, B. Z., Xiang, Z., Ewing, G. J., Sermet, Y., & Demir, I. (2020). A Comprehensive Review of Deep Learning Applications in Hydrology and Water Resources. *ArXiv*. https://doi.org/10.2166/wst.2020.369

Smith, M. B., Koren, V. I., Zhang, Z., Reed, S. M., Pan, J. J., & Moreda, F. (2004). Runoff response to spatial variability in precipitation: An analysis of observed data. *Journal of Hydrology*, *298*(1–4), 267–286. https://doi.org/10.1016/j.jhydrol.2004.03.039

Sultan, B., Lejeune, Q., Menke, I., Maskell, G., Lee, K., Noblet, M., Sy, I., & Roudier, P. (2020). Current needs for climate services in West Africa: Results from two stakeholder surveys. *Climate Services*, *18*, 100166. https://doi.org/10.1016/j.cliser.2020.100166

Tan, J., & Huffman, G. J. (2019). *Morphing Computing Morphing Vectors for Version 06 IMERG*.

Tan, J., Petersen, W. A., & Tokay, A. (2016). A novel approach to identify sources of errors in IMERG for GPM ground validation. *Journal of Hydrometeorology*, *17*(9), 2477–2491. https://doi.org/10.1175/JHM-D-16-0079.1

Tappan, G. G., Cushing, W.M., Cotillon, S.E., Mathis, M.L., Hutchinson, J.A., Herrmann, S.M., and Dalsted, K. J. (2016). *West Africa Land Use Land Cover Time Series: U.S. Geological Survey data*. U.S. Geological Survey Data Release. https://doi.org/http://dx.doi.org/10.5066/F73N21JF

Usman, M. T., & Reason, C. J. C. (2004). Dry spell frequencies and their variability over southern Africa. *Climate Research*, *26*(3), 199–211. https://doi.org/10.3354/cr026199

Vahid Mohammadi, S. M. (2019). Engineering Tools in the Beverage Industry. In *Engineering Tools in the Beverage Industry*. Elsevier. https://doi.org/10.1016/c2017-0-02377-7

Van Rossum, G. & Drake, F. L. (2009). *Python 3 Reference Manual*. Scotts Valley CA: CreateSpace.

Verdin, K. L. (2017). *Hydrologic Derivatives for Modeling and Applications (HDMA) database: U.S. Geological Survey data release*. https://doi.org/https://doi.org/10.5066/F7S180ZP.

Vogel, P., Knippertz, P., Fink, A. H., Schlueter, A., & Gneiting, T. (2018). Skill of global raw and postprocessed ensemble predictions of rainfall over Northern Tropical Africa. *Weather and Forecasting*, *33*(2), 369–388. https://doi.org/10.1175/WAF-D-17-0127.1

Wilheit, T. T. (1986). Some comments on passive microwave measurement of rain. *Bulletin - American Meteorological Society*, *67*(10), 1226–1232. https://doi.org/10.1175/1520-0477(1986)067<1226:SCOPMM>2.0.CO;2

WWAP. (2000). *Water in a Changing World*. https://doi.org/10.1142/9781848160682_0002

Xiao, C., Chen, N., Hu, C., Wang, K., Xu, Z., Cai, Y., Xu, L., Chen, Z., & Gong, J. (2019). A spatiotemporal deep learning model for sea surface temperature field prediction using time-series satellite data. *Environmental Modelling and Software*, *120*, 104502. https://doi.org/10.1016/j.envsoft.2019.104502

Xie, P., & Xiong, A. Y. (2011). A conceptual model for constructing high-resolution gauge-satellite merged precipitation analyses. *Journal of Geophysical Research Atmospheres*, *116*(21). https://doi.org/10.1029/2011JD016118

Yang, T., Sun, F., Gentine, P., Liu, W., Wang, H., Yin, J., Du, M., & Liu, C. (2019). Evaluation and machine learning improvement of global hydrological model-based flood simulations. *Environmental Research Letters*, *14*(11), 114027. https://doi.org/10.1088/1748-9326/ab4d5e

Yaro, J. A. (2013). The perception of and adaptation to climate variability/change in Ghana by small-scale and commercial farmers. *Regional Environmental Change*, *13*(6), 1259–1272. https://doi.org/10.1007/s10113-013-0443-5

Appendix I

In this Appendix the technical characteristics of the DL models are presented.

