Crop yield probability density forecasting via quantile random forest and Epanechnikov Kernel function.

Samuel Asante Gyamerah[a], Philip Ngare[b], Dennis Ikpe[c]

[a] *Pan African University, Institute for Basic Sciences, Technology, and Innovation, Kenya*
[b] *University of Nairobi, Kenya*
[c] *Michigan State University, USA*

Abstract

A reliable and accurate forecasting method for crop yields is very important for the farmer, the economy of a country, and the agricultural stakeholders. However, due to weather extremes and uncertainties as a result of increasing climate change, most crop yield forecasting models are not reliable and accurate. In this paper, a hybrid crop yield probability density forecasting method via quantile regression forest and Epanechnikov kernel function (QRF-SJ) is proposed to capture the uncertainties and extremes of weather in crop yield forecasting. By assigning probability to possible crop yield values, probability density forecast gives a complete description of the yield of crops. A case study using the annual crop yield of groundnut and millet in Ghana is presented to illustrate the efficiency and robustness of the proposed technique. The proposed model is able to capture the nonlinearity between crop yield and the weather variables via random forest. The values of prediction interval coverage probability and prediction interval normalized average width for the two crops show that the constructed prediction intervals cover the target values with perfect probability. The probability density curves show that QRF-SJ method has a very high ability to forecast quality prediction intervals with a higher coverage probability. The feature importance gave a score of the importance of each weather variable in building the quantile regression forest model. The farmer and other stakeholders are able to realize the specific weather variable that affect the yield of a selected crop through feature importance. The proposed method and its application on crop yield dataset is the first of its kind in literature.

Keywords: climate change, crop yield uncertainty, crop yield forecasting, quantile random forest, kernel density estimation, Epanechnikov kernel

1. Introduction

The agriculture sector is seen as one of the biggest emitter of greenhouse gases and concurrently a major sector that is affected by climate change. In reviewing the factors that affect crop growth, productivity, and yield, [1] indicated that soil moisture, the

☆correspondence: saasgyam@gmail.com

availability of soil nutrients, and solar radiation are the top three factors that limit the growth of crops and hence limit the yield of crops. Changes in the surface temperature, humidity, and rainfall affects the moisture content of the soil and the level of nutrients in the soil. Hence, there is a direct effect of climate change on crop growth, productivity and yield. It can therefore be stated that variations in the yield of are mostly affected by the change in weather. Even more is the effect of the uncertainties in the pattern of weather both between and within planting seasons on the crop production and yield. This has significantly affected the yield of most crops causing economic and food security risks in most developing and under-developed countries. Crop yields are less predictable than ever before because of the direct effect of weather events and changing weather patterns caused by climate change.

Weather variables are difficult to control especially for small-holder farmers in most developing and under-developed countries and mostly have great impact on the farming activities of these farmers. For this reason, an effective and reliable insurance is needed to hedge farmers and stakeholders from the peril of weather uncertianties. Traditional insurance for agricultural risk management is not patronized in most developing countries because of high premiums, loss adjustments, moral hazards, adverse selections, and complex information requirements [2]. However, weather derivatives and index based insurance such as area-yield and weather index insurance are seen as effective risk management tools in the agricultural sector for both small and large scale farmers in developing/under-developed countries. Accurate forecasting of crop yields is a principal component for the ratemaking process in the derivative and index-insurance markets. An accurate and mathematically tractable crop yield forecasting model, especially for crops with high out-of-sample forecasting propensity is important for the farmer, policy-makers, the government, field managers, and industry players in decision making process [3]. On the part of the government, this will enable an effective planning to avoid food shortage and if possible governments can arrange for food imports rather than seeking for emergency food assistance. For the industry players like the insurance and financial sector, crop yield forecasting helps in measuring crop loss' in advance. Consequently, fair premium rates and pricing of agricultural index insurance and weather derivatives can be determined. The farmer however is able to measure the future uncertainties of the farm produce and make effective plans for a set of possible outcomes and in particular precision farming. Crop yield forecasting is important for trade development policies and other humitarian assistance connected to food security.

To improve the performance of the methods used in crop yield forecasting, a lot of research have been done in recent decades. A number of literature based on statistical models have been used to predict the yield of crops [4, 5, 6]. These literatures have serve as a substitute to process-based models, which always involve a comprehensive data on the conditions of the soil, cultivar, and management. [7] used different time series models (simple and double exponential smoothing, Damped-Trend Linear Exponential Smoothing and autoregressive moving averages (ARMA)) to predict maize yield in five communities in Ghana. The authors concluded that the ARMA model was more robust than the other time-series models. However, time series models such as moving averages, simple and double exponential smoothing, quadratic and linear regression perform poorly in predicting crop yields [8, 9]. These statistical predictions suffer from different sources of error like variations in weather variables. In most statistical methods, there are

always little or no interaction between the features for prediction. However, crop yield and weather variables are highly nonlinear and there are interactions between weather variables. Hence, using statistical methods will be computationally costly and may not lead to an optimal performance especially when there are cases of extreme events. Alternative to statistical models, are the emergence of machine learning (ML) techniques such as random forest, support vector regression, and neural networks. These ML techniques are able to capture the nonlinearity of crop yields and its predictors [10, 11, 12]. Hybrid methods between statistical and machine learning approach are seen to improve the accuracy of predictions in most forecasting problems [13, 14].

Generally, most statistical and machine learning literatures on crop yield forecasting are based on point forecasts [5, 7, 10, 11]. Point forecasts give an estimate of the future crop yield for each time horizon and do not convey any information about the uncertainty of the predictions. Different from point forecasting is interval prediction. Interval prediction tries to build a well-calibrated lower and upper bounds of the future unknown predictions with a prescribed probability $(1 - \tau)$ called the confidence level. Due to the increased uncertainty of weather in recent years as a result of climate change [15, 16], point and interval prediction are not able to predict the yield of crops accurately. The uncertainties of weather variables directly influence the development of crops and hence the quantity and quality of crop yields. It is therefore imperative to quantify the probable uncertainties associated with crop yield forecasts. Contrary to point and interval prediction, probability density forecasting. Probability density forecasting gives a new approach to solve this forecasting problem in the midst of uncertainties. Probability density forecasting quantifies the uncertainty and give estimates of the complete probability distribution of the future crop yield. Despite the importance of probability density forecasting, there is no empirical evidence to crop yield forecasting in literature.
In probability density forecasting, Kernel density estimation (KDE) is very important in the density estimation process. KDE is a non-parametric method of estimating the distribution of a dataset without prior assumptions of the datasets. Appropriate choice of bandwidth for a kernel density estimator is of crucial importance to the density function of random variables. To obtain a complete crop yield probability density curve, Epanechnikov kernel function and solve-the-equation plug-in approach of Sheather and Jones (SJ) bandwidth selection method are combined with QRF model. Our proposed method (from hence QRF-SJ) will help to obtain a complete conditional probability density in different time horizons by selecting a suitable bandwidth and kernel function

Quantile regression (QR) can be used to to construct a nonparametric probability density forecasting. Given one or more covariates, QR generalizes the theory of a univariate quantile to a conditional quantile. Because of the robustness of QR in handling outliers in explained measurements, it is widely used for regression analysis in the areas of econometrics and statistics [17]. Conventional linear QR is however unable to deal with complex non-linear problems [18]. To explore non-linear functions for QR, [19] proposed a quantile random forest (QRF) model, which combines the advantages of random forest and quantile regression models. In furtherance to the application of Meinshausen proposed model, [14] proposed a hybrid semi-parametric quantile regression forest to estimate the non-linear relationship in multi-period value-at-risk. They concluded that their proposed method was more accurate compared to common distributions like normal distribution. In the area of medicine, [13] applied quantile regression forest to Cancer Cell Line Ency-

clopedia (CCLE) dataset to give a point and interval prediction. The quantile regression forest improved the accuracy of prediction of drug response. However, these literatures [e.g. 13] are not able to obtain the probability density functions of the future response variable in a single QRF model. Hence the need to combine quantile regression forest and kernel density estimation to estimate the probability density functions of future crop yields.

