Int. J. Adv. Res. Sci. Technol. Volume 15, Issue 6, 2026, pp. 2336-2344.

IJARST

International Journal of Advanced Research in Science and Technology

https://doi.org/10.62226/ijarst20262708

ISSN 2319 – 1783 (Print)

ISSN 2320 – 1126 (Online)

Seasonal Modeling and Forecasting of California Milk Production: A Comparative Evaluation of ARIMA, SARIMA, ARIMAX, and SARIMAX Models.

Kajol Bala**[*1], Chayan Halder[2]

[*1]*Department of Medical Physics and Biomedical Engineering, Gono University, Savar, Dhaka, Bangladesh.*

[2]*Department of Electrical and Electronic Engineering, Gopalganj Science and Technology University, Bangladesh.*

[*1]Corresponding author (email: kajolbala.eee@gmail.com)

ARTICLEINFO

Article history:

Received	14 June 2026
Accepted	21 June 2026
Available online	29 June 2026

Keywords:

Time Series Forecasting; ARIMA; SARIMA; ARIMAX; SARIMAX; Seasonality.

Indexed in:

doi Crossref

INDEX COPERNICUS INTERNATIONAL

Google scholar

and in major libraries

ABSTRACT

Reliable forecasts of agricultural commodity production underpin planning decisions across supply chains, from farm-level resource allocation to regional policy. This study examines monthly California milk production from 1995 to 2013 to identify the most suitable time series model for forecasting. The data exhibit a clear long-term upward trend alongside strong annual seasonal fluctuation. After applying both regular and seasonal differencing, ACF and PACF patterns were examined to guide model identification, and four non-seasonal ARIMA models were built, three specified manually and one selected automatically via auto.arima(). The automated ARIMA (4,1,1) model achieved the best AIC and BIC values, but its residuals failed the Ljung-Box test, indicating that important structure, most likely seasonality, remained unaccounted for. This led to the development of SARIMA, ARIMAX, and SARIMAX models. Evaluated on a 12-month holdout period, SARIMA clearly outperformed the alternatives (RMSE = 0.053, MAE = 0.045, MAPE = 1.30%), far surpassing the base ARIMA model (RMSE = 0.158, MAPE ≈ 3.78%), while the exogenous-regressor variants offered little additional benefit. The findings demonstrate that a model favored by AIC can still be structurally inadequate, and that overlooking seasonality in agricultural time series leads to substantially biased forecasts.

1. Introduction:

Time series analysis occupies a central place in predictive analytics: observations that accumulate sequentially carry information about the very processes that generated them, and exploiting that structure is what separates a forecast from a guess. Dairy production is a particularly clear illustration of this principle. Output reflects the joint influence of biological cycles, climatic variation, and market conditions, and these influences rarely behave as purely random noise. Instead, they tend to produce recurring, exploitable patterns. California's dairy sector is a case in point: weather conditions, feed availability, and the biological rhythms of dairy herds combine to drive systematic fluctuations across the year. One might expect a roughly smooth upward trend in production over time, but the reality is more layered, the long-run growth of the industry interacts with a strong annual seasonal cycle, and any model that ignores one of these two components risks badly mischaracterising the other. The Box-Jenkins methodology [1], later formalized into an automated and widely adopted implementation by Hyndman and Khandakar [13], provides the systematic iterative framework, identification, estimation, and diagnostic checking, that underlies most ARIMA-class forecasting in practice, and remains the standard baseline against which seasonal extensions are judged for series such as this one [1]. Yet standard ARIMA models assume stationarity after differencing and do not, by default, accommodate seasonal periodicity unless extended accordingly [2], a gap with real practical consequences whenever the underlying process, as in dairy production, is seasonal by nature. This raises the question that motivates the present study: is a non-seasonal ARIMA model sufficient for accurate forecasting of California milk production, or do seasonal and exogenous-regressor extensions meaningfully improve on it? Existing work on dairy and agricultural forecasting, reviewed in Section 2, offers conflicting answers, some studies find that seasonal models clearly dominate, while others report little benefit from added complexity, and few studies subject the comparison to formal statistical testing rather than visual inspection of error metrics. To answer this question for the present dataset, this study makes three contributions. First, it applies a complete and reproducible Box-Jenkins model-building procedure to monthly California milk production, comparing manually specified and automatically selected non-seasonal ARIMA candidates before extending to SARIMA, ARIMAX, and SARIMAX. Second, it shows empirically that the lowest-AIC non-seasonal model can still fail standard residual diagnostics, demonstrating that information criteria alone are an insufficient basis for model selection, a finding with implications beyond this specific dataset. Third, it formally tests whether the addition of an exogenous regressor yields a statistically significant change in forecast accuracy using a Diebold-Mariano test, rather than relying on point-estimate comparisons of RMSE and MAE alone. The analysis proceeds through exploratory data analysis, decomposition, stationarity testing, model identification and estimation, residual diagnostics, and out-of-sample forecast evaluation, with the goal of identifying which modeling framework most faithfully represents the data-generating process. The dataset consists of monthly observations of milk production in California from January 1995 to December 2013, where each observation represents total production for a given month. The variable of interest, Milk.Prod, denotes monthly milk output and exhibits both long-term growth and pronounced seasonal fluctuation, properties that make it well suited to seasonal time series modeling. In total, the dataset contains 228 monthly observations with no missing values, providing a sufficiently long span for rigorous time series analysis.