For the main CNN classifier architecture examined, the input images are first manipulated by three convolutional blocks, consisting of two convolutional layers featuring zero padding and one average pooling or max pooling layer each. This constitutes the feature extraction part of the CNN which derives spatial information from the image. In a next step, the output of the feature extraction part is flattened, meaning that it is converted into a 1-D array, and the data is then processed by two dense and two dropout layers. The output layer features a sigmoid activation function mapping the output in [0, 1]. Using a decision threshold this sigmoid output is then used to perform the binary classification.

CNNs are most commonly used to analyze visual imagery and image classification and are the optimal choice for processing 2D-images. The input to the CNNs features 32 pixels by 32 pixels images in three different channels (32, 32, 3). The CNN architecture tested consists of three convolutional blocks (2 convolutional layers and 1 max pooling layer for each block) with zero-padding and ReLU as activation function with 'he_normal' as kernel initializer (Kaiming et al., 2015). The fully-connected layer flattens the output of the last max pooling layer. Next, the classifier consists of two dropout layers, a dense layer with 16 neurons and a ‘tanh’ activation function and an output dense layer with one neuron activated by a sigmoid function. As expected, the model features 26.409 trainable parameters, way less than the

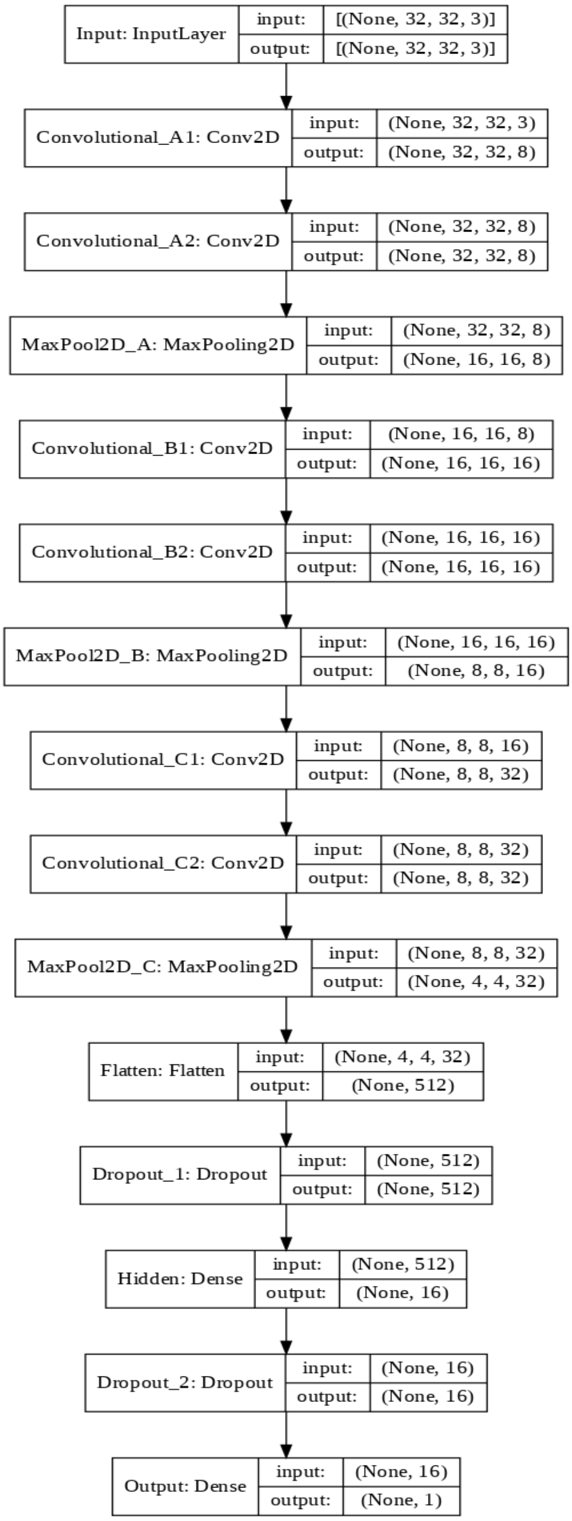


Figure A1: Architecture of the proposed CNN classifier. The 3 convolutional blocks (2 convolutional layers and 1 max pooling layer each), the flattening layer and the classifier part of the CNN can be distinguished.

proposed FNN. An overview of the CNN architecture is depicted in Figure A1.

Other features were also be tested as part of the convolutional blocks, such as different kernel sizes, number of filters, same or valid zero-padding, batch normalization and average pooling instead of max pooling.

Also, a FNN is tested for the classification task. The number and type of hidden layers, as well as the overall architecture of the FNN, is defined after trial-and-error processes and is a subject closely related to the spatiotemporal characteristics of the data used. The potential of a Feed-Forward Neural Network is being examined, which consist of a 1024-node input layer (when dealing with 32x32 pixels images), three dense layers with 512, 128 and 16 nodes each and with ReLU as activation function and a one-node output layer with a sigmoid activation function, bounding the output between zero (dry day) and one (wet day). The overall architecture of the suggested FNN is presented in Figure A2.

