

Article

Model Ensembles of Artificial Neural Networks and Support Vector Regression for Improved Accuracy in the Prediction of Vegetation Conditions and Droughts in Four Northern Kenya Counties

Chrisgone Adede ^{1,2,*} , Robert Oboko ¹, Peter W. Wagacha ¹ and Clement Atzberger ³ 

¹ School of Computing and Informatics, University of Nairobi (UoN), P.O. Box 30197, GPO, Nairobi 00100, Kenya; roboko@uonbi.ac.ke (R.O.); waiganjo@uonbi.ac.ke (P.W.W.)

² National Drought Management Authority (NDMA), Lonrho House-Standard Street, Box 53547, Nairobi 00200, Kenya

³ Institute of Geomatics, University of Natural Resources and Life Sciences, Vienna (BOKU), Peter Jordan Strasse 82, A-1190 Vienna, Austria; clement.atzberger@boku.ac.at

* Correspondence: adedekris@gmail.com

Received: 18 September 2019; Accepted: 28 November 2019; Published: 8 December 2019



Abstract: For improved drought planning and response, there is an increasing need for highly predictive and stable drought prediction models. This paper presents the performance of both homogeneous and heterogeneous model ensembles in the satellite-based prediction of drought severity using artificial neural networks (ANN) and support vector regression (SVR). For each of the homogeneous and heterogeneous model ensembles, the study investigates the performance of three model ensembling approaches: (1) non-weighted linear averaging, (2) ranked weighted averaging, and (3) model stacking using artificial neural networks. Using the approach of “over-produce then select”, the study used 17 years of satellite data on 16 selected variables for predictive drought monitoring to build 244 individual ANN and SVR models from which 111 models were automatically selected for the building of the model ensembles. Model stacking is shown to realize models that are superior in performance in the prediction of future drought conditions as compared to the linear averaging and weighted averaging approaches. The best performance from the heterogeneous stacked model ensembles recorded an R^2 of 0.94 in the prediction of future (1 month ahead) vegetation conditions on unseen test data (2016–2017) as compared to an R^2 of 0.83 and R^2 of 0.78 for ANN and SVR, respectively, in the traditional approach of selection of the best (champion) model. We conclude that despite the computational resource intensiveness of the model ensembling approach, the returns in terms of model performance for drought prediction are worth the investment, especially in the context of the continued exponential increase in computational power and the potential benefits of improved forecasting for vulnerable populations.

Keywords: ensemble; support vector regression; artificial neural networks; drought risk management; drought forecasting; ensemble member selection

1. Introduction

Droughts are temporary and recurrent events characterized by the absence of precipitation over an extended period of time [1]. Usually, one distinguishes between meteorological, hydrological, agricultural, and socio-economic droughts, as documented in UNOOSA [2]. The need for drought monitoring is commonly accepted and is well understood in the context of losses arising from drought

occurrence and the need for planned action. Losses from past droughts are documented, for example in Government of Kenya [3] and Cody [4] with a detailed review of a range of impacts in Ding et al. [5].

Common approaches for drought monitoring have been described in detail in Rembold et al. [6], AghaKouchak et al. [7] and Mishra & Singh [8] amongst others. Drought monitoring usually happens in the context of drought early warning systems (DEWS) that are increasingly either near real time or ex-ante (predictive). Well known near real time (NRT) systems include for example the univariate system of BOKU (University of Natural Resources and Life Sciences, Vienna) [9] and the Famine Early Warning Systems Network (FEWSNET) [10], both using Moderate Resolution Imaging Spectroradiometer (MODIS) vegetation data. The US drought monitor [11] is a well-known multi-variate system.

Predictive systems are either based on a single variable/index, a multi-variable index or multiple indices (variables). Single variable indexes, especially those based on the standardized precipitation index (SPI), are for example deployed in Ali et al. [12] and Khadr [13]. The single variables approach mostly predicts one type of drought mostly meteorological. The single variable indices are generally easy to interpret as compared to multi-variable (super-index) indices like in the case of the Enhanced Combined Drought Index (ECDI) in Enenkel [14] and Multivariate Standardized Drought Index (MSDI) in Hao & AghaKouchak [15]. While ECDI integrates four input datasets (rainfall, soil moisture, land surface temperature, and vegetation status), the MSDI integrates the Standardized Precipitation Index (SPI) and the Standardized Soil Moisture Index (SSI) for drought characterization. Increasingly, the use of multiple variables in predictive drought systems is gaining strength. In those systems, multiple indices are used to build predictive systems. Examples include the use of 11 variables and indices in Tadesse et al. [16]. In the work of Adede et al. [17], 10 variables are used in the prediction of future vegetation conditions. Adede et al. [17] further documents the limitation of the methods applied in common predictive systems including: the use of single techniques and the loss of opportunity in the selection of single models even in multivariate systems.

The increasing popularity of predictive DEWS is based on their ability to aid stakeholders to react before a crisis occurs especially in the light of increased damages suffered from droughts. The significant limitation of drought forecasting is the chaotic behavior of the atmospheric system and its interaction with the land surface. The proliferation of multiple indices and variables for drought monitoring has led to the need for multi-variate models that are both highly predictive and stable over time to support proactive drought risks management (DRM) initiatives. One way to improve stability and accuracy of models is model ensembling.

Variouly, model ensembling is defined as the formulation of multiple individually trained models and the subsequent combination of their outputs [18–21]. In this sense, model ensembling is akin to the innate human nature of seeking multiple opinions in decision making. Model ensembling, therefore, aims to produce more accurate and more stable predictors that arise from the ability to average out noise and with that achieve better generalizability [20,22]. According to Mendes-Moreira et al. [23], model selection has three advantages: (1) reduces computational costs, (2) increases prediction accuracy, and (3) avoids the multi-collinearity problem.

The common issues in model ensembling include (i) the process for over-production of ensemble base models, (ii) the selection of the base models for ensemble membership, and (iii) the combination of the outputs of the ensemble members. At the heart of the over-production of base models is hyper-parameter tuning. Hyper-parameter tuning is critical in instances when automation of model building and selection is a key objective like in Nay et al. [24] in which three hyper-parameters needed to be tuned for a gradient boosted machine (GBM). The issue of the selection of ensemble membership deals with the question of which sub-set of the models from the model over-production process offer the best predictive power. The problem of ensemble membership selection is for example documented in Partalas et al. [25], Re & Valentini [21] and Reid [26]. The distinct approach to ensemble member prediction includes greedy search [25], ensemble pruning [21] and statistical ensembling methods in Escolano et al. [27]. Greedy search takes local optimal decisions when changing the size of a current set and differs from ensemble pruning that uses both statistical and semi-definite programming

approaches to determine best ensemble sizes. On the other hand, the statistical ensemble method uses resampling to estimate the accuracy of individual members and multiple comparisons to choose ensemble membership.

Related work on ensemble modeling includes the use of bagging and boosting that both aim to generate strong learners from weak learners [20,28]. While bagging reduces variance by averaging predictions from multiple sub-sets of the training data, boosting increases model performance by sequentially learning models. Apart from bagging and boosting, an increasingly common approach to model ensembling is stacking. Stacking a meta/super learner is used to combine weak base predictors to reduce generalization error. Dzeroski & Zenko [29] documents stacking as having a performance at least comparable to the selection of the best classifier. The study in Belayneh et al. [28], for example, uses both bagging and boosting, and a single index in drought prediction using wavelet transforms, while Ganguli & Reddy [30] used the copula method on support vector regression (SVR) to simulate ensembles of drought forecasts. On the contrary, the systems in Wardlow et al. [31] and Tadesse et al. [32], for example, use multiple indices in the forecasting of future vegetation conditions.

Model ensembles can either be homogeneous, and thus based on the same technique [33], or heterogeneous and hence multi-technique [26]. Theoretically, due to the diversity of their membership as a result of different base learners [26], heterogeneous models should outperform homogeneous model ensembles. The superiority of the heterogeneous ensemble is for example empirically shown in the diagnosis of cancer [34]. The objective of this study is to investigate the performance of both homogeneous and heterogeneous multi-variate model ensembles in the prediction of vegetation conditions using ANN and SVR as the chosen case study techniques. The model ensembles are realized using three different ensemble approaches: non-weighted averaging, weighted averaging and ANN driven model stacking and their performance evaluated against that of the champion models.

2. Material and Methods

2.1. Study Area

The study area (Figure 1) covers an area of over 215,000 km² in Northern Kenya. The area covers four counties that are classified as arid counties within Kenya's arid and semi-arid lands (ASALs). The study area is characterized by both low rainfall and low vegetation cover with an average normalized difference vegetation index (NDVI) of below 0.4 except for May and November where NDVI peaks at 0.43. The rainfall averages around 250 mm for Turkana, Marsabit and Mandera and a little higher at around 370 mm in Wajir. Kenya experiences a bimodal rainfall pattern with two rainy seasons in March–May (MAM) and October–December (OND). Across the counties in the study area, 5–6 months are considered “wet”. Additional details are documented in Klisch & Atzberger [9], Adede et al. [17] and De Oto et al. [35].

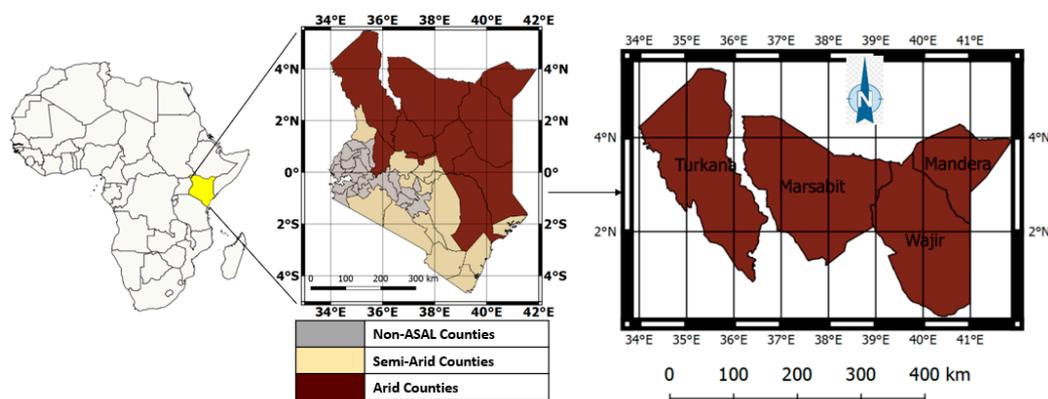


Figure 1. The study area (to the right) and its location within Kenya. The inset (left) provides the location of Kenya in Africa while the map of Kenya (center) shows the grouping of the 47 Kenyan counties into arid and semi-arid lands (ASAL) and non-ASAL.

2.2. Geodata

The study uses 22 variables and indices derived from rasterized information about vegetation, precipitation, temperature, and evapotranspiration. The variables are grouped into three dataset-categories: vegetation, precipitation and water balance datasets. The data used in the study covered the period March 2001 to December 2017 and are as described in Table 1.

Table 1. The study base datasets: categories, sources and description. The sources are grouped into three classes: (1) precipitation-related indices, (2) vegetation-related indices, and (3) water balance indices.

Base Dataset	Source	Description
Vegetation		
Normalized Difference Vegetation Index (NDVI)	LPDAAC Didan [36,37]	Combination of both MODIS Terra (MOD13Q1) & MODIS Aqua (MCD13Q1) using the Whittaker smoothing approach [9]
Precipitation		
Rainfall Estimates (RFE)	RFE from both TAMSAT [38] and CHIRPS [39]	TAMSAT version 3.0 product & CHIRPS version 2.0 product aggregated and spatially sub-set by BOKU
Water Balance		
Land Surface Temperature data (LST)	LPDAAC [40]	MODIS Terra Land Surface Temperature/Emissivity 8-Day L3 Global 1km SIN Grid V006 product (MOD11A2)
Evapotranspiration (EVT)	LPDAAC [41]	MODIS/Terra Net Evapotranspiration 8-Day L4 Global 500 m SIN Grid V006 (MOD16A2).
Potential Evapotranspiration (PET)	LPDAAC [41]	MODIS/Terra Net Evapotranspiration 8-Day L4 Global 500 m SIN Grid V006 (MOD16A2)
Standardized Precipitation-Evapotranspiration (SPEI)	SPEI Global Drought Monitor [42]	The standardized difference between precipitation and potential evapotranspiration [43].

LPDAAC = Land Processes Distributed Active Archive Center; MODIS = Moderate Resolution Imaging Spectroradiometer; TAMSAT = Tropical Applications of Meteorology using Satellite; CHIRPS = Climate Hazard Group InfraRed Precipitation; BOKU = University of Natural Resources and Life Sciences, Vienna.

2.3. Predictor and Target Variables

The above datasets are transformed to indices/variables that are then used for the predictive study following two rescaling approaches: (i) the calculation of relative range differences and (ii) the standardization approach.

The relative range difference is calculated as shown in Equation (1).

$$RR_h(c, i) = 100 * \left(\frac{[X(c, i) - MIN(c, i)]}{[MAX(c, i) - MIN(c, i)]} \right) \quad (1)$$

where RR_h is the scaled relative range difference, $X(c, i)$ the current value, $MIN(c, i)$ the historical minimum, and $MAX(c, i)$ the historical maximum.

The standardization approach involves fitting the base dataset to an appropriate probability distribution so that the mean of the transformed variable is zero and the standard deviation is equivalent to one. The necessary computations are outlined for example in Vicente-Serrano et al. [43] and WMO [44].

Both the relative range and standardised transformations are done at the pixel level (c) for each time step (i) prior to aggregation. The variables/indices used in this predictive modelling study are from the datasets listed in Table 1 and are described in Table 2.

