ISEIS

Journal of Environmental Informatics 37(1) 1-15 (2021)

Journal of Environmental Informatics

www.iseis.org/jei

# Using Satellite Remote Sensing and Machine Learning Techniques towards Precipitation Prediction and Vegetation Classification

D. Stampoulis[1 *], H. G. Damavandi[1], D. Boscovic[2], and J. Sabo[1]

[1] *Future H2O, Office of Knowledge Enterprise Development, Arizona State University, Tempe, AZ 85281, USA*
[2] *Center for Assured and Scalable Data Engineering, Arizona State University, Tempe, AZ 85281, USA*

Received 07 Sep 2018; revised 26 Mar 2019; accepted 10 Apr 2019; published online 25 Jan 2020

**ABSTRACT.** The spatial distribution, magnitude and timing of precipitation events are being altered globally, often leading to extreme hydrologic conditions with serious implications to ecosystem services, water, food and energy security, as well as the welfare of billions of people. Motivated by the pressing need to understand, from a hydro-ecological perspective, how the dynamic nature of the hydrologic cycle will impact the environment in water-stressed regions, we implemented a novel approach that predicts precipitation spatio-temporal trends over the drought-burdened region of East Africa, based on other major hydrological components, such as vegetation water content (VWC), soil moisture (SM) and surface temperature (ST). The spatial patterns and characteristics of the inter-relations among the four aforementioned hydrologic variables were investigated over regions of East Africa characterized by different vegetation types and for various precipitation intensity rates during 2003 ~ 2011. To this end, we analyzed multi-year satellite microwave remote sensing observations of SM, ST, and VWC (derived from Naval Research Laboratory's WindSat radiometer) as well as their response to precipitation patterns (derived from NASA's TRMM 3B42 V7). We categorized precipitation into four bins (ranges) of intensity and trained five different state-of-the-art machine learning models for each of these categories. The models were then applied to predict the spatio-temporal precipitation dynamics over this complex region. Specifically, the Random Forest and Linear Regression models outperformed the others with the normalized mean absolute error being less than 27% for all of the categories. The characteristics of the predicted precipitation were in turn used to classify vegetation regimes in East Africa. Our results indicate significant discrepancies in the performance of the models with varying values in the predicting skill as well as their ability to accurately classify vegetation into different types. Our predictive models were able to forecast the three vegetation regimes, i.e., Forest/Woody Savanna, Savanna/Grasslands and Shrubland, with precision rate of at least 81% for all of the aforementioned precipitation bins.

*Keywords:* machine learning, linear regression, passive microwave remote sensing, precipitation, random forest, soil moisture, surface temperature, vegetation water content

## 1. Introduction

Water is the key environmental parameter that provides an inter-connectedness among ecosystem components. Varying precipitation (P) patterns, vegetation, and soil moisture (SM) dynamics, are the linkages in a natural environment that determine the complexity in such systems (Dunbar et al., 2001; Porporato et al., 2002; Asbjornsen et al., 2011). In regions where water is constantly or seasonally limited, its spatial and temporal distribution determines the phenology and sustainability of vegetation regimes (McVicar et al., 2012) and thus influences the regional hydro-climatology and biotic composition (Wolff et al., 2011; D'Odorico et al., 2012; Grimm et al., 2013; Fisher et al., 2014). Any variations in the water's availability or timing can significantly impact ecological processes and services, food, water, and energy security, economic prosperity, and even create tension between riparian countries that have different water needs at different times of the year.

Climate variability and the associated intensification of the hydrologic cycle are altering the spatial distribution, magnitude and timing of precipitation events, often leading to extreme hydrologic conditions, such as droughts or floods. Several studies (Boko et al., 2007; Christensen et al., 2007; Müller et al., 2011; Faramarzi et al., 2013; Müller et al., 2014) show that regions where water is constantly or seasonally limited, will be disproportionally impacted in future climates, jeopardizing crop and livestock production, fish stocks and fisheries. Moreover, the increasing water demands due to the rising population further exacerbate the problem, by fundamentally changing the water supply in many regions across the globe, with severe implication to natural habitats and the welfare of billions of people. Accurately assessing and quantifying P dynamics at the global scale is therefore highly critical. However, this need becomes imperative in regions that are topographically and climatologically very diverse, and which are usually characterized by a signif-

* Corresponding author. Tel.: +1 480 9653312.
*E-mail address:* dstampou@asu.edu (D. Stampoulis).

ISSN: 1726-2135 print/1684-8799 online
 doi:10.3808/jei.202000427

icant paucity of in-situ P data. Water-stressed regions or countries typically have limited resources for mitigation and adaptation, while their economies often depend primarily on rain-fed agricultural systems. One such geographical domain is the region of East Africa. The only way to measure P over this topographically complex domain is via remote sensing from space. However, despite its advantages, satellite-derived P has its own limitations. Specifically, satellite P products are characterized by limited temporal coverage (the average life span of a satellite ranges from 5 to 10 years). It becomes therefore evident that alternative methods of precipitation estimation are of paramount importance, as they can complement or enhance existing precipitation estimation techniques (satellite, airborne or ground-based i.e., radar-derived or in-situ).

Accurate prediction of P has always been crucial in hydrological research, since it plays a key role in weather forecasting which could also serve as a promising tool to determine the early warnings of severe weather events. Forecasting P is a complex process since it depends on other factors such as surface temperature (ST), humidity and pressure which are highly time-dependent and vary in space. Thus, the need to build a robust mathematical model to ensure the accurate prediction of P is of vital importance.

A joint initiative between the National Severe Storms Laboratory (NSSL) of National Oceanic Atmospheric Administration (NOAA), Aviation Weather Research Program of the Federal Aviation Administration, the Salt River Project, and the National Weather Service (NWS) Office of Hydrology Development led to the National Mosaic and Multi-sensor QPE (Quantitative Precipitation Estimation) project, or NMQ, offering a real-time quantitative precipitation estimation. In addition to NMQ, a variety of statistical methods have been explored in the literature to forecast the precipitation rates such as the ARIMA (Autoregressive Integrating Moving Average) model, Generalized Linear Model (GLM), Markov model, gray theory-based prediction model, among others. Geetha et al. (2015) explored the capacity of regressive integrated moving average (ARIMA) model to predict the rainfall of a coastal region in India for 2009 ~ 2013 with the potential predictors of temperature, dew point, wind speed, maximum temperature, minimum temperature and visibility. Graham et al. (2017) described the Box-Jenkins time series seasonal ARIMA approach for prediction of rainfall on monthly scales in the district of Allahabad with temporal coverage of 1985 ~ 2015. Stern et al. (1983) fitted the non-stationary Markov chains to the occurrence of rain, and gamma distributions with parameters which vary with the time of year to the rainfall amounts towards an effort to develop robust numerical methods for rainfall prediction. Chandler et al. (2002) illustrated the use of generalized linear models (GLMs) to test the changes in the rainfall pattern of South Galway region of western Ireland. Mangaraj et al. (2012) fitted a 2-state Markov chain probability model to the collected daily rainfall data in Orissa state of India towards an attempt to study the pattern of rainfall occurrence. Ingsrisawang et al. (2010) employed the use of three statistical methods, i.e., First-order Markov Chain, Logistic model, and Generalized Estimating Equation (GEE) in modeling the rainfall prediction over the eastern part of Thailand for Meteor and GPCM datasets, obtained from Thai Meteorological Department (TMD) and Bureau of the Royal Rain Making and Agricultural Aviation (BRRAA), with temporal coverage of 2004 ~ 2008. Moreover, Ho et al. (2015) developed an effective flood forecasting system for midsize rural watersheds where grey rainfall forecasting technique was adopted based on existing hourly rainfall data. These studies have yielded promising results in accurately forecasting precipitation. However, certain deficiencies exist, which call for further scientific scrutiny. The prediction error of extremum is larger in the ARIMA model, the GLM model assumes a particular base relation between target observations and the predictors, and the Markov model and gray model-based forecasting model are primarily suitable for the exponential growth of precipitation rates (Nelder et al., 1972; Du et al., 2018). Machine learning models have recently emerged as powerful fast big data processing unit to overcome these shortcomings, while being equipped with advanced optimization modules.

Limited studies have proposed novel methodologies to classify the vegetation types. Beon et.al. (2017) proposed a modified approach to map vegetation in Saemangeum - an estuarine tidal flat on the coast of the Yellow Sea in South Korea using multi-temporal downscaled images. This study used co-kriging methods to downscale Landsat imagery to the resolution of a RapidEye image, providing an effective method for creating an accurate vegetation map, which is essential for monitoring and managing the ecosystem of the reclaimed Saemangeum area. Additionally, Xu et.al. (2007) proposed a three-step method, combining vegetation and environmental factors and the feature extracted from remote sensing images, to classify vegetation types in Beijing suburb area. Gilmore et.al. (2008) examined the effectiveness of using multi-temporal satellite imagery, field spectral data, and LiDAR top of canopy data to classify and map the common plant communities of the Ragged Rock Creek marsh, located near the mouth of the Connecticut River.

In this work, on the other hand, we examine the potential of fully automated data-driven methods backed by advanced optimization techniques to disentangle the underlying interrelations of the investigated hydrological and hydro-meteorological components to forecast the precipitation rate and, ultimately, classify the vegetation regime. Note that the automation characteristics and the fast-processing knowledge discovery of the data driven models, and machine learning methods in particular, lends itself to miscellaneous applications as the inherent interrelations of the data points are merely learnt by the optimization, without any manual or external human effort. Additionally, the statistical strategies require an assumption to set the base model type (for example the order of non-linearity relation of predictors and the target). Contrarily, the advanced nonparametric machine learning models such as the Random Forest can unravel any mapping function through information inference of the data, without any prior assumption over the data, and thus lead to an efficient alternative to the statistical methods. Moreover, the traditional statistical models are limited in the number of input variables that can be effectively combined for precipitation prediction. Advances in machine learning has now cir-

cumvented this issue as the number of input variables can be arbitrary large (Mitchell et al., 1997).

Hydrologist often refer to "calibration" as an inevitable procedure to tune the model's parameters so that the simulated flow resembles the observed flow data (Singh et al., 2002). This is accomplished by properly adjusting the parameters involved. While, this technique would necessarily improve the reliability of the model, certain challenges associated with calibration have been reported in literature (Sorooshian et al., 1983; Beven et al., 2001, 2006; Madsen et al., 2003); limitations such as the physical distortion caused by an incorrect parameter tuning, and of course, the tedious computational time. We could associate the training phase of the learning models to the calibration stage of the conventional hydrological modeling. In the training stage, the parameters of the machine are incrementally optimized to extract the complex relation of the input and output with an end goal of a better prediction in testing phase. To this end, we provide the machine with a huge bulk of miscellaneous independent input data. The training stage, benefits from advanced optimizers (such as Stochastic Gradient Descent) to uncover physical underpinnings while evading the time-costly computations, leading to a nimble and optimized learning module. The optimal hyper-parameters of the best model, trained in the training phase, would then be imparted to the testing phase to evaluate the performance of the trained machine.