Due to the nonlinearity of crop yield and its predictors (weather variables), we propose a forecasting model by combining quantile random forest and kernel density estimation. The contributions of this paper are: 1) We implemment a comprehensive probabilistic crop yield forecasting method based on quantile random forest and Kernel density estimation (QRF-SJ). The full conditional probability density curve of future crop yields are illustrated and all the observed crop yield values are within the forecasted probability density curve. 2) Two interval prediction evaluation metrics (prediction intervals coverage probability and prediction interval normalized average width) are used to assess the performance of the proposed QRF-SJ. 3) We demonstrated the superiority and feasibility of the proposed QRF-SJ model using groundnut and millet as case studies. 4) The feature/variable importance (a score that gives the effectiveness of each feature in predicting the yield of the crop) is presented. This gives information to agricultural stakeholders about the imporatant features that affect the yield of a specific crop.

The rest of the paper is organized as follows: section 2 explains the mathematical background of quantile regression, random forest, quantile random forest, and kernel density estimation. The model evaluation metrics for point prediction and prediction intervals are presented in section 3. In section 4, we consider a case study to show the effectiveness and superiority of the proposed method using crop yield dataset. The conclusion and future work are outlined in section 5.

2. Quantile Random Forest based on Kernel Density Estimation

This section provides a comprehensive explanation used in developing the probabilistic crop yield forecasting. Generally, three steps are used for the probabilistic crop yield forecasting. Firstly, the dataset are divided into a training and testing dataset. In the second phase, the training dataset is used to train the quantile regression forest (QRF) model. The QRF model is then used to predict the testing data on different quantiles. In the final step, the probability density function are obtained by using kernel density estimation with Epanechnikov kernel function and SJ bandwith selection. Our model has not been applied in other field of research or in the agriculture sector so far. It is therefore the first of its kind in literature.

2.1. Quantile Regression (QR)

Conventional linear regression models make a summary of the average relation betweeen explanatory variables $X = [X_1, X_2, \cdots, X_k]'$ and a response variable Y depending on the conditional mean function $\mathbb{E}(Y|X)$. It gives a partial estimate of the relationship, as it might be needed in recounting the relationship of distinct points in the conditional distribution of Y. Contrary to the conventional linear regression, QR gives the quantiles of the conditional distribution of Y as a function of X [20]. That is, QR provides much detail information about the distribution of Y than conventional linear regression

model. By using QR, we can make a good inference on the distribution of the predicted values. Machine learning techniques that are based on quantile regression such as the quantile random forest have an extra advantage of been able to predict non-parametric distributions. A QR problem can be formulated as;

$$q_Y(\tau \mid X) = \mathbf{X}_i'\boldsymbol{\beta}_\tau \tag{1}$$

where $q_Y(\tau \mid \cdot)$ is the conditional τ_{th} quantile of crop yield variables Y, X are the explanatory variables or regressors, and $\beta_\tau = [\beta_\tau(0), \beta_\tau(1), \cdots \beta_\tau(k)]'$ is a vector of values of quantile τ. By minimizing the loss function of a specific τ_{th} quantile, vector of values can be evaluated,

$$\begin{aligned}\min_{\beta} \sum_{i=1}^{N} \rho_\tau(Y_i - \mathbf{X}_i'\boldsymbol{\beta}) =& \min_{\beta}\Big[\sum_{i:Y_i \geq \mathbf{X}_i'\boldsymbol{\beta}} \tau \mid Y_i - \mathbf{X}_i'\boldsymbol{\beta} \mid + \sum_{i:Y_i < \mathbf{X}_i'\boldsymbol{\beta}} (1-\tau) \mid Y_i - \mathbf{X}_i'\boldsymbol{\beta} \mid \Big], \\ =& \min_{\beta}\Big[\sum_{i} \mid \tau - \mathbf{1}_{y_i < \mathbf{X}_i'\boldsymbol{\beta}} \mid (Y_i - \mathbf{X}_i'\boldsymbol{\beta})\Big]\end{aligned} \tag{2}$$

Where $\mathbf{1}$ is the indicator function, N is the size of the sample data, and $\mathbf{X}_i = (x_{1i}, x_{2i}, x_{3i}, \cdots, x_{ki})$ are the independent variables. Now, consider the distribution of a discrete random variable Y_i with a less-than-well-behaved density, then the conditionbal density function at the τ_{th} quantile given x_i is defined as

$$q_\tau(x) = \inf\{y : F_\tau\big(y \mid X = x\big) \geq \tau\} \tag{3}$$

where $F_\tau\big(y \mid X = x\big)$ is the distribution function for Y_i conditional on X_i.

2.2. Random Forest (RF)

RF is a binary tree machine learning algorithm and a non-parametric method for regression and classification problems. The aim of RF is to predict the square integrable random response $Y \in \mathbb{R}$ by computing the regression function $c(x) = \mathbb{E}\big[Y|\mathbf{X} = \mathbf{x}\big]$. Assume a training dataset $D_n = \{\big(\mathbf{X}_i, y_i\big)_{i=1}^{n} | \mathbf{X}_i \in \mathbb{R}^M, y \in \mathbb{R}\}$ of an observed dataset is randomly selected from an (unknown) probaility distribution $(\mathbf{x}_i, y_i) \sim (\mathbf{X}, Y)$. We seek to use D_n to build an estimate. Where n is the total number of training samples and M is the total number of features.

Suppose θ is the parameter that determines a specific splitting node of RF regression trees. Let $T(\theta)$ be the decision tree under consideration. Consider the conditional distribution of Y given $X = x$ depending on the decision tree and the event that x can be determined at a point on the decision tree R. If there is one and only one leaf node which satisfies x and is represented as $\ell(x, \theta)$ for the decision tree $T(\theta)$, then the prediction of a single tree $T(\theta)$ for a point x in the observed data is the average over the observed values in $\ell(x, \theta)$. The weight vector $w_n(x, \theta)$ for the total observation in $\ell(x, \theta)$ is given as

$$w_n(x, \theta) = \frac{\mathbf{1}_{\{X_n \in R_{\ell(x,\theta)}\}}}{\{p : X_p \in \Omega_{\ell(x,\theta)}\}}. \tag{4}$$

Where $\sum_{i=1}^{n} w_n(x,\theta) = 1$, and the prediction of the single tree $Y \mid X = x$ is the weighted average of true observation $Y_i (i = 1, 2, \cdots, n)$,

$$\hat{\vartheta}(x) = \sum_{i=1}^{n} w_n(x,\theta) Y_n \tag{5}$$

RF uses the average prediction of k individual trees, each built with an i.i.d. vector θ_i $i = 1, 2, 3, \cdots, k$ to approximate $\mathbb{E}(Y \mid X = x)$. Denote $w_n(x)$ as the average of of $w_n(\theta)$ over the ensemble of trees,

$$w_n(x) = \frac{1}{k} \sum_{i=1}^{K} w_n(x, \theta_i) \tag{6}$$

Then, the prediction of RF is

$$\hat{\vartheta}(x) = \sum_{n=1}^{N} w_n(x) Y_n \tag{7}$$

RF estimates the conditional mean of Y, given $X = x$, by weighting the sum of all the observations. The weight is larger when the conditional distribution of Y given $X = X_n$, is identical to the conditional distribution of Y given $X = x$ [21].

RF depends on some parameters for optimal performance. The number of tress (ntree) to grow and the number of variables that is sampled as candidates for each split (mtry). For regression problems, mtry $= \frac{M}{3}$, where M=number of features for prediction. Apart from using RF for quantile regression forest, we shall use RF for feature importance and partial dependence plots (PDP). For the feature importance, we use the percentage mean decreasing accuracy ("%IncMSE") to know the importance of each of the features in building the prediction model. The PDP illustrates how the RF model predictions are affected by each feature assuming the rest of the features in the RF model are controlled.

