ORIGINAL RESEARCH ARTICLE

A COMPARATIVE ANALYTICAL STUDY OF MANY REGRESSION MODEL APPROACHES, ARIMA MODEL AND A HYBRID MODEL FOR FORECASTING AREA, PRODUCTION, AND PRODUCTIVITY OF COCONUT IN KERALA.

ABSTRACT.

The study assesses coconut production area, production, and productivity forecasting models in Kerala. For the area of coconut production, the ARIMA model is selected because to its accuracy and normality of residual plots. With greater MAPE, RMSE, and R^2 values, the polynomial regression model offers the greatest fit for productivity and production. Due to its high R^2 and MAPE values, a hybrid model developed combining the best-fitting polynomial and ARIMA models provides more accurate data representation.

Keywords.*ARIMA, Forecasting, Mathematical Modeling,Regression*

Comment [ES1]: Add more to explain the full work.
Motivation of work should add

Comment [ES2]:

Comment [ES3]: the greater value not imply the best model

Comment [ES4]: cite all references in the paper

INTRODUCTION

Coconut production in Kerala, also known as the land of coconuts, holds significant importance in the region. The state of Kerala, located on the southern tip of the Indian subcontinent, experiences long growing seasons and benefits from the monsoon rains that flood the land. This favorable climate allows for frequent coconut crops, with each tree yielding approximately 20 to 30 coconuts every 40 days. Coconuts are deeply ingrained in the culture and daily life of Keralites. They are utilized for various purposes, including food, water, toddy (a local alcoholic beverage), mats, roofs, cooking oil, and rope production. Coconuts play a vital role in celebrations, religious offerings, and auspicious occasions. Kerala has historically been a major producer of coconuts, accounting for around 68% of India's total production by the late

1970s. Today, Kerala still contributes approximately 45% of India's coconut production, with most of the coconut cultivation concentrated in the southern Indian states, including Kerala and its neighboring states. However, coconut production in Kerala faces challenges, with one major threat being 'Root Wilt Disease'. Coastal regions with sandy soils and fertile interior places are ideal for coconut cultivation in Kerala, resulting in higher yields. Major cities like Trivandrum, Kochi, Kollam, Alappuzha, Ponnani, and Tanur are known for their high coconut cultivation. Although large coconut plantations are rare, more than 95% of coconut trees in Kerala are grown in the front and back yards of homesteads, with an average land holding of 20-25 cents per household and about 15 coconut trees. In recent years, coconut production in Kerala has experienced a decline, with the financial year 2020-21 witnessing a production of below 7,000 million nuts compared to 8,452 million nuts in FY18. Despite this decline, coconut products continue to be manufactured and have both domestic and export markets.

It is essential for several reasons to predict coconut production in Kerala, India. Coconut is crucial to Kerala's economy, making a substantial contribution to the production of jobs, revenue, and numerous businesses. Effective planning for coconut-related industries, such as oil extraction and coir manufacturing, depends on accurate estimates. Since it is a staple, production predictions guarantee a steady supply for consumption, lowering the risk of hunger. Accurate projections help with trade planning because coconut goods are also exported. These projections are used to determine government policies, subsidies, and resource distribution. Forecasts aid in preparing for weather impacts in areas susceptible to climatic change. They assist both producers and consumers by promoting research and environmentally friendly practices, stabilizing markets by lowering price swings. In essence, forecasting coconut production is vital for economic growth, employment, and agricultural sustainability in Kerala.

The purpose of this research is to evaluate the predictive effectiveness of several regression model techniques, such as linear regression, multiple regression, and polynomial regression, in forecasting crucial variables connected to coconut cultivation in Kerala. It also aims to explore the ARIMA (AutoRegressive Integrated Moving Average) model's usefulness in capturing temporal patterns and trends in coconut cultivation and generating reliable forecasts. In addition, the article intends to create a hybrid forecasting model that combines the capabilities of regression and ARIMA models to improve prediction accuracy. It seeks to examine and compare

the predicting accuracy of these various models utilizing important performance measures through a comparison study.

MATERIALS AND METHODS

The approach used in this study is based on historical data from 1956 to 2021 for the area of coconut cultivation,coconut production, and productivity in Kerala. Forecasting is done using linearregression, non-linear regression, ARIMA models, and a hybrid model. The measures of variability, regression formulae, MAPE values, $R^2$and predicted values for the models were calculated using SPSS and R software.

