Climate Change in Albania and Its Effect on Cereal Yield

L. Basha, E. Gjika

Open Science Index, Environmental and Ecological Engineering Vol:18, No:2, 2024 publications.waset.org/10013492.pdf

***Abstract*—**This study is focused on analyzing climate change in Albania and its potential effects on cereal yields. Initially, monthly temperature and rainfalls in Albania were studied for the period 1960-2021. Climacteric variables are important variables when trying to model cereal yield behavior, especially when significant changes in weather conditions are observed. For this purpose, in the second part of the study, linear and nonlinear models explaining cereal yield are constructed for the same period, 1960-2021. The multiple linear regression analysis and lasso regression method are applied to the data between cereal yield and each independent variable: average temperature, average rainfall, fertilizer consumption, arable land, land under cereal production, and nitrous oxide emissions. In our regression model, heteroscedasticity is not observed, data follow a normal distribution, and there is a low correlation between factors, so we do not have the problem of multicollinearity. Machine learning methods, such as Random Forest (RF), are used to predict cereal yield responses to climacteric and other variables. RF showed high accuracy compared to the other statistical models in the prediction of cereal yield. We found that changes in average temperature negatively affect cereal yield. The coefficients of fertilizer consumption, arable land, and land under cereal production are positively affecting production. Our results show that the RF method is an effective and versatile machine-learning method for cereal yield prediction compared to the other two methods: multiple linear regression and lasso regression method.

***Keywords*—**Cereal yield, climate change, machine learning, multiple regression model, random forest.

I. Introduction

In recent decades, climate change has generated worldwide attention. July 2021 has earned the undesirable credit as the world's hottest month ever recorded in all the world. The combined land and ocean-surface temperature was 0.93 °C above the 20th-century average of 15.8 °C. The last 10 years were the ocean's warmest decade since at least the 1800s, where 2020 saw the highest global sea level, risen nearly 178 mm over the past 100 years [1].

Climate change already affects human health directly and impacts food production. Authors in [1] examined trends and observed variability in extreme climatic occurrences, such as extreme daily temperatures, extreme daily rainfall amounts, unusually warm monthly temperatures and their impact in excessive loss of life, excessive economic or monetary losses, or both. In [2], authors applied a global vegetation and hydrology model, to quantify the contributions of changing precipitation, temperature, atmospheric CO_2 content, land use and irrigation to worldwide trends in 20^{th} century river discharge. Numerous studies have analyzed temperature, rainfall, and other factors that determine how the climate changes in various locations. In [3], the authors examine how frequently incidents with absolutely extreme temperatures occur in Argentina. They fitted a generalized extreme value distribution, where calculations are made for the years 1956 to 2003 to determine the year's highest (lowest) maximum and minimum temperatures as well as return values. Reference [4] estimated temperature change between April 1, 1954, and March 31, 2013 and examined annual extreme temperatures in the Amur River watershed.

Therefore, many efforts have used historical records to obtain statistical models for climacteric variables such as, temperature and rainfall. In [5], authors describe current patterns in climate that have been seen for global mean air temperature, global sea level, and carbon dioxide concentration. Time series models have grown in popularity thanks to their ability to recognize trends. Employing a spatio-temporal analysis, [6] examined the trends, variability, and spatial patterns of temperature and rainfall. Using geographic information systems and an autoregressive integrated moving average time series model, they performed simulations of recent temperature and precipitation data. Authors in [7] performed time series trend analysis to inspect the change of rainfall and temperature in northcentral Ethiopia using gridded monthly precipitation data. A three-day rainfall pattern to increase the potential of statistical data to be replicated is proposed in [8]. In order to demonstrate the applicability of the suggested disaggregation method, probability distribution and L-moment statistics were examined by [8].

Climate change has a high impact on food production. Therefore, it is crucial to find a solution for the major issue of accurate yield prediction. More than 70% of the low-income population worldwide depends on agriculture and natural resources for a living [9]. Cereals are a good source of minerals, sugars, proteins, vitamins, and micronutrients, which are necessary for the body's efficient functioning. Developing nations may experience greater hardship as a result of the price increase since they use more cereals annually (166 kg per person) than industrialized nations (133 kg per person) and are less resilient [9]. Also, the Ukraine crisis brought about food shortages for the world's poorest citizens. 30% and 20%, respectively, of the world's wheat and maize exports come from Ukraine and the Russian Federation. Cereals are additionally regarded as a solid source of calories. In rich and developing nations, respectively, 60% and 30% of calories are obtained from cereals. The increase in worldwide agricultural

L. Basha is with Department of Applied Mathematics, Faculty of Natural Science, University of Tirana, Albania (e-mail: lule.hallaci@fshn.edu.al).

E. Gjika is with School of Mathematics and Statistics, Carleton University, Ottawa, Canada (e-mail: eraldagjika@cmail.carleton.ca).

Open Science Index, Environmental and Ecological Engineering Vol:18, No:2, 2024 publications.waset.org/10013492.pdf

productivity has already been slowed by climate change. According to a recent study [10], anthropogenic climate change (ACC) has had a negative influence on worldwide agricultural total factor productivity since 1961 by around 21%, with warmer regions like Africa seeing a bigger impact (about 34%) than cooler regions like Europe and Central Asia (about 7.1%).

Numerous studies have demonstrated the impact of a wide range of meteorological conditions on agricultural yields in various seasons and places, see [11]-[13]. The authors in [14] investigate the susceptibility of barley, maize, and wheat to fluctuations in growing season precipitation as well as socioeconomic adaptation capacity indicators including literacy and poverty rates at both national and subnational scales in Morocco. To assess the impact of climate change on corps yield, the two main modules used include process-based modeling and statistical modeling. In [15], the authors analyze how well statistical models can forecast how crop yields will change in response to changes in temperature and precipitation for close to 200 sites in sub-Saharan Africa. They used process-based model CERES-Maize to simulate historical yields, and then fit statistical regressions to the simulated data. Reference [16] investigates how climate change affects agricultural output by using a combination of qualitative descriptive, quantitative statistical analysis and multiple linear regression analysis. The uncertainty of estimates of global temperature impact on crop yields was analyzed, by using two different multi-model ensemble approaches [17], with process-based crop simulation models. The estimated loss in worldwide wheat yield with a 1 °C rise in temperature is between 4.1% and 6.4%. A regression analysis model to evaluate the precision and efficacy of yield estimates for the Indian rice crop between various environmental factors including temperature, rainfall, and other factors, is studied in [18]. Artificial Neural Networks (ANNs) and multiple linear regression were used in Ardabil to estimate grain yield, see [19]. Authors in [20] looked at data-driven methodologies to evaluate linked hypotheses for greater scientific comprehension and improved predictive modeling, and they found that nonlinear regression models outperform linear methods. Using multiple linear regression as a standard for both global and regional agricultural yield estimates, assess the effectiveness of RF regression [21]. RF methods are widely used, for their advantages, for predicting ecosystem or crop productivity, see [22]-[24]. To investigate the effects of elevated CO_2 concentration ([CO_2]), elevated temperature, and drought stress on the yield and nutritional attributes of C3 cereals, authors in [25] undertook a meta-analysis of the collection of current literature. The effectiveness and viability of various local-scale statistical crop yield forecasting systems are examined in [26]. Four types of grain (wheat, barley, maize, and rice) are taken into account in two different agro climatic regions of Italy. Six meteorological and remote sensing indicators are taken into consideration while building models of various levels of complexity, and they are used as potential predictive factors. A comprehensive, farm-level dataset covering the years 2001 to 2015 was used to derive yield information at three different spatial aggregation scales.

According to climate change predictions, the Mediterranean region will have 25-30% less precipitation in the latter half of the 21st century [27]. Also, the seasonality of rainfall is far more significant. In reality, the anticipated Mediterranean rainfall deficit should have a significantly greater influence on summer precipitation than on winter precipitation. But the biggest factor influencing crop growth in the Mediterranean is winter rain. Also, it is believed that one of the key elements restricting food output globally is drought. Wheat crop responses to water constraint are influenced by a number of variables, including genetics, stress duration and intensity, and plant developmental stage. The occurrence of drought in winter during the early phases of the crop cycle has recently been documented [28], despite the fact that rainfall during winter has historically been ample and correlates with the lowest evapotranspiration rates.