2.3. Quantile Random Forest (QRF)

Conventional RF predict values in individual leaf node, which is considered as the sample mean in the leaf node. This can lead to biasness extreme values in the data samples can be over- or under-estimated. To improve th accuracy of the prediction when there are extreme values in the sample dataset, the median can be used. Hence, the median is used for point prediction in QRF model.

QRF is a robust, non-linear, and non-parametric regression method based on random forests method for determining conditional quantiles [19]. QRF gives an approximation of the complete conditional distribution. Just like RF, QRF is a set of binary regression trees. However, for each leaf node of the tree, QRF evaluates the estimated distribution $F(y \mid X = x) = P(Y \leq y) \mid X = x) = \mathbb{E}(1_{\{Y \leq y\}} \mid X = x)$ as alternative to only the mean of Y values in RF. Given a probability p, the quantile $q_\tau(X)$ is evaluated as $\hat{q}_\tau(X = x_{new}) = \inf\{y : \hat{F}(y \mid X = x_{new}) > \tau\}$. The quantiles provide a comprehensive information on the distribution of Y as a function of the predictands (X) than only the conditional mean. For interval prediction,

$$\Big[q_{\tau_l}(X), q_{\tau_u}(X)\Big] = \Big[\inf\{y : \hat{F}(y \mid X = x) \geq \tau_l\}, \inf\{y : \hat{F}(y \mid X = x) \geq \tau_u\}\Big] \tag{8}$$

where $\tau_l < \tau_u$ and $\tau_u - \tau_l = \alpha$, α is the probability that the predicted value fall within y to lie in the interval $[q_{\tau_l}(X), q_{\tau_u}(X)]$.

We define an approximation to the acummulated conditional probability $\mathbb{E}(\mathbf{1}_{\{Y \leq y\}} \mid X = x)$ by the weighted mean of all the observations of $\mathbf{1}_{\{Y \leq y\}}$ as,

$$F(y \mid X = x) = \sum_{n=1}^{N} w_n(x) \mathbf{1}_{\{Y_n \leq y\}}, \tag{9}$$

where $w_n(x)$ is the same weights as in random forests. By plugging $\hat{F}\big(y \mid X = x\big)$ into 3, the estimate $\hat{q}_\tau(x)$ of the conditional quantiles $q_\tau(x)$ are derived,

$$\hat{q}_\tau(x) = \hat{F}^{-1}(\tau) = \inf\{y : \sum_{n=1}^{N} w_n(x) \mathbf{1}_{\{Y_n \leq y\}} \geq p\} \tag{10}$$

2.4. Kernel density estimation using Epanechnikov Kernel function

Kernel density estimation (KDE) is a non-parametric method of estimating the probability density function (pdf) or regression functions. KDE is basically used for data smoothing. A Kernel density estimator at x for an observed independent and identically distributed (i.i.d.) and data $\mathbf{X} = (X_1, X_2, \cdots, X_n)$ drawn from an unknown distribution with an unknown density $f_X(x)$, is

$$\hat{f}(x; b) = \frac{1}{Nb} \sum_{i=1}^{N} K\Big(\frac{X_i - x}{b}\Big) \tag{11}$$

where N is the sample size, K is the Kernel function, $h > 0$ is the smoothing parameter also called bandwidth. The Kernel function is non-negative and is defined as $\int_{-\infty}^{\infty} K(x)dx = 1$. Gaussian, Rectangular, Uniform, Cosine, Epanechnikov, and Quartic are but some common examples of kernel functions used in literatures. Different results are obtained depending on the type of Kernel function used. In this study, we use, we use the Epanechnikov Kernel to build our QRF-SJ model. The Epanechnikov Kernel is defined as:

$$K(u) = \begin{cases} \frac{3}{4}(1 - u^2) & \mid u \mid \leq 1, \\ 0 & \text{otherwise.} \end{cases} \tag{12}$$

Where $u = \frac{X_i - x}{b}$. when building the QRF-SJSTE model. The choice of the Epanechnikov kernel is motivated because it has the lowest (asymptotic) mean square error (MSE) [22, 23].

Solve-The-Equation Plug-In Approach of Sheather and Jones (SJ)

Bandwidth determines the smoothness of the kernel density plot and is comparable to the binwidth in a histogram. The selection of a proper bandwidth is the most difficult problem in obtaining a good KDE [23]. A larger value bandwidth value causes over smoothing and a very small bandwidth value causes under smoothing. To get better results of kernel density estimator, this paper uses Sheather and Jones (SJ) solve-the-equation (SJ) bandwidth selector [see 24, 25] for estimating the bandwidth parameter.

To quantify the accuracy of the kernel density estimator, the asymptotic mean integrated squared error (AMISE) is used. AMISE is an approximation of mean integrated squared error (when $n \to \infty$ and $b = b(n) \to 0$) of $\hat{f}(x)$,

$$AMISE(\hat{f}_b(x)) = (nb)^{-1}R(K) + b^4 R(f'')\Big(\int x^2 K/2\Big)^2 \tag{13}$$

where the notation $R(g) = \int g^2(x)dx$ for a function g, $\int x^2 K = \int x^2 K(x)dx$, and f'' is the second derivative of f. The first and second term in equation 13 are the integrated variance and integrated squared bias respectively. A very small h results in a large integrated variance and a very large h results in a large integrated squareed bias.

Figure 1 is the schematic structure of the QRF-SJ model.

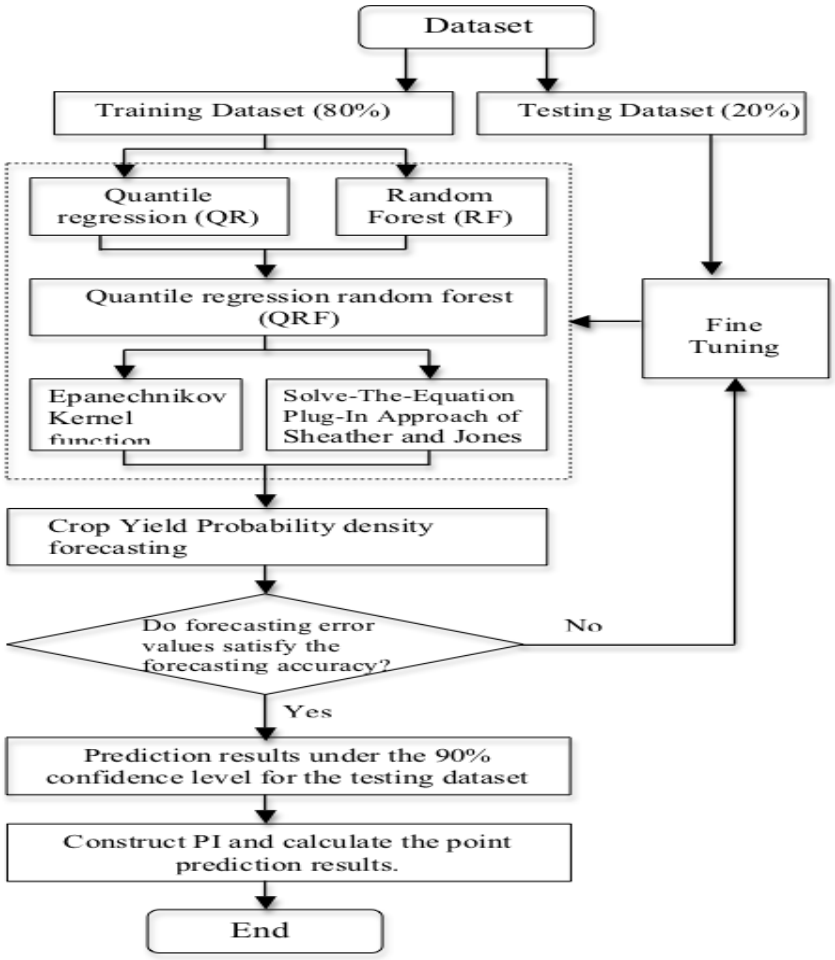

Figure 1: The flowchart of QRF-SJ probability density forecasting model

3. Model Evaluation Metrics

3.1. Evaluating point prediction errors

We use root mean squared error (RMSE), mean absolute percentage error (MAPE), mean squared error (MSE), and level of Accuracy to compare the performance of different forecasting models for point/deterministic forecasting.
RMSE estimates the residual betweeen the observed and the predicted values. The smaller the RMSE, the better the model.