Secondary data from 1956 to 2021 were collected from the official website of the Directorate of Economics and Statistics to predict the area of coconut cultivation,coconut production, and productivity in Kerala. (Figure 1)

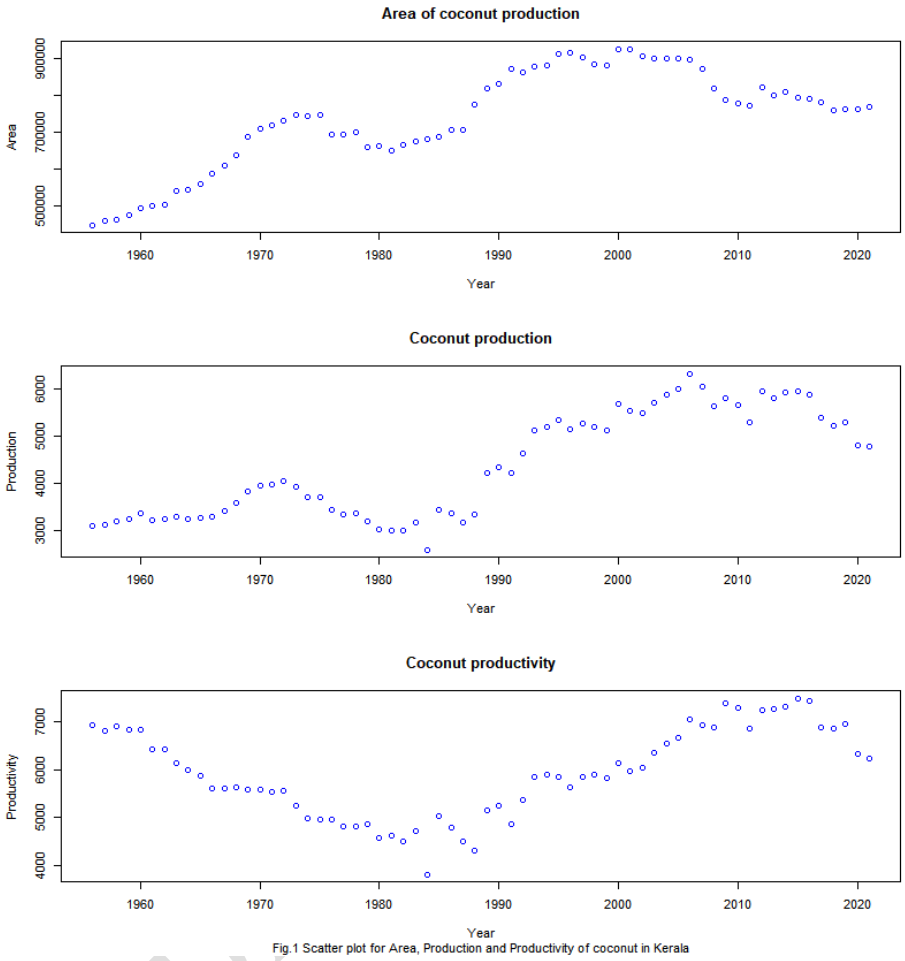

Figure 1: Scatter plot for area of coconut cultivation,coconut production, and productivity in Kerala

AUTO REGRESSIVE INTEGRATED MOVING AVERAGE (ARIMA) MODEL

An Autoregressive Integrated Moving Average (ARIMA) model[2] is a model for statistical analysis that is used to examine time series data and forecast future patterns. It is essentially a linear statistical technique and works best for modelling time series data's because of how easily it was developed and applied. Moving average models and autoregressive models are combined in ARIMA models.The AR models forecast the values of a variable x_t based on a

number p of previous values of the same variable number of autoregressive delays x_{t-k}, $1 \leq k \leq p$and incorporate a random disturbance e_t.. The MAmodels generate predictions of a variable x_t based on a number q of previous disturbances of the same variable prediction errors of past values e_{t-k}, $1 \leq k \leq q$. Combining the above two models AR(p) and MA(q) yields more adaptable models known as ARMA (p,q). Equation of the model are as follows;

General ARIMA(p, d, q) Model:ARIMA(p, d, q):

$$Yt^d = c + \varphi_1 Y_{t-1} + \varphi_2 Y_{t-2} + \ldots + \varphi_p Yt - p + \varepsilon t + \theta_1 \varepsilon t - 1 + \theta_2 \varepsilon t - 2 + \ldots + \theta q \, \varepsilon_{t-q} \quad (1)$$

Where $Yt^d$is the differenced time series at order ,θ_i 's are the autoregressive coefficients,$\varphi_i's$are the autoregressive coefficients, $\varepsilon_t$is the white noise error term at time , and c is a constant.

Box and Jenkins presented in 1976 a difference process characterized by an order of integration parameter d for transforming non-stationary time series into stationary time series. This mathematical transformation converts ARMA p, q models for non-stationary transformed time series to ARIMA p, d, q models, or Autoregressive Integrated Moving Average models.