Table 2. Variables used in the study to predict vegetation condition index (VCI). Near infrared (NIR) and Red are the spectral reflectances in near infrared and red spectral channels of MODIS satellite.

No	Variable	Variable Description	Index Calculation
1	NDVIdekad	NDVI for last dekad of the month	$NDVI = (NIR - Red) / (NIR + Red)$
2	VCIdekad	VCI for the last dekad of the month	Transformed NDVI based on Equation (1)
3	VCI1M	VCI aggregated over 1 month	Transformed NDVI based on Equation (1)
4	VCI3M	VCI aggregated over the last 3 months	Transformed NDVI based on Equation (1)
5	TAMSAT_RFE1M	TAMSAT RFE aggregated over 1 month	TAMSAT RFE version 3 product (in mm) [38]
6	TAMSAT_RFE3M	TAMSAT RFE aggregated over the last 3 months	TAMSAT RFE version 3 product (in mm) [38]
7	TAMSAT_RCI1M	TAMSAT Rainfall Condition Index (RCI) aggregated over the last 3 months	TAMSAT RFE based on Equation (1)
8	TAMSAT_RCI3M	TAMSAT RCI aggregated over the last 3 months	TAMSAT RFE based on Equation (1)
9	TAMSAT_SPI1M	TAMSAT Standardized Precipitation Index (SPI) aggregated over the last 1 month	TAMSAT RFE transformed to a normal distribution so that $SPI_{mean\ c,i} = 0$ [44].
10	TAMSAT_SPI3M	TAMSAT SPI aggregated over the last 3 months	TAMSAT RFE transformed to a normal distribution so that $SPI_{mean\ c,i} = 0$ [44]
11	CHIRPS_RFE1M	CHIRPS RFE aggregated over 1 month	CHIRPS RFE version 3 product (in mm) [39].
12	CHIRPS_RFE3M	CHIRPS RFE aggregated over the last 3 months	CHIRPS RFE version 3 product (in mm) [39].
13	CHIRPS_RCI1M	CHIRPS RCI aggregated over the last 1 month	CHIRPS RFE based on Equation (1)
14	CHIRPS_RCI3M	CHIRPS RCI aggregated over the last 3 months	CHIRPS RFE based on Equation (1)
15	CHIRPS_SPI1M	CHIRPS SPI aggregated over the last 1 month	CHIRPS RFE transformed to a normal distribution so that $SPI_{mean\ c,i} = 0$ [44].
16	CHIRPS_SPI3M	CHIRPS SPI aggregated over the last 3 months	Same as Index No. 15
17	LST1M	LST aggregated over 1 month	Average LST over the last one month
18	EVT1M	EVT aggregated over 1 month	Average MODIS EVT over the last month
19	PET1M	PET aggregated over 1 month	Average MODIS PET over the last month
20	TCI1M	Temperature Condition Index (TCI) aggregated over 1 month	MODIS LST based on Equation (1)
21	SPEI1M	Standardized Precipitation Evapotranspiration Index (SPEI) aggregated over 1 month	Follows the standardization approach on the difference between precipitation (P_i) and potential evapotranspiration (PET_i) using the logistic probability distribution
22	SPEI3M	SPEI aggregated over the last 3 months	Same as Index No. 21

Note: Variables 1–4 are vegetation indices while variables 5–16 are two sets of precipitation indices from TAMSAT (5–10) and CHIRPS (11–16), respectively. The study methodology is designed to select between TAMSAT and CHIRPS for the modeling process. Variables 17–22 are commonly used together with the vegetation and precipitation indices in predictive drought modeling and are used in this study as water balance variables.

The NDVI dataset in Table 1 (variable 1) is smoothed in NRT using a modified Whittaker smoothing algorithm [9,45]. The study in Atkinson et al. [46] documents the performance of the Whittaker smoother against four others. The NDVI dataset is directly sourced from BOKU University, Vienna. Those smoothed and gap-filled NDVI values are used to calculate the three remaining vegetation datasets (variables 2–4) in Table 2. All vegetation datasets are calculated at the pixel level prior to aggregation in both time and space. The vegetation condition index (VCI) variables are calculated following the relative range formula in Equation (1). The precipitation datasets from both CHIRPS and TAMSAT are also calculated at the pixel level prior to aggregation. While the RCI-based datasets (variables 7, 8, 13, and 14) follow Equation (1), the SPI variables (variables 9, 10, 15, and 16) have the base dataset (RFE) fitted to a probability distribution then transformed to a normal distribution so that the SPI has a mean

of zero (0) and a standard deviation of one (1) as recommended by WMO [44]. The SPEI datasets are also standardized following the logistic probability distribution [42,43].

2.4. The Learning Scenario

To be able to develop predictive drought monitoring models, the phenomenon of drought needs to be well defined. Following the literature review, we considered the following key characteristics in its definition: spatial coverage, severity and duration. With these key concepts, we formulate the predictive drought monitoring problem using Equation (2).

$$D_{(i,j)} = f(x_1, x_2, x_3 \dots x_n) \quad (2)$$

where $D_{(i,j)}$ is a quantification of drought severity (intensity) for a spatial extent i at time j , and f is a function that accepts a set of n ($n \geq 1$) variables and transforms them to approximate the real value for drought severity $D_{(i,j)}$. The n variables $x_1, x_2, x_3 \dots x_n$ are predictor variables that are used in drought monitoring. We interpret f as machine learning (ML) technique generated functions using artificial neural networks (ANN) and support vector regression (SVR) as case study techniques. The study in Atkinson & Tatnall [47], for example, provides an introduction to the use of ANN, particularly the backpropagation algorithm in remote sensing. A valuable review of support vector machines (SVM) in remote sensing can be found in Mountrakis et al. [48].

The learning scenario in this study therefore involves the need to define $D_{(i,j)}$ for all the four counties in the study area and to over-produce and select the appropriate f 's to be selected for model ensembling.

2.5. Definition of Drought ($D_{(i,j)}$)

Given that drought severity, $D_{(i,j)}$, cannot be directly quantified, the practice is to use proxy index(es) in its quantification. The most common proxies are the SPI and the VCI. Based on McKee et al. [49], the SPI values are typically between -3 and $+3$ with higher values indicating wet conditions, and vice versa. The VCI [9,50,51], on the other hand, is a relative range normalization of the NDVI in the 0–1 range (both end points included) that is generally scaled to the 0–100 range. Values below 0 and beyond 100, respectively, are possible if new minima or maxima occur.

From the properties of both the SPI and VCI, we chose the VCI as the basis of definition of drought for four reasons:

- with a range of 0 to 100, VCI is easy to interpret.
- as an index, VCI is indicative of agricultural drought, which is a later stage drought as compared to meteorological drought indicated by SPI.
- VCI is more directly related to food and fodder availability in the study area compared to SPI.
- VCI is a measured quantity as opposed to the SPI that is, for the case of this study, a modelled quantity.

The task of drought prediction is therefore expressed as the task of predicting future (1 month ahead) VCI values aggregated over 3 months (VCI3M) using lagged values of the studied predictor variables. The aggregation over 3 months was chosen as this better reflects drought severity and duration.

2.6. Handling of Duplicate Precipitation Datasets and Variable Selection

The study processed a duplicate set of precipitation variables from TAMSAT and CHIRPS, respectively (see Tables 1 and 2). To avoid including redundant variables from the same group, and in order to test which of the two precipitation datasets is most predictive, a variable selection was performed in this study. The aim was to select between TAMSAT and CHIRPS variables for the prediction of future drought intensity as defined by VCI3M.

Multiple methods were used in the selection between the two datasets. Test for normality using both density plots and Shapiro Wilk test informed the choice of methods for variable selection. Subsequently, correlational analysis, step-wise regression, Akaike information criterion (AIC), Relative importance (RImp) of variables, and a modelling approach to variable selection were used to establish which datasets between TAMSAT and CHIRPS to use in the modeling process.

2.7. Modelling Methodology

With the drought prediction problem formulated as in Equation (2), the modelling methodology is reduced to the search for all the f 's built from ANN and SVR that approximate drought severity ($D_{(i,j)}$). A subset of these f 's deemed to offer the best approximation of ($D_{(i,j)}$) based on an objective performance measure are then chosen and their outputs combined to predict future ($D_{(i,j)}$) values.

2.7.1. Model Building

Bagging, as a method of model ensembling, is used as a standard to build both the ANN and SVR models in a setup in which the training dataset (2001–2015) is resampled as indicated in the model building process (Figure 2). The main steps in the model building process are sampling, actual model building and model performance evaluation for both ANN and SVR techniques. Not indicated in the model building process is normalization of the variables prior to building of the models.

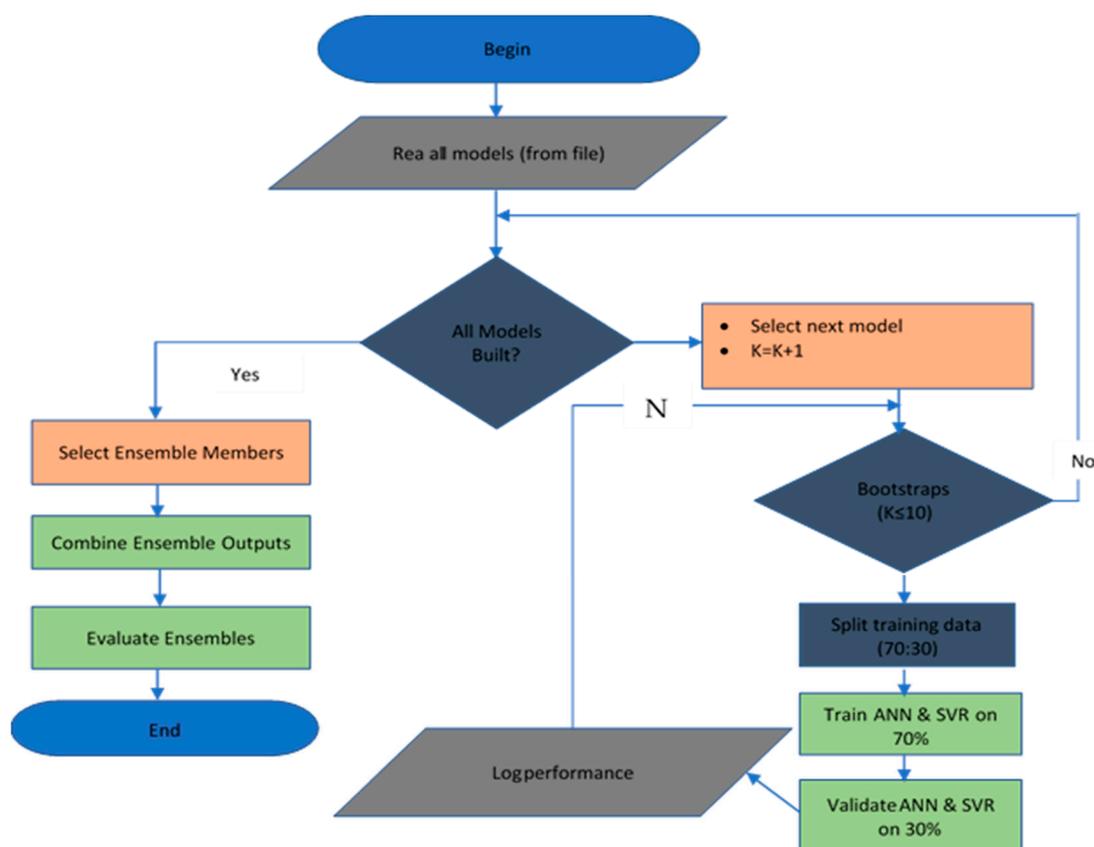


Figure 2. Model building process from model over-production to model selection for ensemble membership. The actual model building process using both ANN and SVR are preceded by a model space reduction process undertaken by two distinct steps. First, the formulation of assumptions and second, the use of cut-off criteria on models considered predictive enough to be included in the ensembles. Not included in this scheme is the fact that all variables were normalized prior to modeling to ensure the input variables were all at a comparable range.

2.7.2. Model Space Reduction & the Ensemble Membership

After the selection between TAMSAT and CHIRPS variables, the study has 16 variables. The initial cardinality of the modelling space, with VCI3M as the target variable, is a massive 65,535 models from all the 16 lagged predictors.

To achieve a reduction in the cardinality of the model space, we group the variables into categories and follow this up with assumptions. The 16 variables are grouped into three categories:

- vegetation (VCI3M, NDVIDekad, VCI1M & VCIdkad),
- precipitation (RFE1M, RFE3M, RC11M, RC13M, SPI1M & SPI3M) all from TAMSAT, and,
- water balance variables (LST1M, EVT1M, PET1M, TCI1M, SPE11M & SPE13M).

The grouping of variables is for example validated for soundness in Adede et al. (2019). The process of model space reduction is presented in Figure 3.