Albania is still quite susceptible to climate change and fluctuation. The nation is vulnerable to a wide range of environmental dangers, including hydro-meteorological dangers, geophysical dangers including earthquakes, harsh weather conditions, and temperature changes. In [29], author analyses the heat wave phenomena over two important Albanian cities. In [30], the analysis focuses on the urban heat island phenomenon in Tirana during the summer, demonstrating that maintaining an unchanged airflow and implementing a high density of vegetation with a well-defined and dense crown layer can have a substantial positive impact on summer air temperature and enhance human thermal comfort. Seasonality patterns of climacteric variables and their importance in country energy production by hydropower plants are analyzed in [31] and [32]. They have used statistical models and also machine learning models to perform accurate prediction models using average temperature and rainfalls. Likewise, there have been significant changes in Albania's agriculture since 1991, which supports the majority of the country's residents (about 60% of the estimated total population). Only 24% of Albania's land is used for agriculture, and while though it no longer accounts for the majority of the country's GDP, it still made up about 21% in 2019 [33].

The objective of this paper is to study the climate change in Albania and its impact on cereal yield. Here we evaluate the performance of linear and nonlinear methods to understand if weather indices are relevant in cereal yield modeling. We also calculate whether nonlinear regression models outperform linear techniques in capturing cereal yield variability.

II. Methodology

In cases of using reliable and sufficient data for model design, statistical methods, compared to process-based methods, provide the possibility of simple but reasonable predictions. In order to overcome the weakness and disadvantages of traditional regression methods, statistical methods provide access using machine learning algorithms, in our case we use RF algorithm.

A. Multiple Linear Regression Analysis

A statistical method known as multiple linear regression employs a number of explanatory variables to forecast the result of a dependent variable. To simulate the linear relationship

Open Science Index, Environmental and Ecological Engineering Vol:18, No:2, 2024 publications.waset.org/10013492.pdf

between the explanatory and response variables, multiple linear regression is used as:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \varepsilon \tag{1}$$

where for $i=n$ observations; y_i is the response variable; x_i are the independent variables; β_0 is the intercept; β_p is the slope coefficients for each independent variable; ε is the model's error term [33], [34]. In the multiple regression model, the response variable and independent variables have a linear relationship. The residuals should have a normally distributed distribution because the independent variables do not have a lot of correlation with one another.

Statistical models can simply be modified to diverse agro-climatic regions and are adaptable enough to ingest large amounts of data. Its estimation, however, frequently relies on regression techniques, which can make it susceptible to the well-known shortcomings and limitations of such models. Multiple linear regression models, for instance, have succeeded when the number of predictors is minimal [35]. Reference [36] demonstrated that when it comes to explaining yield variability, linear models frequently outperform more complicated statistical models. Unfortunately, ordinary least squares estimation is sometimes prone to overfitting issues when a large number of potential explanatory factors are taken into account, which reduces the predictive strength of these methods. More complex estimate methods, like ridge or lasso regressions, can be helpful in these situations, especially when there are many variables relative to the number of data and/or when the inputs are highly correlated [37].

B. Lasso Regression

In order to improve simpler models with smaller coefficient values during training, an extension of linear regression necessitates adding penalties to the loss function. Popular regularized linear regression with an L1 penalty is called Lasso Regression, which adds a penalty equal to the absolute value of the magnitude of coefficients. In this method, the model generates a model with a small number of coefficients, some of which may be zeroed out and dropped from the model. Greater penalties provide coefficient values that are closer to zero, which is great for creating more straightforward models.

For those variables whose predictor power, in response variables, is very small, or which do not contribute much, the method has the effect of shrinking the coefficients. The lasso involves estimating β as the solution to the penalized least-squares problem.

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \frac{1}{2} \|y - X\beta\|_2^2 + \lambda \|\beta\|_1 \tag{2}$$

where $\|\beta\|_1$ is the ℓ_1 norm of the coefficient vector:

$$\|\beta\|_1 = \sum_{j=1}^{p} |\beta_j| \tag{3}$$

and λ denotes the amount of shrinkage. The bias increases with increase in λ and variance increases with decrease in λ [38], [39].

C. RF Regression

Used frequently in regression and classification issues is the supervised machine learning method known as RF [40]. Building a sizable sample of de-correlated trees and averaging them is a significant refinement of bagging. It is a method that makes use of ensemble learning and combines numerous weak classifiers to offer answers to challenging issues. Enhancing the decrease in bagging variance by reducing inter-tree correlation, without increasing variation, is the concept of RFs. Random selection of input variables enables the achievement of this objective in the tree growth process. When B of these trees $\{T(x;\Theta_b)\}_1^B$ have grown, the following is the RF predictor:

$$\hat{f}_{rf}^{B}(x) = \frac{1}{B} \sum_{b=1}^{B} T(x;\Theta_b) \tag{4}$$

where Θ_b characterizes the *bth* RF tree in terms of split variables, cutpoints at each node, and terminal-node values. RFs algorithm uses out-of-bag (OOB) samples to measure the predictive power of each factor. The OOB samples are sent down the tree as the bth tree grows and the accuracy of the prediction is also reported. The accuracy is once more computed, after the random permutation of the *jth* variable within the OOB samples. The amount of samples that reach the node divided by the entire number of samples gives the node probability.

Large datasets can be handled with the RF algorithm. The bootstrapped sample and the random subset of features, for some nodes, may yield an invariant feature space, which presents problems for the model when the data are relatively sparse. The possibility of overfitting using RF should also be taken into consideration.

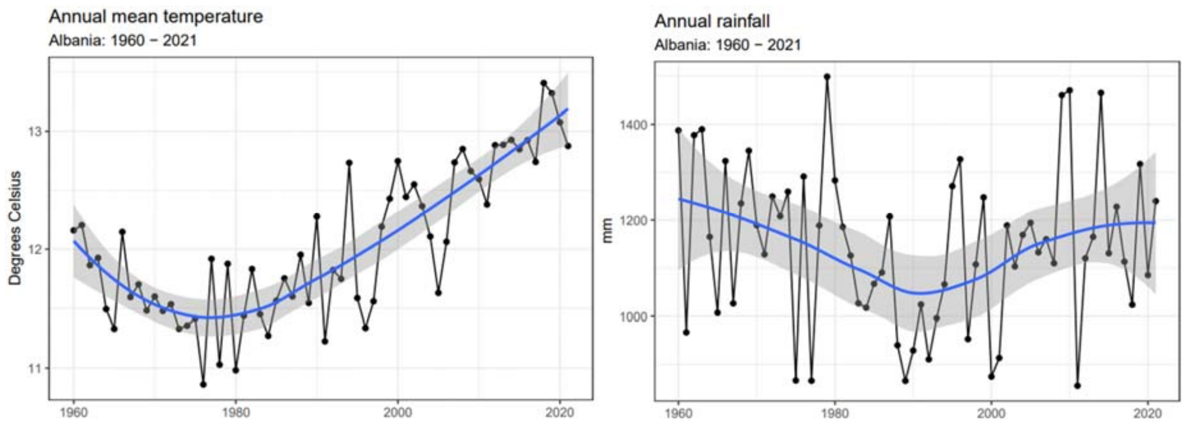
III. Results and Discussions

The objective of this paper is to study the impact of climate change on cereal yield in Albania. In the first step, the average temperature and rainfall in Albania were studied for the period 1960-2021. In the second step for the same period, the yield of cereals in Albania is studied. We have also taken into consideration some external factors which are considered important, in addition to temperature and precipitation, such as: fertilizer consumption (kilograms per hectare of arable land); arable land (hectares); land under cereal production (hectares) and nitrous oxide emissions (thousand metric tons of CO_2 equivalent). Linear and nonlinear methods were used to understand if weather indices and the other factors are relevant in cereal yield modeling. Further we evaluated by measure of performance the accuracy of nonlinear regression models and linear approaches at capturing cereal yield variability.