The identification, estimation, diagnosis, and forecasting stages of the ARIMA model building strategy are iterative. The ability to identify a model can be done using the pattern of the data, time series plot, and acf and pacf. After selecting the tentative model, the parameters are calculated and their statistical significance is evaluated.Choose a new model, estimate parameters, test if not stationarity, and diagnose if residuals are low and white noise. If strong correlation persists, find, estimate, and diagnose a new model.

REGRESSION METHOD

The area, production, and productivity of coconut are expected to follow the past trend in univariate regression analysis. The regression modelsthat were applied in this investigation are illustrated below.

Linearregression model :

$$y = a + bt + \varepsilon_t (2)$$

Exponential regression model :

Comment [ES5]: Before using short form, use full one in first point.

Comment [ES6]: Add more details about regression models(assumptions)

$$y = ab^t + \varepsilon_t \quad (3)$$

Logarithmic regression model :

$$y = a + b\ln(t) + \varepsilon_t \quad (4)$$

Polynomial regression model :

$$y = a + b_1 t + b_2 t^2 + b_3 t^3 + \cdots . + \varepsilon_t \quad (5)$$

where y is the area/production/productivity of coconut in Kerala, a is the intercept, b denotes the derivatives, t stands for the explanatory variable (year), and ε_t denotes the error.

A hybrid model will be formed as follows

$$y = a\alpha(t) + b\gamma(t) + \varepsilon(t) \quad (6)$$

Where y is the area/production/productivity of coconut in Kerala, a and b represents constant or intercept,α is the best fitted ARIMA model for y is the area/production/productivity,γ is the best fitted regression model for y area/production/productivity, t -year and ε error term.

EVALUATION OF FORECASTS

Comment [ES7]: Add about R2

Comment [ES8]: Topic should change. These formula are used select best model.

To assess the performance of the included models, the Mean Absolute Percentage Error (MAPE)[1] is used. The performance of the model is improved by a decreased MAPE. MAPE is calculated using the following formula.

$$MAPE(\%) = \frac{1}{n}\sum_{i=1}^{m}\left(\frac{|e_i|}{y_i}\right) \times 100\% \quad (7)$$

where y_i represents the observed value, e_i represents the residue, and m denotes the number of data points. According to the Lewis scale, MAPE value below 10% implies a highly accurate forecast, while a MAPE of 10% to 20% implies a good prediction, MAPE between 20% and 50% implies a reasonable prognosis, and MAPE more than 50% implies an inaccurate forecast.

The Akaike Information Criterion (AIC) is frequently utilised for model evaluation in the context of ARIMA models[3]. While considering the trade-off between model complexity (number of parameters) and fit quality, it serves as a gauge for the model's goodness of fit. The AIC penalises models with excessive parameters to prevent overfitting. The negative twice the

log-likelihood and the negative twice the number of parameters is added to determine a model's AIC value. A model that balances the complexity of the model with its goodness of fit has a lower AIC value. To prevent overfitting, the AIC penalises models with more parameters.The model that should be used from a set of candidate models for a dataset is the one with the lowest AIC value.

An autocorrelation function (ACF) plot is a graphical depiction of the autocorrelation of a time series. Autocorrelation measures the link between a variable's present value and its past values. The ACF plot depicts the correlation coefficient along thevertical-axis and the lag on the horizontal-axis. It aids in examining how a time series compares to a delayed version of itself over a range of time.

A Q-Q plot is used to determine whether a piece of data conforms to a particular theoretical distribution. It compares the observed data quantiles to the theoretical distribution quantiles. If the points deviate from the straight line, it suggests a difference in the distributional shape between the observed data and the theoretical distribution. Outliers indicate data points that significantly deviate from the expected distribution.

The residual vs fitted plot is a scatter plot that displays the residuals (vertical axis) against the fitted values (horizontal axis) in a regression analysis. It is commonly used to examine the assumptions and identify potential issues in a regression model. It can help detect heteroscedasticity, non-linearity, outliers, randomness, heteroscedasticity, and non-linearity. Outliers are observations that significantly deviate from the overall pattern of the data, and should be investigated further to determine if they are influential or not.

Comment [ES9]: Repeating same characteritics

RESULT AND DISCUSSION

The models for area of coconut cultivation, coconut production, and coconut productivityforecasting in Kerala are assessed in this section. This section evaluates the forecasting models utilized in this investigation. Following that, the results are displayed.

The time series for the area of coconut cultivation, coconut production, and coconut productivityis far away from being stationery, as seen in Fig. 1. Still, a unit-root test was run to ensure. The test returned a p-value larger than 0.05, indicating that the null hypothesis was not

rejected.