$$RMSE(y_i, \hat{y}_i) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \tag{14}$$

MAPE gives the average of the absolute percentage errors. An optimal model has the lowest MAPE.

$$MAPE(y_i, \hat{y}_i) = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100\% \tag{15}$$

Just like RMSE and MAPE, the smaller the MSE, the better the model. Unlike MAPE, MSE is greatly influenced by outliers.

$$MSE(y_i, \hat{y}_i) = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \tag{16}$$

Accuracy is used to measure the precision of the model in predicting the observed dataset. The higher the accuracy values, the better the prediction model. It is given as

$$Accuracy(y_i, \hat{y}_i) = 1 - \frac{1}{N}\sum_{i=1}^{N}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \tag{17}$$

$y_i, \hat{y}_i$ are the actual and the predicted values of the crop yield.

3.2. Uncertainty of Prediction Intervals

Different metrics are used to evaluate the prediction intervals for the results obtained from the probability density forecasting; prediction intervals coverage probability (PICP) and prediction interval normalized average width (PINAW).
PICP is the percentage of the testing data that fall in the interval specified by the upper bound U_i and the lower bound L_i of the prediction interval (PI). A larger PICP indicates that most of the forecasted data fall within the PI. Generally, the value of the PICP should be greater than the nominal confidence level.

$$PICP = \frac{1}{N}\sum_{i=1}^{N} c_i$$

where N is the total number of years over the period of forecasting and c_i is a Boolean variable define as

$$c_i = \begin{cases} 1, & if \;\; y_i \in [L_i, U_i] \\ 0 & if \;\; y_i \notin [L_i, U_i] \end{cases}$$

If the quality of the forecast depends only on the PICP, the coverage probability can be artificially improved by increasing the range between the upper and the lower bound. However, a larger interval width is empirically not informative. To better evaluate the quality of the PIs, the width of the PI's must be measured. A narrow PI gives more information to the forecaster than a wider PI. Therefore, a normalized metric, PINAW which measures the average width of the PIs can be use. PINAW is defined as:

$$PINAW = \frac{1}{NR}\sum_{i=1}^{t}(U_i - L_i)$$

where R is the range of the underlying targets (difference between minimum and maximum targets)

4. Numerical Results

. To demonstrate the feasibility and suitability of the proposed QRF-SJ method, historical crop yield[1] data from 2000 to 2016 for different crops in Tamale metropolitan[2] (Northern region) were obtained from the Statistics, Research and Information Directorate (SRID) of the Ministry of Food and Africulture, Ghana. The Northern region of Ghana is much drier[3] as compared to the southern part of Ghana and agriculture contributes more than 75% of the economic activities in the metropolis. The region is considered to be the food basket of Ghana. Cowpea, cassava, groundnut, maize, millet, sorghum, rice, and yam are the major crops grown in this region. For the purpose of this research, two of the crops (groundnut and millet) are selected as a case study to evaluate the performance of our proposed model. These selected crops can be used as a proxy to create an area-yield index insurance instrument for the insurance sector.

Station based daily sunlight, humidity, precipitation, minimum temperature, maximum temperature and average temperature from 2000 to 2016 are obtained from the Ghana Meteorological Service. The k-nearest neighbors (KNN) algorithm was used for imputing missing data points in the datasets. KNN locates the k closest neighbors to the observed dataset with the missing data point and imputes the data point based on the non-missing data points in the neighbors. The datasets are then averaged to the same size as the crop yield dataset. Because of the scaling sensitivity of the inputs fed into most forecasting techniques, the variables for the inputs are set into an identical scale. The scale used is the min-max normalization which was set to be in the interval $[0, 1]$. The normalization is given as:

$$I_{NORM} = \frac{I - I_{MIN}}{I_{MAX} - I_{MIN}}$$

where I_{NORM} is the normalized numerical value; I_{MIN}, I_{MAX} is the minimum and maximum values of the inputs respectively.

[1] crop yield is defined as the harvested production of a crop per unit of the harvested area and is measured in metric ton per hectare (Mt/Ha).

[2] The capital town of the Tamale metropolitan is Tamale, which also happens to be the regional capital of the Northern region.

[3] This is because of its close proximity to the Sahara, and the Sahel.

To validate the model, we divided the dataset into a training (80%) and testing dataset (20%). That is, the dataset from 2000-2013 are selected as the training dataset and 2014-2016 are selected as the testing dataset.

Crop	Mean	Std	Min	Max	Skewness	Variance
Groundnut	1.22	0.50	0.50	1.90	0.05	0.25
Millet	1.19	0.30	0.72	1.70	0.10	0.09

Table 1: Summary Statistics of Groundnut and Millet Yield

4.1. Empirical results and analysis of the Models

In order to prove the superiority of the QRF, it is compared with some popular forcasting techniques like Radial Basis Neural Network (NN), Generalized Linear Model (GLM), Support Vector Regression with linear (SVR(linear)) and radial basis (SVR(radial)) kernel function. For the QRF, both the predicted median and mean of the crop yield are used as the point prediction for the testing dataset. The evaluation metrics using RMSE, MAPE, MSE, and Accuracy of these methods are given in Table 2. Figures 2 and 3 show the visual performance of the evaluation metrics of the forecasting techniques. Comparative to the four benchmark methods, it is evident that QRF (both mean and median prediction) performed better in predicting the yield of groundnut and millet. Quantitatively, the RMSE, MAPE, MSE, and Accuracy of predicting the yield of groundnut are $27.05\%, 24.25\%, 7.32\%$ and 62.73% respectively. The RMSE, MAPE, MSE, and Accuracy of millet yield prediction are $1.77\%, 1.02\%, 0.03\%$ and 97.37% respectively. It can be observed that the QRF median prediction is the same as the mean prediction. Generally, QRF is optimal in predicting the yield of all the two crops. For this reason, we conclude that QRF is the best forecasting technique for predicting the yield of the two crops as compared to the other benchmark forecasting techniques. This motivated us to use QRF for our probability density forecasting.

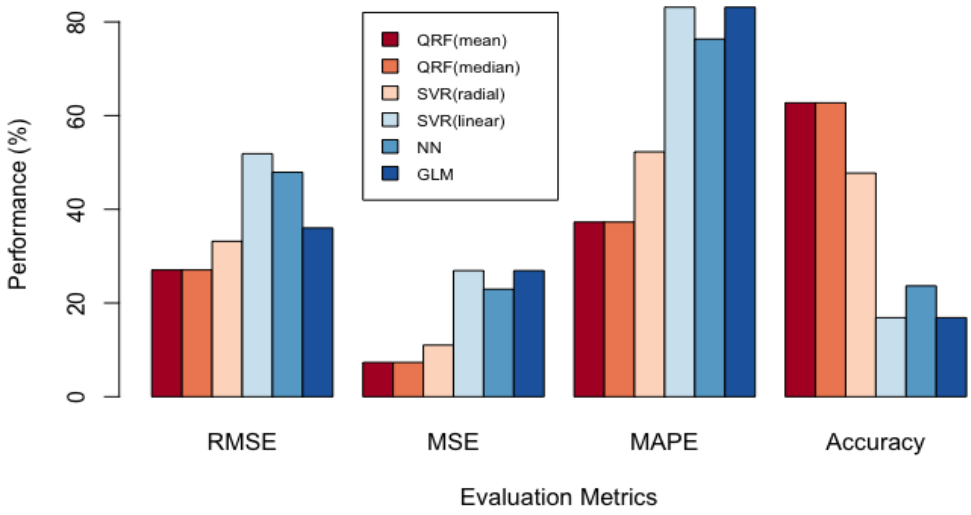

Figure 2: Bar chart of RMSE, MAPE, MSE, Accuracy of groundnut yield responses and predicted values by QRFs, SVR(radial), SVR(linear), NN, and GLM. QRF (mean): mean prediction of crop yield given weather features using QRF; QRF (median): median prediction of crop yield response using QRF; SVR(radial): prediction of crop yield using SVR radial basis kernel function; SVR(linear): prediction of crop yield using SVR linear kernel