A. Temperature and Rainfall

The topography of Albania affects the country's climate. Albania's average annual temperature from 1960 to 2021 is 12.02 °C and mean annual rainfall is 1485 mm, with steady rainfall occurring throughout the year. The Albanian alps have the most annual precipitation, averaging 2800-3000 mm, while the southeast has lower annual precipitation, averaging just approximately 1000 mm. In Saranda, to the south, the average yearly temperature ranges from 17.6 °C to 7 °C, Vermosh in the North. Lowland regions are distinguished by a consistent mean temperature range of 14 °C to 16 °C. The three cities with the coldest recorded temperatures were Sheqeras (25.8 °C), Voskopoj (25.6 °C), and Bize (34.7 °C), while with the highest were Kuçov (43.9 °C), Roskovec (42.8 °C), and Iflig (42.4 °C). Annual average rainfall is 1430 mm.

For the period 1960-2021, Fig. 1 shows the regional variance of the measured (a) annual average temperature and (b) yearly rainfall. Since the 1960s, the average annual temperature in Albania has increased by 1.2 °C, and the average annual precipitation has decreased slightly (albeit statistically insignificantly). The lowest annual temperature, 10.6 °C is in 1977, while the highest, 13 °C, is in 2018. Year 2012 registers the smallest amount of rainfall with annual rainfall of 718.1 mm, while the largest amount of rainfall was recorded in 1980, with annual rainfall of 1301.3 mm. In the last 60 years, the average amount of rainfall in Albania is 1018.7 mm.



(a) (b)

Fig. 1 Observed annual mean (a) temperature and (b) rainfall for Albania, 1960–2021

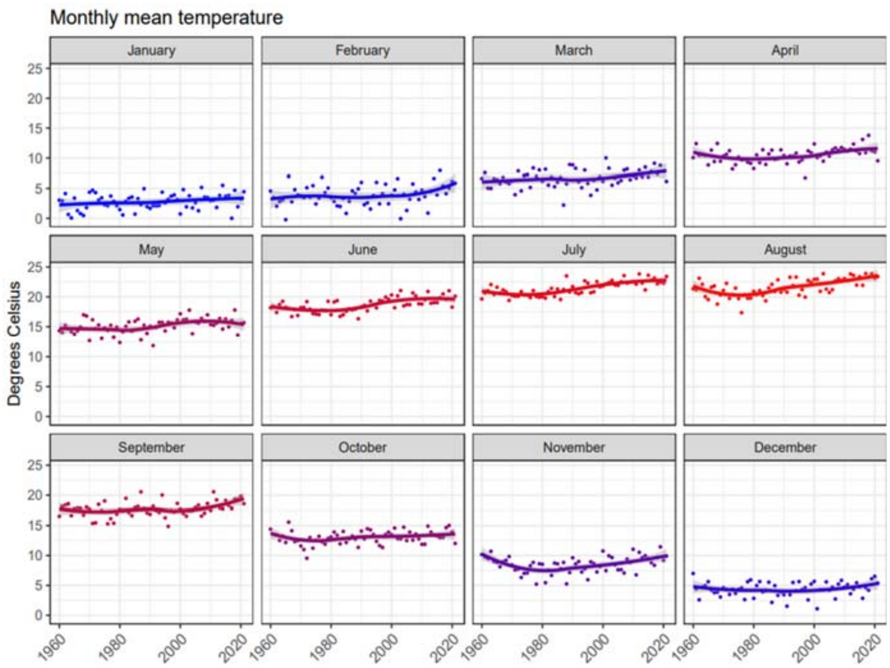


Fig. 2 Average monthly temperature for Albania, 1960–2021

Fig. 2 displays the typical monthly temperatures in Albania, which range from 21 °C (July-August) to 2 °C (January). There has been evidence that temperature rises are more pronounced in the summer months of June (18.8 °C), July (21.4 °C), August (21.6 °C), and September (17.6 °C). Also, April (10.5 °C) and November (8 °C) also are experiencing temperature rise.

The colder months (October to March) have the most precipitation (66% of the total), with November recording the

Open Science Index, Environmental and Ecological Engineering Vol:18, No:2, 2024 publications.waset.org/10013492.pdf

Open Science Index, Environmental and Ecological Engineering Vol:18, No:2, 2024 publications.waset.org/10013492.pdf

highest total with 129.3 mm. The months with the most rain are November and December, whereas the months with the least rain are July, with 39.2 mm and August, with 42.1 mm, of rainfall, respectively. In the last 20 years there is a decrease in rainfall for the months of September, November and December, Fig. 3.

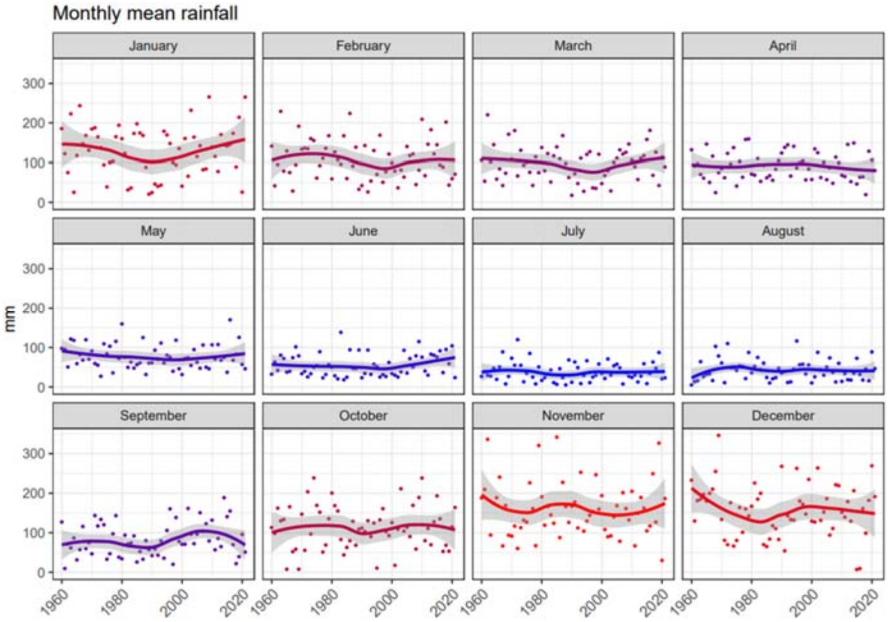
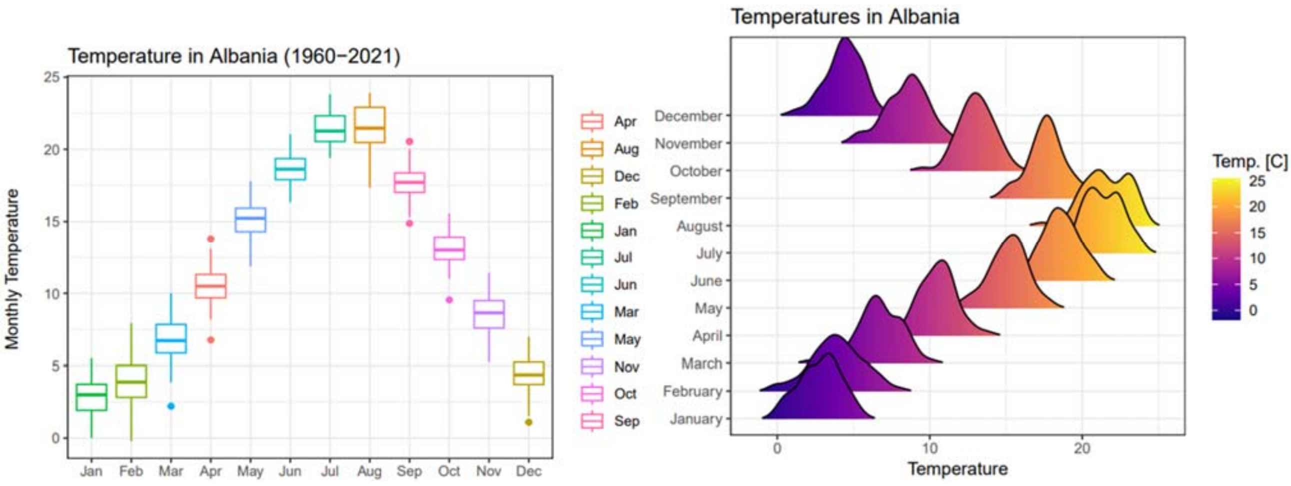