Therefore, an appropriate design of parallel-distributed learning framework not only accelerates the information inference, and hence, outperforms the conventional hydrological or statistical methods in terms of computational cost, but also promises to ease the problem of modeling while the model is solely learnt through data, without any prior assumptions.

Several studies have focused on the implementation of machine learning techniques in P prediction. Sumi et al. (2012) studied the data-driven machine learning method to predict the P level for daily and monthly rainfall of the Fukuoka city in Japan where a hybrid multi-modal method using the Artificial Neural Network (ANN), the K-nearest Neighbor (KNN), Multivariate adaptive regression splines (MARS) and Support vector Regression (SVR) and the Root-Mean-Square-Error (RMSE) of the prediction is presented. Khan et al. (2006) examined the potential of Support Vector Regression (SVR) and Multilayer Linear Perception (MLP) in predicting lake water levels. Specifically, water level data for Lake Erie from 1918 to 2001 were used in training the two models to predict this property for the following 12 months; although the evaluation results (using root-mean-square and correlation coefficient) were promising in both cases, the intra-model comparison showed that the MLP outperformed SVR. Kenabatho et.al. (2015) explored the artificial neural network (ANN) and Multiplicative Autoregressive Integrated Moving Average (MARIMA) models to predict the rainfall in Botswana. Hybrid models such as the combination of wavelet transform and artificial neural network (WANN) has been proposed by Kim et.al. (2003) where predictive models to predict the Conchos River Basin were proposed and evaluated. In addition, Karran et al. (2014) compared the use of four different models, i.e., ANN, SVR, wavelet-ANN, and wavelet-SVR in a Mediterranean, Oceanic, and Hemiboreal watershed.

Despite the increasing number of studies focusing on the use of machine learning models, little has been done regarding the utilization of microwave remote sensing observations of major hydrologic components in a natural system in the aim of predicting P patterns. SM is the "core of the hydrological cycle" (Noy-Meir, 1973; Eagleson, 1978; Federer, 1979; Eagleson, 1982), and is characterized by a "cause and consequence" relationship with the regional vegetation (Rodriguez-Iturbe, 2000). As one of the most important and dynamic variables in water-stressed regions, SM impacts land surface energy flux, the interaction of land surface with the atmosphere, and a suite of hydrological processes. Moreover, the interactions between SM and vegetation water content (VWC), especially in arid or semi-arid regions, are substantial, resulting in strong interdependence and significant feedbacks between SM dynamics and land-atmosphere water. Furthermore, P is the main climatic driver of vegetation dynamics (Stampoulis et al. 2014; Stampoulis et al., 2016), while (ST) is another environmental variable that plays a critical role in the aforementioned feedbacks and interactions. Therefore, jointly examining the dynamics of P, VWC, SM, and ST can provide a more holistic and integrated characterization of the regional hydrologic regime.

The current study jointly uses daily passive microwave remote sensing observations of the above variables, i.e., VWC, SM and ST for the 2003 ~ 2011 period, provided by WindSat, a satellite-based polarimetric microwave radiometer, to predict the spatio-temporal patterns of P in the highly complex water-stressed region of East Africa. In this work daily passive microwave remote sensing observations of P, derived from NASA's Tropical Rainfall Measuring Mission (TRMM) were also used for the purpose of training the various machine learning models. We perform various analyses such as accuracy, precision, F-1 score, Receiver Operating Characteristics (ROC) and the area under this curve to evaluate the performance of our vegetation classifier. A comprehensive introduction on these parameters can be found at Fawcett (2006). The primary objectives of this study are to: 1) investigate the efficiency and skill of various machine learning models in predicting P trends over East Africa, and 2) classification of different vegetation types based on the predicted precipitation level. To the best of our knowledge this is the first eco-hydrological study that jointly uses multi-year daily microwave remote sensing observations of VWC, SM and ST to predict the highly complex P trends in the significantly diverse region of East Africa using state-of-the-art machine learning models. This study also offers significant insight into a new vegetation classification method based on precipitation intensity levels.

## 2. Study Region

### 2.1. Topography and Land Use

Equatorial East Africa (EA) (100N ~ 100S / 300E ~ 500E) is the study area which encompasses the countries of Ethiopia, Somalia, Kenya, Uganda, Tanzania, and southeastern Sudan. Defined by the Great Rift Valley, EA is characterized by landscapes with high relative relief in close proximity to the ocean (Pik, 2011, Figure 1a). Several orographic features of varying

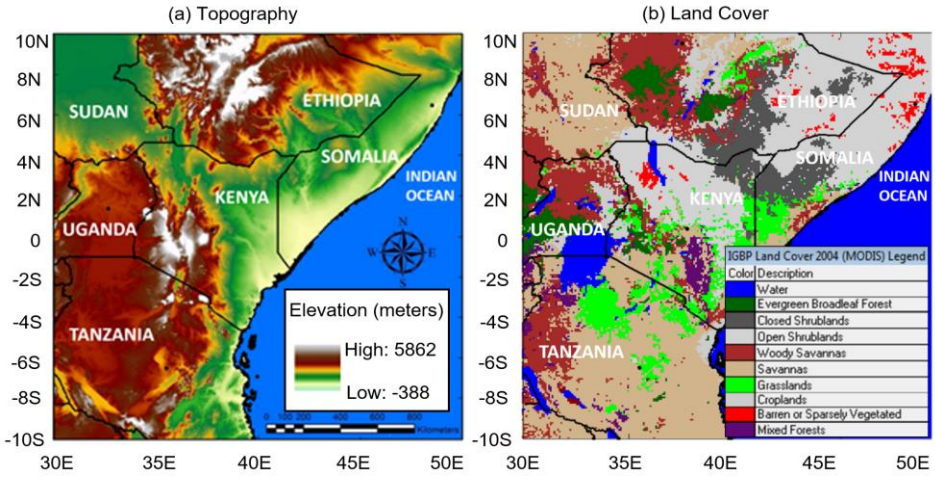


**Figure 1.** Maps of (a) topography and (b) land-cover categories in East Africa.

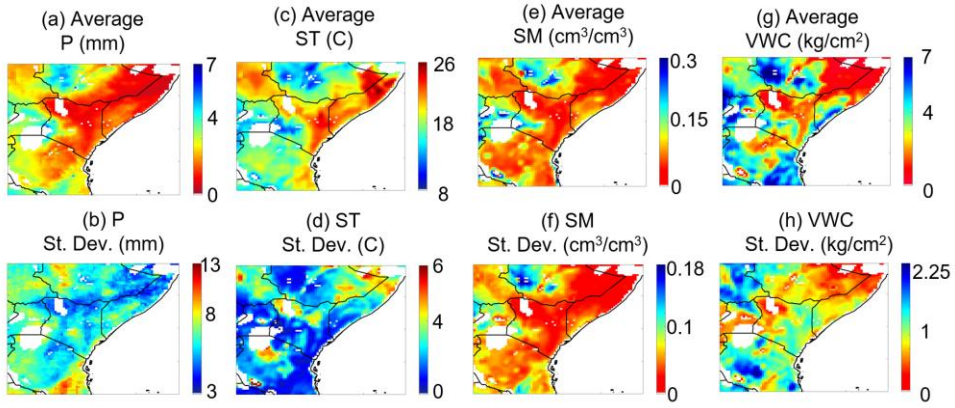


**Figure 2.** Maps of (left) average and (right) standard deviation values of (a ~ b) precipitation derived from TRMM 3B42, as well as (c ~ d) surface temperature (e ~ f) soil moisture, and (g ~ h) vegetation water content, derived from WindSat's respective algorithms over East Africa for 2003 ~ 2011.

sloping relief, large inland lakes, and widely spaced deserts or semi-arid sites make this region one of the most topographically diverse areas of the continent (Figure 1a), with an enormous effect on its climatology (Nicholson et al., 1990; Nicholson, 1996; Conway et al., 2005). Figure 1b shows the different vegetation categories of EA. For the most part, the study area is characterized by Forest, Woody Savanna, Savanna, Grasslands, and Shrublands. Savanna and Shrublands occupy 36 and 35% of the study area respectively, while woody savanna regions represent 10% of EA. Grasslands and forested regions account for 8 and 6% of the total area respectively. Other land-cover categories, such as croplands, mixed forest, and barren land occupy smaller regions that appear sporadically in the study area, and therefore only the five aforementioned categories were used for the classification of vegetation types.

For the purposes of this analysis, we grouped certain vegetation categories using several geographical and physiological factors. The entire northeastern region of EA (light- and dark-grey regions in Figure 1b) is represented by Shrublands. Moreover, Savanna and Grasslands are not only physiologically and phenotypically similar vegetation types (McPherson, 1997; Anderson et al., 2007) but they also appear sporadically in the same geographical region, i.e., central and northern Tanzania (Figure 1b). Therefore, these two vegetation types will be deemed as one category. Furthermore, areas characterized by Forest or Woody Savanna always appear in tandem; however, this vegetation regime is represented by two major geographically apart regions, i.e., dark-green and brown regions in central-western Uganda and western Ethiopia (Figure 1b). These two regions are jointly assessed, as spatiotemporal analyses of VWC and SM behavior over both regions were performed using WindSat VWC and SM observations and showed very similar responses.

All analyses for this study were conducted at the 1/4 degree spatial and daily temporal resolution, and only 0.25 deg pixels with homogeneous vegetation types were used for the classification method. Land-cover type homogeneity was determined by implementing a threshold value of 60% as the minimum number of the finer resolution land-cover pixels with common vegetation type within each 0.25 deg pixel.

### 2.2. Climatology

Although EA lies within the tropical latitudes, it exhibits a complex pattern of regional climatic proles (Nicholson, 1996), owing to the combination of large-scale tropical controls, such as the Intertropical Convergence Zone (ITCZ) that migrates bi-annually across the region (Nicholson, 1996; Wolff et al., 2011), the existence of various surface water bodies, high relative relief, and maritime influences (Nicholson, 2000; Verschuren et al., 2000). Because of the ITCZ, parts of the study area experience a bimodal P regime that brings rainy seasons from March to May (namely "long rains") and from October to December (namely "short rains") (Kabanda et al., 1999). The bimodal regime, however, changes gradually into a single season with increasing distance from the Equator (Conway et al., 2005). The major sources of moisture flux into the region are the monsoonal wind systems, the flow of which is significantly modified inland by the various topographical patterns (Ogallo, 1988), resulting in high spatial and temporal variations in P (Figure 2a, b). Similarly, ST in the region varies greatly in space; Somalia, eastern Kenya, southeastern Ethiopia, South Sudan, and parts of Tanzania are remarkably hotter than the rest of the study domain, while north-eastern/northwestern regions are subject to greater temporal T variations (Figure 2c). EA is also characterized by great hetero-geneity in VWC (Figure 2g) and to a lesser extent in SM (Figure 2e). Temporal variations of SM and VWC also change significantly in space, indicating that the region is extremely complex both topographically and climatologically (Figure 2f, h).