Training this FNN will be computationally-heavy since it features 592.545 trainable parameters. However, if using multiple pixels around the gauge, it is really likely that the ANN will feature convolutional layers.

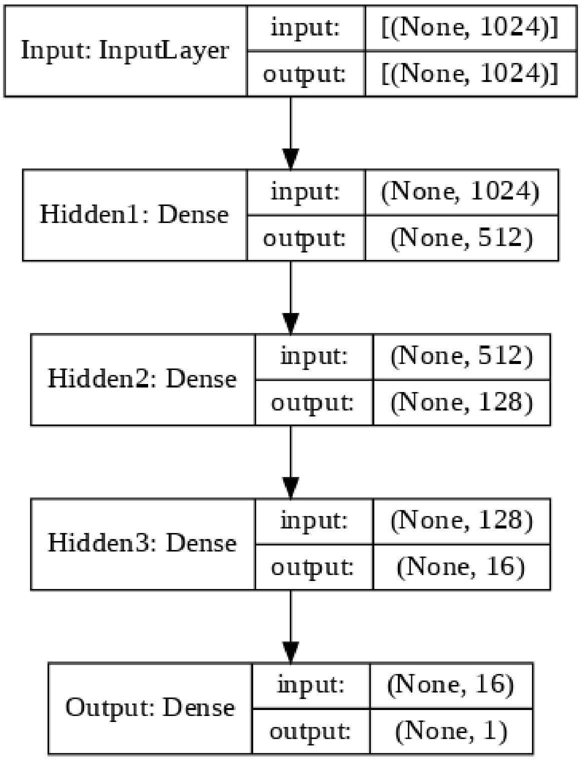


Figure A2: Architecture of the proposed classifier FNN.

Appendix II

Apart from the classifier ANNs built from scratch, several famous state-of-art classifiers are going to be tested for this task. These models were trained to perform multi-class classification (approximately 1.000 different classes) using the ImageNet dataset. ImageNet is a dataset of over 15 million labeled high-resolution images belonging to roughly 22,000 categories. The images were collected from the web and labeled by human labelers using Amazon's Mechanical Turk crowd-sourcing tool.

ImageNet consists of variable-resolution images. However, the input of the models is 224 by 224 pixels, using three channels (224, 224, 3). Hence, the satellite images are resized to target height and width.

For each model, the already trained parameters of the feature extraction part were frozen. Simultaneously, the classifier was discarded and a binary classifier was put in place. Then, the model was trained for the satellite images, with only the trainable parameters of the new classifier being updated.

The MobileNetV2 architecture is based on an inverted residual structure where the input and output of the residual block are thin bottleneck layers opposite to traditional residual models which use expanded representations in the input an MobileNetV2 uses lightweight depthwise convolutions to filter features in the intermediate expansion layer. The whole network has 3.538.984 trainable parameters. The architecture of MobileNetV2 contains the initial fully convolution layer with 32 filters, followed by 19 residual bottleneck layers.

Resnet50 features 23.587.712 trainable parameters. The network can take the input image having input size of 224 x 224 x 3. ResNet50 architecture performs the convolution and max-pooling using different kernel sizes in four different stages.

InceptionV3 features 21.802.784 trainable parameters and is made up of symmetric and asymmetric building blocks, including convolutions, average pooling, max pooling, dropouts, and fully connected layers. Batch-normalization is used extensively throughout the model and applied to activation inputs, while loss is computed via the softmax function. This model also incorporated the property of label smoothing, a type of regularizing component added to the loss formula that prevents the network from becoming too confident about a class.

The performance of the three fine-tuned state-of-art classifiers on the total dataframe was assessed. Overall, their performance was not exceptional. This can be due to the fact that they were initially built to perform multiclass classification in RGB images and not binary classification in satellite images. The best performing one was MobileNetV2 with performance close to the CNN that we built from scratch. On the contrary, the ResNet50 model only reproduced dry days. The performance of the three models can be seen analytically in the table below:

Table B1: Performance of state-of-art classifiers on the total dataframe

	MobileNetV2	ResNet50	InceptionV3
Total params:	2,340,033	23,718,913	21,933,985
Trainable params:	82,049	131,201	131,201
Accuracy:	0.75	0.70	0.73
F1-score: Wet days	0.51	0.00	0.50
F1-score: Dry days	0.82	0.85	0.80