Then, we found that ARIMA (0,2,1), ARIMA (1,2,1), ARIMA (0,2,2), and ARIMA (1,2,2) all had at least one significant coefficient among all possible combinations of second differenced ARIMA models. The ARIMA (0,2,1) model is preferred for modelling the area of coconut cultivation in Kerala, consideringthe MAPE value AIC, $R^2$and the normality of residual plots. In a comparable study, ARIMA (0,1,0) and ARIMA (0,1,3) were found to be the idealfits for coconut production and productivity in Kerala. Table (1) displays the MAPE, RMSE, and R^2 values of fitted ARIMA models. The residual plots, acf plots, and pacf plots for several ARIMA models are shown in Figure (2).

Table 1. ARIMA models for area of coconut cultivation,coconut production, and productivity in Kerala

	Fitted model	MAPE	RMSE	AIC	R^2
Area	ARIMA (0,2,1)	19.63188	21079.7	1463.97	0.9758
Production	ARIMA (0,1,0)	4.361268	268.6475	914.66	0.9421
Productivity	ARIMA (0,1,3)	3.612401	276.1907	925.65	0.9130

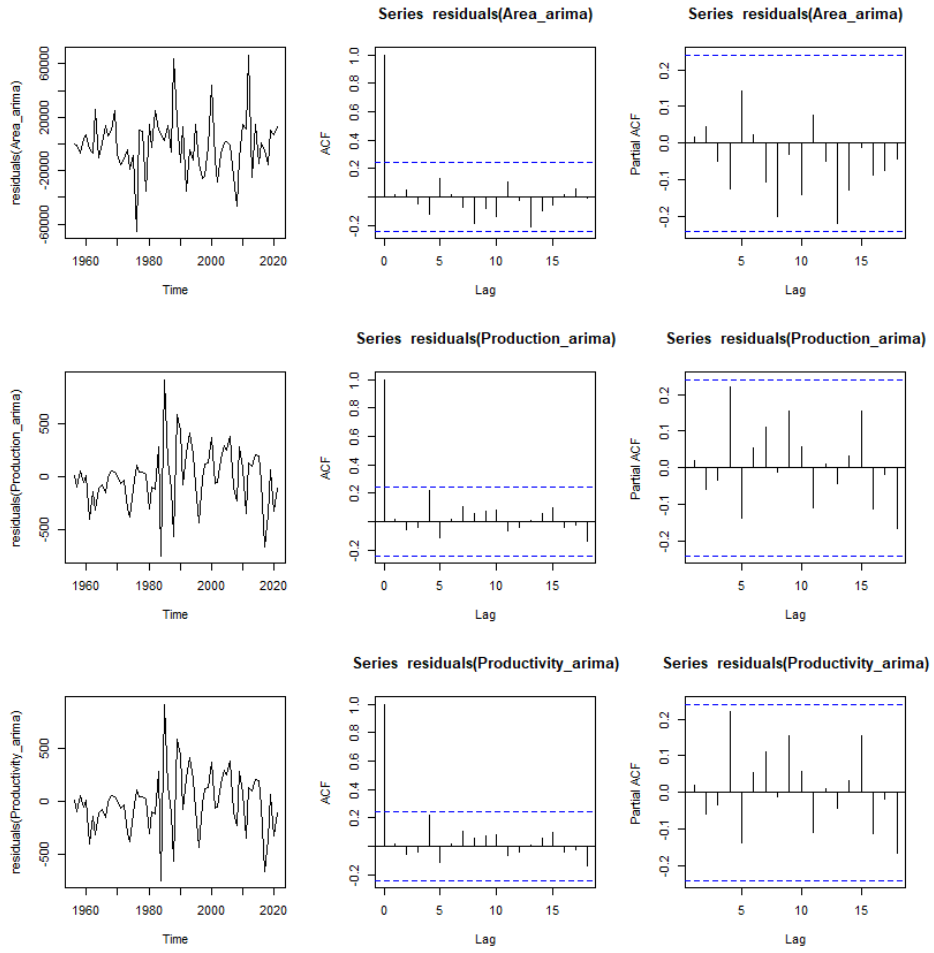

Figure.2: the residual plots, acf plots, and pacf plots for ARIMA models

Aunivariate regression analysis was performed on the area of coconut cultivation, coconut production, and coconut productivity in Kerala. Linear, exponential, log, and polynomial models are created with year as the explanatory variable and area, production, and productivity as the response variables. Table (2) shows the modelequation, R^2, and MAPE value for all the regression models of area of coconut cultivation, coconut production, and coconut productivity in Kerala. The residual vs. fitted and Q-Q plots of several regression models are shown in

Figures (3), (4), and (5).