Fig. 3 Average monthly rainfall for Albania, 1960–2021



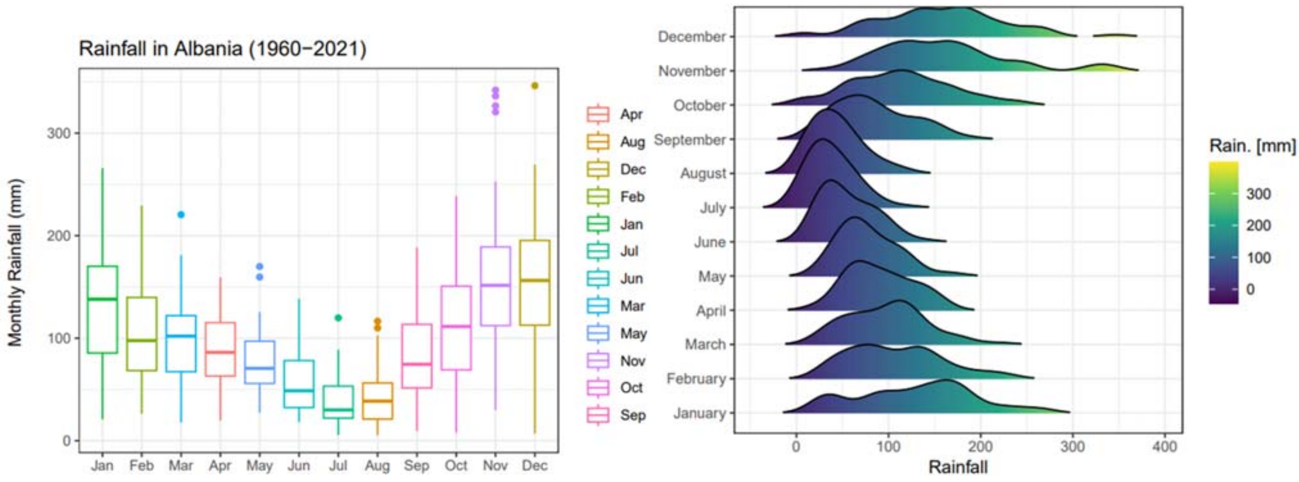
(a) (b)

Fig. 4 (a) Boxplots and (b) Ridgeline plot of variations of monthly mean temperature

Fig. 4 compares the annual cycles of the monthly mean temperature and the ridgeline plot. The lower and higher quartiles are shown in boxes. Both the median temperature and the median rainfall are shown as horizontal lines in each box. The minimum and maximum temperatures noted for each month are shown by vertical lines that extend from each box. All of the estimates show a consistent picture of increased variability in the monthly mean temperature for July and August. Variability for temperature is also larger in the latter 20 years in November and noticeably smaller in March, but no other consistent seasonal patterns are evident.

Fig. 5 compares the annual cycles of the monthly mean rainfall. Variability for mean rainfall is large in January, October, November and December. The minimum and highest amounts of rain that were observed for that month are shown by vertical lines that extend from each box. Ridgeline plot displays an augmented density plot, based on kernel methods, exactly on Gaussian kernel, for rainfall. Additionally, they support the comparison of temperature and precipitation variations between months by highlighting potential variations in variability and distribution shape.

Open Science Index, Environmental and Ecological Engineering Vol:18, No:2, 2024 publications.waset.org/10013492.pdf



(a) (b)

Fig. 5 (a) Boxplots and (b) Ridgeline plot of variations of monthly mean rainfall

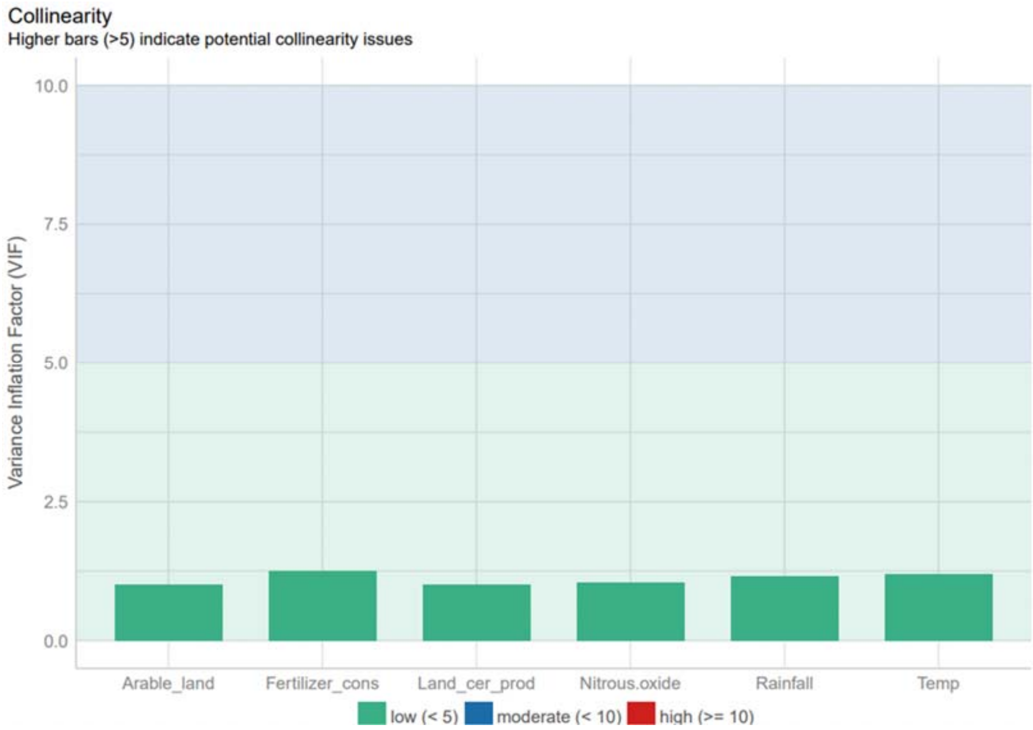


Fig. 6 Collinearity between variables

B. Cereal Yield

In Albania, the average value from 1960 to 2021, for cereal yield, was 2847 kg per hectare, with a minimum of 845 kg in 1961 and a maximum of 4950 kg in 2013. The statistics for cereal in 2018 was 45,600 100 kg/ha. This represents a reduction from the prior value for 2017 of 48.100 100 kg/ha. The agriculture industry in Albania is still underdeveloped despite the country's pleasant climate, fertile terrain, and an abundance of water resources. This is partly because of the country's fragmented land, small plot sizes, and minimal mechanization.

Building multivariate linear regression models, lasso method, and RF approaches are the next steps in this work to determine whether weather indices, average temperature, average rainfall and other parameters such as: fertilizer consumption, arable land, land under cereal production, and nitrous oxide emissions, are important when estimating cereal yield.

1. Multivariate Linear Regression Model Results

Before building a linear regression model, the problem of multicollinearity in data must be studied. The standard errors of the regression coefficients become extremely sensitive to even small changes in the model due to multicollinearity. A rule of thumb to detect multicollinearity is that when variance inflation

Open Science Index, Environmental and Ecological Engineering Vol:18, No:2, 2024 publications.waset.org/10013492.pdf

factor (VIF) is greater than 10, then there is a problem of multicollinearity. From Fig. 6, we can conclude that there is a low correlation, with VIF which vary from 1 to 1.26, and it is not severe enough to warrant corrective measures.

The significance of each regression coefficient in the multiple linear regression models is examined using a t-test with a 95% confidence level. The t-stat value must be greater than 1.96 or less than 1.96, respectively, in order for the multivariate regression linear relationship between cereal yield and each independent variable (temperature, rainfall, fertilizer consumption, arable land, land under cereal production, and nitrous oxide emissions) to be significant at the 95% level. Initially the multiple regression model is constructed taking into account all variables.