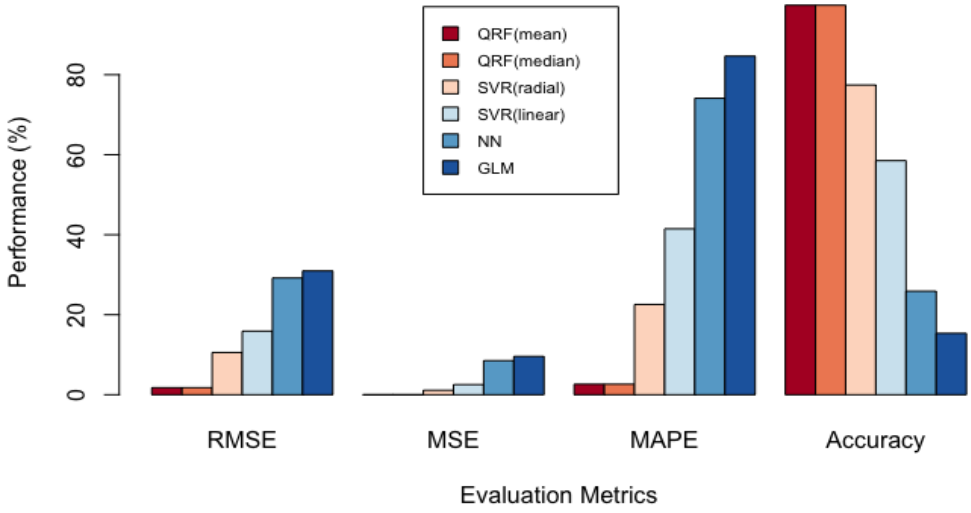

Figure 3: Bar chart of RMSE, MAPE, MSE, Accuracy of millet yield responses and predicted values by QRFs, SVR(radial), SVR(linear), NN, and GLM. QRF (mean): mean prediction of crop yield given weather features using QRF; QRF (median): median prediction of crop yield response using QRF; SVR(radial): prediction of crop yield using SVR radial basis kernel function; SVR(linear): prediction of crop yield using SVR linear kernel

Method	**Groundnut**				**Millet**			
	RMSE	**MSE**	**MAPE**	**Accuracy**	**RMSE**	**MSE**	**MAPE**	**Accuracy**
QRF(mean)	27.05	7.32	37.27	62.73	1.77	0.03	2.63	97.37
QRF(median)	27.05	7.32	37.27	62.73	1.77	0.03	2.63	97.37
SVR (radial)	33.18	11.01	52.27	47.73	10.56	1.11	22.57	77.43
SVR (linear)	51.87	26.91	83.13	16.87	15.90	2.53	41.48	58.52
NN	47.91	22.95	76.34	23.66	29.18	8.52	74.12	25.88
GLM	36.02	12.97	58.25	41.75	30.95	9.58	84.65	15.35

Table 2: Evaluation metrics using RMSE (%), MAPE (%), MSE (%), and Accuracy (%) of point prediction via QRFs, SVR(radial), SVR(linear), NN, and GLM.

To show the satisfactory performance of the proposed QRF-SJ model, PICP and PINAW are used as the evaluation metrics. The performance of the model is presented in Table 3. It is clear from the table that the constructed PI cover the target values with perfect probability. This is very important for effective decision making process. For both groundnut and millet, the PICP was evaluated to be 100%. The PINAW of groundnut is however smaller than that of millet. Considering the high variance of the

yield of groundnut in table 1, the probabilistic performance of our proposed method is sufficient.
To study the performance of the prediction in an intuitive way, a visual representation of the prediction results for the two crops are presented. Figures 4 and 5 give the point forecast and the prediction interval for the QRF-SJ of groundnut and millet respectively. To assess the performance of point forecast for the QRF-SJ model in a quantitative way, the model performance metrics over the predicted period for the two crops are presented in table 4. It is clear from figures 4 and 5 that the target values lie within the prediction interval. We can therefore conclude that the proposed model captures the uncertainty of the two crop yields accurately.
A visual representation of the probability density curve for the predictions based on QRF-SJ for the yield of groundnut and millet is presented in figures 6 and 7 respectively. The actual crop yield for the specific year are presented in orchid, blue, and red dots for 2014, 2015, and 2016 respectively. In figures 6 and 7, the respective actual crop yields of groundnut and millet for each of the predicted year fall within the predicted region of the forecast distribution. The probability density curve gives the complete probability distribution of the future crop yield and hence the uncertainty of the forecasting can be quantified.

Eventhough there are different kernels in kernel density estimation that can capture the nonlinearity of the crop yield and weather variables, we chose to use a relatively simple and conventional kernel, the cosine kernel.

Crop	**Confidence level (%)**	**PICP (%)**	**PINAW (%)**
Groundnut	90	100	12.65
Millet	90	100	16.76

Table 3: Prediction Interval evaluation metrics of QRF-SJ method

Crop	**Confidence level (%)**	**MAPE (%)**	**RMSE (%)**	**MSE (%)**	**Accuracy (%)**
Groundnut	90	27.50	28.947	13.96	76.59
Millet	90	1×10^{-14}	0.72	7.83	98.03

Table 4: Evaluation metrics using MAPE, RMSE, and MSE of point prediction via QRF-SJ for testing data

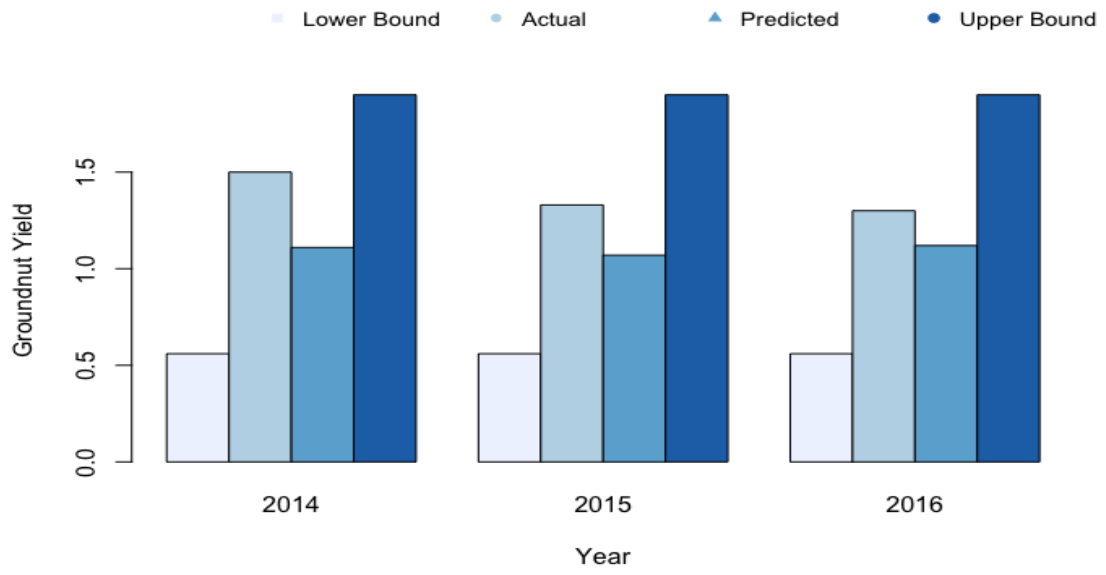

Figure 4: Prediction results based on QRF-SJ probability density forecasting model from 2014-2016 for the yield of groundnut