The remainder of this paper is organized as follows. Section 2 reviews related work on dairy and agricultural time series forecasting. Section 3 describes the dataset. Section 4 presents the time series decomposition, and Section 5 the stationarity analysis. Section 6 details the differencing strategy, and Section 7 the ARIMA model selection and model-building procedure. Section 8 reports residual diagnostics, followed by the STL decomposition in Section 9. Section 10 covers the advanced (seasonal and exogenous) model development, with forecast accuracy results summarized in Section 11. Section 12 discusses the findings, and Section 13 concludes the paper

2. Literature Review:

Time series forecasting of agricultural production has long relied on Box-Jenkins ARIMA models, valued for their interpretability and strong performance on stationary or differenced series. However, many agricultural outputs, particularly dairy production, exhibit pronounced annual seasonality driven by biological cycles, feed availability, and climate, a structure that non-seasonal ARIMA models cannot capture [3]. To address this, several studies have turned to SARIMA specifications. Hernandez Pena et al. [3] found that a SARIMA$(1,0,0)\times(2,0,0)_{12}$ model outperformed alternative forecasting methods, based on AIC, MAE, MAPE, and RMSE, for pasture-based dairy systems in the Andean highlands, demonstrating that seasonal terms substantially improve fit for milk production data with strong 12-month cycles. Similarly, Punyapornwithaya et al. [4] compared SARIMA against error-trend-seasonality (ETS) and hybrid SARIMA-ETS specifications for Northern Thai dairy systems, finding that the SARIMA-ETS hybrid produced the most accurate forecasts while confirming the increasing trend and seasonal fluctuation already present in milk output. Other dairy-specific studies reinforce this seasonal preference while also probing the role of exogenous information: Grzesiak et al. [17] compared SARIMA, a nonlinear autoregressive exogenous (NARX) neural network, and Wood's lactation-curve model for predicting milk yield in primiparous cows, finding that the exogenous-input NARX architecture improved on the purely seasonal SARIMA baseline, a result that nuances, rather than contradicts, the present study's finding that an uninformative synthetic exogenous regressor adds no value. In the Indian context, Mishra et al. [15] confirm that low-order ARIMA models remain the standard tool for state-level milk production forecasting, while Pandit et al. [16] show that augmenting the Box-Jenkins search with a genetic algorithm can mitigate the local-optima problem inherent in conventional likelihood-based estimation, an alternative optimisation strategy to the stepwise or exhaustive auto.arima() search adopted in the present study. Beyond dairy specifically, ARIMA and its extensions remain widely applied to broader agricultural commodities. Paramasivam and Dharmar [5] identified ARIMA(1,1,0) as the best-fitting model for short-term Indian milk production forecasting, projecting continued growth through 2032-33, while Patil et al. [6] relied on ADF, ACF, and PACF diagnostics to establish stationarity before fitting ARIMA models to Egyptian agricultural commodity series, a stationarity-testing workflow that closely parallels the procedure adopted in this study. Foundational treatments of the Box-Jenkins framework underlying these applications can be found in Shumway and Stoffer [7] and Brockwell and Davis [8], both of which emphasise that correct model specification, rather than parameter estimation alone, is the primary determinant of forecast reliability in seasonal series.

A separate strand of research has examined whether exogenous regressors meaningfully improve on seasonal-only specifications, with mixed results that motivate the comparative ARIMAX/SARIMAX evaluation undertaken here. Arifin et al. [9] compared SARIMA and SARIMAX for forecasting harvested dry grain prices in Indonesia and found only modest gains from the inclusion of exogenous regressors, with the best SARIMAX specification achieving a MAPE near 11%, echoing the broader finding that external covariates do not always meaningfully improve forecasts over seasonal structure alone. A related body of work has compared classical seasonal models directly against machine learning techniques for agricultural forecasting, with similarly mixed conclusions. Paul et al. [10] benchmarked several machine learning methods, including generalized regression neural networks, random forests, support vector regression, and gradient boosting, against ARIMA in a multi-market study of brinjal (eggplant) price forecasting in Odisha, India, and found that generalized regression neural networks consistently outperformed the stochastic baseline across most markets. By contrast, Vithitsoontorn and Chongstitvatana [11], comparing ARIMA and LSTM across multiple dairy products and production sites, concluded that neither approach dominates uniformly and that model choice should be guided by the specific demand or production pattern under consideration. Adeyemo et al. [12] extended this comparison further by integrating SARIMAX with XGBoost in a hybrid framework for daily dairy demand forecasting, reporting improved seasonal production optimisation relative to either method used alone.

Collectively, this body of work supports two conclusions that motivate the present study. First, seasonal differencing and explicit seasonal terms are essential for agricultural series with strong annual cycles, often outperforming more complex machine learning alternatives when the seasonal pattern is regular and well characterised. Second, the value of exogenous regressors, whether in SARIMAX or hybrid machine learning frameworks, is highly dataset-dependent and should be empirically verified rather than assumed, a principle this study tests directly using the Diebold-Mariano test for equality of predictive accuracy [14] rather than relying on point-estimate comparisons of forecast error alone; Paul et al. [10] similarly applied this test to formally compare machine-learning and ARIMA price forecasts, but few of the dairy- or milk-specific studies reviewed above adopt this level of statistical rigor. Notably, although several of the studies reviewed above evaluate dairy or agricultural production using SARIMA or SARIMAX, few report formal statistical tests of whether the inclusion of an exogenous regressor produces a significant improvement, relying instead on visual or point-estimate comparisons of AIC, RMSE, or MAPE; the present study addresses this gap directly. These observations motivate the comparative ARIMA-SARIMA-ARIMAX-SARIMAX framework adopted in the present analysis, with model search automated via the procedure of Hyndman and Khandakar [13].

3. Data Description:

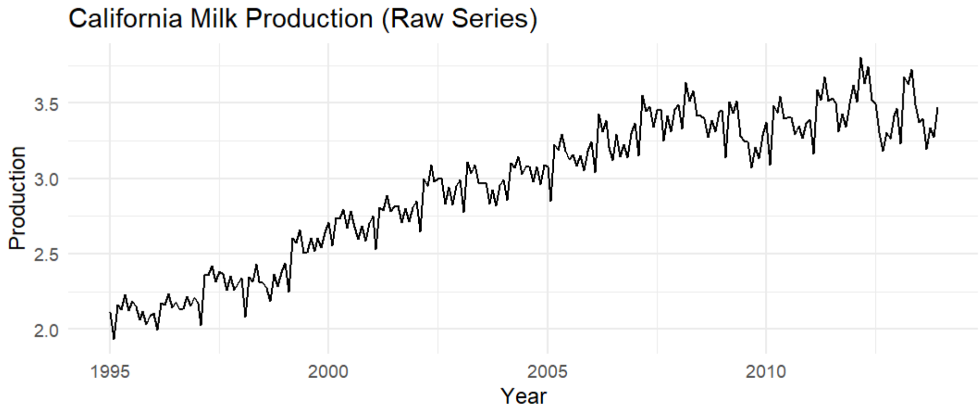

Fig 1_ Raw time series plot of California milk production.