Table 2. Regression models for area of coconut cultivation,coconut production, and productivity in Kerala

	Regression Model	Model Equation	MAPE	R^2
Area	Linear	-9893738.8+ 5346.7t	10.54783	0.585
	Exponential	$0.09499365206\ e^{0.0079698t}$	99.9981	0.5983
	Log	-80231509+ 10660799log(t)	10.50569	0.5884
	Polynomial	$738214+827490t-539128t^2-134862t^3$	5.486873	0.8513
Production	Linear	48.614t -92285.593	11.92	0.7073
	Exponential	$0.0000010465\ e^{0.0111286t}$	99.67	0.6953
	Log	-729578+ 96636logt	11.96	0.7069
	Polynomial	$4383.32+ 7523.79t+ 448.28t^2-2888.61t^3 - 2457.67t^4+ 969.07t^5+ 1016.50t^6 -1422.12t^7 - 522.24t^8 + 1129.58t^9$	03.68	0.9593
Productivity	Linear	-32222.497+19.186t	12.30	0.1447
	Exponential	$10.98\ e^{0.0031574t}$	99.84	0.1284
	Log	-281583+ 37855log(t)	12.32	0.1423
	Polynomial	$5929.35 + 2969.38\ t+ 5324.26\ t^2-3210.62t^3 - 2195.70\ t^4$	03.11	0.9322

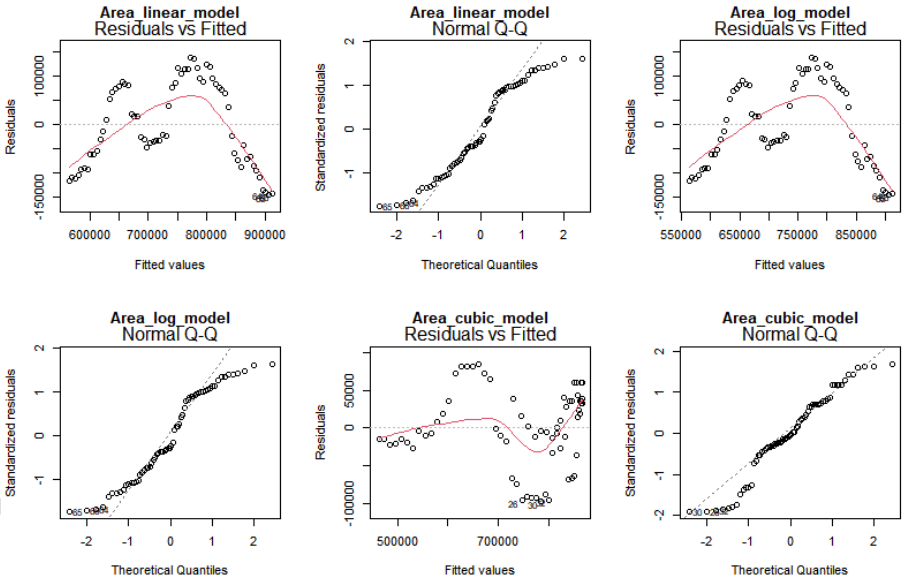

Figure.3. residual vs. fitted and Q-Q plots of various regression models for Area of coconut production in Kerala

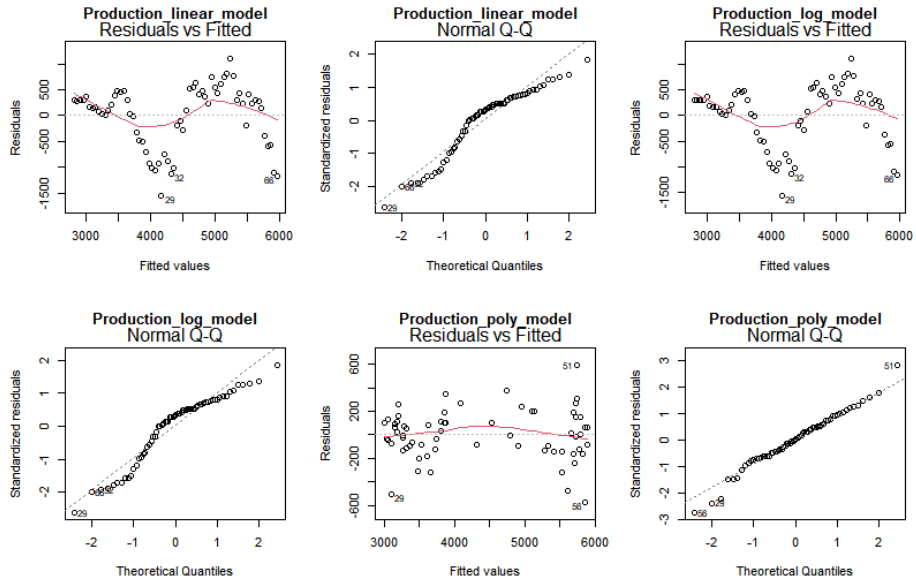

Figure.4. residual vs. fitted and Q-Q plots of various regression models for production of coconut in Kerala