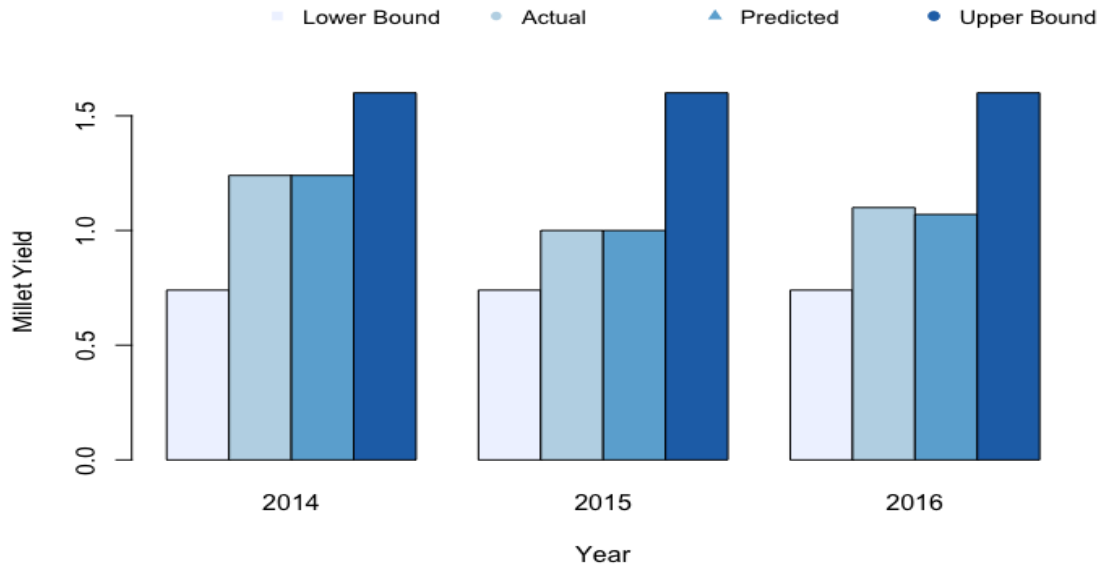

Figure 5: Prediction results based on QRF-SJ probability density forecasting model from 2014-2016 for the yield of millet

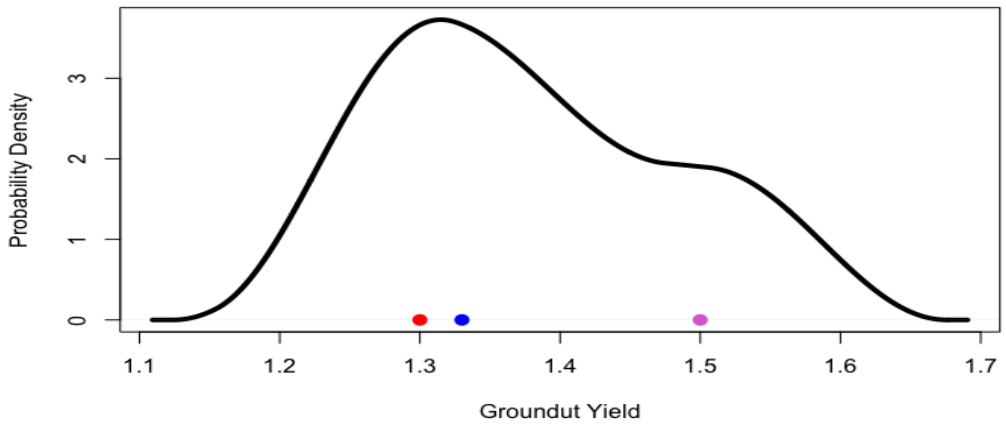

Figure 6: probability density curve based on QRF-SJ from 2014-2016, the dots on the x-axis represents the actual values of the crop yield.

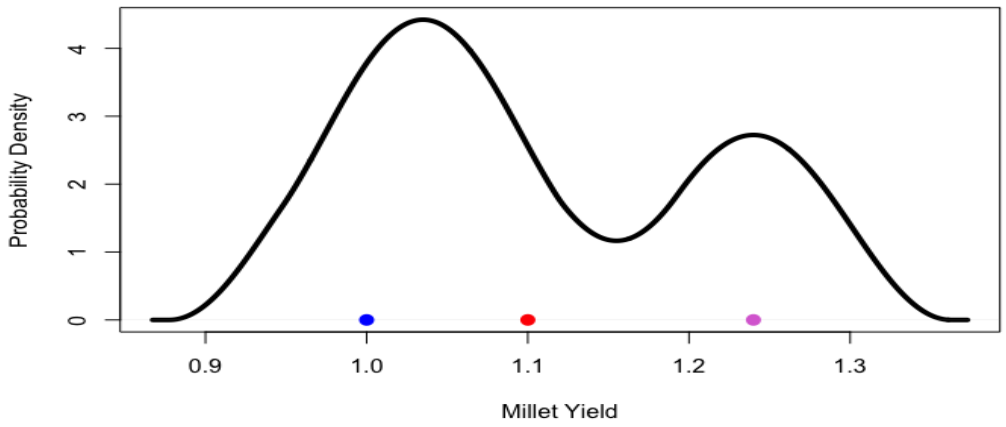

Figure 7: probability density curve based on QRF-SJ from 2014-2016, the dots on the x-axis represents the actual values of the crop yield.

4.2. Feature Importance

The level of variable importance measures according to random forest is shown in Table 5. The higher the percentage increase in mean square error (%IncMSE) of a feature, the higher the importance of that feature in the prediction model.

From the feature importance measure in Table 5, average temperature, minimum temperature, and rainfall are the three most important features among the six features that

affect the yield of groundnut. Humidity, rainfall, and average temperature are the top three features that influences the yield of millet. The amount of sunshine does not have a lot of effect on the yield of groundnut. Maximum temperature do not have a lot of effect on the yield of both crops. Generally, the average temperature do have a lot of effect on the yield of these two crops. The partial dependence plots (PDP) in Figure 8 and 10 show the marginal effect of the top three important features of the yield of groundnut and millet respectively.

From the PDP of groundnut, an increase in the amount of average and minimum temperatures result in a decrease in the yield of groundnut. However, an increase in the amount of rainfall, increases the yield of groundnut. An aincrease in the amount of rainfall increases the yield of groundnut. An increasing relative humidity and minimum temperatures decreases the yield of millet. In both crops, the yield is minimum when the minimum temperature os around 23.2°. From the PDPs, it can be concluded that the yield of groundnut and millet increases as the amount of rainfall increases to about 100mm. The yield of groundnut and millet decreases as the minimum temperature inceases from about 22° to 23.2°. Figures 9 and 11 show the three-dimensional (3-D) partial dependence of the top 3 ranked features from feature importance measures of Random Forests model for groundnut and millet respectively. From table 5, it is clear that all the six weather features do influence the yield of all the two crops.

Feature	**Groundnut**		**Millet**	
	Rank	**%IncMSE**	**Rank**	**%IncMSe**
Sunshine	6	0.101	4	0.120
Humidity	4	0.144	1	0.210
Rainfall	3	0.162	2	0.198
AvgT	1	0.175	3	0.121
MaxT	5	0.125	5	0.114
MinT	2	0.171	6	0.115

Table 5: Rank corresponds to variable importance measure determined by Random Forest (RF) model for each crop dataset. AvgT, MaxT, MinT represent average, maximum and minimum temperature respectively.