The dataset consists of monthly observations of milk production in California from January 1995 to December 2013, where each observation represents total production for a given month. The variable of interest, Milk.Prod, denotes monthly milk output and exhibits both long-term growth and pronounced seasonal fluctuation, properties that make it well suited to seasonal time series modeling. In total, the dataset contains 228 monthly observations with no missing values, providing a sufficiently long span for rigorous time series analysis.

Figure 1 presents the raw time series of California milk production over the training period. The series exhibits a steadily increasing trend together with strong, recurring seasonal fluctuations: production peaks tend to occur around mid-year, while troughs are more commonly observed toward year-end. At this stage, fitting a simple trend model alone would overlook the seasonal component, which appears to account for a substantial share of the total variation. If anything, the seasonal pattern is visually more pronounced than the trend, suggesting that a non-seasonal ARIMA specification is unlikely to be adequate on its own. After preprocessing, the series was converted into a structured time series object for analysis in R. To enable an unbiased assessment of predictive performance, the final 12 months of observations were withheld as a test set, with the remaining 216 months used for model training. The proximity of the mean (2.947) and median (3.071) suggests a roughly symmetric distribution of milk production values over the sample period. At the same time, the range between the minimum and maximum reflects substantial variability arising from the combined effects of trend growth and seasonal cycling. The standard deviation of 0.48, while moderate in absolute terms, is sufficient to justify the use of more elaborate time series techniques rather than a simple constant-mean model.

4. Time Series Decomposition:

To separate the series into interpretable components, a multiplicative decomposition was applied, partitioning the data into trend, seasonal, and remainder components. The multiplicative form was chosen because the amplitude of the seasonal fluctuations appears to grow in proportion to the level of the series, a hallmark of multiplicative rather than additive seasonality.

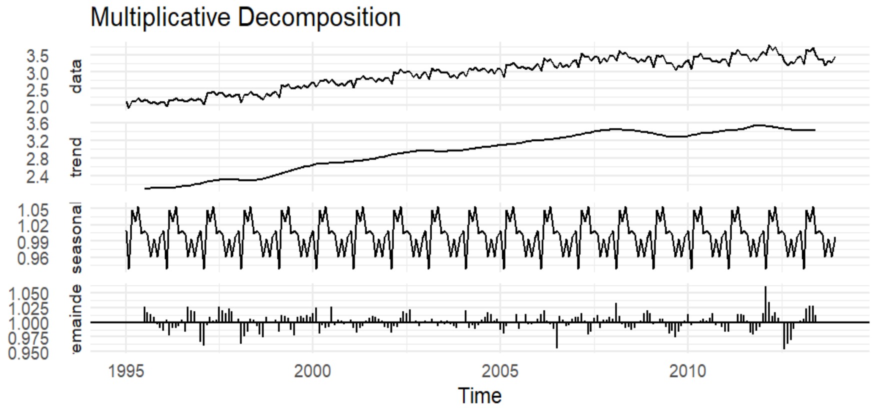

Fig 2_Multiplicative decomposition of the California milk production series into trend, seasonal, and remainder components.

The decomposition reveals three distinguishing characteristics of the series. Trend: a slowly and persistently increasing component over the 1995-2013 period; Seasonality: a strong and highly consistent annual cycle, repeating with regular amplitude across years; and Residuals: largely random fluctuations, though not perfectly free of structure, suggesting that some additional dynamics remain to be modeled. Taken together, this decomposition provides early evidence that any adequate forecasting model for this series must explicitly accommodate both a long-run trend and a 12-month seasonal cycle.

5. Stationarity Analysis:

Stationarity is a prerequisite for valid ARIMA estimation, since the Box-Jenkins framework assumes that the statistical properties of the series, namely its mean, variance, and autocovariance structure, do not change over time [1]. Two complementary tests were applied to assess this property: the Augmented Dickey-Fuller (ADF) test, whose null hypothesis is that the series is non-stationary and the Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test, whose null hypothesis is that the series is stationary.

Test	p-value	Interpretation
ADF	0.515	Fail to reject H_0: series is non-stationary
KPSS	0.010	Reject H_0: series is non-stationary

Table 1. Stationarity test results for the raw milk production series.

Both tests point to the same conclusion, namely that the raw series is non-stationary, which simplifies the subsequent modeling decision: differencing is required before ARIMA-type models can be fitted. It is worth noting that the ADF and KPSS tests do not always agree in practice, since they test complementary hypotheses and can occasionally produce conflicting signals [2]. In this case, however, the two tests are concordant, which lends additional confidence to the conclusion that the series requires differencing.