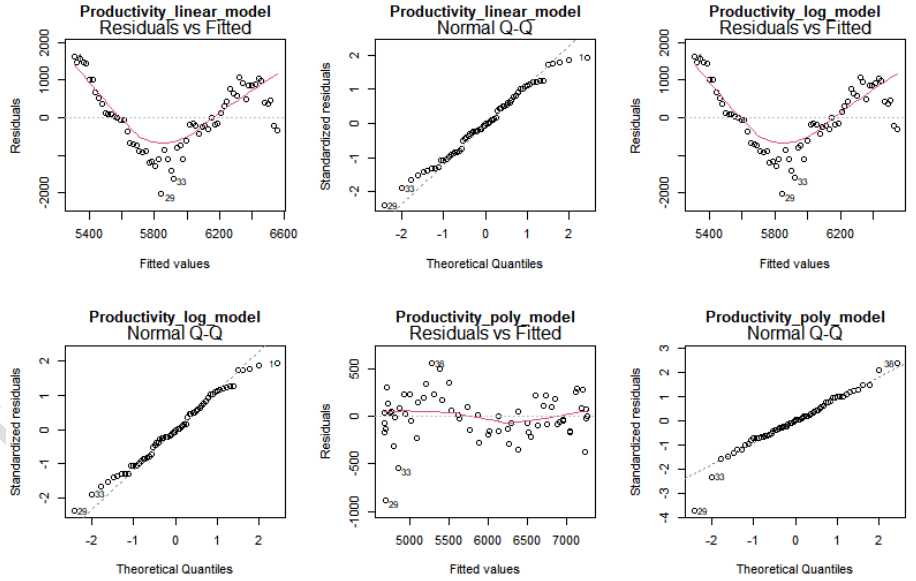

Figure 5. residual vs. fitted and Q-Q plots of various regression models for production of coconut in Kerala

A direct elimination of the exponential model is possible due to the high MAPE value. Linear and log models have nearly similar adjusted R^2 and MAPE values. It is clear from the MAPE and adjusted R^2 values that the polynomial regression model provides the best fit for area of coconut cultivation, coconut production, and coconut productivity in Kerala. The residual and Q-Q plot given in in fig 2 also indicates that polynomial regression model is a better model than others.

In the residual vs. fitted plot of the polynomial model, unlike the other two models, the spread of residuals is approximately the same across the x-axis, indicating homoscedasticity. The normal QQ plot for the area models exhibits S-shaped deviations for both the linear and log models. It implies that the tails of one dataset are heavier than those of the other. This implies that one dataset contains extreme values that the other dataset does not. The polynomial model, however, has a normal Q-Q plot that is almost a straight line, making it more accurate than the other two models. The same logic suggests that the polynomial model is better suited regression model for coconut production and productivity in Kerala

Comparing the MAPE value, RMSE value and the R^2 value given in **Table 1**, we found that ARIMA (0,2,1) is the best fit for the area of coconut production in Kerala. For production and productivity, the polynomial regression model was found to be better than their ARIMA models. **Table 3** provides predicted values for coconut production, area, and productivity in Kerala for the years 2022, 2023, and 2024 with a 95% confidence interval based on the best fit models of regression and ARIMA.

Comment [ES10]: Error values should be minimum for best model

The MAPE value and R^2 value of the hybrid model, which was created using the best-fitting polynomial and ARIMA models of coconut area, production, and productivity in Kerala, are shown in Table 4. The corresponding MAPE values for area, production, and productivity are 1.78, 3.56, 3.01, and 0.9799, 0.9664, 0.9395, respectively. It is evident from the MAPE and R^2 values that the hybrid model offers a better model than the corresponding polynomial or ARIMA model. This conclusion is also supported by the residual vs. fitted plot and QQ plot shown in Figure6

Comment [ES11]: Correct this sentence. Something is missing.

Table 3: Forecast values for coconut production, area, and productivity in Kerala for the years 2022, 2023, and 2024 with a 95% confidence interval based on the best fit models of regression and ARIMA

	Model	Year	Predicted value	95% LB	95% UB
Area	Regression	2022	713934.4	597895.5	829973.3
		2023	696759.0	577658.5	815859.5
		2024	678541.1	555821.7	801260.4
Production	Regression	2022	4877.871	3926.765	5828.977
		2023	5304.246	3601.424	7007.069
		2024	6282.488	3316.873	9248.102
Productivity	Regression	2022	6016.973	5434.3413	6599.605
		2023	5686.732	5066.3831	6307.080
		2024	5308.169	4639.7477	5976.591
Area	ARIMA	2022	767523.3	724872.4	810174.3
		2023	766237.7	696765.0	835710.4
		2024	764952.0	668416.2	861487.9
Production	ARIMA	2022	4788	4257.426	5318.574
		2023	4788	4037.655	5538.345
		2024	4788	3869.018	5706.982
Productivity	ARIMA	2022	6252.104	5693.591	6810.617
		2023	6120.883	5471.509	6770.257
		2024	6080.661	5313.554	6847.768