We have built four different multivariate linear regression model:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon \qquad (5)$$

where y_i is the cereal yield; and x_i are all the explanatory variables: temperature, rainfall, fertilizer consumption, arable land, land under cereal production, and nitrous oxide emissions.

- The first model *modlinear:* includes all the variables.
- The second model *modlinear1:* does not include rainfall.
- The third model *modlinear2:* does not include nitrous oxide emissions.
- The fourth model *modlinear3:* does not include rainfall and nitrous oxide emissions.

According to the t-test results, the nitrous oxide emissions had no significant linear relationship with cereal yield. The explanatory variables for a multiple-regression model are then chosen using a stepwise regression procedure.

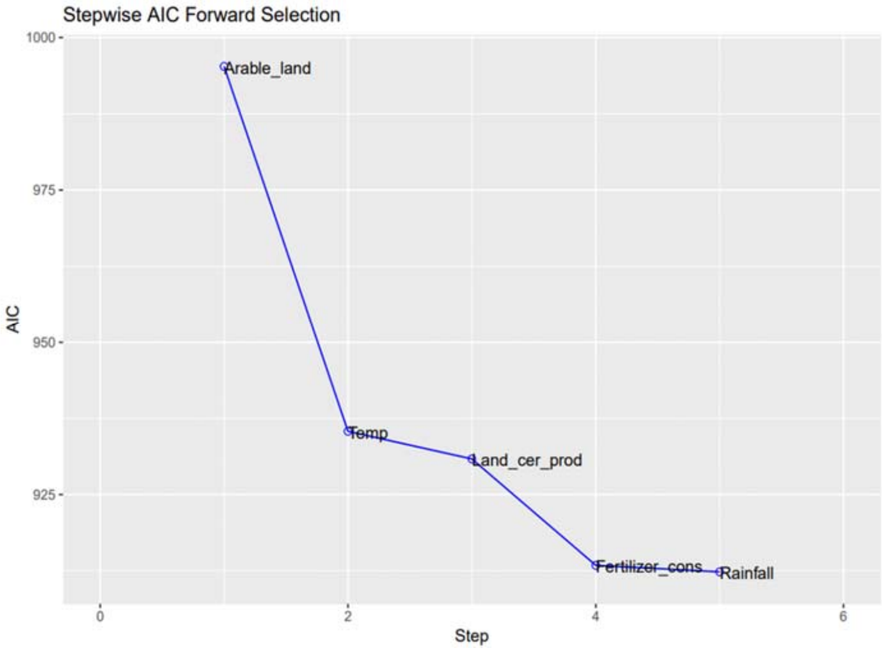


Fig. 7 Stepwise forward selection

A forward-selection rule begins with no explanatory factors and adds variables one at a time depending on the most statistically significant variable, until there are no more statistically significant variables. Stepwise forward selection is presented in Fig. 7, where also here we can conclude that the nitrous oxide emissions variable has been dropped from the model.

The results of stepwise regression procedure are also displayed in Table I.

TABLE I
STEPWISE REGRESSION PROCEDURE

Variable	AIC	R-Sq	Adj.R-Sq
Arable land	995.258	0.67397	0.66854
Temperature	935.375	0.87984	0.87576
Land under cereal production	930.844	0.89185	0.88625
Fertilizer consumption	913.37	0.921	0.91546
Rainfall	912.304	0.92481	0.9181

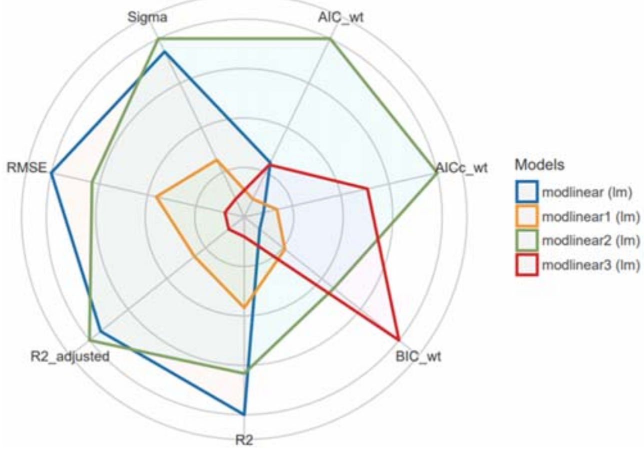


Fig. 8 Comparison of multivariate linear regression model

The t-test outcomes in this study can be explained using the final model as follows: Temperature is a key indicator of climate change, and the computed value for this variable was -258.7 with a significance level of 0.024. The quantity of such significance is less than the 5% level when using the 0.05 significance threshold, indicating that a change in temperature has a negative impact on cereal yield. The coefficients of fertilizer consumption, arable land and land under cereal production, show that these variables are positively affecting production.

Accuracy of these prediction models are measured using R2, adjusted R2, Root Mean Square Error (RMSE), AIC, BIC, Fig. 8. The model without nitrous oxide emissions is the best model, from multivariate linear regression models, with an AIC value 912.3.

Following we have studied the heteroscedasticity test and normality test. The goal of the heteroscedasticity test is to determine whether there is variance inequality between the residuals of different observations in regression models. Visually, if there are specific patterns, a fan or cone shape in the residual plot, it indicates the presence of heteroscedasticity.

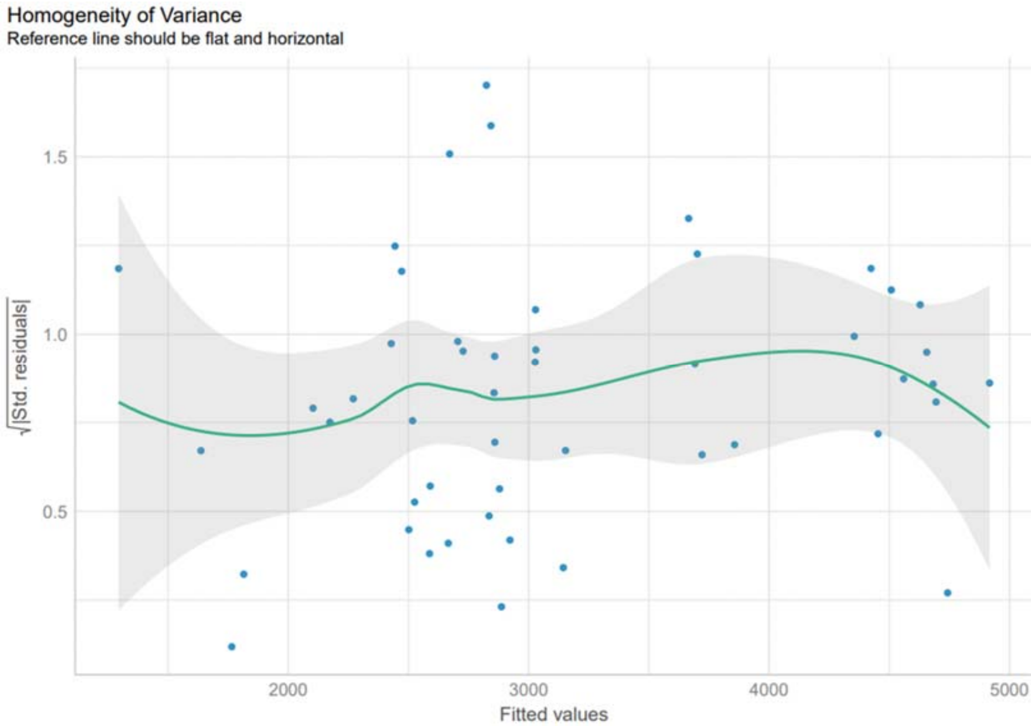


Fig. 9 Heteroscedasticity test

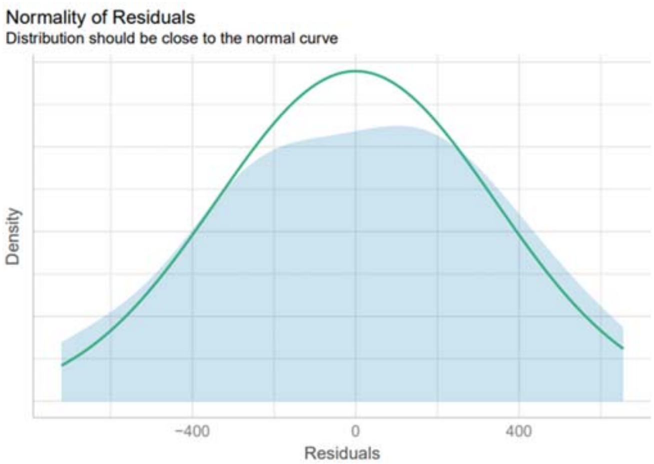


Fig. 10 Normality tests

Fig. 9 shows the results of the heteroscedasticity test, and we can infer that there is no evidence of heteroscedasticity in our regression model. Error variance appears to be homoscedastic ($p = 0.943$). Therefore, we can agree that the heteroscedasticity classic assumption test is satisfied. Additionally, it is valuable for research.