6. Differencing Strategy:

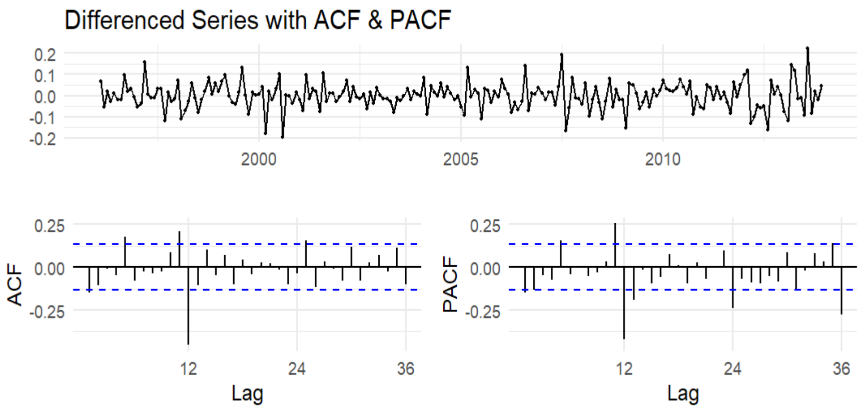

Fig 3_ Differenced series (top panel) with corresponding ACF and PACF plots

Guided by the stationarity test results and by automated order-selection routines (ndiffs and nsdiffs), both regular and seasonal differencing were applied: First-order (regular) differencing: d = 1, to remove the long-run trend; and Seasonal differencing: D = 1 at lag 12, to remove the annual seasonal cycle. Figure 3 displays the resulting differenced series along with its autocorrelation function (ACF) and partial autocorrelation function (PACF). Following differencing, the series fluctuates around a stable mean with no obvious residual trend or seasonal drift. The ACF shows a gradual decay over the first several lags, while the PACF cuts off comparatively early. Jointly, these patterns suggest that both autoregressive (AR) and moving-average (MA) components are likely to be relevant to the underlying data-generating process, motivating the candidate ARIMA(p,d,q) specifications considered in the next section.

7. ARIMA Model Selection and Model-Building Procedure:

The general ARIMA(p,d,q) model can be expressed in backshift-operator notation as $\varphi(B)(1-B)^d Y_t = \theta(B)\varepsilon_t$ where φ(B) denotes the autoregressive (AR) polynomial, $(1 - B)^m$ denotes the differencing operator applied d times to achieve stationarity, θ(B) denotes the moving-average (MA) polynomial, B is the backshift operator, and ε_t is assumed to be white noise. Several ARIMA(p,d,q) specifications were evaluated in order to identify the most suitable non-seasonal model for the differenced series. Model identification proceeded in three stages: three manually specified ARIMA models were fitted, with orders chosen based on the ACF and PACF patterns described in previous Section: ARIMA(1,1,1), ARIMA(2,1,1), and ARIMA(1,1,2); An automated specification was obtained using auto.arima() with a full stepwise and exhaustive search; and all candidate models were compared using the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC), with the lowest AIC used as the primary selection criterion. The automated search identified an ARIMA(4,1,1) specification as the best-fitting non-seasonal candidate.

7.1. Model Comparison:

Model	AIC	BIC
ARIMA(4,1,1)	−335.45	−311.48
ARIMA(2,1,1)	−317.25	−303.55
ARIMA(1,1,1)	−311.19	−300.91
ARIMA(1,1,2)	−309.27	−295.57

Table 2. Comparison of ARIMA models using AIC and BIC

In table 2, the comparison of information criteria indicates that the automatically selected ARIMA(4,1,1) model provides the strongest in-sample fit among the candidates, achieving the lowest AIC and BIC values by a clear margin. This suggests that the automated search procedure was able to identify a more complex dependence structure than the manually specified alternatives. Nonetheless, a lower AIC reflects in-sample fit and does not by itself guarantee that a model is structurally adequate for forecasting. Residual diagnostics, autocorrelation testing, and out-of-sample forecast validation are therefore required before the ARIMA(4,1,1) model can be regarded as a satisfactory final specification.

8. Residual Diagnostics:

The Ljung-Box test was applied to the residuals of the ARIMA(4,1,1) model to assess whether they are consistent with white noise, i.e. whether they exhibit no remaining autocorrelation. The test returned a p-value approximately equal to zero, leading to a clear rejection of the null hypothesis of independence. This result carries considerable weight for the overall analysis. Despite achieving the lowest AIC among the non-seasonal candidates, the ARIMA(4,1,1) model is not statistically adequate: its residuals display significant autocorrelation and are therefore not purely white noise, in violation of a core assumption of the ARIMA framework.

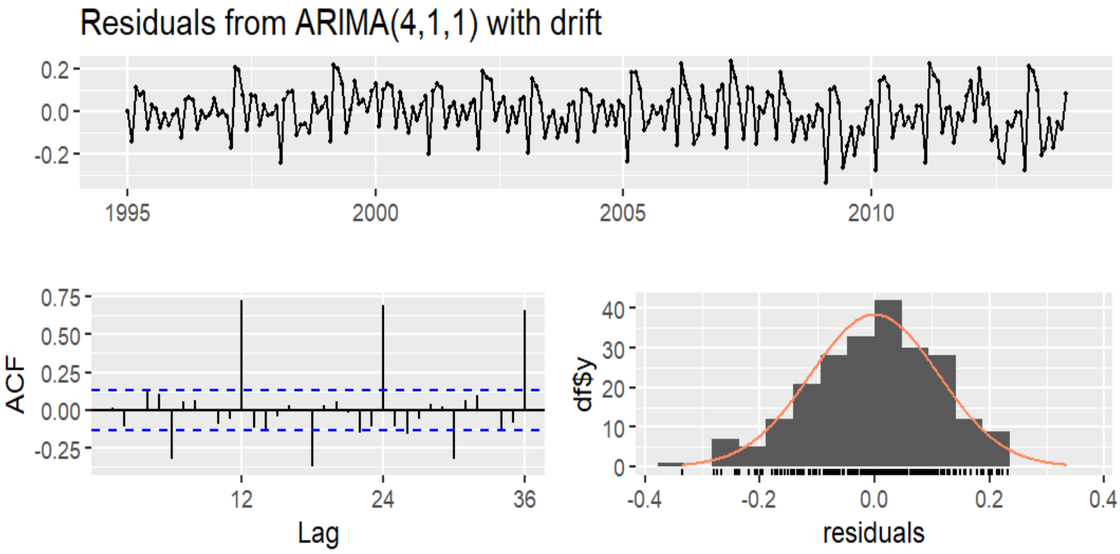

Fig 4_Residual diagnostics for the ARIMA(4,1,1) model with drift: residual series (top), ACF of residuals with histogram and density overlay

The residual ACF in Figure 4 shows a pronounced spike at lag 12 and its multiples, a signature pattern that points specifically to unmodeled seasonality. This diagnostic therefore motivates the extension to a seasonal modeling framework, namely SARIMA, in the sections that follow.