## 3. Data

### 3.1. Precipitation

The region of EA is characterized by a severe paucity of in-situ P data (Dinku et al., 2007), and thus, the only way to measure P over this topographically complex domain is via remote sensing from space. The satellite P product used in this study is derived from a joint mission between NASA and Japan Aerospace Exploration Agency (JAXA) and named Tropical Rainfall Measuring Mission (TRMM) Multi-satellite Precipitation Analysis (TMPA), specifically 3B42 V7, which is a gauge-adjusted (over land only) product (Huffman et al., 2007). This product is the combination of two sub-products, the microwave and the microwave-calibrated infrared. The final product has a relatively fine spatial (0.25 deg) and temporal resolution (3 hourly) and is available both as post-analysis (3B42 V7) where the 3-hourly passive microwave/infrared estimates are adjusted using monthly gauge comparisons, as well as in real time (3B42 RT) without the gauge correction. The P product used in this study is TMPA 3B42 V7 (covering all areas 500N ~ 500S for 1998 ~ 2014) at the daily scale and for the 2003 ~ 2011 period. Due to the sparseness of gauges in the study region, the estimated P is characterized by relatively high uncertainty (Tian et al., 2010; Dinku et al., 2010; Behrangi et al., 2014).

### 3.2. Vegetation Water Content, Soil Moisture, and Surface Temperature

WindSat was developed by the Naval Research Laboratory (NRL) primarily to provide the Navy with the much needed ocean surface wind vector measurements; however, it also measures other environmental parameters such as SM, ST, and VWC. Daily observations of VWC, volumetric SM, and ST were provided by the physically-based land algorithm of the NRL's Wind-Sat radiometer for 2003 ~ 2011. Its algorithms simultaneously retrieve VWC, SM, and ST using polarized 10.7-, 18.7-, and 37-GHz channel measurements (Li et al., 2010; Turk et al., 2014). The algorithm's approach is among the few multi-channel algorithms (Njoku et al., 1999, 2003; Owe et al., 2001, 2008) that add the 37-GHz channels. The Single Channel Algorithm (SCA) has also been using the 37 GHz channel to correct for factors that affect the retrieval (Mladenova et al., 2014). Sensitivity studies (Li et al., 2010; Turk et al., 2014) showed that the 37-GHz channels can offer significant SM sensitivities under low vegetation conditions. In another study, Parinussa, Holmes, and De Jeu (2012) derived surface SM from WindSat using C or X-band brightness temperature observations according to the Land Parameter Retrieval Model (LPRM) and validated the retrieved SM using in situ observations in Europe and Australia; WindSat SM retrieval was found to have a consistent response to changing environmental conditions, consistent temporal behavior, and the ability to capture the daily variation of SM.

WindSat automatically accommodates nonlinear transitions, such as that between significant SM sensitivity over desert to high ST sensitivity over vegetated land (Li et al., 2010). The WindSat land algorithm uses Sensor Data Records resampled to a global cylindrical Equal-Area Scalable Earth Grid (EASE-Grid) (Brodzik et al., 2002) of 25 km for further SDR data processing and land retrieval. The land algorithm bins the swath data onto the EASE-Grid and composes different orbits into separate daily ascending (evening passes) and descending (early morning passes) les. For this study, WindSat data were resampled via (nearest neighbor) interpolation to a regular 0.25 deg grid, and only descending passes were used, to ensure smaller retrieval errors, as the differences between effective land surface and vegetation temperatures are at the daily minimum.

## 4. Methodology

### 4.1. Data Splitting

There are two major precipitation types: stratiform and convective. The first type is characterized by low precipitation rates (usually long steady rain at low rates) while the latter is the typical summer storm (short duration and intense rainfall). Moreover, although there is no consensus on a fixed precipitation rate threshold that clearly defines the limit between these two types, several studies (e.g., Liu et al., 2013) consider the rate range of 4 ~ 6mm/hr as that threshold value. Therefore, we used the value of 5 mm/hr as our designated threshold. Further, and for the purposes of this study, we divided these two precipitation regimes ($P \leq 5$ and $P \geq 5$) in two subcategories for each one, as in "light- and heavy-stratiform" and "light- and heavy convective", to achieve a better representation of rainfall characterization over East Africa. As such, we have divided our precipitation data into four bins shown in Table 1 and for each bin we trained a predictive model.

### 4.1. Missing Values Interpolation

Due to a couple temporary WindSat instrument failures, the recorded values for VWC, SM and ST are missing for those time periods. For each of these missing values at some time *t*, we perform a linear interpolation between two adjacent non-missing values before and after *t*. Figure 3 shows the original vegetation time series for one cell as well as the interpolated version for a short time interval.

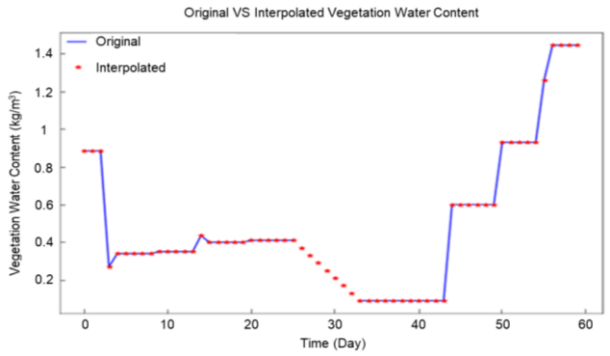


**Figure 3.** Original vegetation water content time series versus the interpolated version.

### 4.3. Prediction of Precipitation Rate Using Other Hydrological Components

There are indeed many environmental, as well as human-induced parameters affecting precipitation regionally or globally. However, this study aims at using only major hydrological (soil moisture, vegetation water content) and hydrometeorological (surface temperature) components to predict precipitation over the region of East Africa. All of the aforementioned variables are derived from satellite remote sensing techniques and therefore, although other parameters, such as topography, do affect the aforementioned correspondence, we do not account for non-hydrological or non-hydro-meteorological variables in predicting precipitation, as this would be beyond the scope of this study. Figure 4 shows the work flow of P prediction using the other strictly hydrological or hydrometeorological parameters. This process was carried out on a cell-by-cell basis Figure 5). Starting from the top left, we capture the VWC, SM, and ST as the features to predict the P rate of that particular cell. Henceforth, we split the P rates for all nine years (i.e., 3287 days) into a training and a testing session. The model is trained via training examples and then the optimal learning hyper-parameters are sent into the testing session to predict the testing examples. In this study, we have used 80% of the examples (i.e., 2629 days) for training the model and have tested the model against the remaining 658 examples. Each cell contains four time series for VWC, SM, ST and P for the under-study time interval i.e., 2003 ~ 2011, leading to 3287 data points (or days) per each variable. It is worth noting that we first shuffle and then split the time series into 80% for training and 20% for testing, leading to two disjoint and absolutely random sets. Therefore, the training and testing data points are randomly selected discrete data points. With this choice of data splitting, we reduce the effect of temporal autocorrelation between the data points, as they are chosen in an absolutely random and shuffled manner. In the end, the predicted results are cross-checked via a human expert. From a technical perspective, we developed a function *F*() that can map observations of VWC, SM, and ST into one precipitation rate with some error level denoted as *w*():

$$P = F(SM, VWC, ST) + w(t) \tag{1}$$

It is worth noting that the training and testing examples were used seperately and no training example would be used in the testing session. Since the number of days with low precipitation rate were considerably high, potentially biasing our learning model towards predicting very low precipitation rate for almost any combination of feature set, we categorized our dataset based on the precipitation rates into four levels shown in Table 1. Via this approach, we were able to build a model for each category independently and predict the precipitation rate with low error rates.

**Table 1.** Dividing the precipitation rates into four levels, indicating the light/heavy stratiform and convective precipitation patterns

| Precipitation Level (P) | Precipitation Pattern | Precipitation Level |
|---|---|---|
| $0 < P \leq 2$ | Light Stratiform | Level 0 ~ 2 |
| $2 < P \leq 5$ | Heavy Stratiform | Level 2 ~ 5 |
| $5 < P \leq 10$ | Light Convective | Level 5 ~ 10 |
| $P > 10$ | Heavy Convective | Level 10 |

### 4.4. Models

To identify the best predictive model that estimates the precipitation rate using the hydrological components, we trained five state-of-the-art machine learning models and evaluated their performances individually. We examined linear regression, nearest neighborhood regression, random forest, support vector regression and multilayer perception as the predictive model. In this section we briefly describe the mathematical derivations of these learning models.

#### 4.4.1. Nearest Neighborhood

This model is the simplest predictive model, as it predicts the observed value for one testing set of feature vector as a linear combination of the observed values for the nearest feature vectors in feature space. In this work, we have regressed a new sample as the mean of the nearest three examples in feature space, where our criterion was the distance in Euclidean space of feature vectors:

$$P(x^*) = \sum_{i=1}^{k} \alpha_i P(x_i)$$
$$x^* = [VWC^* \;\; SM^* \;\; ST^*] \tag{2}$$

where $x_i$ is the *i*-th nearest neighbor of *x* in feature domain and

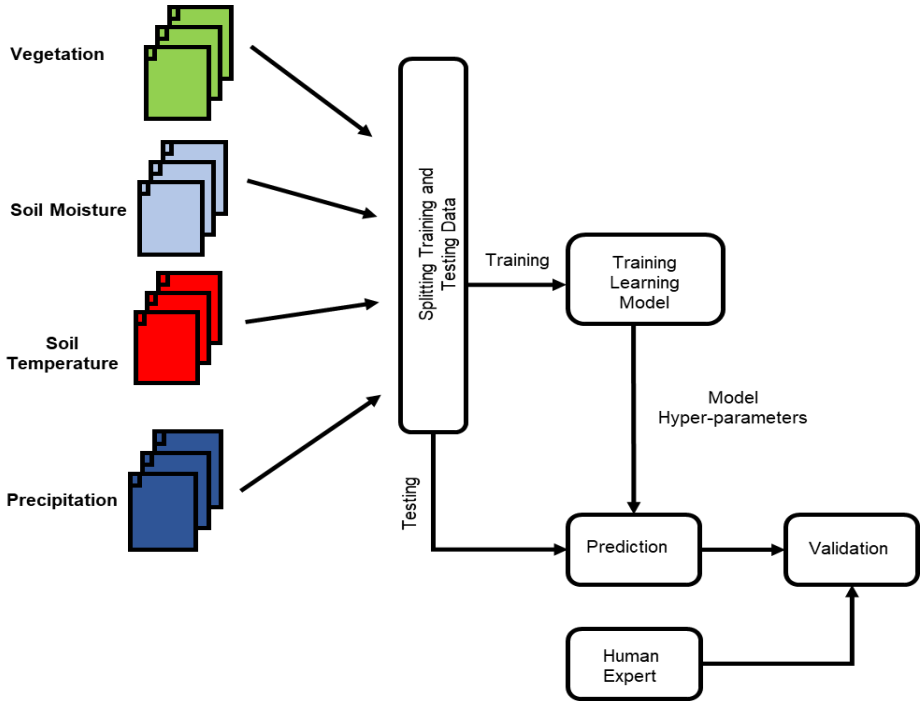


**Figure 4.** The block-diagram indicating the flow of learning methodology to predict the precipitation rate using vegetation water content, soil moisture and the soil temperature. Note that we split the white region into 80 ×80 cells and train a model for each cell.