The normality tests are an addition to the graphical normality examination. Here we conduct the Shapiro-Wilk Normality test. This test is based on the correlation between the data and the corresponding normal scores. Fig. 10 displays the results of the normality test graphically, demonstrating that the conventional probability plot chart has a normal chart pattern. Residuals appear as normally distributed ($p = 0.750$). Both numerical and graphical presentation show that data follow a normal distribution.

2. Lasso Method Results

Lasso method is a technique that has received a great deal of interest. In this method, as the smoothing parameter change, the sample path of the estimates moves continuously to zero. The natural log of is plotted on the x-axis of Fig. 11 with the mean square error (MSE) on the y-axis. The number of variables

Open Science Index, Environmental and Ecological Engineering Vol:18, No:2, 2024 publications.waset.org/10013492.pdf

Open Science Index, Environmental and Ecological Engineering Vol:18, No:2, 2024 publications.waset.org/10013492.pdf

present at each site is indicated across the top. Mean squared error rises in correlation with the shrinking parameter, and finally factors are dropped from the model.

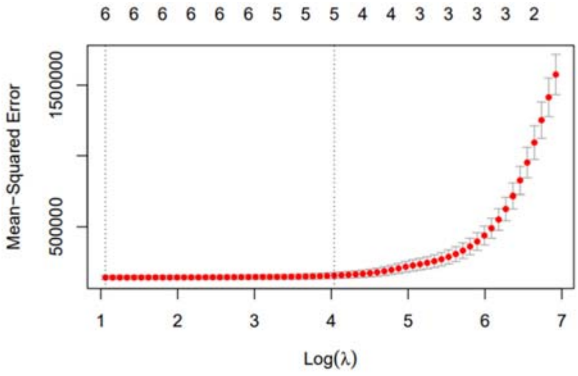


Fig. 11 Shrinkage parameter

The lambda that is closest to the y-axis and within one standard error of the minimum is shown by the second dotted line. This lambda produces the lowest MSE. The best value in this example was 2.0798 and the best value one standard error above was 59.23. The calculated coefficients for this value of were fairly similar to the real coefficients that were utilized to generate the dependent variable. By entirely reducing the coefficients to zero, lasso regression has the possibility of removing predictors from the model. In our case no variable has been dropped from the model.

Our data have been divided into train (80%) and test set (20%). We often do not incorporate our data onto the training set. A frequent data partitioning technique is k-fold cross-validation, which repeatedly separates and resamples our data. The best model is then selected after training on these samples. Using our data, we employed 10-fold cross-validation. We can adjust the fraction parameter, which determines how strongly the shrinkage operates in the optimization model, using lasso regression. Fig. 12 shows the model throughout all of the tuning parameters.

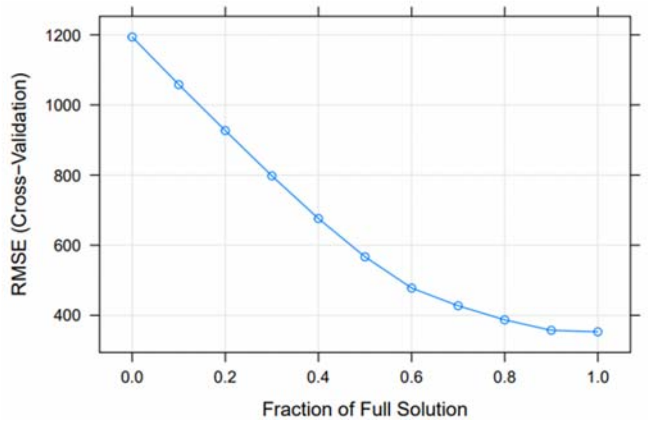


Fig. 12 Model throughout all of the tuning parameters

3. RF Results

We developed and used RF, a machine learning technique based on binary trees, for cereal yield estimation, as a regression tool. We created a customized RF model to modify the parameters in order to find the optimal set of parameters for our model and to compare the results of different parameter combinations, where we found that the best parameters for our model are 500 trees and 110 maximum nodes. The effectiveness of the predictor variables is measured by how much node impurity was reduced when they were chosen for the splits. The higher the value of increase node purity, the more significant the feature.

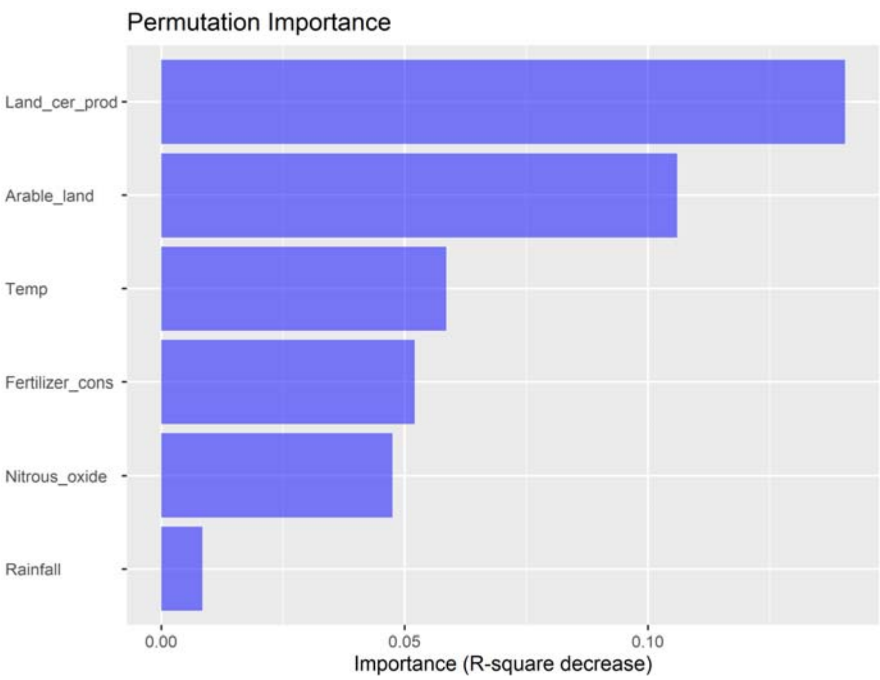


Fig. 13 Variable importance plots

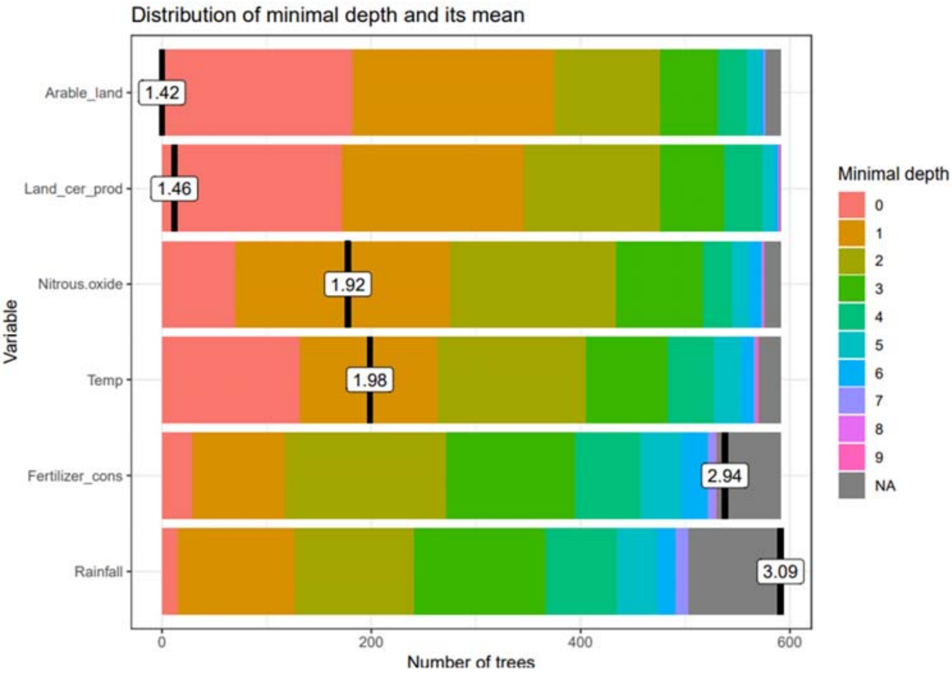


Fig. 14 Distribution of minimal depth

The plot of Fig. 13 shows the variable importance. We can clearly see that temperature has significant importance in the behavior of the cereal yield. The OOB data are used to compare average node MSEs before and after the permutation.