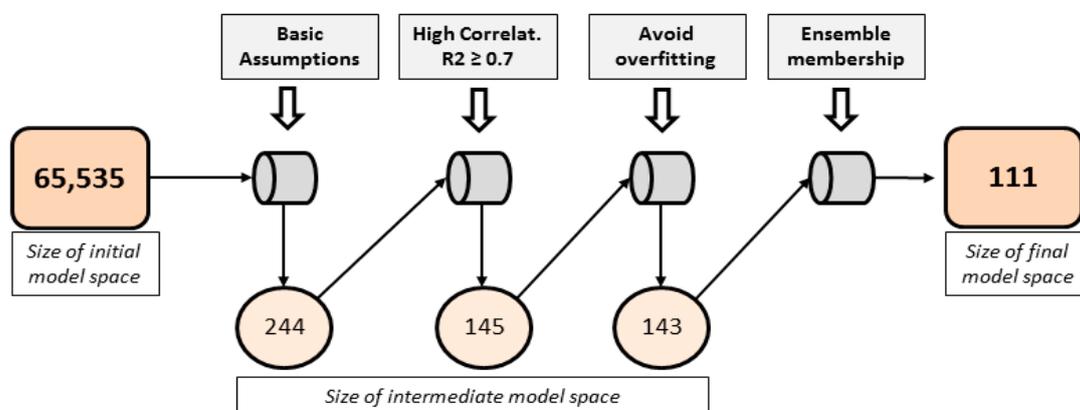


Figure 3. Illustration of the model space reduction process. The assumption of including only one variable of a type (precipitation, vegetation and water balance) reduces the set of possible models from 65,535 to 244. The selection of models with higher performance ($R^2 \geq 0.7$) and the subsequent elimination of overfit models further reduces the model space to 143 models. The set of 143 models is subjected to the ensemble membership selection process using a greedy version of iterative elimination of models from the ensemble as long as no significant loss in performance is recorded. In this way, a final model space with 111 models is achieved.

All the 244 models were built using both the ANN and SVR techniques. A further reduction was achieved by only retaining the 145 models that had $R^2 \geq 0.7$ from both the ANN and SVR techniques. Two further models were removed as they were indicating overfitting problems—here characterized to have more than a 3% loss in performance (ANN) between training and validation. This reduced the model space from 145 to 143 ANN models that were the basis of the search for the smallest sub-set of the ensemble membership that is most predictive. Even though model selection was done using R^2 , it was the case that adjusting for the possible influence of the number of variables in the models through the use of adjusted R^2 resulted in the same ranking of the models. Alternative model selection using root mean squared error (RMSE) also results in a similar ranking of the models.

The ensemble membership was realized from an experimental process. The experimental process involved the ranking of all the models by descending R^2 , the iterative elimination of the lowest ranked models in batches of five from the ensemble, and a recalculation of the performance of the ensemble as measured by R^2 . The elimination was stopped when a significant reduction in performance was realized. After such a drop-in performance, the last batch of five were eliminated and added to the model one at a time to the best former set of models. The membership was then chosen as the least membership for which R^2 is greedily maximized. This process led to the final 111 models (Figure 3).

2.7.3. Sampling Process

To build the requisite set of models, the dataset (March 2001 to December 2017) was partitioned into in-sample and out-sample data in an approach similar to that used in Adede et al. (2019) and Nay et al. (2018). The in-sample data was, subsequently, repeatedly and randomly split into training and validation datasets in the ration of 70:30 for each iteration of the model building process. The partitioning of the dataset (Figure 4) has the in-sample data covering the period 2001 to 2015, whereas the out-sample that was exclusively used for model testing covers the period 2016–2017.

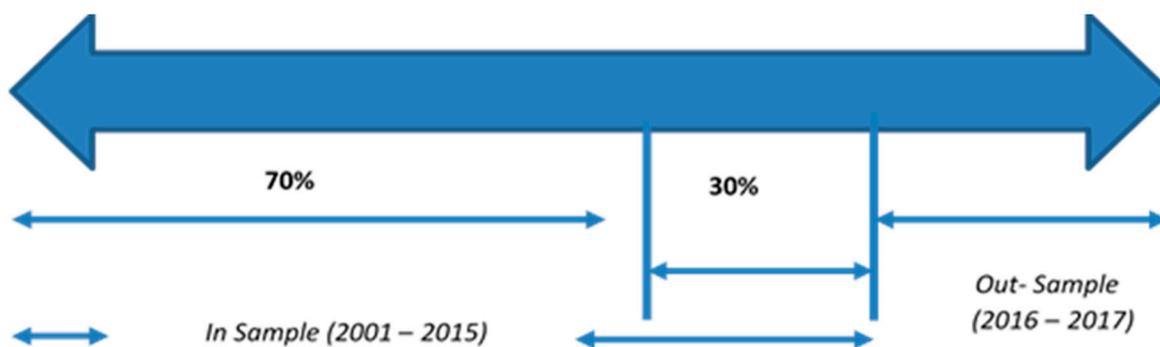


Figure 4. Sampling approach for model building (training and validation) and model testing. Only data from 2001 to 2015 was used for model building. Models were only tested using data from 2016–2017. The in-sample was randomly split (70:30) for model building and validation.

2.7.4. ANN and SVR Model Building Process: Model Parameters and Model Assessment

The 244 models formulated after the assumptions were subjected to an automated brute-force process that built, assessed and logged the performance of each of the models, both for the ANN and for the SVR case study techniques (Figure 2).

The choice of the model hyper-parameters for the ANN process followed the rule of thumb in Huang [52] together with an experimental process. The selected ANN had a three-layer fully connected architecture (excluding input layer) with a formation of 2–5–3–2 (12 nodes and 2 hidden layers). The ANN were trained using resilient backpropagation (RPROP) using the logistic activation function. RPROP has relatively fast convergence speed, accuracy and robustness without the requirement for parameter tuning as documented in Riedmiller & Braun [53] and in Chen & Lin [54]. For SVR, we did an initial run that realized the best performance of all the models at an epsilon of 0.2 and a cost parameter of 32. The search for the single optimal configuration for the SVR technique followed the grid search approach using the statistical computing software R.

Model performance evaluation was done using the coefficient of determination (R^2). Error measures like mean absolute error (MAE), root mean squared error (RMSE) and mean absolute percentage error (MAPE) were also calculated.

2.8. Homogeneous and Heterogeneous Model Ensembling Approaches

The different models built were selected and their scores combined in two distinct approaches: homogeneous ensembles in which only models from one technique are combined and heterogeneous ensembles that combines the outputs from both ANN and SVR techniques.

For both the homogeneous and heterogeneous ensembles, three methods of ensembling (Figure 5) were investigated:

- simple averaging (non-weighted) that assumes equal weights for the individual models,
- rank weighted averaging that uses performance in validation as weights and,
- model stacking that uses an ANN perceptron to learn model weights.

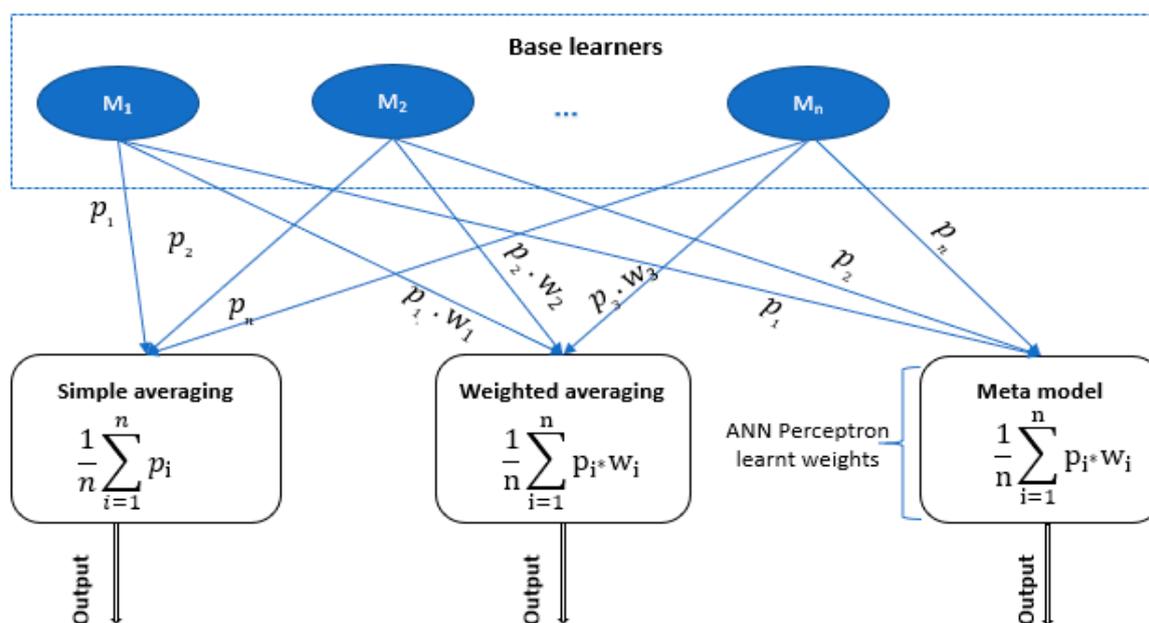


Figure 5. Schema of the model ensemble approaches for simple averaging (left), weighted averaging (center) and model stacking (right). In the simple averaging approach, the models have equal weights and so the model outputs are non-weighted. In the weighted averaging approach, the performance of the models in the validation dataset is used to assign weights to the model. The model stacking approach uses the outputs of the individual models as inputs to a meta model (ANN perceptron) that is then used to tune weights assigned to the individual models.

In the weighted averaging approach, the prediction of each model is weighted by the performance (R^2) of the model in the validation dataset as a proportion of the sum of the R^2 of all the models in the ensemble. Although the model stacking approach uses the same procedure as that of weighted averaging to scale the model weights, the process of realizing the weights is different. In model stacking, the prediction of each model in the validation dataset is used as an input to a perceptron that is then used to learn the set of weights to the inputs that realizes the best performance for the ensemble in the estimation of future vegetation conditions.

2.9. Performance Evaluation of the Model Ensembles

The process for evaluating the performance of the model ensembles had the following steps:

- the selection of a common measure of performance.
- the identification of the base models for comparison.
- scenarios of performance of the best individual models as compared to model ensembles.

Performance in regression was evaluated using R^2 while performance in classification was evaluated using accuracy. The best ANN and SVR models were used as the base models. Performance in classification was analyzed following the five vegetation deficit classes defined on VCI3M as used in Klisch & Atzberger [9], Adede et al. [17], Klisch et al. [55], and Meroni et al. [56] (2019). This classification is presented in Table 3.

Table 3. Drought classes used to assess the performance in classification. The drought classes quantify the vegetation deficit as described in Klisch & Atzberger [9], Adede et al. [17], Klisch et al. [55], and in Meroni et al. [56]. VCI3M is the 3-monthly Vegetation Condition Index (VCI) from filtered and gap-filled MODIS NDVI data.

VCI3M		Description of Class	Drought Class
Limit Lower	Limit Upper		
≤0	<10	Extreme vegetation deficit	1
10	<20	Severe vegetation deficit	2
20	<35	Moderate vegetation deficit	3
35	<50	Normal vegetation conditions	4
50	≥100	Above normal vegetation conditions	5

3. Results

3.1. Selection between TAMSAT & CHIRPS

As a basic design choice, we decided to use either TAMSAT or CHIRPS variables for drought prediction. The choice between TAMSAT and CHIRPS was undertaken using multiple methods, taking into account that the non-SPI datasets of both TAMSAT and CHIRPS are not normally distributed. The choice was informed by the following four indicators:

- the spearman's correlation on the 1-month lag of the variables with VCI3M
- Akaike information criterion (AIC)
- relative importance of variables as partitioned by R^2
- using SVR and general additive model (GAM) techniques

From the multiple selection metrics, even though both precipitation datasets were competitive in drought prediction, TAMSAT generally produced better-ranked variables and was the dataset chosen for model building. The single highest correlation was observed for 3-monthly TAMSAT SPI. The spearman's correlation on the 1-month lag of all variables with VCI3M as the predicted variable returned the results in Table 4 for each of the indicators for each of the TAMSAT and CHIRPS datasets. Between the TAMSAT and CHIRPS variable pairs, the TAMSAT variables are more highly correlated to drought severity as compared to CHIRPS variables, except for RCI1M.

Table 4. Spearman's correlation of VCI3M against 1-month lag of TAMSAT/CHIRPS rainfall indicators. In bold, the higher value for each comparison.

Data Set	RFE1M	RFE3M	RCI1M	RCI3M	SPI1M	SPI3M
TAMSAT	0.23	0.39	0.33	0.64	0.38	0.64
CHIRPS	0.10	0.26	0.34	0.53	0.34	0.52

Furthermore, for the SPI variable pairs, TAMSAT offers better agreement with VCI3M which is in itself an advantage given the popularity of the SPI in drought monitoring. The Akaike information criterion (AIC) is generally used to estimate the quality of a model relative to others, while taking into account the degrees of freedom. AIC had the 3-month aggregates of both SPI and RCI from TAMSAT as the best predictors as presented in Table A4 (Appendix A.3). These results were confirmed by the relative importance of variables as partitioned by R^2 that also ranked SPI3M and RCI3M from TAMSAT ahead of the other variables as shown in Table A5 (Appendix A.3). A final selection of variables using SVR and general additive model (GAM) techniques posted the result in Figure 6. The top performers in each case were two TAMSAT variables: SPI3M and RCI3M.

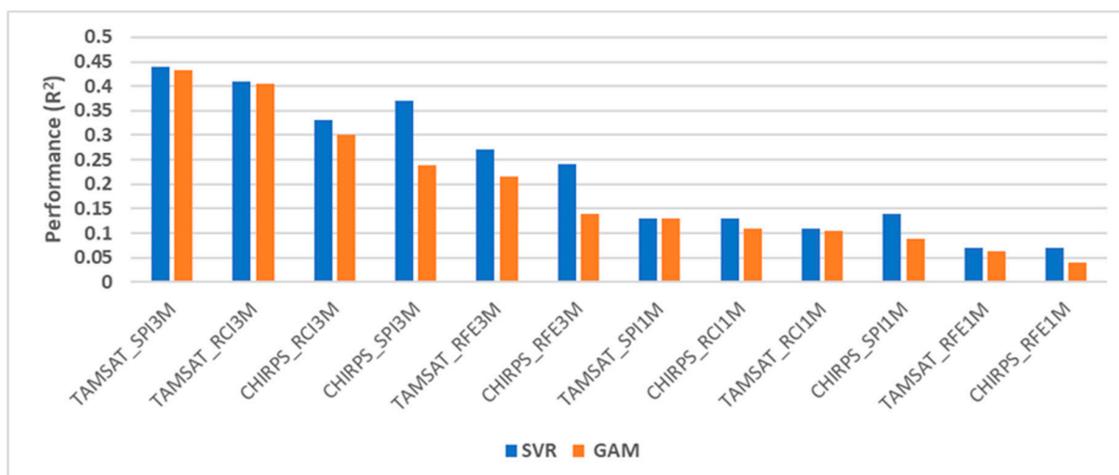


Figure 6. R^2 for SVR and GAM models for variable selection. The presented R^2 is between drought severity (VCI3M) and the precipitation variables of either TAMSAT or CHIRPS. The single variable models were developed with the same configurations using the in-sample datasets covering the period 2001–2015.