$\alpha_i$ is determined by how close $x$ and $x_i$ are in feature space. In case $\alpha_i = 1/K$, then $P(x^*)$ is the average of the nearest neighbors.

4.4.2 Linear Regression

Linear regression seeks the linear relationship between the observed values disturbed by some noise level, $\varepsilon$, and the potential predictive variables. Mathematically, given the dataset of size $K$, $\{y_i; x_i\}_{i=1:k}$, linear regression searches for a matrix $W$, which maps the predictor variables into the observations:

$$Y = W^T X + \varepsilon$$
$$W^{*T} = \arg\min \| Y - W^T X \|^2 \quad (3)$$

where $W^{*T}$ is the linear relationship between $X$ and $Y$ with the minimum prediction error among all valid transformations of $W$. In case the noise level is negligible, the relation between the observation matrix ($Y$) and the matrix of predictors is given by:

$$W^T = YX^T (XX^T)^{-1} \quad (4)$$

Note that, this regression model is suitable to learn a simple but general model.

4.4.3. Random Forest

Random forest is an ensemble of learning model, constructed by a multitude of decision trees and the regressed value would be a linear combination (i.e., the mean or median) of the predicted values by each tree. Given the dataset of size $K$, we randomly extract $N$ ($< K$) examples, fit a tree to these training models, and the final predicted value would be the linear combination of the outcome values, out of these trees. This procedure would lead to a more robust model. A single tree would be highly sensitive to the noise level, but an ensemble of the trees and taking the average of them, would decrease the variance of the model, and hence a more robust model is built.

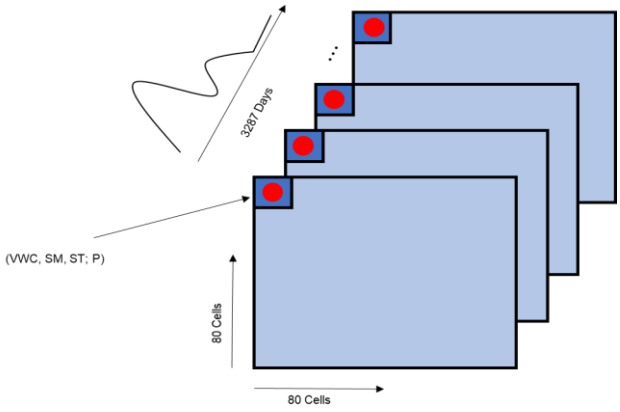


**Figure 5.** The whole region of interest (ROI) shown for the 3287 days corresponding to the nine years of data used in this work. We have split each day into 80 by 80 cells. For each cell, we have three features including vegetation water content (VWC), soil moisture (SM), soil temperature (ST) as well as

the corresponding precipitation value (P) for the 3287 recorded days. We use the 80% of the recorded data for each cell to train a model to predict the precipitation rate for the remaining days.

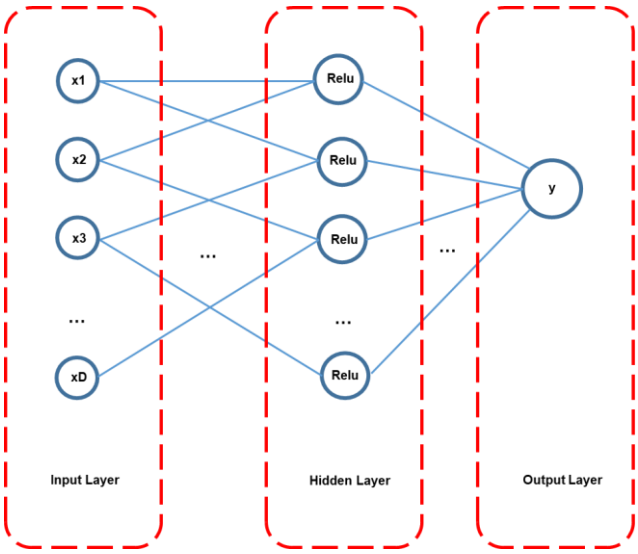


**Figure 6.** The schematic of a multilayer perception (neural network) which maps an input of size D into a single output *y*.

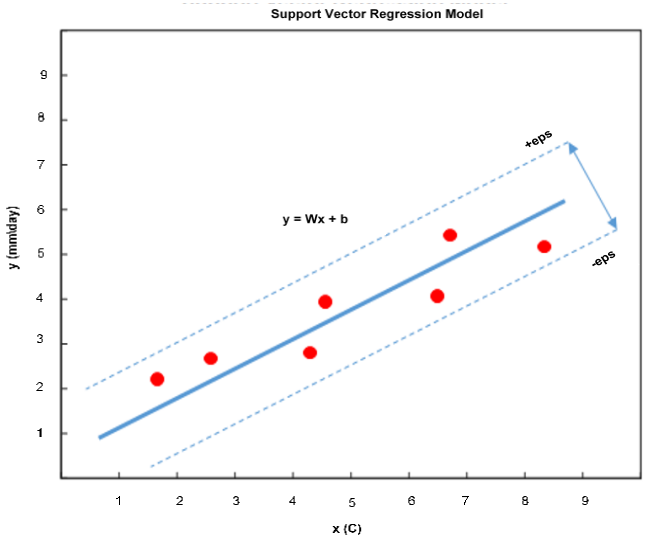


**Figure 7.** The schematic of the support vector regression model.

4.4.4. Multi-layer Perceptron (MLP)

A class of feed-forward neural networks consisting of three fully-connected layers or more, i.e., input, hidden and output layers (Figure 6). MLP is a $R^D \rightarrow R^L$ transformer, where $D$ and $L$ are the input and output sizes, respectively. It would learn a non-linear transformation function like $G$ to map the input into a space where they are linearly separable (classification mode) or they are regressed to a single value (regression mode). It also consists of an activation function which maps the weighted inputs into an output. In this work we have used *Relu* (*Relu* (*x*) = *max* (*0*, *x*)) as the activation function. We have adopted the Adam optimizer to update the weights in the network. Interested researchers are referred to (Bello et al., 2017).

4.4.5. Support Vector Regression

Support vector machine (SVM) can be used in regression mode, maintaining the variables searching for the maximal margin criterion. From a mathematical perspective, we developed a function $f(x)$, with at most having $\varepsilon$-deviation from the target $y$. Here we individualize the hyperplane which maximizes the margin (refer to Figure 7):

$$\min \frac{1}{2} \| W \|^2$$
$$y_i - Wx_i - b \le \varepsilon$$
$$Wx_i + b - y_i \le \varepsilon \quad (5)$$

**Table 2.** Comparison between Different Machine Learning Models in Terms of Normalized Mean Absolute Error (NMAE) and Standard Deviation of Error of the Predicted *P* Level

| Machine Learning Model | Level | NMAE | STD |
|---|---|---|---|
| Random Forest | 0 ~ 2 | 0.26 | 0.164 |
| | 2 ~ 5 | 0.157 | 0.093 |
| | 5 ~ 10 | 0.13 | 0.078 |
| | 10 | 0.154 | 0.14 |
| Support Vector Regression | 0 ~ 2 | 0.286 | 0.199 |
| | 2 ~ 5 | 0.172 | 0.114 |
| | 5 ~ 10 | 0.143 | 0.094 |
| | 10 | 0.142 | 0.152 |
| Linear Regression | 0 ~ 2 | 0.26 | 0.158 |
| | 2 ~ 5 | 0.155 | 0.091 |
| | 5 ~ 10 | 0.13 | 0.076 |
| | 10 | 0.15 | 0.131 |
| K-Nearest Neighborhood (K = 3) | 0 ~ 2 | 0.283 | 0.197 |
| | 2 ~ 5 | 0.169 | 0.11 |
| | 5 ~ 10 | 0.141 | 0.091 |
| | 10 | 0.169 | 0.157 |
| Multi-layer Perceptron (NN) | 0 ~ 2 | 0.699 | 0.265 |
| | 2 ~ 5 | 0.332 | 0.164 |
| | 5 ~ 10 | 0.259 | 0.147 |
| | 10 | 0.176 | 0.125 |

*Random forest and Linear Regression out-perform the other three machine learning models with the lower normalized mean absolute error for most of the precipitation levels.

## 5. Results

In order to evaluate the performance of each model's skill to predict the precipitation rate, we definne the normalized mean absolute error as:

$$NMAE = (\sum_{i=1}^{N} | \hat{P}^i - P^i |) / \max(P) \quad (6)$$

where $P$ and $\hat{P}$ are the actual and predicted precipitation rates with size $N$, respectively. Note that this criterion would sug-

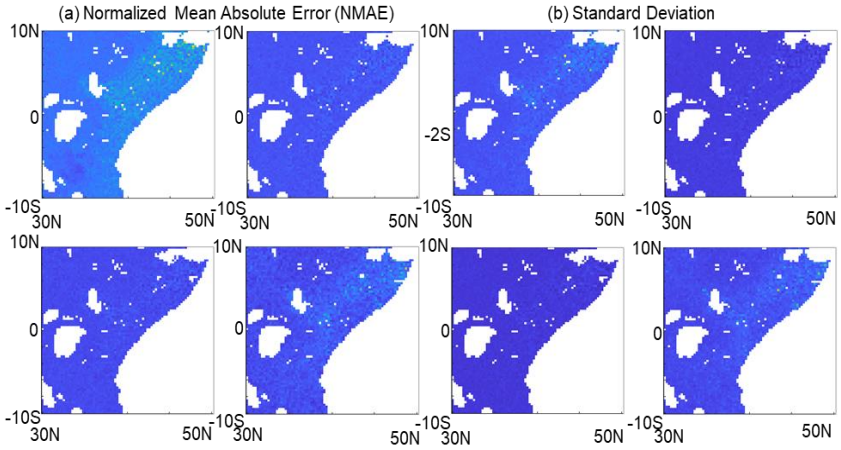


**Figure 8.** (a) Normalized mean absolute error (four panels on the left) and its (b) standard deviation surfaces (four panels on the right) in prediction of precipitation for the four precipitation levels.