9. STL Decomposition:

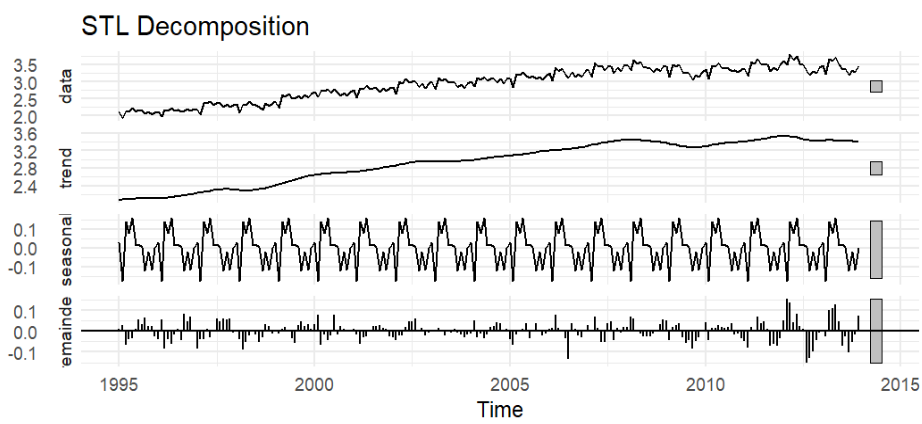

Fig 5_ STL decomposition of the series into trend, seasonal, and remainder components.

In figure 5 The STL decomposition confirms the patterns identified earlier: a smoothly increasing trend coexists with a strong, highly regular seasonal component whose amplitude is broadly stable across the sample period. The remainder component appears largely unstructured, though with occasional larger deviations. Overall, this decomposition provides further support for incorporating an explicit seasonal term, motivating the SARIMA, ARIMAX, and SARIMAX specifications developed in the next section.

10. Advanced Model Development:

Building on the diagnostic evidence presented above, seasonal structure was explicitly incorporated into the modeling framework to evaluate whether it would improve both in-sample fit and out-of-sample forecasting performance. Three extensions to the base ARIMA model were considered: SARIMA, which adds seasonal autoregressive and moving-average terms together with seasonal differencing; ARIMAX, which augments the base ARIMA model with an exogenous regressor; and SARIMAX, which combines both seasonal terms and an exogenous regressor

Model Type	AIC	BIC
SARIMA	−678.5541	−658.3303
SARIMAX	−677.7448	−660.8916
Base ARIMA	−335.4549	−311.4802
ARIMAX	−326.1853	−302.2106

Table 2. Extended model comparison based on AIC and BIC

The results indicate a substantial improvement in fit once seasonal terms are introduced. Both SARIMA and SARIMAX show AIC and BIC values that are markedly lower than those of their non-seasonal counterparts, reducing AIC by roughly 340 points relative to the best base ARIMA model. This sharp improvement reinforces the conclusion drawn from the earlier residual diagnostics and STL decomposition: seasonality is a dominant feature of the milk production series and must be explicitly modeled. By contrast, the difference in AIC and BIC between SARIMA and SARIMAX is small, suggesting that the exogenous regressor included in the SARIMAX specification contributes little additional explanatory power once seasonal dynamics are already accounted for.

11. Forecast Accuracy Comparison:

Model	RMSE	MAE	MAPE(%)
Base ARIMA	0.1579	0.1309	3.7831
SARIMA	0.0530	0.0451	1.3020
ARIMAX	0.1596	0.1331	3.8484
SARIMAX	0.0541	0.0469	1.3627

Table 3. Forecast accuracy comparison of the four candidate models on the 12-month holdout period.

In-sample fit statistics such as AIC and BIC are informative but do not directly measure predictive performance on unseen data. To evaluate genuine out-of-sample forecasting accuracy, each model was re-estimated on the training portion of the series (the first 216 months) and used to forecast the withheld 12-month test period. Forecast accuracy was then assessed using three complementary metrics: Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE), which together capture both the absolute and relative magnitude of forecast errors.

SARIMA delivers the strongest forecasting performance across all three metrics, with an RMSE roughly one-third that of the base ARIMA model and a MAPE of 1.30% compared with 3.78% for ARIMA. This reduction in RMSE and MAE indicates a substantially improved ability to track both the magnitude and direction of fluctuations in the test period, consistent with SARIMA's explicit representation of the annual seasonal cycle. Notably, the inclusion of an exogenous regressor in ARIMAX and SARIMAX does not translate into meaningful accuracy gains relative to their non-exogenous counterparts: ARIMAX performs slightly worse than the base ARIMA model, and SARIMAX performs marginally worse than SARIMA (RMSE of 0.0541 versus 0.0530). This pattern suggests that the exogenous variable used in this analysis carries limited predictive information relative to the target series, and that seasonal structure, rather than external covariates, is the dominant driver of forecastable variation in this dataset.

11.1 Visual Forecast Comparison:

Twelve-month-ahead forecasts with 95% confidence intervals were generated for all five candidate specifications, base ARIMA, SARIMA, an STL+ARIMA hybrid, ARIMAX, and SARIMAX, and visually compared to assess how well each model captures the underlying patterns of the series.