From the analysis above, the study decided to choose the TAMSAT datasets over the CHIRPS datasets. For the remainder of the document, only TAMSAT-related results will be presented.

3.2. Correlation between 3-Monthly Vegetation Condition Index and the Individual Predictor Variables

The results of the correlation analysis between the selected drought indicator (i.e., the 3-monthly vegetation condition index VCI3M) and the individual predictor variables from MODIS and TAMSAT are summarized in Table 5. As indicated in Table 5, the variables TCI, LST and PET are negatively correlated to VCI3M, while all others have a positive correlation to VCI3M. The correlation between the predictor variables is provided in Figure A1 (Appendix A.4). The correlation between vegetation variables is generally high. The problem of multi-collinearity is however avoided by the assumption to use only one variable from each of the groups in individual models.

Table 5. Correlation between the lagged predictor variables and future vegetation conditions (VCI3M). The precipitation variables are those derived from TAMSAT.

Lagged Variable	Correlation with Drought Severity (VCI3M)
TCI1M_lag1	−0.58
LST1M_lag1	−0.45
PET1M_lag1	−0.34
NDVIDekad_lag1	0.16
SPEI1M_lag1	0.19
RFE1M_lag1	0.23
SPEI3M_lag1	0.28
RCI1M_lag1	0.33
SPI1M_lag1	0.38
RFE3M_lag1	0.39
EVT1M_lag1	0.59
RCI3M_lag1	0.64
SPI3M_lag1	0.64
VCI3M_lag1	0.82
VCI1M_lag1	0.88
VCIdekad_lag1	0.89

3.3. Performance of ANN and SVR in the Training Dataset (2001–2015)

Using ANN and SVR, 244 prediction models for 1-month ahead VCI3M values were trained using the different combinations of the study variables. Out of the 244 models subjected to the ANN model training process, 145 models (~59%) have an $R^2 \geq 0.7$ in the validation dataset (Figure 7, top). Only a few models (~6%) were found to have been overfit as indicated by the gray bars and only two for models with $R^2 \geq 0.7$.

Similar findings were obtained using SVR for drought prediction (Figure 7, center). In terms of model overfitting, 21 of the 244 SVR models (~9%) are overfitted, however, they are confined to models with $R^2 \leq 0.5$. Like the case for ANN (Table A2 in Appendix A.2), there is no occurrence of overfitting in the top 30 (Table A3 in Appendix A.2) and top 100 models ordered by descending R^2 in the validation dataset.

The analysis of the performance of the pairings of the 244 ANN and SVR models is presented in Figure 7 (bottom). The ANN and SVR techniques turned out to be competitive in model validation: 127 models (52%) posted similar performance (no difference in R^2 of the validation data set), with ANN outperforming SVR in 105 pairings (43%) and SVR outperforming ANN in 12 pairings (5%).

Given that no models with $R^2 \geq 0.7$ were overfit from the SVR process, the choice of models for model ensembling (Figure 3) was therefore a function of models from the ANN process. The selection of the appropriate models for ensemble membership was thus from the 143 ANN models paired with the corresponding SVR models of the same formula.

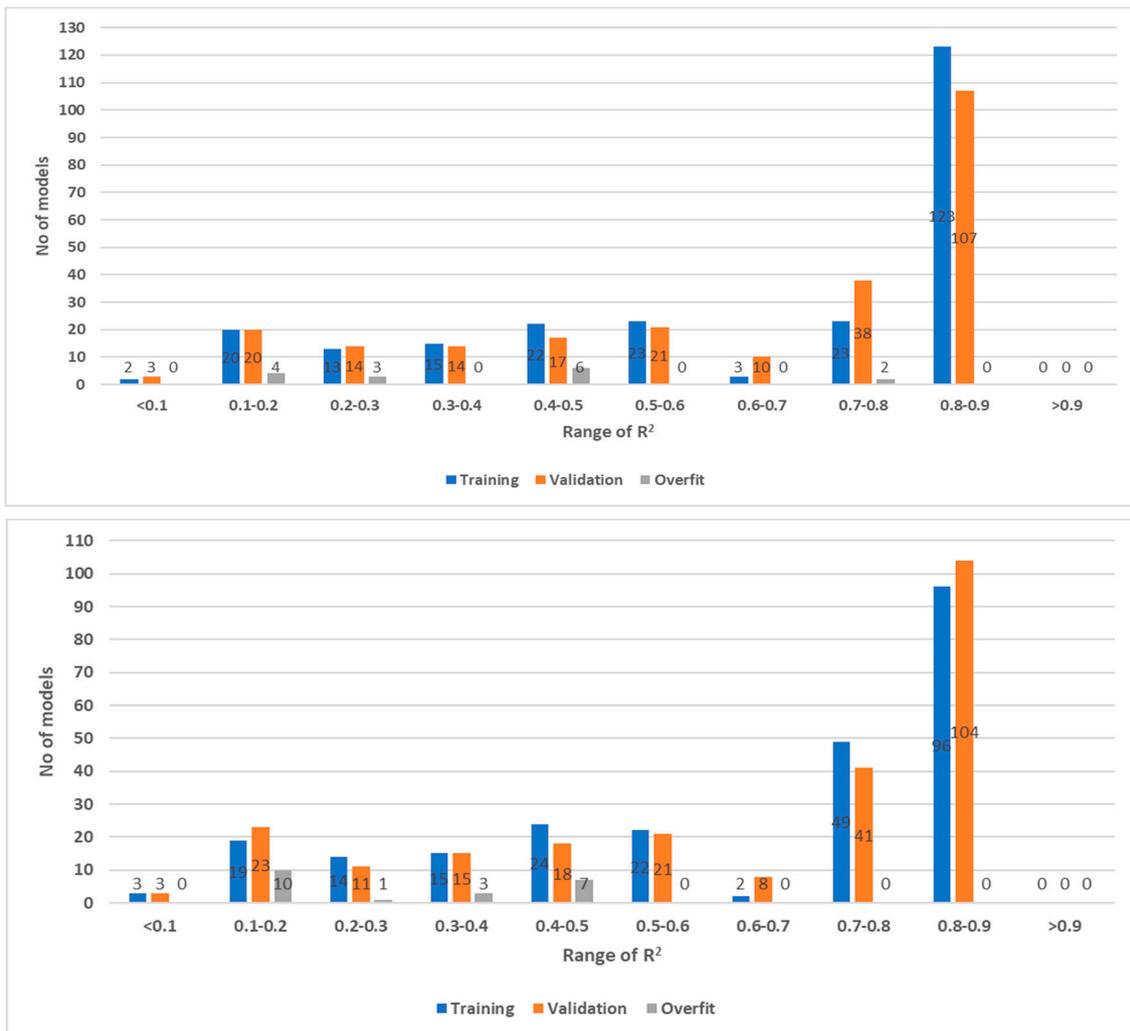


Figure 7. Cont.

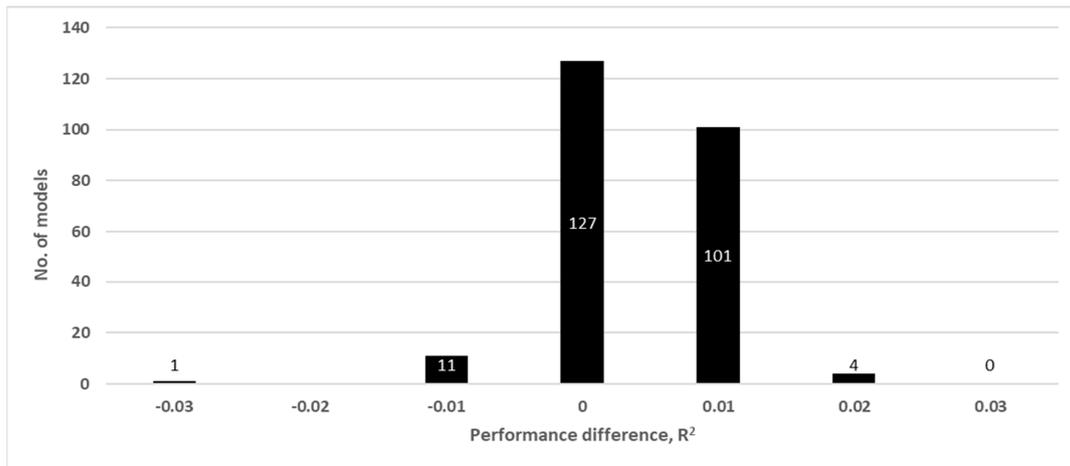


Figure 7. Performance of ANN (top) and SVR models (center) in the prediction of VCI3M grouped by performance (R^2). Models indicating significant overfitting problems are highlighted in gray. The model training uses the in-sample data (2001–2015) and follows on training with 70% and validating on 30% of the data as described in Section 2.7.3 on sample selection. (bottom) Performance difference between ANN and SVR model pairings in the validation dataset. The performance of the ANN and SVR models are rounded off to two decimal places prior to the calculation of the performance difference. The zero (0) difference represents 127 cases for which the SVR model equals the ANN models in performance. To the left of zero difference are the cases (12) in which SVR models outperform ANN models while to the right are cases (105) for which ANN models are superior to SVR models.

3.4. Selection of Ensemble Membership

Ensemble membership, also model pruning, is the selection phase of model ensembling. From the 143 models that had an $R^2 \geq 0.7$ and were not overfit, the construction of the ensemble membership faced the key question: Is there a smaller ensemble size that would perform the same, if not better than all the 143 models combined? This question was answered following the process described in Section 2.7.2 on model space reduction and the ensemble membership selection. The corresponding results are shown in Figure 8.

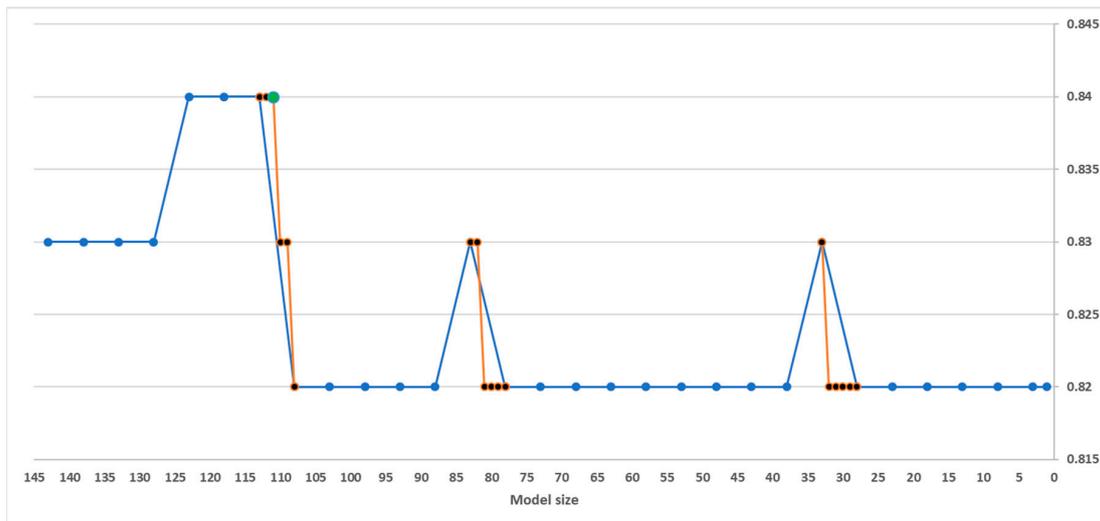


Figure 8. Ensemble membership selection showing the reduction from 143 models to 111 models (green dot) in the ensemble using the back-forward selection procedure. The models are eliminated in batches of five (blue lines) but in instances where a drop of performance is realized, a forward selection beginning with the last smallest ensemble size is done by the addition of one model at a time (orange lines in the plot). The values are rounded off to maximum two digits.

The elimination of the ranked models in batches of five (blue line) and the forward selection in single units (orange line) realize the best smallest performances of the ensemble membership at a total of 111 models (highlighted as green dot). For the ensemble membership, we hence opted for the size of 111. Such an ensemble size provides a good trade-off between computational complexity and performance. Note that in Figure 8 we also continued the process down towards a single variable in order to verify that no other model would outperform the ensemble membership size.

3.5. Performance of the Best ANN and the Best SVR Models in Out-of-Sample Data (2016–2017)

In training, two different models have been identified as the best performing (champion) models for each of the ANN and SVR techniques. Both the ANN and SVR champion models posted an R^2 of 0.86 in the validation dataset. The performance in training was also a comparable R^2 of 0.81 for the ANN to 0.85 for the SVR model.

The performance of the ANN and SVR techniques in the out-of-sample dataset using the 96 data points for the period 2016–2017 (Figure 4) is a contradiction of the performance in model validation. The ANN champion posted a relatively stable performance of an R^2 of 0.82 (4% loss in performance) while the SVR champion lost performance to record an R^2 of 0.78. The performance in the test data at county level for both the ANN and the SVR is as provided in Table 6.