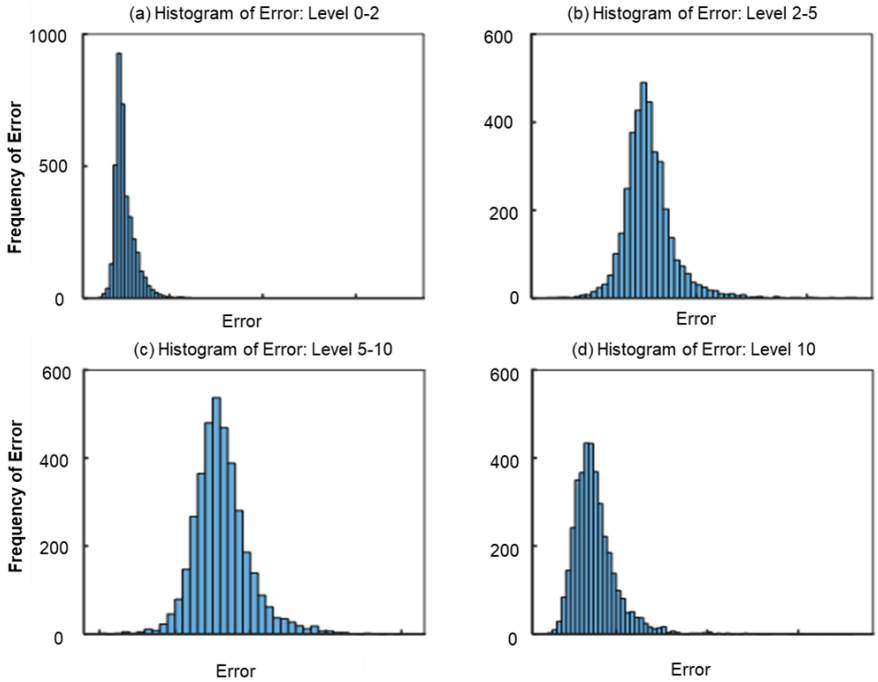


**Figure 9.** Histogram of normalized mean absolute error (NM-AE) of precipitation prediction using random forest model for four precipitation levels of (a) Level 0 ~ 2, (b) Level 2 ~ 5, (c) Level 5 ~ 10, and (d) Level 10. Note: In all cases, the error does not exceed 0.2.

gest the percentage of error with respect to the maximum value in the actual precipitation records. We compute this value for each cell and the average of these values would be treated as the performance of a model. The standard deviation of the errors for each cell, suggest how well the model has predicted the precipitation rate across different cells. Table 2 shows the normalized mean absolute error and the standard deviation of the predicted precipitation value for five different methods, Random Forest (using 100 trained decision trees), Support Vector Regression ($\varepsilon$ = 0.2), Multi-layer Perceptron Neural Network (with 32 hidden layers, Relu as the activation function and Adam method as the optimizer) and Linear Regression. Evidently, random forest and linear regression have outper-formed the other methods having the least NMAE for most of the precipitation levels. The NMAE surface as well as error standard deviation have been plotted in Figure 8. The maximum NMAE is around 1.5 in the first category, where this number is less than 0.6 in all other three categories. Water bodies in the figure are shown as white regions, as this study focuses on precipitation that occurs over land only. Figure 9 illustrates the histogram of error for the random forest model (one of the best predictive models in Table 2) for all four precipitation levels. Note that, in

all of the cases, the error does not exceed 0.2, indicating the high performance of the model. As can be seen in Figure 10, apart from the MLP, the other four models exhibit similar performance in terms of NMAE for all four categories, revealing the robustness of these learning models.

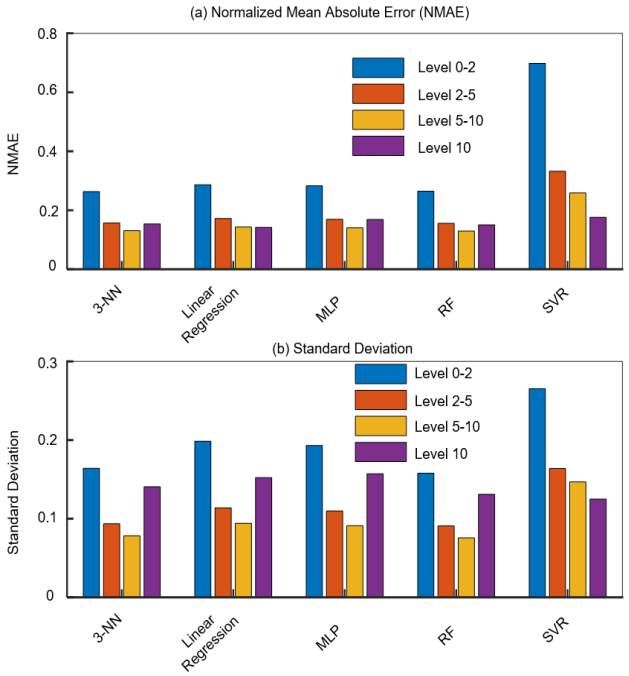


**Figure 10.** (a) Normalized mean absolute error and its (b) standard deviation for each of the four precipitation levels compared across different learning models.

## 6. Classification of Different Vegetation Types

In this section, we attempt to associate the P level with different vegetation types, i.e., Forest/Woody Savanna, Savanna/Grasslands and Shrublands. In Figure 11, we show the average of the P level for the regions containing these vegetation types in three different curves. We calculated the spatial average of the P level for each region and hence we have 3278 data points, one for each day in the nine-year period used in this study. As shown in Figure 11, on average, more P occurs over the Forest/Woody Savanna regions, while less is observed over Savanna/Grasslands. Shrublands observe the least P. In the previous section, we trained a model for each cell using 80% of the P data within each cell and predict the remaining 20% based on the trained model and reported the normalized MAE as well as its standard deviation. Hence, for each cell we have a time series containing actual 2629 days (80% of 3287 days) and 658 predicted data points. In this section, we train a random forest model to predict the vegetation type via taking the average of the P levels for each day but just using the actual 2629 days and evaluate either the remaining 658 can predict the vegetation type. We report the Receiver Operating Characteristics (ROC) and the area under its curve, accuracy, precision, recall and F-1 score as the classification performance parameters. In Table 3, we also calculated the most important feature in predicting the P level at the regions containing each of the vegetation group. Clearly, (ST) has been the most relevant feature to predict the P level in Savanna/Grasslands as well as Shrublands, where (SM) contributed more in the Forest/Woody Savanna regions. Figure 12 presents the most important feature to predict the P level across all cells in the region of interest. Evidently, the feature with the highest contribution is ST followed by SM and VWC. This pattern could be attributed to the fact that in water-limited regions there is strong interdependence and significant feedbacks between ST and surface moisture, resulting in substantial land-atmosphere water and energy fluxes. Furthermore, temperature is a very crucial parameter that affects rainfall distribution. It is noteworthy that under "temperature" there are other parameters included as well, such as wind (caused by temperature differences) which are considered as important predictors for P. Figure 13 shows the P map and the regions with the three vegetation types, i.e., Forest/Woody Savanna, Savanna/Grasslands and Shrublands.

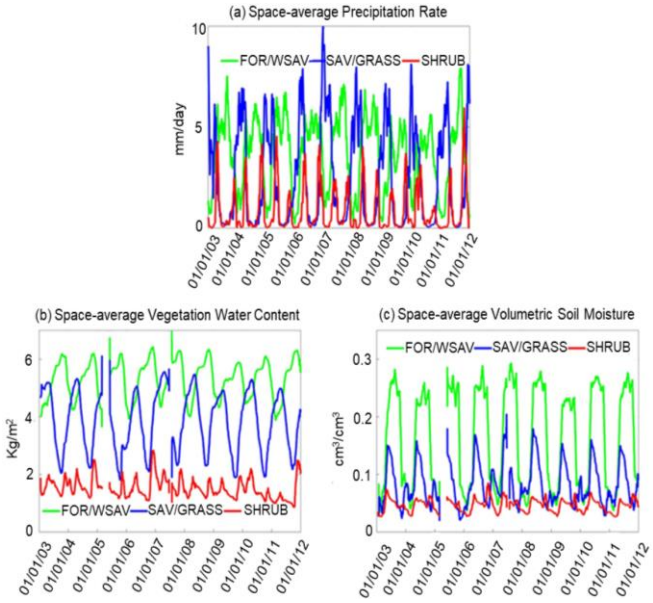


**Figure 11.** Time series of the smoothed spatially-averaged daily measurements of (a) P rate, (b) VWC and (c) SM for the three major land-cover categories of East Africa. Note: Gaps in the time series are due to temporary instrument failure.

**Table 3.** The most Important Feature in Precipitation Prediction over Each of the Vegetation Regimes

| Precipitation Level (P) | Savanna / Grasslands | Forest / Woody Savanna | Shrublands |
|---|---|---|---|
| Level 0-2 | SM | SM | ST |
| Level 2-5 | ST | VWC | ST |
| Level 5-10 | ST | ST | ST |
| Level 10 | ST | SM | ST |

*As shown, (ST) has been the most important feature to predict P over the Savanna/Grasslands as well as Shrublands whereas (SM) was characterized by the highest contribution in predicting P over Forest/Woody Savanna regions.

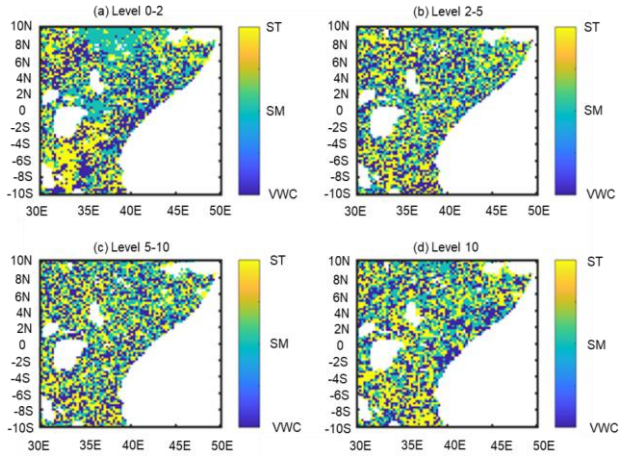


**Figure 12.** Feature importance plot for the precipitation prediction for the four precipitation levels of (a) Level 0 ~ 2, (b) Level 2 ~ 5, (c) Level 5 ~ 10, and (d) Level 10, for all 80 × 80 cells.

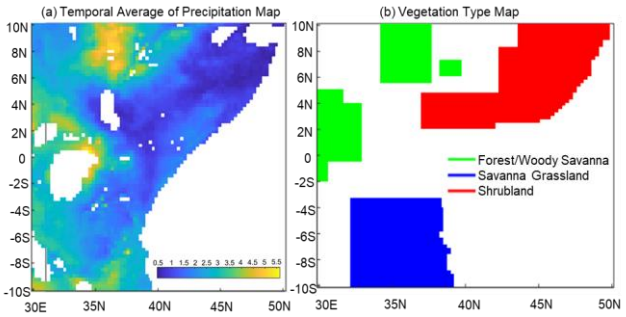


**Figure 13.** The figure indicating (a) the temporal average precipitation map and (b) the vegetation type map for the region of interest.