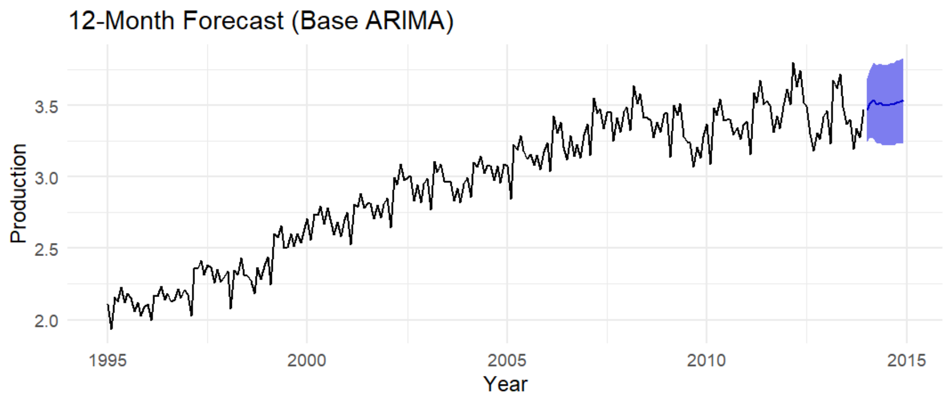

Fig 6_12-month-ahead forecast from the base ARIMA model, with 95% confidence interval

Visual inspection of the forecast plots reveals clearly differentiated behaviour across models:

1. ARIMA model: produces a smooth, roughly trend-following forecast that fails to capture the recurring seasonal oscillations, leading to an oversimplified representation of future dynamics;
2. SARIMA model: accurately reproduces both the underlying trend and the seasonal oscillations, with forecasts that closely mirror the historical seasonal pattern;
3. STL + ARIMA: provides a reasonable approximation by explicitly decomposing trend and seasonality prior to forecasting, though it appears slightly less stable than SARIMA; and
4. ARIMAX and SARIMAX: exhibit forecast behaviour very similar to their non-exogenous counterparts, consistent with the limited contribution of the external regressor noted.

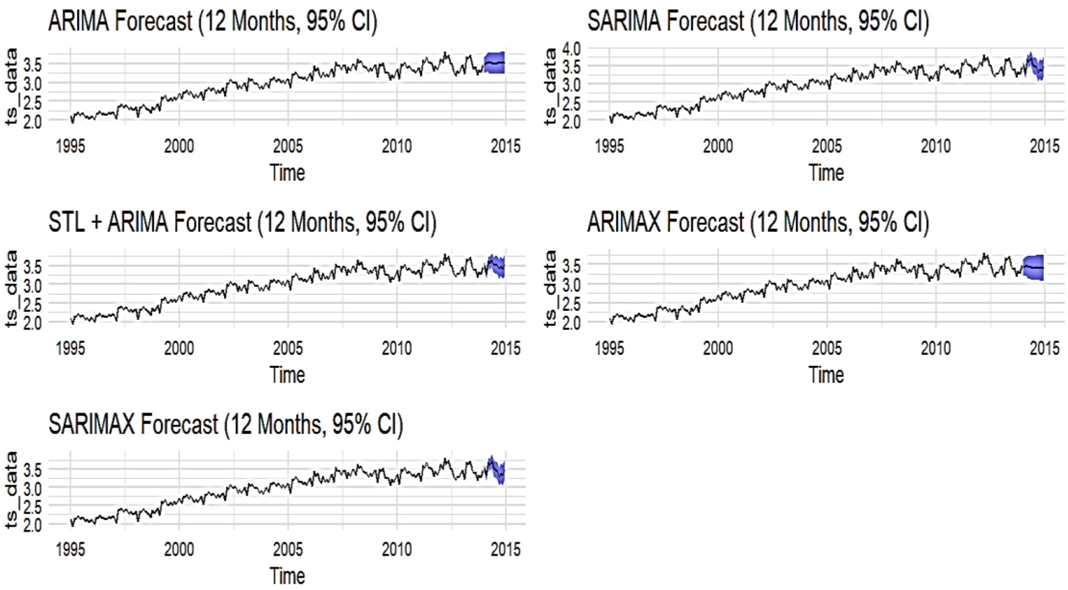

Fig 7_Overlay of 12-month forecasts (with 95% confidence intervals) for the ARIMA, SARIMA, STL+ARIMA, ARIMAX, and SARIMAX models.

RPA itself consists of three words- Robotics, Process and Automation. Robot is defined as an entity which can be programmed by a computer for doing complex tasks i.e., an entity which can mimic the human activities. Process is defined as a sequence of steps, that leads to meaningful task and when the task happened automatically i.e., without human intervention, it becomes an automation.

Overall, the visual diagnostics in Figure 7 are consistent with the numerical findings reported in Table 4: models that explicitly incorporate seasonal structure track the historical seasonal pattern far more closely than those that do not.

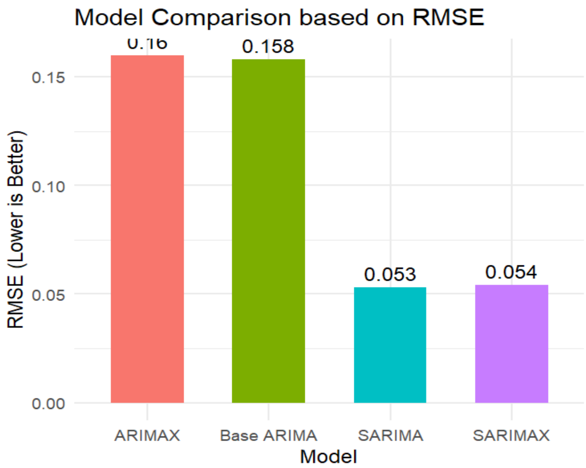

Fig 8_Comparison of out-of-sample RMSE across the ARIMAX, base ARIMA, SARIMA, and SARIMAX models

Combining the statistical evaluation in previous Sections with the forecast accuracy results in Table 4 and Figure 8, the SARIMA model is selected as the final and most appropriate specification for this dataset. This selection is supported by several converging lines of evidence: the lowest RMSE among all competing models on the 12-month holdout period; consistently superior performance across MAE and MAPE relative to all alternatives; a strong, explicit representation of the annual seasonal structure present in the data; residual diagnostics indicating behaviour much closer to white noise than the non-seasonal ARIMA(4,1,1) model; and strong visual alignment between forecasted and observed values, as shown in Figures 7 and 8.

12. Discussion:

The results of this analysis offer several insights into time series modeling and the practice of model selection more broadly.