Table 6. Performance (R^2) of the champion models for each of the counties in the study area on the out-of-sample dataset covering the period 2016–2017. Note that the models are non-county specific. Also reported is the overall performance across all four counties.

	Mandera	Marsabit	Turkana	Wajir	Overall
ANN	0.79	0.79	0.86	0.79	0.82
SVR	0.70	0.77	0.88	0.71	0.78

The results at county level indicate the consistency of ANN champion in out-performing the SVR champion except for Turkana where SVR outperforms ANN. Given that the models built are non-county specific, the performance of the best ANN and SVR models across the counties remains acceptable at an $R^2 \geq 0.79$ and $R^2 \geq 0.71$ for ANN and SVR, respectively. In the following, we use the R^2 from Table 6 as the baseline performance of the best model approach to model selection. These values will be used as the basis for comparison with the performance of both the homogeneous and heterogeneous model ensembles.

3.6. Performance of Homogeneous Model Ensembles in out-of-Sample Data (2016–2017)

From the 111 models selected for the investigation of model ensembles, we build both homogeneous models of ANN and SVR independently. The performances of the homogeneous ANN and SVR models were then investigated in both regression and classification. For each technique, results from the three ensembles approaches of non-weighted, weighted and stacked are presented.

3.6.1. Performance of Homogeneous Ensembles in Regression

The performance of the homogeneous model ensembles in regression is provided in Tables 7 and 8 for ANN and SVR techniques respectively. For both ANN and SVR techniques, the model stacking approach to the formulation of model ensembles offers the best improvement as compared to the best (champion) model approach. In general, across the counties, the weighted together with the non-weighted approaches to model ensembling provide competitive performance to the best model approach, but underperform compared to the model stacking. For all four counties, the stacked homogeneous ANN model yields higher R^2 on the out-of-sample test data set, compared to the SVR stack. The differences are particularly strong for Mandera and Wajir.

Table 7. Performance (R^2) of the ANN homogeneous model ensembles for each county. Each approach has the results derived from the non-weighted, weighted and stacked approaches of model ensembling. For comparison, the ANN champion model is also included (top row). In bold, the best results.

Approach	Mandera	Marsabit	Turkana	Wajir	Overall
ANN Champion	0.79	0.79	0.86	0.79	0.82
ANN Homogeneous Simple Average	0.78	0.86	0.88	0.80	0.84
ANN Homogeneous Weighted Average	0.79	0.86	0.88	0.81	0.85
ANN Homogeneous Stacked	0.93	0.87	0.89	0.93	0.91

Table 8. Performance (R^2) of the SVR homogeneous model ensembles for each county. Each approach has the results derived from the non-weighted, weighted and stacked approaches of model ensembling. For comparison, the SVR champion model is also included (top row). In bold, the best results.

Approach	Mandera	Marsabit	Turkana	Wajir	Overall
SVR Champion	0.70	0.77	0.88	0.71	0.78
SVR Homogeneous Simple Average	0.71	0.80	0.87	0.73	0.80
SVR Homogeneous Weighted Average	0.71	0.80	0.87	0.73	0.80
SVR Homogeneous Stacked	0.88	0.85	0.88	0.88	0.88

3.6.2. Performance of Homogeneous Ensembles in Classification

The summary of the classification performance of both the ANN and SVR homogeneous ensembles is presented in Tables 9 and 10, respectively.

Table 9. Classification accuracy for the ANN homogeneous ensembles. Each approach has the results derived from the non-weighted, weighted and stacked approaches of model ensembling. For comparison, the ANN champion model is also included (top row). In bold, the best results.

Approach	Mandera	Marsabit	Turkana	Wajir	Overall
ANN Champion	0.71	0.75	0.71	0.67	0.71
ANN Homogeneous Simple Average	0.67	0.83	0.67	0.63	0.70
ANN Homogeneous Weighted Average	0.67	0.79	0.67	0.63	0.69
ANN Homogeneous Stacked	0.79	0.88	0.75	0.71	0.78

Table 10. Classification accuracy for the SVR homogeneous ensembles. Each approach has the results derived from the non-weighted, weighted and stacked approaches of model ensembling. For comparison, the SVR champion model is also included (top row). In bold, the best results.

Approach	Mandera	Marsabit	Turkana	Wajir	Overall
SVR Champion	0.58	0.75	0.83	0.58	0.69
ANN Homogeneous Simple Average	0.63	0.83	0.67	0.63	0.69
ANN Homogeneous Weighted Average	0.63	0.83	0.71	0.63	0.70
ANN Homogeneous Stacked	0.79	0.88	0.75	0.71	0.78

Using the three approaches to the building of model ensembles, for both the ANN and SVR approaches, it is clear that the homogeneous ensembles are superior to the traditional champion model selection approach.

3.7. Performance of Heterogeneous Model Ensembles in out-of-Sample Data (2016–2017)

The performance of the heterogeneous model ensembles of ANN and SVR was assessed both in regression and classification. Like in the case of homogeneous models, we use the champion models (ANN and SVR) as the base models for the evaluation. Given that the predictions for the models were

averaged for each model across the techniques, 222 models for each input data point in the test data were used.

3.7.1. Performance of Heterogeneous Ensembles in Regression

The performance of the heterogeneous models is presented in Table 11, at county level with the champion ANN and champion SVR as the base models. The heterogeneous stacked approach in regression is seen to offer the best improvement in performance across each of the counties and on the entire dataset. The improvement in performance by R^2 using the heterogeneous stacked approach compared to the ANN champion as the base ranges from 0.05 (Turkana) to 0.17 (Wajir). The overall regression performance increases by 0.12 units.

Table 11. Performance (R^2) of the heterogeneous model ensembles for each county. Each approach has the results derived from the non-weighted, weighted and stacked approaches to model ensembling. For comparison, the ANN and SVR champion models are also included. In bold, the best results.

Approach	Mandera	Marsabit	Turkana	Wajir	Overall
ANN Champion	0.79	0.79	0.86	0.79	0.82
SVR Champion	0.70	0.77	0.88	0.71	0.78
Heterogeneous Simple Average	0.74	0.82	0.87	0.76	0.82
Heterogeneous Weighted Average	0.76	0.83	0.88	0.78	0.82
Heterogeneous Stacked	0.94	0.94	0.91	0.96	0.94

Given the superiority of the heterogeneous stacking approach, we present in Figure 9a further analysis of their performance in regression based on the out-of-sample data (2016–2017). For each of the four counties, 24 data points are presented showing the observed (in blue) and the predicted (in orange) drought indicator.

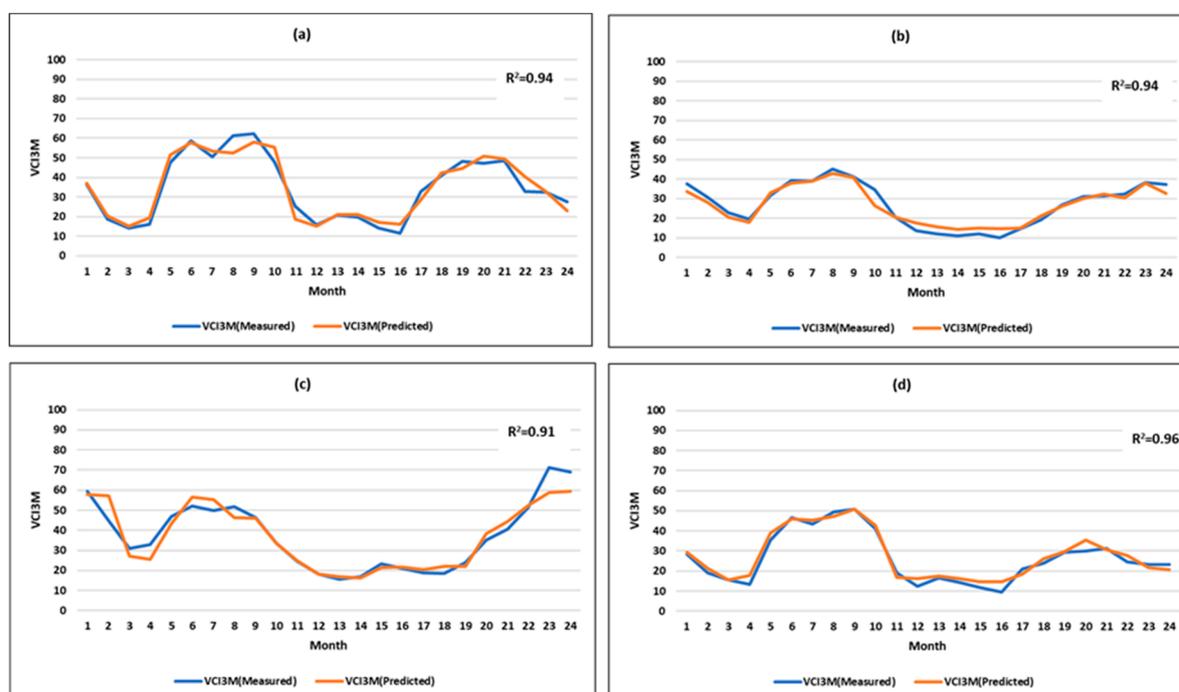


Figure 9. Plot of the actual values of VCI3M versus the values predicted 1 month ahead from the heterogeneous stacked ensembles in the test data over 24 months for (a) Mandera ($R^2 = 0.94$); (b) Marsabit ($R^2 = 0.94$); (c) Turkana ($R^2 = 0.91$) and (d) Wajir ($R^2 = 0.96$). Results are for out-of-sample data (2016–2017).

The heterogeneous stacked model, as indicated in Figure 9, posts an excellent agreement between the measured vegetation conditions as compared to the predicted values with an R² of between 0.91 and 0.96 for the counties.

3.7.2. Performance of Heterogeneous Ensemble in Classification

The classification accuracy of the heterogeneous ensemble as compared to that of the champion models is presented in Table 12.

Table 12. Classification accuracy of the heterogeneous ensemble. Each approach has the results derived from the non-weighted, weighted and stacked approaches to model ensembling. For comparison, the ANN and SVR champion models are also included. In bold, the best results.

Approach	Mandera	Marsabit	Turkana	Wajir	Overall
ANN Champion	71	75	71	67	71
SVR Champion	58	75	83	58	69
Heterogeneous Simple Average	63	83	71	63	70
Heterogeneous Weighted Average	63	83	71	67	71
Heterogeneous Stacked	71	88	79	83	80

The accuracy of the heterogeneous model ensembles in classification shows quite an overall improvement in performance of 9 and 11% for the ANN and SVR processes, respectively (last column). For Turkana county, the SVR champion however slightly outperforms the three approaches in heterogeneous model ensembling.

The corresponding month by month performance of the heterogeneous stacked classifier is as presented in Figure 10 and as visualized in Figure A2 (Appendix A.5). The months in blue indicate when the class prediction is correct while grey indicates non-correct predictions. This illustrates that the heterogeneous stacked ensemble accuracy ranges from a best of 88% in Marsabit to a minimum of 71% for Mandera. Clearly, the performance of the heterogeneous stacked classifiers with an overall accuracy of 80% is superior to that of the best champion model (ANN) that posts accuracy of 71% over the entire test dataset (Table 9). Only for Turkana, the SVR champion performs better compared to the heterogeneous stacked classifier, as highlighted in Table 12.

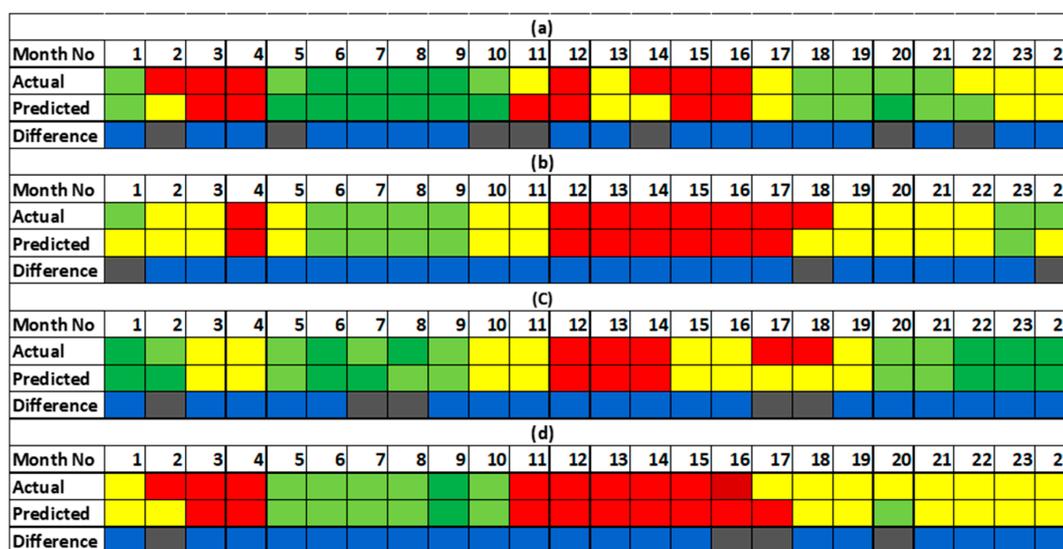


Figure 10. Performance of the heterogeneous stacked ensemble classifier for the each of the counties showing months of difference in grey and those of agreement in blue. Predictions were made 1 month ahead. The classification accuracies are: (a) 71% for Mandera county; (b) 88% for Marsabit county; (c) 79% for Turkana county and; (d) 83% for Wajir county.

Moderate to extreme vegetation deficits correspond to the occurrence of drought events. Their correct prediction forms another important judgment on the utility of the outputs of the modelling approaches. Table 13 presents the performance in the prediction of moderate to extreme vegetation deficits across the different classes for the counties in the study area.