## 7. Receiver Operating Characteristics

A commonly used method to evaluate the performance of a classifier is the rate of true positive rate:

$$TPR = \frac{TP(True\ \ Positive)}{TP(True\ \ Positive)+FN(False\ \ Negative)} \quad (7)$$

as a function of false positive rate:

$$FPR = \frac{FP(False\ \ Positive)}{FP(False\ \ Positive)+TN(True\ \ Negative)} \quad (8)$$

for different cut-off thresholds of the classifier. Such a curve is called Receiver Operating Characteristics (ROC). Each point of the ROC curve represents a sensitivity and specificity for a threshold used by the classifier. As the classifier can reach the maximum true positive rate (= 1) at a lower false positive rate, we have trained a better model. Henceforth, the area under the ROC (AUC) serves as a telling representative for the performance of the classifier. The random classifier would have the same true and positive rate, and hence, has the AUC of half. We fine-tune the classifier's parameters to achieve a better classifier with the AUC of higher than 0.5. Figure 14 shows the ROC of the trained random forest model to predict the three types of vegetation regimes, Forest/Woody Savanna, Savanna/ Grasslands and Shrublands as well as the area under its curve (AUC). As can be seen, in all of the cases, our model has outperformed the random classifier (the dashed black line) in which the average AUC is above 0.62 within the four P levels.

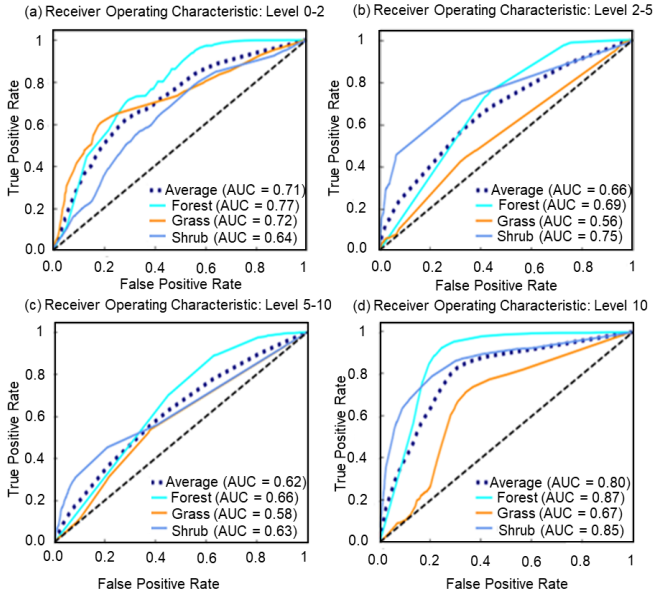


**Figure 14.** The Receiver Operating Characteristics (ROC) and the area under the curve (AUC) for the Random Forest model to predict the vegetation type for the four different P levels of (a) Level 0 ~ 2, (b) Level 2 ~ 5, (c) Level 5 ~ 10, and (d) Level 10 for the Forest/Woody Savanna, Savanna Grassland, Shrubland and the average curve. Note: For better illustration, we have used short format for the three types of vegetation as Forest, Grass and Shrub.

## 8. Conclusions

Inspired by recent advances in artificial intelligence, and machine learning strategies in particular, as a powerful tool to approximate the physical-based hydrological models, this study aims at evaluating the performance of the top state-of-the-art machine learning models to predict the precipitation rate using three potential hydrological predictors, i.e., vegetation water content (VWC), soil moisture (SM) and surface temperature (ST) over the region of East Africa. Although precipitation rates have been estimated using machine learning models in recent studies (Khan et al., 2006; Sumi et. al 2012; Kenabatho et al., 2015), this is the first time, to the best of our knowledge, the three aforementioned hydrological components are explored, within a learning framework, as the potential drivers of precipitation.

In this work, to enhance the prediction accuracy, the investigated variable was divided into four categories based on the

precipitation rates, and a learning model was trained for each category (light-heavy stratiform and light-heavy convective precipitation patterns, refer to Table 1). We reported the prediction performance as the normalized mean square error (NMAE defined in Equation 6) between the observed and predicted precipitation rates. The Random forest and Linear Regression models outperformed the others, achieving the minimum prediction NMAE for most of P levels. Our results present the surface temperature as the main element to forecast the precipitation rate, followed by soil moisture and vegetation water content. We contribute this to the strong correlation between soil moisture and surface temperature in water-limited regions. Such a strong correlation would then lead into water fluxes in the region. Moreover, temperature fluctuation will directly affect the Earth's water cycle, via impacts on evapotranspiration, and changes in the conditions for cloud formation, and will consequently alter precipitation patterns. As such, surface temperature plays a pivotal role in determining the precipitation rate.

All of the variables (VWC, SM, ST) used to predict precipitation in this study are derived from satellite remote sensing techniques and therefore, although topography does affect the correspondence between precipitation and the aforementioned parameters, its influence is, for the most part, partitioned into the effect of (among others as well) temperature, soil moisture and vegetation water content on precipitation. In other words, via the use of these three parameters, we inherently take into account regional geomorphologic characteristics. However, we note that accounting for non-hydrological or non-hydro-meteorological variables in predicting precipitation is beyond the scope of this study.

Additionally, we used our predicted precipitation rates to train a random forest model and classify the three vegetation regimes i.e., Forest/Woody Savanna, Savanna/Grasslands and Shrublands. Receiver Operating Characteristic (ROC), the area under its curve, accuracy, precision, recall as well as the F-1 score were reported (Table 4) to evaluate the performance of this random forest model.

**Table 4.** Area under the ROC Curve (AUC), Accuracy, Precision, Recall and F-1 Score of the Random Forest Classifier to Detect the Three Vegetation Regimes, Forest/Woody Savanna, Savanna/Grasslands and the Shrublands for the Four P levels

| Precipitation Level (P) | AUC | Accuracy | Precision | Recall | F-1 Score |
|---|---|---|---|---|---|
| Level 0 ~ 2 | 0.71 | 0.50 | 0.81 | 0.50 | 0.57 |
| Level 2 ~ 5 | 0.66 | 0.46 | 0.93 | 0.46 | 0.58 |
| Level 5 ~ 10 | 0.62 | 0.45 | 0.94 | 0.45 | 0.59 |
| Level 10 | 0.80 | 0.54 | 0.85 | 0.54 | 0.62 |

Our main premise is based on the fact that we associate unique precipitation trends and characteristics to different vegetation types. Indeed, in a specific geographic location where various vegetation types exist, there is a significant correlation between major precipitation attributes (e.g. total annual accumulation, average rainfall rate, timing of precipitation, etc.) and vegetation type. Clearly, the predicted precipitation is associated with uncertainty and therefore the same applies to the classification of vegetation types using predicted precipitation rates to train our models. However, in this study we showed that there is an immense potential for utilizing AI and machine learning models to explore the inherent relationships of major hydrological components and vegetation, without the need of utilizeing information on physical characteristics and properties. However, we acknowledge that using predicted precipitation inherently propagates uncertainty to our vegetation classification scheme.

The results in this work indicate the prominent capacity of the advanced artificial intelligence (AI) techniques, and machine learning models in particular, to unravel the inherent interrelationships of hydrological components, including but not limited to, vegetation water content, soil moisture, surface temperature and the precipitation rate via developing learning models, without an explicit knowledge of the underlying physical behaviors. These learning models, accompanied by the cutting-edge theories in optimization, have elevated the prediction accuracy since they are capable of acquiring knowledge based upon their prior experience, obtained by mining diverse data resources with dissimilar distributions. Such knowledge acquisition strategy not only hones the prediction skills of these learning models, as we feed them with more and more data, but would also boost the general automation level in hydrological modeling. Furthermore, these AI technologies own miscellaneous hyper-parameters to be tuned, and hence, expedite the computational time and provide the researchers in the field with an efficient alternative for the physical-based hydrological models.

## 9. Limitations and Future Work

In this section, we discuss the current study's limitations from both hydrological and data analytics perspectives. Although satellite remote sensing observations have numerous advantages, they are also characterized by certain limitations. Low frequency of the observations, varying errors in space and time, and relatively low coverage period (typically a few years to a couple of decades) constitute the main disadvantages, which are, of course, reflected in our methodology. No mathematically- or physically-based approach is ever 100% reliable when it comes to representing natural processes. However, based on our results, the presented novel methodology is characterized by a significant potential for unraveling and further describing the most intricate linkages and interactions among the major hydrological components of the different ecosystems in East Africa. Because of the nature of remote sensing products, which have errors that vary in space, the methodological approach presented in this study has an efficiency that varies with different geographical area, as well as with the use of different satellite products. Moreover, the land surface type is assumed to be non-dynamic, especially for the limited time period of the data used in this study. Furthermore, the method presented here is directed towards identifying the three main vegetation regimes of East Africa and, therefore, our findings apply only to this geographic location. The extensibility of this study is how-

ever feasible, depending on the availability of satellite observations and with varying efficiency in identifying the regional main vegetation regimes.

Data-driven models have recently obtained immense applicability in hydrologic modeling as they tackle the conventional shortcomings in physically based models, such as the uncertainty in the estimation of hydrological parameters. Furthermore, the data deluge from meteorological observations leading to a daunting big-data and signal-processing challenge calls for a nimble interpretation scheme to disentangle the intricate relationship of hydrological components without an explicit need to deal with the underlying physical processes. However, certain limitations still apply to data-driven models. A limiting factor in data-driven models and machine learning is the lack of sufficiently clean and homogenous data. While developing suitable learning architecture often remains as the primary challenge in AI, data quality is essential for the algorithms to function as intended. Noisy data, datasets containing tremendous outliers and missing values, are the quintessential drawbacks of a reliable machine learning model. Data governance, integration and exploration are prescribed as the potential solutions to this conundrum. As mentioned throughout the paper, the dataset used herein contained missing values for a number of days due to the temporary Windsat satellite failure. We adopted linear interpolation to impute these missing values. The interoplation operation would then replace the actual observations by the synthetic statistically inferred values. This, however, could deteriorate the performance of our predictive model, while dealing with enormous missing values. This issue often gets plagued as the outliers are also observed in the data. Moreover, a general rule of thumb in machine learning suggests feeding massive amount of data to machine with an aim to enhance the general predictive capability of the model. In this study, however, we used nine years of data for each of the components. Evidently, training the machine learning models with several more years of data would lead to a more general, and hence, better performing model.

Further analyses, including the use of other remote sensing products regarding surface hydrologic properties (e.g. normalized radar cross section, evapotranspiration), can reveal more information about the dynamics of different ecosystem processes, and provide a more integrated understanding of plant and ecosystem responses and behavior during extreme hydrologic conditions, which will undoubtedly provide machine learning models with more information, thus resulting in more representative findings of higher accuracy. As mentioned earlier, we for the purpose of this study studied the direct effect of VWC, SM and ST on the precipitation rate, and subsequently, their indirect effect on the vegetation regime, while a comprehensive study of the potential predictors necessitates a detailed exploration of several parameters including but not limited to the aforementioned components. Evidently, introducing other potential predictors would enhance our predictive model. Note that there exists absolutely no limitation on the number of the to-be-fed features into our learning models, and hence, making our model robust enough to be applicable to miscellaneous scenarios. In terms of further improvement on the machine learning models, a prolific literature on training the machine, utilizing the optimal subset of training data points, has been introduced as active machine learning literature (Cohn et al., 1996; McCallumzy et al., 1998; Brinker et al., 2003; Nguyen et al., 2004). Machine learning models, deployed with sophisticated active learning module, would definitely constitute the future exploratory work of this study.