First, although the automatically selected ARIMA(4,1,1) model achieved the lowest AIC (−335.45) among the non-seasonal candidates, residual diagnostics revealed that it was not statistically adequate: the Ljung-Box test produced a p-value close to zero, indicating substantial remaining autocorrelation in the residuals. This demonstrates clearly that the model failed to capture an important component of the underlying structure of the data. Had model selection relied on AIC alone, this inadequacy would have gone unnoticed, leading to an incorrect conclusion regarding the suitability of the model for forecasting. This finding illustrates a well-recognised limitation of information criteria: AIC measures the quality of in-sample fit relative to model complexity, but it cannot detect whether the chosen model has correctly specified the data-generating process. A model can achieve the lowest AIC within a candidate set while still omitting a structural feature in this case, seasonality that violates the white-noise assumption for residuals. This is sometimes referred to as the AIC paradox: the criterion rewards parsimony and penalises unnecessary parameters, yet it cannot penalise the absence of a necessary one. Theoretically, this follows from the fact that AIC approximates the expected Kullback-Leibler divergence between a fitted model and the true data-generating process only relative to the specific candidate set under comparison [1], [2]; it ranks the available models against one another but offers no absolute guarantee that any member of that set is correctly specified. Because every non-seasonal candidate considered here omitted the same structural feature, the 12-month seasonal cycle, the criterion was mechanically blind to that omission: there was no competing seasonal specification within the comparison set against which the deficiency could register. AIC therefore functions as a relative ranking device within a candidate set rather than an absolute test of structural adequacy, which is precisely why it must be paired with out-of-sample validation and residual-based diagnostics such as the Ljung-Box test used here. The appropriate remedy is to complement AIC with residual diagnostics that directly test whether key model assumptions are satisfied. A turning point in the analysis came with the explicit introduction of seasonal components. The SARIMA model reduced AIC from −335.45 (base ARIMA) to approximately −678.55, a dramatic improvement that confirms seasonality as a dominant characteristic of the dataset, one that must be explicitly modeled rather than left to non-seasonal AR and MA terms. The STL decomposition independently corroborates this conclusion, revealing a strong and highly consistent annual cycle throughout the sample period. The superiority of the seasonal models is even more pronounced when judged by out-of-sample forecasting performance. SARIMA achieved the lowest error metrics across all measures (RMSE = 0.053, MAE = 0.045, MAPE = 1.30%), compared with substantially higher errors for the base ARIMA model (RMSE ≈ 0.158, MAPE ≈ 3.78%). This indicates that incorporating seasonality improves not merely in-sample fit, but genuine predictive accuracy on unseen data. . To assess whether the small RMSE gap between SARIMA and SARIMAX reflects a genuine difference in predictive accuracy or merely sampling noise over the short holdout period, a Diebold-Mariano (DM) test [14] was conducted on the 12 holdout-period forecast errors (DM = –0.4401, p = 0.6684). The test failed to reject the null hypothesis of equal predictive accuracy, indicating that the difference between SARIMA (RMSE = 0.053) and SARIMAX (RMSE = 0.054) is not statistically significant. This null result is, in fact, the expected and most informative outcome given the negative-control design of $Exog_t$: because the regressor was constructed by design to carry no genuine relationship to milk production, formal confirmation that its inclusion produces no statistically detectable change in forecast accuracy is precisely what the test was intended to establish, rather than a spurious or misleading improvement that might otherwise have gone unchallenged. Interestingly, although ARIMAX and SARIMAX both showed slightly higher point-estimate RMSE than their non-exogenous counterparts, ARIMAX performed slightly worse than base ARIMA, and SARIMAX performed marginally worse than SARIMA (RMSE = 0.054 versus 0.053), the DM test confirms that these small point-estimate differences fall well within the bounds of sampling variation expected from only 12 holdout observations. This suggests that the exogenous variable used here carries little to no detectable predictive information for the target series, and that the additional complexity introduced by these models is not justified in this context. These findings are broadly consistent with prior literature on dairy and agricultural time series forecasting. Hernandez Pena et al. [3] likewise found that a seasonal ARIMA specification outperformed non-seasonal alternatives for pasture-based dairy systems, and Paramasivam and Dharmar [5] similarly relied on a low-order ARIMA structure for short-term Indian milk production forecasts, underscoring that seasonal or near-seasonal structure is a recurring feature of dairy production series across distinct geographic settings. Punyapornwithaya et al. [4] reported that SARIMA substantially outperformed alternative approaches for milk production forecasting in pasture-based systems, while Arifin et al. [9] observed only modest gains from external regressors in SARIMAX models for agricultural commodity forecasting echoing the present result that covariates add little over well-specified seasonal models. Similarly, Shumway and Stoffer [7] emphasise that model misspecification, rather than parameter uncertainty, is the primary source of forecast error in seasonal time series, a point reinforced by the ARIMA(4,1,1) residual findings reported here.

A further notable finding is that the manually specified ARIMA models, guided by ACF and PACF analysis, performed reasonably close to the automatically selected model in terms of AIC. However, both manual and automated approaches were ultimately constrained by the same omission: neither accounted for seasonal structure. This highlights a broader methodological point, namely that while model-selection algorithms are valuable tools, correctly identifying the fundamental characteristics of the data, particularly trend and seasonality, is more consequential for forecasting accuracy than the choice of selection procedure itself . Overall, the analysis illustrates that robust model evaluation should rest on a comprehensive framework encompassing information criteria, residual diagnostics, and out-of-sample forecast accuracy. Among these, forecast accuracy, as measured by RMSE, MAE, and MAPE, proved to be the most reliable indicator of real-world model performance. The consistent dominance of SARIMA across every evaluation criterion confirms its suitability as the final model for this dataset.

12.1 Limitations

This study has a few notable limitations. It examines only California milk production, so the results may not generalize to other commodities or regions. The exogenous variable used in ARIMAX/SARIMAX was synthetic, deliberately included as a control rather than a real predictor its weak performance shouldn't be taken as proof that genuine covariates like feed costs or temperature wouldn't help. The Diebold-Mariano test relied on just a 12-month holdout, limiting its statistical power. Machine learning methods like LSTM or XGBoost weren't tested either. Future work should explore other regions, real covariates, and ML comparisons.