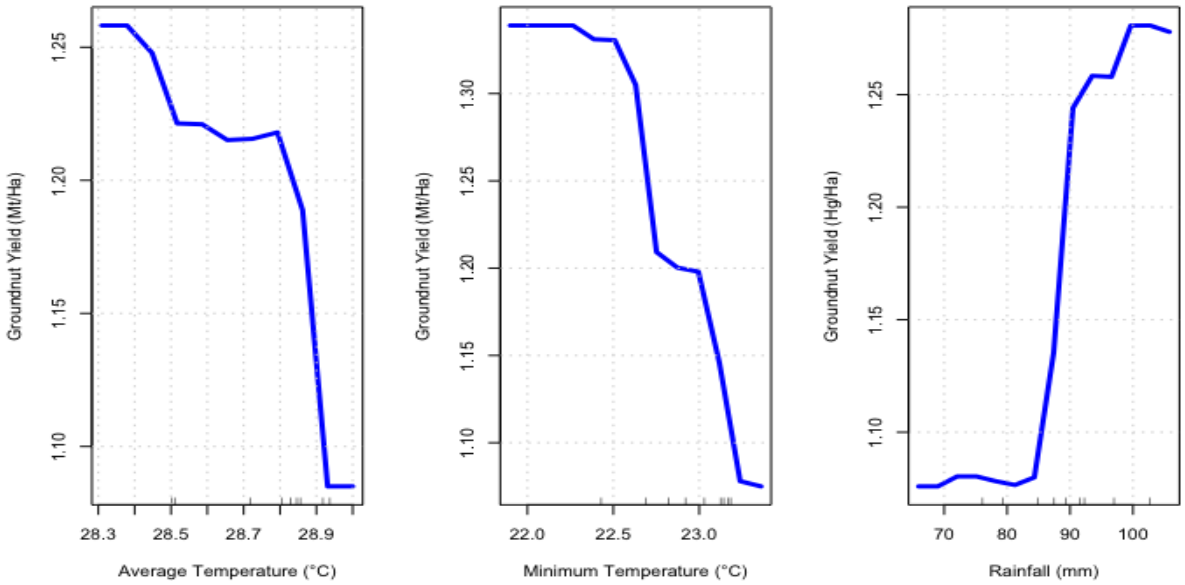

Figure 8: Partial dependence plot of groundnut for the top 3 ranked predictor variable from variable importance measures of Random Forests models.

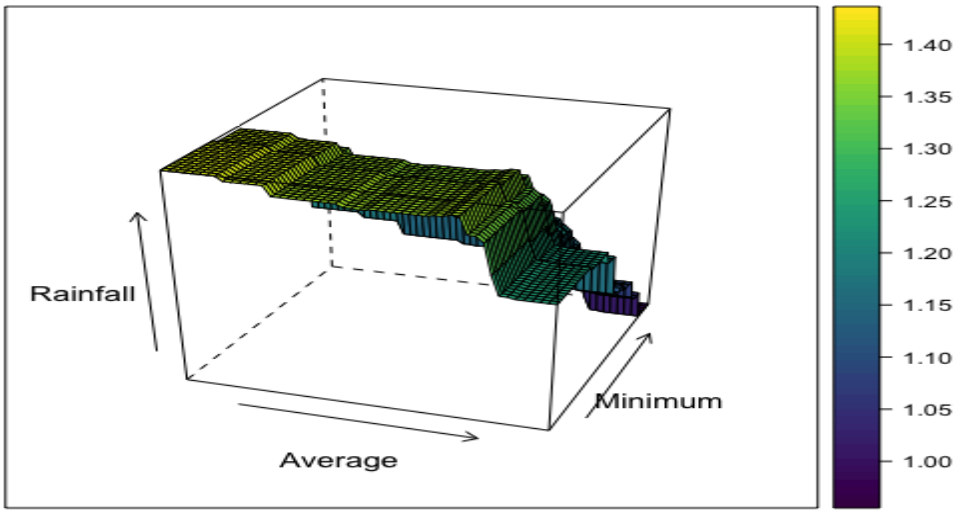

Figure 9: Groundnut: A 3-D surface Partial dependence of Rainfall on Average and Minimum Temperature based on a random forest.

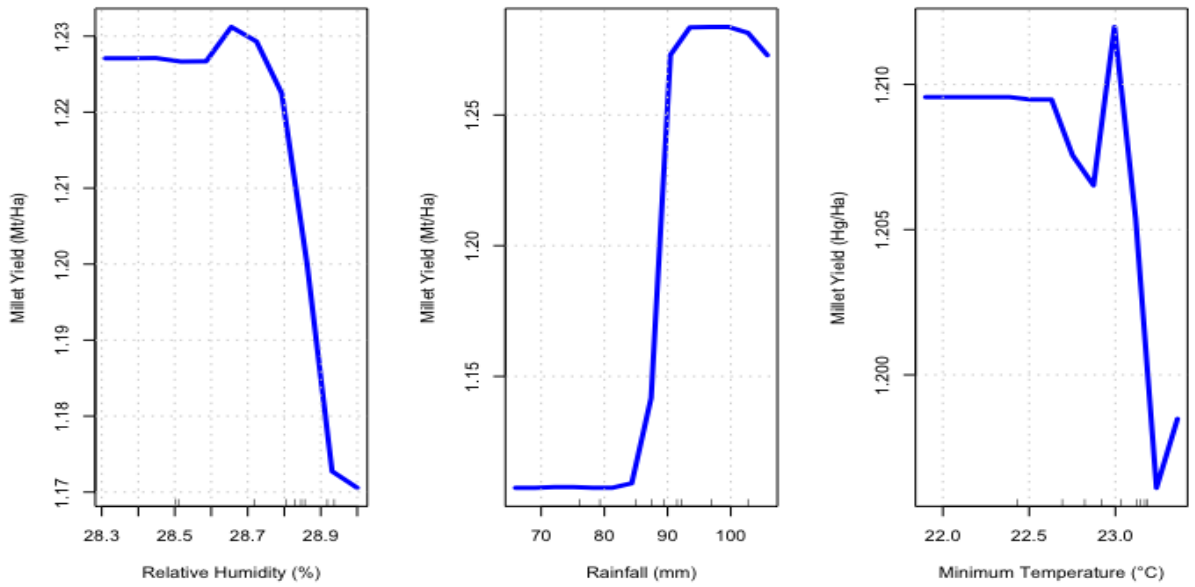

Figure 10: Partial dependence plot of millet for the top 3 ranked predictor variable from variable importance measures of Random Forests models.

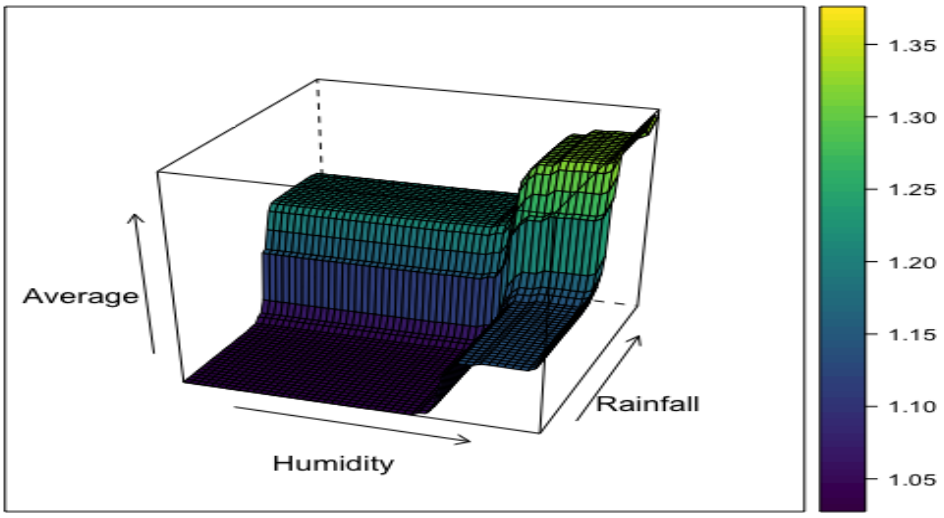

Figure 11: Millet: A 3-D surface Partial dependence of Average temperature on humidity and rainfall based on a random forest

4.3. Empirical Application of crop yield probabilistic density forecasting

4.3.1. QRF-SJ method

Due to the variability of the yield of crops as a result of weather variables, a method that can effectively handle this effect is proposed using quantile random forest and kernel density estimation to study the lower and upper bounds of the predictions. QRF does not only outperform other benchmark methods like SVR, NN, and GLM in point forecasting but also gives useful probability distribution when combined wth kernel density estimation. The main objective of probability density forecasting is to verify if the probability distributions of the crop yields fall wihin the prediction intervals. Forecasting the distributions serves as an indicator for the accuracy of forecast and provides important information for decision making. Crop yield forecasting is an integral factor in precision farming and can promote the expansion of the agricuture sector. Robust and efficient crop yield forecasting model increases the ratemaking framework in crop (re)insurance and weather derivatives market, thereby enhancing the participation of agricultural insurance in the insurance community. During premium calculation and claims liquidation, crop insurance and weather derivatives companies are very functional in crop yield forecasting.