While the advantages of machine learning models have been extensively studied, the reliability implications of using them in cascaded mode is not well understood. One of the major sources of unreliability is the error propagation issue. As mentioned earlier, we forecasted the precipitation rates via VWC, SM and ST, and utilized the predicted precipitation to classify the vegetation regime in a sequential framework. Clearly, the prediction error in precipitation would then impose additional error in the vegetation classification. As part of the further study, we focus our investigations towards fine-tuning the learning parameters of our models to mitigate this issue.

All of the aforementioned efforts will ultimately lead to a more sustainable management of water and carbon resources in future climates.

**Acknowledgement.** This work was supported by the National Science Foundation award (grant: GR10458) and conducted at Future H2O, Office of Knowledge Enterprise Development (OKED) at Arizona State University. The authors would like to acknowledge and appreciate Dr. Francis J. Turk from the Jet Propulsion Laboratory (JPL) and Dr. Li Li from the Naval Research Laboratory (NRL) for providing the WindSat data.

## References

Anderson, R. C., Fralish, J. S., and Baskin, J. M. (2007). *Savannas, barrens, and rock outcrop plant communities of North America*, Cambridge University Press, 2007.

Anosh, G. and Mishra, E. P. (2017) Time series analysis model to forecast rainfall for Allahabad region, *J. Pharmacogn. Phytochem.*, 6, 1418-1421.

Asbjornsen, H., Goldsmith, G. R., Alvarado-Barrientos, M. S., Rebel, K., Van Osch, F. P., Rietkerk, M., Chen, J. Q., Gotsch, S., Tobo ́n, C., Geissert, D. R., Go ́mez-Tagle, A., Vache, K., and Dawson, T. E. (2011). Ecohydrological advances and applications in plant-water relations research: a review, *J. Plant Ecol.*, 4(1-2), 3-22. https://doi. org/10.1093/jpe/rtr005.

Ashworth, N. J. and Wedderburn, R. W. W. (1972). Generalized linear models, *J. Roy. Stat. Soc. Ser. A.*, 135(3), 370-384. https://doi.org/10.2307/2344614.

Behrangi, A., Tian, Y., Lambrigtsen, B. H., and Stephens, G. L. (2014). What does CloudSat reveal about global land precipitation detection by other spaceborne sensors? *Water Resour. Re*s., 50(6), 4893-4905. https://doi.org/10.1002/2013WR014566.

Bello, I., Barret, Z., Vijay, V., and Quoc, V. L. (2017). Neural optimizer search with reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning*, vol 70, pp. 459-468.

Beven, K. (2006). A manifesto for the equifinality thesis, *J. Hydrol.*, 320(1), 18-36. https://doi.org/10.1016/j.jhydrol.2005.07.007.

Beven, K. and Jim F. (2001). Equifinality, data assimilation, and uncertainty estimation in mechanistic modelling of complex environmental systems using the GLUE methodology, *J. Hydrol.*, 249(1), 11-29. https://doi.org/10.1016/s0022-1694(01)00421-8.

Boko, M., Niang, I., Nyong, A., Vogel, C., Githeko, A., Medany, M.,

Osman-Elasha, B., Tabo, R., Yanda, P., Adesina, F., Agoli-Agbo, M., Attaher, S., Bounoua, L., Brooks, N., Dubois, G., Githendu, M. W., Hilmi, K., Misselhorn, A., Morton, J., Obioh, I., Ogbonna, A., Ouaga, H. N., Vincent, K., Washington, R., and Ziervogel, G. (2007). Africa, in M.L. Parry (Eds.), *Climate change 2007: Impacts, adaptation and vulnerability. Contribution of working group II to the fourth assessment report of the intergovernmental panel on climate change*, Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, pp. 433-467, 2007.

Brinker, K. (2003). Incorporating diversity in active learning with support vector machines, *Proceedings of the 20th international conference on machine learning* (ICML-03), Washington, DC, USA, pp. 59-66, 2003.

Chandler, R. E. and Wheater, H. S. (2002). Analysis of rainfall variability using generalized linear models: a case study from the west of Ireland, *Water Resour. Res.*, 38(10), 10-1. https://doi.org/10.1029/2001WR000906

Christensen, J. H., Hewitson, B., Busuioc, A., Chen, A., Gao, X., Held, I., and Whetton, P. (2007). Regional climate projections, in S. Solomon (Eds.), *Climate Change 2007: The physical science basis. Contribution of working group I to the fourth assessment report of the intergovernmental panel on climate change*, Cambridge Univer-sity Press, Cambridge, United Kingdom and New York, NY, USA, pp. 849-940, 2007.

Cohn, D. A., Ghahramani, Z., and Jordan, M. I. (1996). Active learning with statistical models, *J. Artificial Intel. Res*., 4, 129-145. https://doi.org/10.1613/jair.295.

Conway, D., Allison, E., Felstead, R., and Goulden, M. (2005). Rainfall variability in East Africa: implications for natural resources management and livelihoods: One contribution of 24 to a Discussion Meeting Atmosphere-ocean-ecology dynamics in the Western Indian Ocean, *Philos. Trans. R. Soc. Lond.*, 363(1826), 49-54. https://doi.org/10.1098/rsta.2004.1475.

Dinku, T., Ceccato, P., Cressman, K., and Connor, S. J. (2010). Evaluating detection skills of satellite rainfall estimates over desert locust recession regions, *J. Appl. Meteorol. Climatol.*, 49(6), 1322-1332. https://doi.org/10.1175/2010JAMC2281.1.

Dinku, T., Ceccato, P., Grover-Kopec, E., Lemma, M., Connor, S. J., and Ropelewski, C. F. (2007). Validation of satellite rainfall products over East Africa's complex topography, *Int. J. Remote Sens.*, 28(7), 1503-1526. https://doi.org/10.1080/01431160600954688.

D'Odorico, P. and Bhattachan, A. (2012). Hydrologic variability in dry-land regions: impacts on ecosystem dynamics and food security, Philos. *Trans. R. Soc. Lond.*, 367(1606), 3145-3157.https://doi.org/10.1098/rstb.2012.0016.

Du, J. L., Liu, Y. Y., and Liu, Z. J. (2018). Study of Precipitation Forecast Based on Deep Belief Networks, *Algorithms*, 11(9), 132.

Dunbar, M. J. and Acreman, M. C. (2001). Applied hydro-ecological science for the twenty-first century, *Hydro-ecology: Linking hydrology and aquatic ecology*, International Association of Hydrological Sciences, UK, 266, 1-17, 2001.

Eagleson, P. S. (1978). Climate, soil, and vegetation: 1. Introduction to water balance dynamics, *Water Resour. Res.*, 14(5), 705-712. https://doi.org/10.1029/WR014i005p00705.

Eagleson, P. S. (1982). Ecological optimality in water-limited natural soil-vegetation systems: 1. Theory and hypothesis, *Water Resour. Res.*, 18(2), 325-340. https://doi.org/10.1029/WR018i002p00325.

Faramarzi, M., Abbaspour, K. C., Vaghefi, S. A., Farzaneh, M. R., Zeh-nder, A. J., Srinivasan, R., and Yang, H. (2013). Modeling impacts of climate change on freshwater availability in Africa, *J. Hydrol.*, 480, 85-101.https://doi.org/10.1016/j.jhydrol.2012.12.016.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Lett.*, 27(8), 861-874. https://doi.org/10.1016/j.patrec.2005.10.010.

Federer, C. A. (1979). A soil-plant-atmosphere model for transpiration and availability of soil water, *Water Resour. Res*., 15(3), 555-562. https://doi.org/10.1029/WR015i003p00555.

Fisher, J. B. and Andreadis, K. M. (2014). Drought: Roles of precipitation, evapotranspiration, and soil moisture, in Y. Wang (Ed.), *Encyclopedia of natural resources: Air*, Taylor and Francis, New York, USA, pp. 1015-1017, 2014.https://doi.org/10.1081/E-ENRA-120047659.

Geetha, A. and Nasira, G. M. (2016). Time-series modelling and forecasting: Modelling of rainfall prediction using ARIMA model, *Int. J. Soc. Sys. Sci.*, 8, 361-372. https://doi.org/10.1504/IJSSS.2016.081411.

Gilmore, M. S., Wilson, E. H., Barrett, N., Civco, D. L., Prisloe, S., Hurd, J. D., and Chadwick, C. (2008). Integrating multi-temporal spectral and structural information to map wetland vegetation in a lower Connecticut River tidal marsh, *Remote Sens. Environ.*, 112 (11), 4048-4060. https://doi.org/10.1016/j.rse.2008.05.020.

Grimm, N. B., Chapin, F. S., Bierwagen, B., Gonzalez, P., Groffman, P. M., Luo, Y. Q., Melton, F., Nadelhoffer, K., Pairis, A., Raymond, P. A., Schimel, J., and Williamson, C. E. (2013). The impacts of climate change on ecosystem structure and function, *Front. Ecol. Environ.*, 11(9), 474-482. https://doi.org/10.1890/120282.

Ho, J. Y. and Lee, K. (2015). Grey forecast rainfall with flow updating algorithm for real-time flood forecasting, *Water*, 7(5), 1840-1865. https://doi.org/10.3390/w7051840.

Huffman, G. J., Bolvin, D. T., Nelkin, E. J., Wolff, D. B., Adler, R. F., Gu, G. J., Yang, H., Bowman, K. P., and Stocker, E. F. (2007). The TRMM multisatellite precipitation analysis (TMPA): Quasi-global, multiyear, combined-sensor precipitation estimates at fine scales, *J. Hydrometeorol.*, 8(1), 38-55. https://doi.org/10.1175/JHM560.1.

Ingsrisawang, L., Supawadee, I., Pramote, L., Premjai, T., Song, K., Prasert, A., and Warawut, K. (2010). Applications of statistical methods for rainfall prediction over the Eastern Thailand. *Proceedings of the Multi Conference of Engineers and Computer Scientists*, pp. 17-19.

Kabanda, T. A. and Jury, M. (1999). Inter-annual variability of short rains over northern Tanzania. *Clim. Res.*, 13(3), 231-241. https://doi.org/10.3354/cr013231.

Karran, D. J., Morin, E., and Adamowski, J. (2014). Multi-step streamflow forecasting using data-driven non-linear methods in contrasting climate regimes, *J. Hydroinf.*, 16(3), 671-689. https://doi.org/10.2166/hydro.2013.042.