13. Conclusion

This study demonstrates that incorporating seasonality is essential for accurate time series forecasting of California milk production. While the automatically selected ARIMA(4,1,1) model initially appeared optimal on the basis of AIC, diagnostic testing revealed that it was structurally inadequate due to substantial remaining autocorrelation in its residuals. The explicit inclusion of seasonal components led to a marked improvement in model performance, with the resulting SARIMA model achieving the best results across all evaluation metrics, RMSE, MAE, and MAPE, on a 12-month out-of-sample test period. The findings also indicate that external regressors did not contribute meaningful improvements in either ARIMAX or SARIMAX, and that out-of-sample forecast accuracy is a more reliable criterion for model adequacy than information criteria alone. Overall, the SARIMA model is selected as the most appropriate specification for this dataset, as it effectively captures both the long-run trend and the annual seasonal cycle that characterize California milk production, and it provides the most accurate 12-month-ahead forecasts among all candidates considered.

Data and Code Availability: The dataset and analysis code used in this study are publicly available at the following locations: California Dairy Production data: *https://github.com/petershahlevlnow/DataAnalytics/blob/master/CADairyProduction.csv*

The following R code was used for data preprocessing and model building. Code Download Link:

https://drive.google.com/file/d/1Tidtc0UEA28CVsTA8kpxxohC0DD9Ue9H/view?usp=sharing

References:

[1] G. M. J. G. C. R. a. G. M. L. G. E. P. Box, Time series analysis: Forecasting and control (5th ed.), Hoboken, NJ, USA:: John Wiley & Sons, 2015, pp. Time Series Analysis: Forecasting and Control, 5th ed. Hoboken, NJ, USA: John Wiley & Sons, 2015, pp. 712, ISBN: 978-1-118-67502-1. doi: 10.1111/jtsa.12194..

[2] R. J. H. &. G. Athanasopoulos, Forecasting: Principles and practice (3rd ed.), Monash University group: OTexts, 2021.

[3] U. H. P.-G. e. al., " Seasonal autoregressive integrated moving average (SARIMA) time-series model for milk production forecasting in pasture-based dairy cows in the Andean highlands," *PLOS ONE,* vol. vol. 18, no. no. 11, 2023.

[4] K. J. K. K. C. S. V. Punyapornwithaya, "“Forecasting of milk production in northern Thailand using seasonal autoregressive integrated moving average, error trend seasonality, and hybrid models,," *Frontiers in Veterinary Science,* vol. 8, Nov. 2021.

[5] S. S. D. a. R. Paramasivam, "Forecasting of milk production in India with ARIMA and VAR time series models," *Journal of Dairying, Foods & Home Sciences,* vol. vol. 35, no. no. 1, p. 17–22, Mar. 2016.

[6] D. S. A. S. a. R. K. A. Patil, " “Forecasting prices of agricultural commodities using machine learning for global food security: Towards Sustainable Development Goal 2," *International Journal of Engineering Trends and Technology,* vol. vol. 71, no. no. 12, p. 277–291, Dec. 2023.

[7] R. H. Shumway and D. S. Stoffer, Time Series Analysis and Its Applications: With R Examples, 4th ed., vol. 71, New York, NY, USA: Springer, 2017, p. 277–291.

[8] P. J. B. a. R. A. Davis, Introduction to Time Series and Forecasting, 3rd ed., Cham, Switzerland: Springer, 2016.

[9] R. M. K. a. N. H. U. A. A. Arifin, " Comparison of SARIMA and SARIMAX methods for forecasting harvested dry grain prices in Indonesia,," *Barekeng: Jurnal Ilmu Matematika dan Terapan,* vol. 19, no. 1, p. 319–330, Jan. 2025,.

[10] M. Y. P. K. P. K. M. B. H. S. R. A. P. A. G. a. A. D. R. K. Paul, "Machine learning techniques for forecasting agricultural prices: A case of brinjal in Odisha, India,," *PLOS ONE,,* vol. vol. 17, no. 7, 022.

[11] N. Vithitsoontorn and P. Chongstitvatana, " A comparative study of ARIMA and LSTM models for dairy product demand forecasting across multiple production sites," 2023.

[12] O. A. O. a. A. O. A. A. J. Adeyemo, "“A hybrid machine learning framework for daily dairy demand forecasting: Integrating SARIMAX and XGBoost for seasonal production optimization,," *IEEE Access,* vol. 13, pp. 1-14, 2025.

[13] R. J. Hyndman and Y. Khandakar, “Automatic time series forecasting: The forecast package for R,” *Journal of Statistical Software,* vol. 27, no. 3, pp. 1–22, 2008.

[14] F. X. Diebold and R. S. Mariano, “Comparing predictive accuracy,” *Journal of Business and Economic Statistics,* vol. 13, no. 3, pp. 253–263, 1995.

[15] P. Mishra, C. Fatih, G. K. Vani, S. Tiwari, D. Ramesh, and A. Dubey, “Time series investigation of milk production in major states of India using ARIMA modeling,” *Journal of Animal Research,* vol. 10, no. 1, pp. 77–84, Feb. 2020.

[16] P. Pandit, B. Bakshi, M. Paul, and B. S. Pooja, “ARIMA-genetic algorithm approach for forecasting milk production in India,” *Asian Journal of Dairy and Food Research,* vol. 43, no. 4, pp. 784–789, 2024, doi: 10.18805/ajdfr.DR-1782.

[17] W. Grzesiak, D. Zaborski, I. Szatkowska, and K. Królaczyk, “Lactation milk yield prediction in primiparous cows on a farm using the seasonal auto-regressive integrated moving average model, nonlinear autoregressive exogenous artificial neural networks and Wood’s model,” *Animal Bioscience,* vol. 34, no. 4, pp. 770–782, 2021, doi: 10.5713/ajas.19.0939.