Table 13. Performance in the prediction of moderate to extreme droughts using the heterogeneous stacked ensemble compared to the best ANN and SVR models. In bold, the best results.

County	ANN Champion	SVR Champion	Heterogeneous Stacked Ensemble
Mandera	0.62	0.46	0.69
Marsabit	0.71	0.71	0.94
Turkana	0.75	0.00	0.83
Wajir	0.72	0.61	0.78
Overall	0.70	0.69	0.82

The heterogeneous stacked ensemble is shown to offer the best performance in the prediction of moderate to extreme vegetation deficit with accuracies of 82% compared to 69–70% for ANN and SVR champions. At the same time, the accuracies are more consistent over space.

4. Discussion

From the results realized in this study (Section 3), several key findings can be highlighted. The investigation of correlation shows higher correlation between the vegetation datasets and drought severity as compared to the precipitation and other datasets. This is explainable given that we define drought severity in terms of vegetation deficit, therefore, making vegetation conditions good predictors. Rainfall anomalies, on the other hand, may or may not lead to significant changes in vegetation conditions, depending for example on available water resources. For the same reason, many studies focusing on agricultural drought prefer observing directly the vegetation conditions, without relying on rainfall data, which are anyhow often sparse and/or associated with high uncertainties [57].

Theoretically, heterogenous ensembles are expected to outperform both champion models and homogenous ensembles in both regression and classification. This, in the most part, is the case in this study as is also in Petrakova et al. [24]. An overall R^2 of 0.94 in regression and an accuracy of 80% in classification makes heterogenous model ensembling superior to both the homogenous ensembles and champion model approach to prediction of future vegetation conditions. The superiority of the ensembles against champion models especially, as indicated by overall prediction accuracy, is however not guaranteed for generalization across the spatial units. The case in Table 12 has the ANN champion and the SVR champion outperforming the heterogenous ensembles in classification accuracy for Mandera and Turkana respectively. This observation makes for the caution in use of model ensembles as in some tasks, loss of performance can be realized. Similar findings were observed in the estimation of software efforts in Elish [58] and Kocaguneli et al. [59]. The study in Elish [58] had a voting ensemble model derived from five techniques outperforming single models of the techniques in only three out of five datasets. Kocaguneli et al. [59] saw heterogenous ensembles realize accuracies that were far from outperforming single learners.

The building of models that are not specific to each spatial unit of analysis provides an opportunity to scale this approach to the prediction of vegetation conditions across multiple spatial extents. Together with the illustrated generalizability with time (on future test data), this makes for a highly generalizable approach to predictions. Models fine-tuned for each spatial unit separately, were for example reported in Nay et al. [24].

The superiority in performance between ANN and SVR and indeed between any two machine learning techniques is one in which the jury is still out. In this study, the ANN technique is shown to generally outperform SVR in predictive performance in the ratio of 43% to 5% and a tie in 52% of

the cases respectively. This superiority of the ANN to SVR is also observed in the prediction of pipe burst rates in Shirzad et al. [60]. Other studies like that in Mokhtarzad et al. [61] document the SVR to be superior to the ANN. Depending on the application of the techniques, it remains non-clear to pick out one as superior to the other. What is clear in this study, however, is that the ANN champion model outperforms the SVR champion in the unseen data even though they remain competitive in model training.

The ANN technique is particularly documented to be more susceptible to overfitting in cases of large networks. This is however not the case in this study that documents overfitting as more pronounced in SVR as compared to ANN (9% to 6%). This result should be interpreted also in the context that overfitting remains confined to only models with $R^2 \leq 0.5$ in SVR as compared to the ANN technique that has two overfit models with $R^2 \geq 0.7$. The reduced occurrence of overfitting can be attributed to the use of an adequate sample size in the model training process, thus avoiding the high dimension, low sample size (HDLSS) learning scenarios as documented in Liu et al. [62].

5. Conclusions

The output from the drought prediction are linked to the end-of-month drought monitoring results to provide the forecasts of the month ahead. Currently, the results are used within Kenya's National Drought Management Authority (NDMA) as an internal decision tool, especially to support drought response planning to ensure timely drought response.

The traditional approach to model selection ends up with one champion model, usually selected based on their performance in the validation dataset. We demonstrate that this approach is prone to loss of model performance. The building of model ensembles, on the other hand, not only guarantees stability but ensures an increased accuracy in model performance. In order to realize such ensembles, we found that the use of a simple backward-forward approach in the selection of models for ensemble membership was viable as it reduced the model space to a total of 111 models from a set of 143 models.

The model ensembling approaches investigated in the study included non-weighted averaging, weighted averaging and model stacking as applied to both homogeneous and heterogeneous model ensembling approaches. Empirically, it was shown that heterogeneous ensembles combining models from SVR and ANN family which are generally more robust as compared to homogeneous ensembles. Also, model stacking is indicated to be the surest way to realize model ensembles that are better in performance as compared to the champion model approach. In fact, it is empirically shown that a loss in performance could be suffered when linear or weighted averaging approaches are used, especially when the models in the ensembles are selected based on a common performance metric.

The performance of the models learnt using the heterogeneous model stacking approach is noted to be robust both in terms of 1-month ahead regression and classification. Moreover, although the ensemble models were not trained specifically for different administrative units, they generalized well for the four individual units even when testing on unseen out-of-sample test data. This is a key finding since it implies that the approach is robust enough to learn a single model applicable in the prediction across multiple administration units.

We nevertheless advise evaluating the use of more techniques in the model ensembles and the building of many more ensembles using different ensemble sizes to settle the question of performance of model ensembles fully. To further increase the utility of drought prediction, we also recommend to study more extended forecasting periods (up to 3 months) and to estimate how much this would degrade the prediction skill.

Author Contributions: Conceptualization, Chrisgone Adede; Formal analysis, Chrisgone Adede; Investigation, Chrisgone Adede; Methodology, Chrisgone Adede, Robert Oboko, Peter W. Wagacha and Clement Atzberger; Supervision, Robert Oboko, Peter W. Wagacha and Clement Atzberger; Validation, Robert Oboko, Peter W. Wagacha and Clement Atzberger; Visualization, Chrisgone Adede; Writing—original draft, Chrisgone Adede, Robert Oboko, Peter W. Wagacha; Writing—review & editing, Clement Atzberger.

Funding: This research received no direct external funding. The data used in the study was however, partly funded by the European Commission’s funding under a grant contact to the Institute for Surveying, Remote Sensing and Land Information, University of Natural Resources and Life Sciences (BOKU).

Acknowledgments: Our appreciation to the National Drought Management Authority for providing the data from the operational drought monitoring system and meeting the cost of publication. We are also grateful to Luigi Luminari for the continued discussion of the ideas of the paper towards shaping it to have outputs applicable in an operational drought monitoring environment. The helpful contribution of the editors and reviewers are also acknowledged.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A.

Appendix A.1. The Test for Normality for Precipitation Datasets

The results of the Shapiro–Wilk test for normality for the TAMSAT and CHIRPS precipitation datasets is presented in Table A1.

Table A1. Shapiro–Wilk test on normalized CHIRPS and TAMSAT datasets.

No	Variable	<i>p</i> -Value
1	TAMSAT_RFE1M	0.0000
2	CHIRPS_RFE1M	0.0000
3	TAMSAT_RFE3M	0.0000
4	CHIRPS_RFE3M	0.0000
5	TAMSAT_RCI1M	0.0000
6	CHIRPS_RCI1M	0.0000
7	TAMSAT_RCI3M	0.0000
8	CHIRPS_RCI3M	0.0000

With the null hypothesis corresponding to the sample drawn belonging to a normally distributed population, we reject normality for all cases in Table A1 where the *p*-value ≥ 0.05 at $\alpha = 0.05$. The non-SPI variables were not normally distributed, thereby justifying the use of non-parametric methods in their analysis.

Appendix A.2. The Top 30 ANN and SVR Models

Tables A2 and A3 provide a listing of the top 30 models from each of the ANN and SVR techniques with an indication of overfitting as less of an occurrence amongst the top models from each of the techniques.

Table A2. Performance of the top 30 ANN models in training and validation ordered by descending R² in the validation dataset.

No	Model	R ² Training	R ² Validation	Overfit Index	Overfit
1	VCI3M_lag1 + TAMSAT_RCI3M_lag1 + SPEI1M_lag1	0.81	0.86	0.05	No
2	VCIdekad_lag1 + TAMSAT_RFE1M_lag1 + TCI1M_lag1	0.87	0.86	-0.01	No
3	VCIdekad_lag1 + TAMSAT_RCI3M_lag1 + TCI1M_lag1	0.87	0.86	-0.01	No
4	VCIdekad_lag1 + TAMSAT_SPI3M_lag1 + TCI1M_lag1	0.87	0.86	-0.01	No
5	VCI1M_lag1 + TAMSAT_RCI3M_lag1 + TCI1M_lag1	0.87	0.86	-0.01	No
6	VCI3M_lag1 + TAMSAT_SPI3M_lag1 + SPEI1M_lag1	0.82	0.85	0.03	No
7	VCI1M_lag1 + TAMSAT_SPI3M_lag1 + TCI1M_lag1	0.87	0.85	-0.02	No
8	VCIdekad_lag1 + TAMSAT_SPI1M_lag1 + PET1M_lag1	0.87	0.85	-0.02	No
9	VCIdekad_lag1 + TAMSAT_SPI1M_lag1 + TCI1M_lag1	0.86	0.85	-0.01	No
10	VCI1M_lag1 + TAMSAT_SPI3M_lag1 + PET1M_lag1	0.87	0.85	-0.02	No
11	VCIdekad_lag1 + TAMSAT_SPI3M_lag1 + LST1M_lag1	0.86	0.85	-0.01	No
12	VCIdekad_lag1 + TAMSAT_SPI3M_lag1 + PET1M_lag1	0.87	0.85	-0.02	No
13	VCIdekad_lag1 + TAMSAT_RFE3M_lag1 + TCI1M_lag1	0.87	0.85	-0.02	No
14	VCI1M_lag1 + TAMSAT_RFE1M_lag1 + TCI1M_lag1	0.87	0.85	-0.02	No
15	VCI1M_lag1 + TAMSAT_SPI3M_lag1 + LST1M_lag1	0.86	0.85	-0.01	No
16	VCIdekad_lag1 + TAMSAT_RCI3M_lag1 + PET1M_lag1	0.86	0.85	-0.01	No
17	VCI1M_lag1 + TAMSAT_RFE3M_lag1 + TCI1M_lag1	0.86	0.85	-0.01	No
18	VCIdekad_lag1 + TAMSAT_RCI3M_lag1 + SPEI1M_lag1	0.86	0.85	-0.01	No
19	VCI3M_lag1 + TAMSAT_RCI3M_lag1 + SPEI3M_lag1	0.8	0.84	0.04	No
20	VCIdekad_lag1 + TAMSAT_SPI1M_lag1 + LST1M_lag1	0.86	0.84	-0.02	No
21	VCI1M_lag1 + TAMSAT_SPI1M_lag1 + PET1M_lag1	0.86	0.84	-0.02	No
22	VCI1M_lag1 + TAMSAT_SPI1M_lag1 + TCI1M_lag1	0.86	0.84	-0.02	No
23	VCIdekad_lag1 + TAMSAT_RCI3M_lag1 + LST1M_lag1	0.86	0.84	-0.02	No
24	VCI1M_lag1 + TAMSAT_RCI3M_lag1 + PET1M_lag1	0.86	0.84	-0.02	No
25	VCIdekad_lag1 + TAMSAT_RCI1M_lag1 + TCI1M_lag1	0.86	0.84	-0.02	No
26	VCIdekad_lag1 + TAMSAT_SPI3M_lag1 + SPEI1M_lag1	0.85	0.84	-0.01	No
27	VCIdekad_lag1 + TAMSAT_RFE1M_lag1 + LST1M_lag1	0.86	0.84	-0.02	No
28	VCI1M_lag1 + TAMSAT_RCI3M_lag1 + LST1M_lag1	0.86	0.84	-0.02	No
29	VCI1M_lag1 + TAMSAT_RFE1M_lag1 + LST1M_lag1	0.85	0.84	-0.01	No
30	VCIdekad_lag1 + TAMSAT_RFE1M_lag1 + SPEI1M_lag1	0.86	0.84	-0.02	No

Table A3. Performance of the top 30 SVR models in training and validation ordered by descending R² in the validation dataset.