Kenabatho, P. K., Parida, B. P., Moalafhi, D. B., and Segosebe, T. (2015). Analysis of rainfall and large-scale predictors using a stochastic model and artificial neural network for hydrological applications in southern Africa, *Hydrol. Sci. J.*, 60(11), 1943-1955. https://doi.org/10.1080/02626667.2015.1040021.

Khan, M. S. and Coulibaly, P. (2006). Application of support vector machine in lake water level prediction, *J. Hydrol. Eng*., 11(3), 199-205. https://doi.org/10.1061/(ASCE)1084-0699(2006)11:3(199)

Kim, T. W., Valdés, J. B., and Aparicio, J. (2002). Frequency and spatial characteristics of droughts in the Conchos River Basin, Mexico, *Water Int.*, 27(3), 420-430. https://doi.org/10.1080/02508060208687021.

Li, L., Gaiser, P. W., Gao, B. C., Bevilacqua, R. M., Jackson, T. J., Njoku, E. G., Rudiger, C., Calvet, J. C., and Bindlish, R. (2010). WindSat global soil moisture retrieval and validation. *IEEE Trans. Geosci. Remote Sens.*, 48(5), 2224-2241.https://doi.org/10.1109/TGRS.2009.2037749.

Liu, P., Li, C., Wang, Y., and Fu, Y. (2013). Climatic characteristics of convective and stratiform precipitation over the Tropical and Subtropical areas as derived from TRMM PR. *Sci. China Earth Sci*., 56(3), 375-385. https://doi.org/10.1007/s11430-012-4474-4.

Madsen, H. (2003). Parameter estimation in distributed hydrological catchment modelling using automatic calibration with multiple objectives. *Adv. Water Resour.*, 26(2), 205-216. https://doi.org/10.1016/S0309-1708(02)00092-1.

Mangaraj, A. K., Sahoo, L. N., and Sukla, M. K. (2013). A Markov

chain analysis of daily rainfall occurrence at western Orissa of India. *J. Relia. Stat. Stud.*, 6(1), 77-86.

McCallumzy, A. K. and Kama, N. (1998). Employing EM and pool-based active learning for text classification, *Proceedings of the Fifteenth International Conference on Machine Learning*, Madison, Wisconsin, USA, 1998.

McPherson, G. R. (1997). *Ecology and management of North American savannas*, University of Arizona Press, 1997.

McVicar, T. R., Roderick, M. L., Donohue, R. J., and Van Niel, T. G. (2012). Less bluster ahead? Ecohydrological implications of global trends of terrestrial near-surface wind speeds. *Ecohydrol.*, 5(4), 381-388. https://doi.org/10.1002/eco.1298.

Mitchell, T. (1997). *Machine Learning*, McGraw Hill, New York.

Mladenova, I. E., Jackson, T. J., Njoku, E., Bindlish, R., Chan, S., Cosh, M. H., Holmes, T. R. H., de Jeud, R. A. M., Jones, L., Kimball, J., Paloscia, S., and Santi, E. (2014). Remote monitoring of soil moisture using passive microwave-based techniques-Theoretical basis and overview of selected algorithms for AMSR-E, *Remote Sens. Environ.*, 144, 197-213. https://doi.org/10.1016/j.rse.2014.01.013.

Müller, C., Cramer, W., Hare, W. L., and Lotze-Campen, H. (2011). Climate change risks for African agriculture, *Proc. Natl. Acad. Sci.*, 108(11), 4313-4315. https://doi.org/10.1073/pnas.1015078108.

Müller, C., Waha, K., Bondeau, A., and Heinke, J. (2014). Hotspots of climate change impacts in sub-Saharan Africa and implications for adaptation and development, *Global Change Biol.*, 20(8), 2505-25 17. https://doi.org/10.1111/gcb.12586.

Mu-Sup, B., Hwan, C. K., Ok, K. H., Hyun-Kyung, O., and Jong-Chul, J. (2017). Mapping of Vegetation Using Multi-Temporal Downscaled Satellite Images of a Reclaimed Area in Saemangeum, Republic of Korea. *Remote Sens.*, 9(3), 272.

Nguyen, H. T. and Smeulders, A. (2004). Active learning using pre-clustering, *Proceedings of the twenty-first international conference on Machine learning*, Banff, Alberta, Canada, 2004. https://doi.org/10.1145/1015330.1015349.

Nicholson, S. E. (1996). A review of climate dynamics and climate variability in Eastern Africa, *The limnology, climatology and paleoclimatology of the East African lakes*, Springer, Dordrecht, 25-56.

Nicholson, S. E. (2000). The nature of rainfall variability over Africa on time scales of decades to millennia, *Global Planet. Change*, 26(1-3), 137-158. https://doi.org/10.1016/S0921-8181(00)00040-0.

Nicholson, S. E., Davenport, M. L., and Malo, A. R. (1990). A comparison of the vegetation response to rainfall in the sahel and east africa, using normalized difference vegetation index from NOAA AVHRR, *Clim. Change*, 17(2-3), 209-241. https://doi.org/10.1007/BF00138369.

Njoku, E. G. and Li, L. (1999). Retrieval of land surface parameters using passive microwave measurements at 6-18 GHz. *IEEE Trans. Geosci. Remote Sens.*, 37(1), 79-93. https://doi.org/10.1109/36.739125.

Njoku, E. G., Jackson, T. J., Lakshmi, V., Chan, T. K., and Nghiem, S. V. (2003). Soil moisture retrieval from AMSR-E. *IEEE Trans. Geosci. Remote Sens.*, 41(2), 215-229. https://doi.org/10.1109/TGRS.2002.808243.

Noy-Meir, I. (1973). Desert ecosystems: environment and producers, *Annu. Rev. Ecol. Syst.*, 4(1), 25-51. https://doi.org/10.1146/annurev.es.04.110173.000325.

Ogallo, L. J. (1988). Relationships between seasonal rainfall in East Africa and the Southern Oscillation, *J. Climatol.*, 8(1), 31-43. https://doi.org/10.1002/joc.3370080104.

Owe, M., de Jeu, R., and Holmes, T. (2008). Multisensor historical climatology of satellite-derived global land surface moisture, *J. Geophys. Res. (F Earth Surf.)*, 113(1). https://doi.org/10.1029/2007JF000769.

Owe, M., Jeu, R. D. and Walker, J. P. (2001). A methodology for surface soil moisture and vegetation optical depth retrieval using the microwave polarization difference index. *IEEE Trans. Geosci. Remote Sens.*, 39(8), 1643-1654. https://doi.org/10.1109/36.942542.

Parinussa, R. M., Holmes, T. R., and de Jeu, R. A. (2012). Soil Moisture Retrievals From the WindSat Spaceborne Polarimetric Microwave Radiometer, *IEEE Trans. Geo. Remote Sens.*, 50(7), 2683-2694. https://doi.org/10.1109/TGRS.2011.2174643.

Pik, R. (2011). Geodynamics: east Africa on the rise. *Nat. Geosci.*, 4(10), 660. https://doi.org/10.1038/ngeo1274.

Porporato, A. and Rodriguez-Iturbe, I. (2002). Ecohydrology-a challenging multidisciplinary research perspective/Ecohydrologie: une perspective stimulante de recherche multidisciplinaire, *Hydrol. Sci. J.*, 47(5), 811-821. https://doi.org/10.1080/0262666020949298 5.

Rodriguez-Iturbe, I. (2000). Ecohydrology: A hydrologic perspective of climate-soil-vegetation dynamies, *Water Resour. Res.*, 36(1), 3-9. https://doi.org/10.1029/1999WR900210.

Schalko, J. R., (1997). *Artificial neural networks, volume 1*, McGraw-Hill, New York, USA, 1997.

Singh, V. P. and Donald K. F. (2002). *Mathematical models of large watershed hydrology*, Water Resources Publication, 2002.

Sorooshian, S. and Vijai K. G. (1983). Automatic calibration of conceptual rainfall-runoff models: The question of parameter observability and uniqueness, *Water Resour. Res.*, 19(1), 260-268. https://doi.org/10.1029/WR019i001p00260.

Stampoulis, D., Andreadis, K. M., Granger, S. L., Fisher, J. B., Turk, F. J., Behrangi, A., Ines, A. V., and Das, N. N. (2016). Assessing hydro-ecological vulnerability using microwave radiometric measurements from WindSat, *Remote Sens. of Environ.*, 184, 58-72. https://doi.org/10.1016/j.rse.2016.06.007.

Stampoulis, D., Haddad, Z. S., and Anagnostou, E. N. (2014). Assessing the drivers of biodiversity in Madagascar by quantifying its hydrologic properties at the watershed scale, *Remote Sens. Environ.*, 148, 1-15. https://doi.org/10.1016/j.rse.2014.03.005.

Stern, R. D. and Coe, R. (1984). A model fitting analysis of daily rainfall data, *J. Roy. Stat. Soc. Ser. A. (Stat. Soc.)*, 147(1), 1-18. https://doi.org/10.2307/2981736.

Sumi, S. M., Zaman, M. F., and Hirose, H. (2012). A rainfall forecasting method using machine learning models and its application to the Fukuoka city case, *Int. J. Appl. Math. Comput. Sci.*, 22(4), 841-854. https://doi.org/10.2478/v10006-012-0062-1.

Tian, Y. and Peters-Lidard, C. D. (2010). A global map of uncertainties in satellite-based precipitation measurements, *Geophys. Res. Lett.*, 37(24). https://doi.org/10.1029/2010GL046008.

Turk, F. J., Li, L., and Haddad, Z. S. (2014). A physically based soil moisture and microwave emissivity data set for Global Precipitation Measurement (GPM) applications. *IEEE Trans. Geosci. Remote Sens.*, 52(12), 7637-7650. https://doi.org/10.1109/TGRS.2014.2315809.

Verschuren, D., Laird, K. R., and Cumming, B. F. (2000). Rainfall and drought in equatorial east Africa during the past 1,100 years, *Nat.*, 403(6768), 410. https://doi.org/10.1038/35000179.

Wolff, C., Haug, G. H., Timmermann, A., Damsté, J. S. S., Brauer, A., Sigman, D. M., Cane, M. A., and Verschuren, D. (2011). Reduced interannual rainfall variability in East Africa during the last ice age, *Sci.*, 333(6043), 743-747. https://doi.org/10.1126/science.1203724.

Xu, Z. G. and Zhuang, D. F. (2007). The methodology of detailed vegetation classification based on environmental knowledge and remote sensing images, *IEEE International Geoscience & Remote Sensing Symposium*, Barcelona, Spain, 2007. https://doi.org/10.110 9/IGARSS.2007.4423241.

Zhang, J., Kenneth, H., Carrie, L., Steve, V., Brian, K., Ami, A., Suzanne, V. C., Kevin, K., David, K., Ding, F., Seo, D. J., Ernie, W., and Chuck, D. (2011). National Mosaic and Multi-Sensor QPE (NMQ) system: Description, results, and future plans, *Bull. Am. Meteorol. Soc.*, 92.10, 1321-1338. https://doi.org/10.1175/2011BAMS-D-11-00047.1