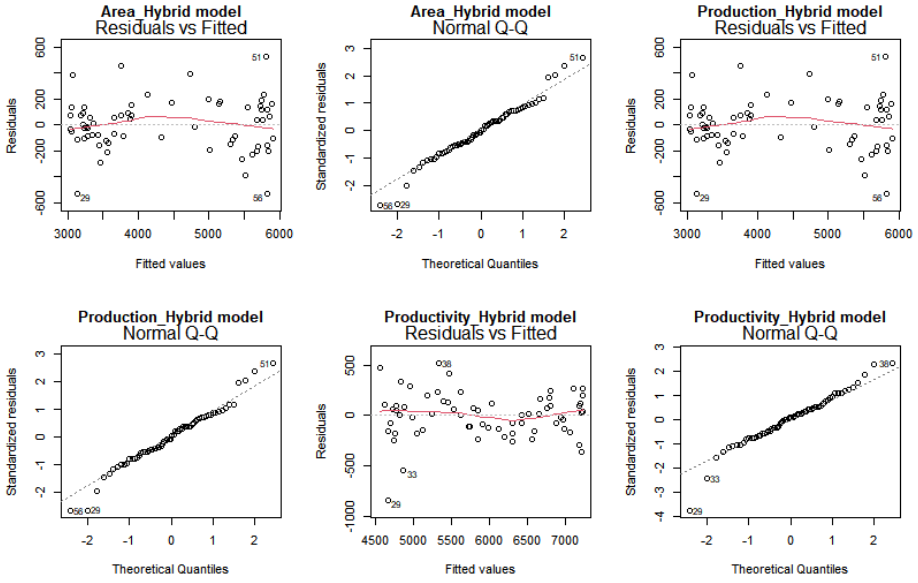

Figure.6. residual vs. fitted and Q-Q plots of Hybrid model for Area of coconut production in Kerala

Table 4. Hybrid models for area of coconut cultivation,coconut production, and productivity in Kerala

	MAPE	R2
Area	1.78	0.9799
Production	3.56	0.9664
Productivity	3.01	0.9395

CONCLUSION

Comment [ES12]: Remove table and figure from this section. Add those to results and discussion section.

There are inherent constraints to predicting coconut production using regression or ARIMA models. The time series data used in ARIMA models are presumed to be stationary, although non-stationary variables like climatic fluctuation and changing agricultural practices may prevent this from being the case. Regression methods rely on linear connections, which may not adequately account for the many non-linear interactions that influence coconut production. Both approaches could have trouble simulating shifting seasonal patterns that are impacted by outside variables like climate change. Furthermore, the effect of climate change on the production of coconuts is not sufficiently taken into consideration, necessitating the addition of climate data or more sophisticated models.

This study evaluates the models for area of coconut cultivation, coconut production, and coconut productivity in Kerala. Best fitted regression model, Arima model and Hybrid model are plotted in figure 7. The ARIMA (0,2,1) model is preferred for modeling area of coconut production in Kerala, considering AIC, residual statistics, and normality of residual plots. A similar study found that ARIMA (0,1,0) and ARIMA (0,1,3) are the best fits for coconut production and productivity in Kerala. Univariate regression analysis was conducted, focusing on year as the independent variable and area, production, and productivity as the response variables. The polynomial regression model provided the best fit for area, production, and productivity in Kerala. The residual and Q-Q plots showed that the polynomial model was more accurate than the linear, exponential, or logarithmic models.Comparing the MAPE value, RMSE

value, and R^2 value, the polynomial regression model was found to be better than their ARIMA models in case of production and productivity. The hybrid model, created using the best-fitting polynomial and ARIMA models of coconut area, production, and productivity in Kerala, offers a better model than the corresponding polynomial or ARIMA model. Table 5 gives the predicted values for coconut production, area, and productivity in Kerala for the years 2022, 2023, and 2024 with a 95% confidence interval based on the Hybrid Models. This conclusion is supported by the residual vs. fitted plot and QQ plot.