No	Model	R ² Training	R ² Validation	Overfit Index	Overfit
1	VCIdekad_lag1 + TAMSAT_SPI3M_lag1 + TCI1M_lag1	0.85	0.86	0.01	No
2	VCIdekad_lag1 + TAMSAT_RCI3M_lag1 + TCI1M_lag1	0.85	0.86	0.01	No
3	VCI3M_lag1 + TAMSAT_RCI3M_lag1 + SPEI1M_lag1	0.81	0.86	0.05	No
4	VCI3M_lag1 + TAMSAT_SPI3M_lag1 + SPEI1M_lag1	0.82	0.85	0.03	No
5	VCI1M_lag1 + TAMSAT_SPI3M_lag1 + TCI1M_lag1	0.85	0.85	0.00	No
6	VCIdekad_lag1 + TAMSAT_SPI3M_lag1 + PET1M_lag1	0.84	0.85	0.01	No
7	VCI1M_lag1 + TAMSAT_RCI3M_lag1 + TCI1M_lag1	0.85	0.85	0.00	No
8	VCIdekad_lag1 + TAMSAT_SPI1M_lag1 + TCI1M_lag1	0.85	0.85	0.00	No
9	VCIdekad_lag1 + TAMSAT_SPI3M_lag1 + LST1M_lag1	0.84	0.85	0.01	No
10	VCIdekad_lag1 + TAMSAT_RFE1M_lag1 + TCI1M_lag1	0.85	0.85	0.00	No
11	VCI1M_lag1 + TAMSAT_SPI3M_lag1 + PET1M_lag1	0.84	0.85	0.01	No
12	VCI1M_lag1 + TAMSAT_SPI3M_lag1 + LST1M_lag1	0.84	0.84	0.00	No
13	VCIdekad_lag1 + TAMSAT_SPI1M_lag1 + PET1M_lag1	0.84	0.84	0.00	No
14	VCI1M_lag1 + TAMSAT_RFE1M_lag1 + TCI1M_lag1	0.85	0.84	-0.01	No
15	VCI1M_lag1 + TAMSAT_SPI1M_lag1 + TCI1M_lag1	0.85	0.84	-0.01	No
16	VCIdekad_lag1 + TAMSAT_RCI1M_lag1 + TCI1M_lag1	0.84	0.84	0.00	No
17	VCIdekad_lag1 + TAMSAT_RCI3M_lag1 + LST1M_lag1	0.84	0.84	0.01	No
18	VCIdekad_lag1 + TAMSAT_RFE3M_lag1 + TCI1M_lag1	0.84	0.84	0.00	No
19	VCIdekad_lag1 + TAMSAT_RFE1M_lag1 + LST1M_lag1	0.84	0.84	0.00	No
20	VCIdekad_lag1 + TAMSAT_SPI1M_lag1 + LST1M_lag1	0.84	0.84	0.00	No
21	VCI3M_lag1 + TAMSAT_RCI3M_lag1 + TCI1M_lag1	0.83	0.84	0.01	No
22	VCIdekad_lag1 + TAMSAT_RCI3M_lag1 + SPEI1M_lag1	0.83	0.84	0.01	No
23	VCIdekad_lag1 + TAMSAT_SPI3M_lag1 + SPEI1M_lag1	0.84	0.84	0.00	No
24	VCIdekad_lag1 + TAMSAT_RCI3M_lag1 + PET1M_lag1	0.83	0.84	0.01	No
25	VCI1M_lag1 + TAMSAT_RCI3M_lag1 + LST1M_lag1	0.83	0.84	0.01	No
26	VCI1M_lag1 + TAMSAT_SPI1M_lag1 + PET1M_lag1	0.84	0.84	0.00	No
27	VCI3M_lag1 + TAMSAT_RCI3M_lag1 + SPEI3M_lag1	0.79	0.83	0.04	No
28	VCI1M_lag1 + TAMSAT_RFE3M_lag1 + TCI1M_lag1	0.84	0.83	-0.01	No
29	VCIdekad_lag1 + TAMSAT_SPI3M_lag1 + EVT1M_lag1	0.83	0.83	0.00	No
30	VCI1M_lag1 + TAMSAT_RCI1M_lag1 + TCI1M_lag1	0.84	0.83	0.00	No

Appendix A.3. Additional Results on Variable Selection

The AIC is presented in Table A4 for the TAMSAT and CHIRPS variables.

Table A4. Akaike information criterion for variables selection.

Variable	Df	Deviance	AIC
TAMSAT_SPI3M_lag1	1	19.596	-683.66
TAMSAT_RCI3M_lag1	1	20.532	-646.53
CHIRPS_RCI3M_lag1	1	24.155	-517.16
CHIRPS_SPI3M_lag1	1	26.279	-450.07
TAMSAT_RFE3M_lag1	1	27.075	-426.32
CHIRPS_RFE3M_lag1	1	29.731	-351.83
TAMSAT_SPI1M_lag1	1	30.078	-342.59
CHIRPS_RCI1M_lag1	1	30.763	-324.68
TAMSAT_RCI1M_lag1	1	30.944	-319.99
CHIRPS_SPI1M_lag1	1	31.456	-306.93
TAMSAT_RFE1M_lag1	1	32.341	-284.85
CHIRPS_RFE1M_lag1	1	33.161	-264.93

The top two variables are shown to be TAMSAT variables, SPI3M and RCI3M, respectively. The same results are achieved by the use of the relative variable importance as presented in Table A5.

Table A5. Importance of variables by Partitioned R².

Variable	Relative Importance
TAMSAT_SPI3M_lag1	0.283
TAMSAT_RCI3M_lag1	0.197
CHIRPS_RCI3M_lag1	0.147
CHIRPS_SPI3M_lag1	0.096
TAMSAT_RFE3M_lag1	0.077
TAMSAT_SPI1M_lag1	0.050
CHIRPS_RFE3M_lag1	0.045
TAMSAT_RCI1M_lag1	0.029
CHIRPS_RCI1M_lag1	0.028
CHIRPS_SPI1M_lag1	0.024
TAMSAT_RFE1M_lag1	0.015
CHIRPS_RFE1M_lag1	0.009

The average relative importance for TAMSAT variables (0.11) shows them to be higher ranked over CHIRPS variables (0.06).

Appendix A.4. Multi-Collinearity of Predictor Variables

An investigation of possible multi-collinearity between the pairs of the predictor variables is provided in Figure A1.

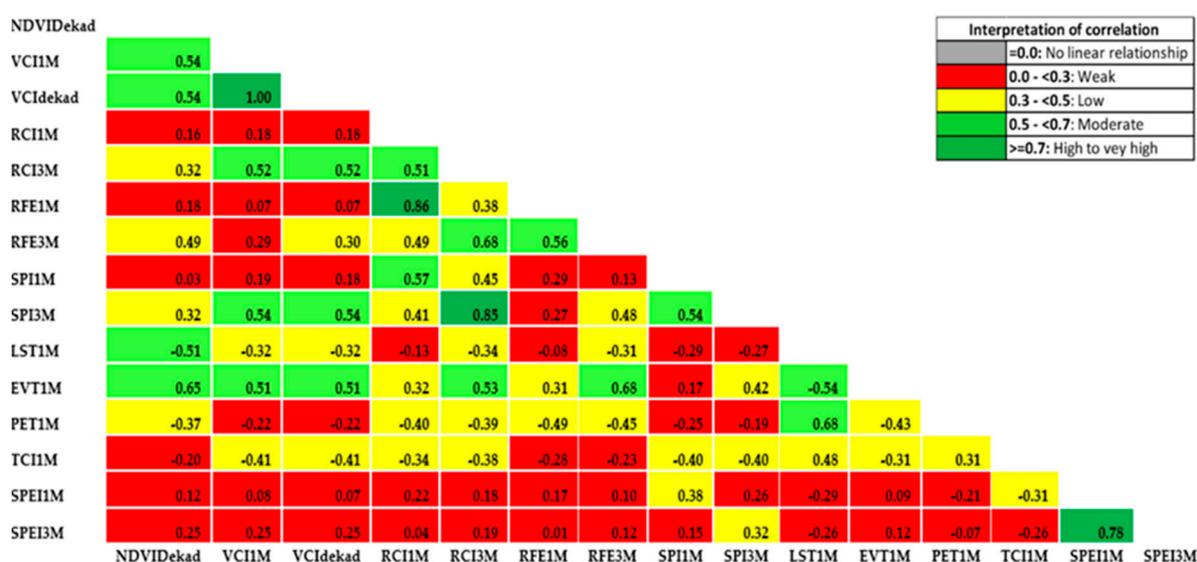


Figure A1. The collinearity matrix for X (predictors) variables. The correlations are categorized such that high correlations are colored green, moderate correlations are colored yellow, and low correlations are colored red.

From the correlation matrix (Figure A1), the following are observed: (1) relatively high correlation (of up to 1 for VCI1M and NDVIDekad) between vegetation datasets. The assumption not to use multiple vegetation variables in the same model is thus justified. (2) SPI and RCI are highly correlated even though, in general, the pairings of precipitation data could be used in the same model. (3) The water balance variables of LST, EVT, PET, TCI, and SPEI have acceptable correlation coefficients with vegetation and precipitation pairings. Grouping the variables in the modeling process into precipitation, vegetation and water balance variables is thus a sound assumption.

Appendix A.5. Maps Illustrating the Performance of the Heterogeneous Stacked Ensemble Classifier

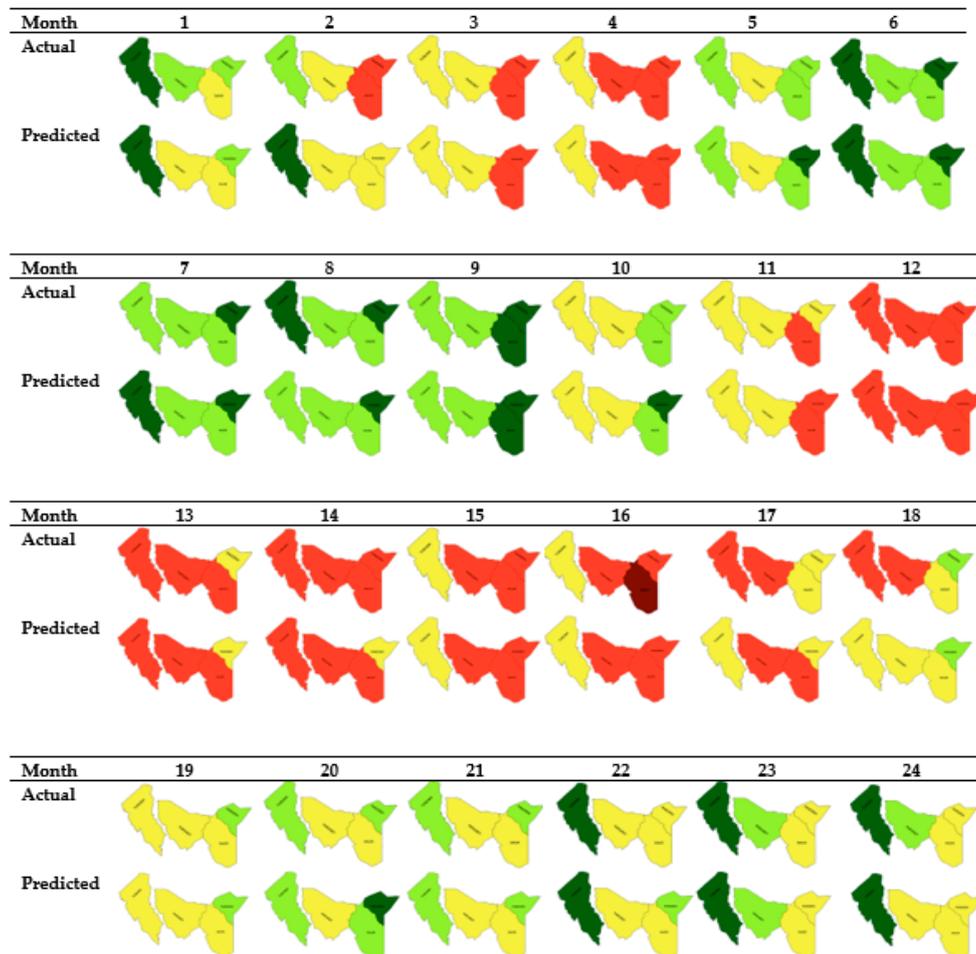


Figure A2. Performance of the heterogeneous stacked ensemble classifier showing the actual vegetation deficit classes against the predicted classes for each of the counties for the 24 months of the test data covering the period 2016–2017. The vegetation deficit classes are: Above Normal (Dark Green), Normal (Green), Moderate (Yellow), Severe (Red) and Extreme (Dark Red).

References

1. Bordi, I.; Fraedrich, K.; Petitta, M.; Sutera, A. Methods for predicting drought occurrences. In Proceedings of the 6th International Conference of the European Water Resources Association, Menton, France, 7–10 September 2005; pp. 7–10.
2. UNOOSA. Data Application of the Month: Drought Monitoring. UN-SPIDER. 2015. Available online: <http://www.un-spider.org/links-and-resources/data-sources/daotm-drought> (accessed on 11 November 2017).
3. Government of Kenya. Kenya Post-Disaster Needs Assessment: 2008–2011 Drought. 2012. Available online: http://www.gfdr.org/sites/gfdr/files/Kenya_PDNA_Final.pdf (accessed on 9 November 2018).
4. Cody, B.A. *California Drought: Hydrological and Regulatory Water Supply Issues*; DIANE Publishing: Collingdale, PA, USA, 2010.
5. Ding, Y.; Hayes, M.J.; Widhalm, M. Measuring economic impacts of drought: A review and discussion. *Disaster Prev. Manag. Int. J.* **2011**, *20*, 434–446. [[CrossRef](#)]
6. Rembold, F.; Atzberger, C.; Savin, I.; Rojas, O. Using low resolution satellite imagery for yield prediction and yield anomaly detection. *Remote Sens.* **2013**, *5*, 1704–1733. [[CrossRef](#)]