The x-axis in Fig. 14 runs from 0 trees to the most trees that might be split by any variable, which in this case is 600 and is reached by all variables depicted. Rainfall variable was ranked lower in his relative variable importance.

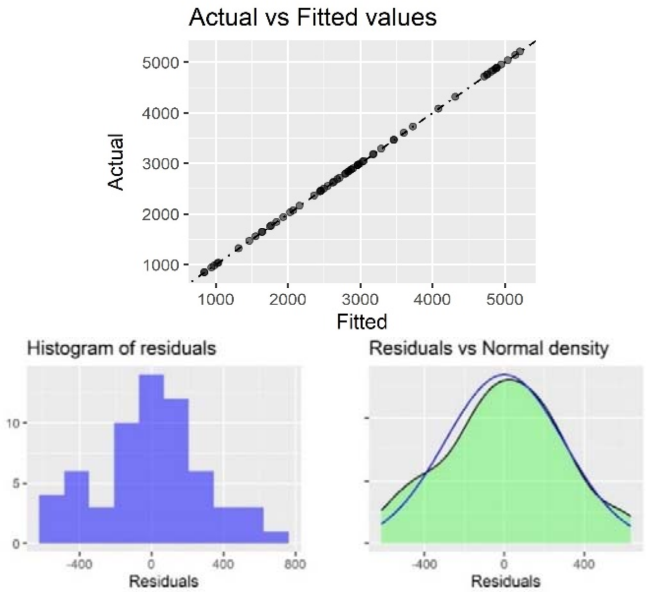


Fig. 15 RF model estimation

Fig. 15 gives the RF model estimation, actual versus fitted values; the histogram of residuals and the normality tests, from which we can conclude that residuals appear as normally distributed, and the model is an appropriate model for our data.

4. Model Performance

In the final phase, we use measures like R-squared and RMSE to assess the effectiveness of linear vs. nonlinear regression models. According to Table II, the RF approach and nonlinear regression method both had marginally superior results.

TABLE II
MODEL PERFORMANCE

Model	R-squared	RMSE
RF	98.30%	160.88
Multiple linear regression	92.59%	338.8
Lasso regression	91.67%	352.98

With an RMSE of 160.88, the RF model had a high concordance between the expected and observed values of the test data, and explained 98.30% of the yield variance. The performance of linear models' benchmark was less satisfactory, compared with RF model. The R-squared values for multiple linear regression and lasso regression were found to be 0.9259 and 0.9167, respectively, and the RMSE values for each were 338.8 and 352.98. The RF model's effectiveness could be ascribed to its ability to capture non-linear interactions and its robustness to data with multicollinearity.

IV. CONCLUSION

This study explores the climate change in Albania and its impact on cereal yield. Since the 1960s, Albania's mean annual temperature has increased by 1.2 °C, and its mean annual precipitation has somewhat decreased. Although it has been noted that temperature rises are more pronounced in the summer, April and November are also seeing increases. In the

Open Science Index, Environmental and Ecological Engineering Vol:18, No:2, 2024 publications.waset.org/10013492.pdf

last 20 years there is a decrease in rainfall for the months of September, November and December.

For cereal yield modeling we have constructed multivariate linear regression models and lasso method. We used RF methods to understand if weather indices, fertilizer consumption, arable land, land under cereal production, and nitrous oxide emissions are relevant in cereal yield modeling. The results obtained showed that, change in temperature and rainfall negatively affects cereal yield while fertilizer consumption, arable land and land under cereal production, positively affecting production.

Our findings further demonstrate the usefulness of RF regression for predicting cereal yield. When predicting crop yield responses, utilizing RF regression has a number of advantages over other methods, such as the conventional multiple linear regression model. The importance of this study is the fact that it is among the first in the country data which apply advanced models, such as RF, to evaluate cereal yield in response to climate variability. The predictions are accurate and the models can be used by institutions for better policy measurements and country investments in the field.

References

[1] D. R. Easterling, J. L. Evans, P. Y. Groisman, T. R. Karl, K. E. Kunkel, P. Ambenje, "Observed variability and trends in extreme climate events: a brief review". *Bulletin of the American Meteorological Society*, vol. 81, no. 3, 2000, pp. 417–425. https://doi.org/10.1175/1520-0477(2000)081<0417:OVATIE>2.3.CO;2

[2] D. Gerten, S. Rost, W. von Bloh, W. Lucht, "Causes of change in 20th century global river discharge". *Geophysical Research Letters* vol.35, no. 20, 2008, pp. 1–5. https://doi.org/10.1029/2008GL035258

[3] M. Rusticucci, B. Tencer, "Observed changes in return values of annual temperature extremes over Argentina", *Journal of Climate*, vol. 21, no. 21, 2008, pp. 5455–5467,DOI: https://doi.org/10.1175/2008JCLI2190.1

[4] B. Yan, Z. Xi, F. Huang, L. Guo, X. Zhang, (2016) "Climate Change Detection and Annual Extreme Temperature Analysis of the Amur River Basin Hindawi Publishing Corporation", *Advances in Meteorology*, vol. 2016, Article ID 6268938, 14 pages, http://dx.doi.org/10.1155/2016/6268938

[5] S. A. Rahmstorf, J. A. Cazenave, J. E. Church, R. F. Hansen, D. E. Keeling, R. C. Parker, J. Somerville, "Recent climate observations compared to projections". *Science* Vol. 316, no 5825,2007. pp. 709 DOI: 10.1126/science.1136843

[6] M. R. Rahman, H. Lateh, "Climate change in Bangladesh: a spatio-temporal analysis and simulation of recent temperature and rainfall data using GIS and time series analysis model". *Theor Appl Climatol* vol.128, 2017, pp. 27–4. https://doi.org/10.1007/s00704-015-1688-3

[7] A. Asfaw, B. Simane, A. Hassen, A. Bantider, "Variability and time series trend analysis of rainfall and temperature in northcentral Ethiopia: A case study in Woleka sub-basin". *Weather and Climate Extremes,* vol. 19, 2018, pp. 29-41, ISSN 2212-0947, https://doi.org/10.1016/j.wace.2017.12.002

[8] H. Park, G. Chung,"A Nonparametric Stochastic Approach for Disaggregation of Daily to Hourly Rainfall Using 3-Day Rainfall Patterns", *Water,* 2020, vol. 12, 2306; https://doi.org/10.3390/w12082306

[9] B. McKevith, "Nutritional aspects of cereals". *Nutr. Bull.* 2004, vol, 29, pp. 111–142. https://doi.org/10.1111/j.1467-3010.2004.00418.x

[10] A. Ortiz-Bobea, T. R Ault, C. M. Carillo, R. G. Chambers, D. B. Lobell, "Anthropogenic climate change has slowed global agricultural productivity growth". *Nat. Clim. Chang.* 2021, vol. 11, pp. 306–312. https://doi.org/10.1038/s41558-021-01000-1