Table 5: Forecast values for coconut production, area, and productivity in Kerala for the years 2022, 2023, and 2024 with a 95% confidence interval based on the Hybrid Models

	Year	Predicted value	95% LB	95% UB
Area	2022	**757355.6**	**753277.0**	**749014.6**
	2023	**718970.8**	**714699.2**	**710171.0**
	2024	**795740.3**	**791854.9**	787858.1
Production	2022	**4862.359**	**5188.429**	**5936.538**
	2023	**4453.166**	**4764.652**	**5413.978**
	2024	5271.553	5612.206	6459.098
Productivity	2022	**6080.780**	**5808.755**	**5528.517**
	2023	**5618.584**	**5337.808**	**5030.656**
	2024	6542.976	6279.701	6026.378

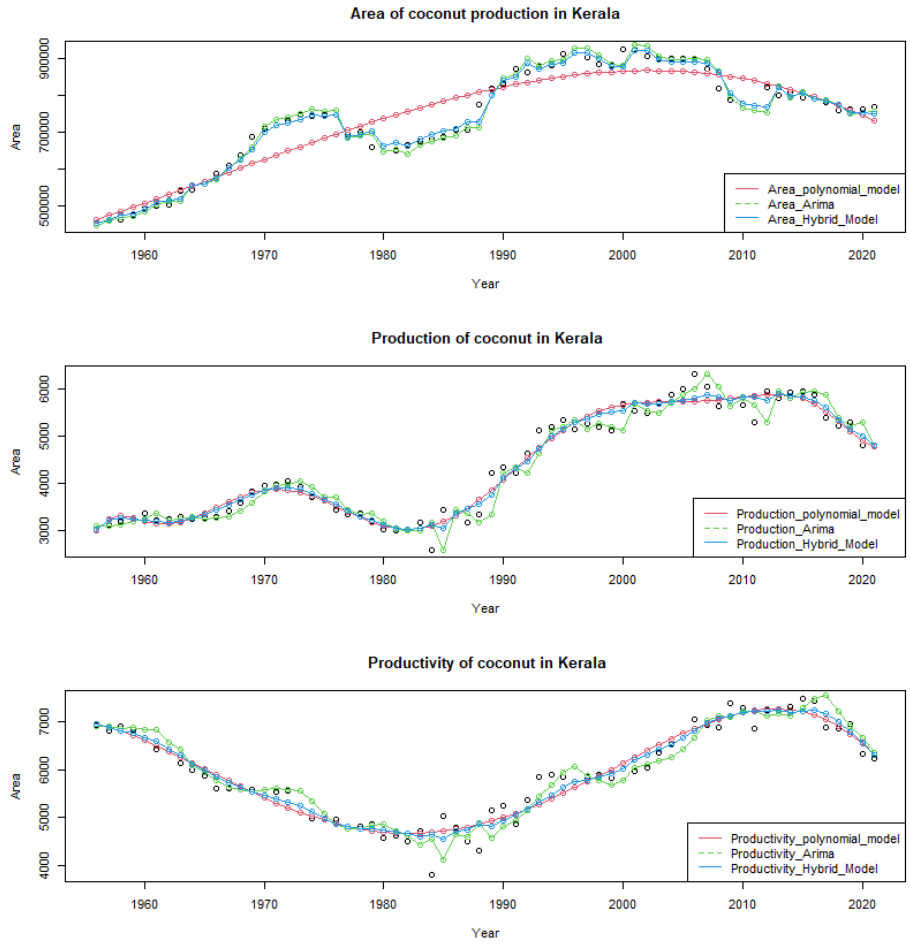

Figure 7: Best fitted models for Area, Production, and productivity of Coconut in Kerala.

REFERENCES

Comment [ES13]: Use one format for references.
Cite all references in the manuscript

1. Bishal Dey, Bidesh Roy, Subir Datta, Taha Selim Ustun. "Forecasting ethanol demand in India to meet future blending targets: A comparison of ARIMA and various regression models", Energy Reports, 2023
2. Box, G.E.P., and G.M. Jenkins. Time Series Analysis: Forecasting and Control, San Francisco, CA.: Holden Day, 1970.

3. Cristina Teresa Lim, Forecasting coconut production in the Philippines with ARIMA model, AIP Conference Proceedings 1643, 86 (2015).
4. Cristina Teresa Lim, Forecasting coconut production in the Philippines with ARIMA model, AIP Conference Proceedings 1643, 86 (2015).
5. Kevin Cullinane. "A short-term adaptive forecasting model for BIFFEX speculation: a Box—Jenkins approach", Maritime Policy & Management, 1992
6. Mark A. Stull. "Design Considerations for a 21st Century Ground Transportation System Based on Value-Capture Financing", Transportation Planning and Technology, 2008
7. P.J. Brockwell, and R.A. Davis, Introduction to Time Series and Forecasting. Springer. 1996.
8. Pradeep Kumar Ganjeer, A. Sahu, M. L. Lakhera. "Predictive Models for Pigeonpea in Northern Hills of Chhattisgarh, India", Advances in Research, 2018