7. AghaKouchak, A.; Farahmand, A.; Melton, F.S.; Teixeira, J.; Anderson, M.C.; Wardlow, B.D.; Hain, C.R. Remote sensing of drought: Progress, challenges and opportunities. *Rev. Geophys.* **2015**, *53*, 452–480. [[CrossRef](#)]
8. Mishra, A.K.; Singh, V.P. A review of drought concepts. *J. Hydrol.* **2010**, *391*, 202–216. [[CrossRef](#)]
9. Klisch, A.; Atzberger, C. Operational Drought Monitoring in Kenya Using MODIS NDVI Time Series. *Remote Sens.* **2016**, *8*, 267. [[CrossRef](#)]
10. Brown, J.; Howard, D.; Wylie, B.; Frieze, A.; Ji, L.; Gacke, C. Application-ready expedited MODIS data for operational land surface monitoring of vegetation condition. *Remote Sens.* **2015**, *7*, 16226–16240. [[CrossRef](#)]
11. Svoboda, M.; LeComte, D.; Hayes, M.; Heim, R.; Gleason, K.; Angel, J.; Rippey, B.; Tinker, R.; Palecki, M.; Stooksbury, D.; et al. The drought monitor. *Bull. Am. Meteorol. Soc.* **2002**, *83*, 1181–1190. [[CrossRef](#)]
12. Ali, Z.; Hussain, I.; Faisal, M.; Nazir, H.M.; Hussain, T.; Shad, M.Y.; Shoukry, A.M.; Hussain Gani, S. Forecasting Drought Using Multilayer Perceptron Artificial Neural Network Model. *Adv. Meteorol.* **2017**, *2017*. [[CrossRef](#)]
13. Khadr, M. Forecasting of meteorological drought using hidden Markov model (case study: The upper Blue Nile river basin, Ethiopia). *Ain Shams Eng. J.* **2016**, *7*, 47–56. [[CrossRef](#)]
14. Enenkel, M.; Steiner, C.; Mistelbauer, T.; Dorigo, W.; Wagner, W.; See, L.; Atzberger, C.; Schneider, S.; Rogenhofer, E. A combined satellite-derived drought indicator to support humanitarian aid organizations. *Remote Sens.* **2016**, *8*, 340. [[CrossRef](#)]
15. Hao, Z.; AghaKouchak, A. Multivariate standardized drought index: A parametric multi-index model. *Adv. Water Resour.* **2013**, *57*, 12–18. [[CrossRef](#)]
16. Tadesse, T.; Demisse, G.B.; Zaitchik, B.; Dinku, T. Satellite-based hybrid drought monitoring tool for prediction of vegetation condition in Eastern Africa: A case study for Ethiopia. *Water Resour. Res.* **2014**, *50*, 2176–2190. [[CrossRef](#)]
17. Adede, C.; Oboko, R.; Wagacha, P.W.; Atzberger, C. A Mixed Model Approach to Vegetation Condition Prediction Using Artificial Neural Networks (ANN): Case of Kenya's Operational Drought Monitoring. *Remote Sens.* **2019**, *11*, 1099. [[CrossRef](#)]
18. Cortes, C.; Kuznetsov, V.; Mohri, M. Ensemble methods for structured prediction. In Proceedings of the International Conference on Machine Learning, Beijing, China, 21–26 June 2014; pp. 1134–1142.
19. Dietterich, T.G. Ensemble methods in machine learning. In *International Workshop on Multiple Classifier Systems*; Springer: Berlin/Heidelberg, Germany, 2000; pp. 1–15.
20. Opitz, D.; Maclin, R. Popular ensemble methods: An empirical study. *J. Artif. Intell. Res.* **1999**, *11*, 169–198. [[CrossRef](#)]
21. Re, M.; Valentini, G. Ensemble Methods. In *Advances in Machine Learning and Data Mining for Astronomy*; Chapman and Hall/CRC: Boca Raton, FL, USA, 2012; pp. 563–593.
22. Güneş, F.; Wolfinger, R.; Tan, P.Y. Stacked ensemble models for improved prediction accuracy. In Proceedings of the SAS Global Forum 2017 Conference, Orlando, FL, USA, 2–5 April 2017; SAS Institute Inc.: Cary, NC, USA, 2017; pp. 1–19.
23. Mendes-Moreira, J.; Soares, C.; Jorge, A.M.; Sousa, J.F.D. Ensemble approaches for regression: A survey. *ACM Comput. Surv. (Csur)* **2012**, *45*, 10. [[CrossRef](#)]
24. Nay, J.; Burchfield, E.; Gilligan, J. A machine-learning approach to forecasting remotely sensed vegetation health. *Int. J. Remote Sens.* **2018**, *39*, 1800–1816. [[CrossRef](#)]
25. Partalas, I.; Tsoumakas, G.; Vlahavas, I. *A Study on Greedy Algorithms for Ensemble Pruning*; Aristotle University of Thessaloniki: Thessaloniki, Greece, 2012.
26. Reid, S. *A Review of Heterogeneous Ensemble Methods*; Department of Computer Science, University of Colorado at Boulder: Boulder, CO, USA, 2007.
27. Escolano, A.Y.; Junquera, J.P.; Vázquez, E.G.; Riaño, P.G. A new Meta Machine Learning (MML) method based on combining non-significant different neural networks. In Proceedings of the ESANN, Bruges, Belgium, 23–25 April 2003; pp. 343–348.
28. Belayneh, A.; Adamowski, J.; Khalil, B.; Quilty, J. Coupling machine learning methods with wavelet transforms and the bootstrap and boosting ensemble approaches for drought prediction. *Atmos. Res.* **2016**, *172*, 37–47. [[CrossRef](#)]
29. Dzeroski, S.; Zenko, B. Is combining classifiers with stacking better than selecting the best one? *Mach. Learn.* **2004**, *54*, 255–273. [[CrossRef](#)]

30. Ganguli, P.; Reddy, M.J. Ensemble prediction of regional droughts using climate inputs and the SVM–copula approach. *Hydrol. Process.* **2014**, *28*, 4989–5009. [CrossRef]
31. Wardlow, B.D.; Tadesse, T.; Brown, J.F.; Callahan, K.; Swain, S.; Hunt, E. *Vegetation Drought Response Index an Integration of Satellite, Climate, and Biophysical Data*; CRC Press/Taylor & Francis: Boca Raton, FL, USA, 2012.
32. Tadesse, T.; Wardlow, B.D.; Hayes, M.J.; Svoboda, M.D.; Brown, J.F. The Vegetation Outlook (VegOut): A new method for predicting vegetation seasonal greenness. *GISci. Remote Sens.* **2010**, *47*, 25–52. [CrossRef]
33. Adhikari, R.; Agrawal, R.K. A homogeneous ensemble of artificial neural networks for time series forecasting. *arXiv* **2013**, arXiv:1302.6210.
34. Petrakova, A.; Affenzeller, M.; Merkurjeva, G. Heterogeneous versus homogeneous machine learning ensembles. *Inf. Technol. Manag. Sci.* **2015**, *18*, 135–140. [CrossRef]
35. De Oto, L.; Vrieling, A.; Fava, F.; de Bie, K.C. Exploring improvements to the design of an operational seasonal forage scarcity index from NDVI time series for livestock insurance in East Africa. *Int. J. Appl. Earth Obs. Geoinf.* **2019**, *82*, 101885. [CrossRef]
36. Didan, K. *MOD13Q1 MODIS/Terra Vegetation Indices 16-Day L3 Global 250m SIN Grid V006 [Data Set]*; NASA EOSDIS Land Processes DAAC: Sioux Falls, SD, USA, 2015. [CrossRef]
37. Didan, K. *MYD13Q1 MODIS/Aqua Vegetation Indices 16-Day L3 Global 250m SIN Grid V006 [Data Set]*; NASA EOSDIS Land Processes DAAC: Sioux Falls, SD, USA, 2015. [CrossRef]
38. Tarnavsky, E.; Grimes, D.; Maidment, R.; Black, E.; Allan, R.P.; Stringer, M.; Chadwick, R.; Kayitakire, F. Extension of the TAMSAT satellite-based rainfall monitoring over Africa and from 1983 to present. *J. Appl. Meteorol. Climatol.* **2014**, *53*, 2805–2822. [CrossRef]
39. Funk, C.C.; Peterson, P.J.; Landsfeld, M.F.; Pedreros, D.H.; Verdin, J.P.; Rowland, J.D.; Pedreros, P. *A Quasi-Global Precipitation Time Series for Drought Monitoring*; US Geological Survey Data Series; U.S. Geological Survey: Reston, VA, USA, 2014; Volume 832.
40. Wan, Z.; Hook, S.; Hulley, G. *MOD11A2 MODIS/Terra Land Surface Temperature/Emissivity 8-Day L3 Global 1km SIN Grid V006 [Data Set]*; NASA EOSDIS Land Processes DAAC: Sioux Falls, SD, USA, 2015. [CrossRef]
41. Running, S.; Mu, Q.; Zhao, M. *MOD16A2 MODIS/Terra Net Evapotranspiration 8-Day L4 Global 500m SIN Grid V006 [Data Set]*; NASA EOSDIS Land Processes DAAC: Sioux Falls, SD, USA, 2017. [CrossRef]
42. Beguería, S.; Vicente-Serrano, S.M.; Reig, F.; Latorre, B. Standardized precipitation evapotranspiration index (SPEI) revisited: Parameter fitting, evapotranspiration models, tools, datasets and drought monitoring. *Int. J. Climatol.* **2014**, *34*, 3001–3023. [CrossRef]
43. Vicente-Serrano, S.M.; Beguería, S.; López-Moreno, J.I.; Angulo, M.; El Kenawy, A. A new global 0.5 gridded dataset (1901–2006) of a multiscalar drought index: Comparison with current drought index datasets based on the Palmer Drought Severity Index. *J. Hydrometeorol.* **2010**, *11*, 1033–1043. [CrossRef]
44. World Meteorological Organization (WMO). Standardized Precipitation Index User Guide. 2012. WMO-No. 1090. Available online: http://www.wamis.org/agm/pubs/SPI/WMO_1090_EN.pdf (accessed on 26 April 2019).
45. Atzberger, C.; Eilers, P.H. Evaluating the effectiveness of smoothing algorithms in the absence of ground reference measurements. *Int. J. Remote Sens.* **2011**, *32*, 3689–3709. [CrossRef]
46. Atkinson, P.M.; Jegathanan, C.; Dash, J.; Atzberger, C. Inter-comparison of four models for smoothing satellite sensor time-series data to estimate vegetation phenology. *Remote Sens. Environ.* **2012**, *123*, 400–417. [CrossRef]
47. Atkinson, P.M.; Tatnall, A.R. Introduction neural networks in remote sensing. *Int. J. Remote Sens.* **1997**, *18*, 699–709. [CrossRef]
48. Mountrakis, G.; Im, J.; Ogole, C. Support vector machines in remote sensing: A review. *ISPRS J. Photogramm. Remote Sens.* **2011**, *66*, 247–259. [CrossRef]
49. McKee, T.B.; Doesken, N.J.; Kleist, J. The relationship of drought frequency and duration to time scales. In Proceedings of the 8th Conference on Applied Climatology, Anaheim, CA, USA, 17–22 January 1993; American Meteorological Society: Boston, MA, USA, 1993; Volume 17, No. 22. pp. 179–183.
50. Kogan, F.N. Remote sensing of weather impacts on vegetation in non-homogeneous areas. *Int. J. Remote Sens.* **1990**, *11*, 1405–1419. [CrossRef]
51. Liu, W.T.; Kogan, F.N. Monitoring regional drought using the vegetation condition index. *Int. J. Remote Sens.* **1996**, *17*, 2761–2782. [CrossRef]

52. Huang, G.B. Learning capability and storage capacity of two-hidden-layer feedforward networks. *IEEE Trans. Neural Netw.* **2003**, *14*, 274–281. [[CrossRef](#)] [[PubMed](#)]
53. Riedmiller, M.; Braun, H. A direct adaptive method for faster backpropagation learning: The RPROP algorithm. In Proceedings of the IEEE International Conference on Neural Networks, San Francisco, CA, USA, 28 March–1 April 1993; Volume 1993, pp. 586–591.
54. Chen, C.S.; Lin, J.M. Applying Rprop neural network for the prediction of the mobile station location. *Sensors* **2011**, *11*, 4207–4230. [[CrossRef](#)] [[PubMed](#)]
55. Klisch, A.; Atzberger, C.; Luminari, L. Satellite-based drought monitoring in kenya in an operational setting. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2015**, *XL-7/W3*, 433–439. [[CrossRef](#)]
56. Meroni, M.; Fasbender, D.; Rembold, F.; Atzberger, C.; Klisch, A. Near real-time vegetation anomaly detection with MODIS NDVI: Timeliness vs. accuracy and effect of anomaly computation options. *Remote Sens. Environ.* **2019**, *221*, 508–521. [[CrossRef](#)]
57. Dinku, T.; Ceccato, P.; Grover-Kopec, E.; Lemma, M.; Connor, S.J.; Ropelewski, C.F. Validation of satellite rainfall products over East Africa’s complex topography. *Int. J. Remote Sens.* **2007**, *28*, 1503–1526. [[CrossRef](#)]
58. Elish, M.O. Assessment of voting ensemble for estimating software development effort. In Proceedings of the 2013 IEEE Symposium on Computational Intelligence and Data Mining (CIDM), Singapore, 16–19 April 2013; pp. 316–321.
59. Kocaguneli, E.; Kultur, Y.; Bener, A. Combining multiple learners induced on multiple datasets for software effort prediction. In Proceedings of the International Symposium on Software Reliability Engineering (ISSRE), Mysuru, India, 16–19 November 2009.
60. Shirzad, A.; Tabesh, M.; Farmani, R. A comparison between performance of support vector regression and artificial neural network in prediction of pipe burst rate in water distribution networks. *KSCE J. Civ. Eng.* **2014**, *18*, 941–948. [[CrossRef](#)]
61. Mokhtarzad, M.; Eskandari, F.; Vanjani, N.J.; Arabasadi, A. Drought forecasting by ANN, ANFIS, and SVM and comparison of the models. *Environ. Earth Sci.* **2017**, *76*, 729. [[CrossRef](#)]
62. Liu, B.; Wei, Y.; Zhang, Y.; Yang, Q. Deep Neural Networks for High Dimension, Low Sample Size Data. In Proceedings of the IJCAI, Melbourne, Australia, 19–25 August 2017; pp. 2287–2293.



© 2019 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).