[11] Q. Mi, X. Li, J. Gao, "How to improve the welfare of smallholders through agricultural production outsourcing: Evidence from cotton farmers in Xinjiang, Northwest China", Journal of Cleaner Production, vol. 256, 2020, https://doi.org/10.1016/j.jclepro.2020.120636

[12] S. Fahad, A. A. Bajwa, U. Nazir, S. A. Anjum, A. Farooq, A. Zohaib, et al. "Crop production under drought and heat stress: Plant responses and management options", *Frontiers in Plant Science,* vol. 8, 2017, pp.1–16. https://doi.org/10.3389/fpls.2017.01147\

[13] Y. Su, S. He, K. Wang, A. R. Shahtahmassebi, L. Zhang, J. Zhang, M. Zhang, M. Gan, "Quantifying the sustainability of three types of agricultural production in China: An emergy analysis with the integration of environmental pollution". *Journal of Cleaner Production*, 2020, https://doi.org/10.1016/j.jclepro.2019.119650

[14] S. Achli, T. E. Epule, D. Dhiba, A. Chehbouni, S. Er-Raki, "Vulnerability of Barley, Maize, and Wheat Yields to Variations in Growing Season Precipitation in Morocco". *Appl. Sci.* 2022, vol. 12, 3407. https://doi.org/10.3390/app12073407

[15] D. B. Lobell, M. B. Burke, "On the use of statistical models to predict crop yield responses to climate change", *Agricultural and Forest Meteorology*, vol. 50, pp. 1443–1452, 2010, doi:10.1016/j.agrformet.2010.07.008

[16] Sumiati, Musdalipa, A. T. Darhyati, A. T. Fitriyah, F. Baharuddin, "The impact of climate change on agricultural production with a cases study of Lake Tempe, district of Wajo, south Sulawesi". *Eurasia J Biosci* vol. 14, 2020, pp. 6761-6771

[17] B. Liu, S. Asseng, C. Müller, et al "Similar estimates of temperature impacts on global wheat yield by three independent methods". *Nature Clim Change* vol. 6, pp. 1130–1136, 2016, https://doi.org/10.1038/nclimate3115

[18] S. Ujjainia, P. Gautam, S. Veenadhari, (2020) "Crop Yield Prediction using Regression Model", *International Journal of Innovative Technology and Exploring Engineering (IJITEE).* vol. 9, no. 10, 2020, DOI: 10.35940/ijitee.J7491.0891020

[19] D. J. Olive, *Multiple linear regression. In Linear Regression*, pp. 17-83. Springer, Cham, 2017.

[20] V. S. Konduri, T. J. Vandal, S. Ganguly, A. R. Ganguly, "Data Science for Weather Impacts on Crop Yield". *Front. Sustain. Food Syst.* Vol. 4, no. 52, 2020. doi: 10.3389/fsufs.2020.00052

[21] J. H. Jeong, J. P. Resop, N. D. Mueller, D. H. Fleisher, K. Yun, E. E. Butler, et al. (2016) "Random Forests for Global and Regional Crop Yield Predictions". *PLoS ONE* vol. 11, no. 6, 2016. https://doi.org/10.1371/journal.pone.0156571

[22] S. Vincenzi, M. Zucchetta, P. Franzoi, M. Pellizzato, F. Pranovi, G. A. De Leo, et al., "Application of a Random Forest algorithm to predict spatial distribution of the potential yield of Ruditapes philippinarum in the Venice lagoon, Italy". *Ecol Model.* 2011; 222(8):1471–8.

[23] O. Mutanga, E. Adam, M. A. Cho, "High density biomass estimation for wetland vegetation using WorldView-2 imagery and random forest regression algorithm". *Int Journal Appl Earth Obs.* vol. 18, 2012, pp. 399–406.

[24] S. Fukuda, W. Spreer, E. Yasunaga, K. Yuge, V. Sardsud, J. Muller, "Random Forests modelling for the estimation of mango (Mangifera indica L. cv. Chok Anan) fruit yields under different irrigation regimes". *Agric Water Manage.* vol. 116, 2013, pp.142–50.

[25] S. Ben Mariem, D. Soba, B. Zhou, I. Loladze, F. Morales, I. Aranjuelo, "Climate Change, Crop Yields, and Grain Quality of C_3 Cereals: A Meta-Analysis of (CO_2), Temperature, and Drought Effects". *Plants* 2021, vol. 10, 1052. https://doi.org/10.3390/plants10061052

[26] D. García-León, R. López-Lozano, A. Toreti, M. Zampieri, "Local-Scale Cereal Yield Forecasting in Italy: Lessons from Different Statistical Models and Spatial Aggregations". *Agronomy,* 2020, vol. *10*, 809. https://doi.org/10.3390/agronomy10060809

[27] F. Giorgi, P. Lionello, "Climate change projections for the Mediterranean region". *Glob. Planet. Chang.* 2008, vol. 63, pp. 90–104. https://doi.org/10.1016/j.gloplacha.2007.09.005

[28] A. C. Russo, C. M. Gouveia, R. M. Trigo, M. L. R. Liberato, C. DaCamara, C. "The influence of circulation weather patterns at different spatial scales on drought variability in the Iberian Peninsula". *Front. Environ. Sci.* 2015, vol. 3, pp. 1–15. https://doi.org/10.3389/fenvs.2015.00001

[29] T. Porja, "Heat Waves Affecting Weather and Climate over Albania". *J Earth Sci Clim Change* vol.4, 2013, 149. Doi: 10.4172/2157-7617.1000149

[30] S. Dervishi, V. Picari, "Analysis of Urban Heat Island Phenomenon and Mitigation Strategies for Tirana, Albania", *Proceedings of the 16th IBPSA Conference Rome, Italy.* 2019. https://doi.org/10.26868/25222708.2019.211334

[31] E. Gjika, L. Basha, A. Ferrja, A. Kamberi, "Analyzing Seasonality in HPP Energy Production and External Variables", ITISE2021-International Conference on Time Series. Granada, 19th-21th July, 2021. Gran Canaria (SPAIN), https://www.mdpi.com/2673-4591/5/1/15

Open Science Index, Environmental and Ecological Engineering Vol:18, No:2, 2024 publications.waset.org/10013492.pdf

[32] E. Gjika, L. Basha, Energy production and consumption relying on climacteric variables (Albania case study), *Finance and Accounting towards Sustainable Development Goals International Conference*, Faculty of Economy, University of Tirana 26 November 2021. https://feut.edu.al/lajmerime/1167-international-conference-finance-and-accounting-towards-sustainable-development-goals-november-26-2021-tirana-albania-2

[33] D. M. Bates, D. G. Watts, Nonlinear Regression Analysis and Its Applications. Wiley,1988.

[34] T. Hastie, R. Tibshirani, J. H. Friedman, The Elements of Statistical Learning: Data mining, Inference, and Prediction. 2nd ed. New York: Springer, 2009

[35] S. M. Quiring, T. N. Papakryiaokou, "An evaluation of agricultural drought indices for the Canadian prairies". *Agric. For. Meteorol.* 2003, 168, pp. 49–62. https://doi.org/10.1016/S0168-1923(03)00072-8

[36] L. Michel, D. Makowski, D. "Comparison of statistical models for analyzing wheat yield time series". *PLoS ONE,* 2013, 8, e78615. https://doi.org/10.1371/journal.pone.0078615

[37] D. Wallach, D. Makowski, J. Jones, F. Brun, "Working with Dynamic Crop Models—Methods, Tools and Examples for Agriculture and Environment" *Academic Press*: Cambridge, MA, USA, 2014.

[38] W. H. Beyer, CRC Standard Mathematical Tables, 31st ed. Boca Raton, FL: CRC Press, pp. 536 and 571, 2002.

[39] S. Kotz, C. B. Read, N. Balakrishnan, B. Vidakovic, *Encyclopedia of Statistical Sciences*, 16 Volume Set, 2nd Edition, Wiley, 9686 Pages

[40] L. Breiman, "Random Forests". *Machine Learning*, vol. 45, 2001. pp. 5-32, https://doi.org/10.1023/A:1010933404324

Open Science Index, Environmental and Ecological Engineering Vol:18, No:2, 2024 publications.waset.org/10013492.pdf