4.3.2. Feature Importance

The feature importance gave a score of the importance of each weather variable in building the quantile regression forest model. By identifying the contributions of these weather variables to crop yields and applying effective forecasting models, there can be effective decision making process' in the agricultural sector. The farmer is able to detect the specific weather variable that affects the yield of crops and can make pragmatic interventions to curtail any uncertainties. The insurance sector is able to realize the speicific weather variable which correlates with the actual farm yield. This will enable insurance company to sell different weather-index insurance products to farmers depending on the weather variable that affect the yield of the crop. Payouts of this weather insurance will be triggered based on the level for which the weather varaible affects the yield of the crop. Base on feature importance, insurers and re-insurers are able to price their insurance product base on the specific climatic factors that affect the yield of a specific crop. They can also merge the most important climatic factors in their pricing models.

5. Conclusion

Crop yield forecasting that considers the uncertainty of weather as a result of changes in climate is of crucial importance to efficient decision making process for the agricultural sector. In this paper, a hybrid crop yield probability density forecasting method that has the potential to draw total conditional probability density curve of future crop yields is presented. In the method, quantile regression forest is used to build the nonlinear quantile regression forecasting model and to capture the nonlinear relationship between the weather variables and crop yields. Epanechnikov kernel function and solve-the equation plug-in approach of Sheather and Jones are employed in the method to construct the probability density forecasting curve. Prediction interval coverage probability and

prediction interval normalized average width are used to evaluate the quality of the prediction intervals constructed by QRF-SJ. The performance and accuracy of the QRF-SJ crop yield forecasting model are evaluated using two real dataset (groundnut and millet yields) as case studies. That is, the results of the predicted crop yield for distinct quantiles are employed as the input of the kernel density estimation. The numerical results give a 100% PICPs and narrow PINAWs. Also, all the observed groundnut and millet crop yield values are located in the probability density curves. The results show the superiority and feasibility of the proposed QRF-SJ model in forecasting the yield of crops in the midst of of weather uncertainties. Using the feature importance, agricultural stakeholders will be able to detect the major weather variables that affect the yield of crops and can make pragmatic interventions to curtail any uncertainties.

Acknowledgements

The first author wishes to thank African Union and Pan African University, Institute for Basic Sciences Technology and Innovation, Kenya, for their financial support for this research.

References

[1] S. S. Rana, R. S. Rana, ADVANCES IN CROP GROWTH AND PRODUCTIVITY, Technical Report, Publication of the Department of Agronomy, CSK Himachal Pradesh Krishi Vishvavidyalaya, Palampur, India, 2014.

[2] S. A. Gyamerah, P. Ngare, D. Ikpe, Regime-switching temperature dynamics model for weather derivatives, International Journal of Stochastic Analysis 2018 (2018).

[3] J. Delincé, et al., Recent practices and advances for amis crop yield forecasting at farm and parcel level: a review., Recent practices and advances for AMIS crop yield forecasting at farm and parcel level: a review. (2017).

[4] W. Shi, F. Tao, Z. Zhang, A review on statistical models for identifying climate contributions to crop yields, Journal of geographical sciences 23 (2013) 567–576.

[5] J. D. Michler, F. G. Viens, G. E. Shively, et al., Risk, agricultural production, and weather index insurance in village south asia, in: International Association of Agricultural Economists 2015 International Conference of Agricultural Economists, Milan, Italy, August, pp. 8–14.

[6] D. B. Lobell, M. B. Burke, On the use of statistical models to predict crop yield responses to climate change, Agricultural and Forest Meteorology 150 (2010) 1443–1452.

[7] A. Choudhury, J. Jones, Crop yield prediction using time series models, Journal of Economics and Economic Education Research 15 (2014) 53.

[8] L. Hong-ying, H. Yan-lin, Z. Yong-juan, et al., Variations trend of grain yield per unit area and fertilizer application systems in liaoning province, System Sciences and Comprehensive Studies in Agriculture 24 (2008) 408–410.

[9] X.-j. ZHANG, X.-l. ZHANG, Application of time series analysis model on total corn yield of shandong province [j], Research of Soil and Water Conservation 3 (2007) 098.

[10] J. H. Jeong, J. P. Resop, N. D. Mueller, D. H. Fleisher, K. Yun, E. E. Butler, D. J. Timlin, K.-M. Shim, J. S. Gerber, V. R. Reddy, et al., Random forests for global and regional crop yield predictions, PLoS One 11 (2016) e0156571.

[11] Y. Everingham, J. Sexton, D. Skocaj, G. Inman-Bamber, Accurate prediction of sugarcane yield using a random forest algorithm, Agronomy for sustainable development 36 (2016) 27.

[12] S. S. Dahikar, S. V. Rode, Agricultural crop yield prediction using artificial neural network approach, International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering 2 (2014) 683–686.

[13] Y. Fang, P. Xu, J. Yang, Y. Qin, A quantile regression forest based method to predict drug response and assess prediction reliability, PloS one 13 (2018) e0205155.

[14] F. Jiang, W. Wu, Z. Peng, A semi-parametric quantile regression random forest approach for evaluating muti-period value at risk, in: Control Conference (CCC), 2017 36th Chinese, IEEE, pp. 5642–5646.
[15] C. Deser, A. Phillips, V. Bourdette, H. Teng, Uncertainty in climate change projections: the role of internal variability, Climate dynamics 38 (2012) 527–546.
[16] J. Reilly, P. H. Stone, C. E. Forest, M. D. Webster, H. D. Jacoby, R. G. Prinn, Uncertainty and climate change assessments, 2001.
[17] P. Barnwal, K. Kotani, Climatic impacts across agricultural crop yield distributions: An application of quantile regression on rice crops in andhra pradesh, india, Ecological Economics 87 (2013) 95–109.
[18] J. Wang, Bayesian quantile regression for parametric nonlinear mixed effects models, Statistical Methods & Applications 21 (2012) 279–295.
[19] N. Meinshausen, Quantile regression forests, Journal of Machine Learning Research 7 (2006) 983–999.
[20] R. Koenker, G. Bassett, Regression quantiles. econometrica 46 33–50, Mathematical Reviews (MathSciNet): MR474644 Digital Object Identifier: doi 10 (1978) 1913643.
[21] Y. Lin, Y. Jeon, Random forests and adaptive nearest neighbors, Journal of the American Statistical Association 101 (2006) 578–590.
[22] V. A. Epanechnikov, Non-parametric estimation of a multivariate probability density, Theory of Probability & Its Applications 14 (1969) 153–158.
[23] M. P. Wand, M. C. Jones, Kernel smoothing, Chapman and Hall/CRC, 1994.
[24] M. C. Jones, J. S. Marron, S. J. Sheather, A brief survey of bandwidth selection for density estimation, Journal of the American statistical association 91 (1996) 401–407.
[25] S. J. Sheather, M. C. Jones, A reliable data-based bandwidth selection method for kernel density estimation, Journal of the Royal Statistical Society: Series B (Methodological) 53 (1991) 683–690.