Time Series Forecasting of Global Wildfire Trends

Kuba Kołpa

28 April 2025

Contents

1 Introduction

Wildfires are among the most destructive natural hazards affecting both ecosystems and human societies. Each year, they burn millions of hectares of land, displacing communities, destroying biodiversity, and degrading air quality. In recent decades, concerns over wildfires have intensified due to their increasing frequency and severity. Climate change is widely seen as a key driver, as rising global temperatures, prolonged droughts, and changing weather patterns create ideal conditions for fires to ignite and spread. In addition to their immediate damage, wildfires contribute significantly to global carbon emissions, which in turn feed back into the warming climate. The resulting consequences of wildfires make them both an immediate hazard and a lasting environmental threat. For these reasons, it is crucial to understand the trends behind wildfire activity and develop models that can help anticipate future patterns.

The aim of this project is to analyze and forecast wildfire activity over time using a wide range of statistical and machine learning techniques. Specifically, we seek to predict trends in wildfire occurrences, the area of land affected, and associated environmental impacts such as carbon emissions. These insights can support more informed environmental planning, resource allocation, and public policy. In addition to producing forecasts, this work compares several time series forecasting methods to evaluate their accuracy and suitability for different types of wildfire-related data.

To support this analysis, we collected several time series datasets from Our World in Data that reflect both wildfire behavior and broader climate trends. These include:

- Weekly area burned by wildfires
- Weekly cumulative area burned
- Weekly cumulative CO_2 emissions from wildfires
- Monthly global average surface temperature
- Monthly global temperature anomalies
- Monthly average surface temperatures by year

The wildfire datasets primarily span from 2012 to 2024 and are reported on a weekly basis, with approximately 680-700 observations each. One dataset, cumulative CO_2 emissions, extends further back to 2003, providing a longer historical context. All wildfire data are derived from satellite monitoring systems and reflect global totals.

The temperature-related datasets are monthly and cover significantly longer periods: the global temperature anomaly data extends back to 1940, and the average surface temperature data begins in 1950, both continuing through 2024. These datasets contain around 900-1,000 observations each and provide a broad picture of long-term climate behavior, seasonal variation, and warming trends. Minimal data cleaning was required, although transformations like differencing or log-scaling were applied in some cases to stabilize variance or address non-stationarity.

This project applies a comprehensive range of forecasting methods. We begin with simple benchmark models, such as the naive and seasonal naive approaches, which offer a baseline for comparison. Next, we explore statistical models including Exponential Smoothing (ETS) and ARIMA, both of which are widely used in time series forecasting due to their flexibility and interpretability. These models allow us to capture level, trend, and seasonal components in the data, with ARIMA offering additional power through autoregressive and moving average terms.

We also incorporate modern forecasting tools that can handle more complex patterns. The Prophet model, developed by Meta, is particularly suited to time series with strong seasonality and trend changes. It is robust to missing data and often requires minimal tuning. In addition, we apply a Neural Network Autoregression (NNAR) model, which uses a feed-forward neural network to learn non-linear relationships in the data. This machine learning approach offers an alternative to traditional statistical models, especially where patterns are not strictly linear or additive.

Together, these forecasting methods allow us to evaluate not only the expected direction of future wildfire trends but also the strengths and limitations of each modeling technique. In the following sections, we present the results of our analysis and compare model performance across the different time series.

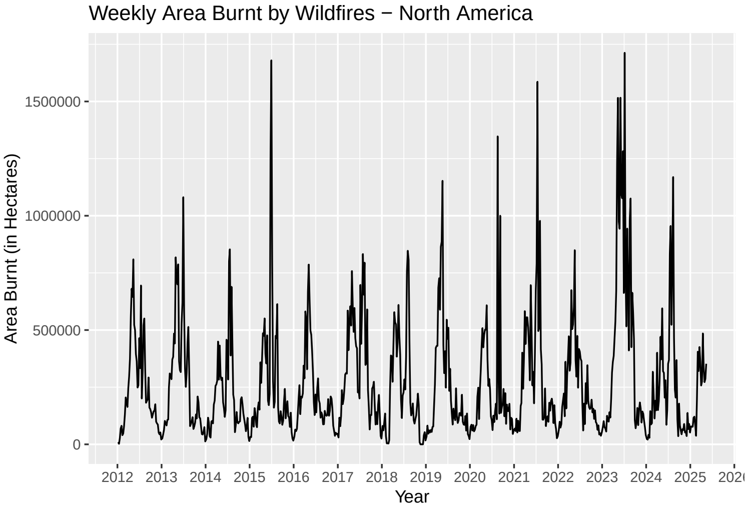
2 Time Series Graphics

To understand the structure of the wildfire and climate datasets, we begin with a series of time series visualizations. These include time plots, seasonal plots, seasonal subseries plots, and autocorrelation function (ACF) plots. Each provides a different lens on the data's behavior over time.

Time plots show the overall trend and fluctuations in each series. Seasonal plots and subseries plots highlight recurring within-year patterns, helping us assess the strength and consistency of seasonality. ACF plots reveal how strongly current values are related to past values, which is useful for identifying autocorrelation and determining the need for differencing.

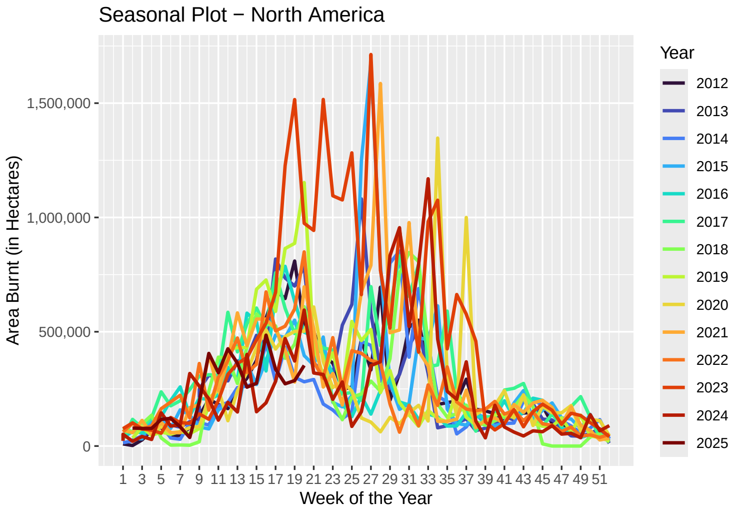
These graphics provide important insights into trends, seasonal cycles, and dependencies within the data. They also help us determine whether the series are stationary and guide the selection of appropriate forecasting methods.

2.1 Time Series Plot



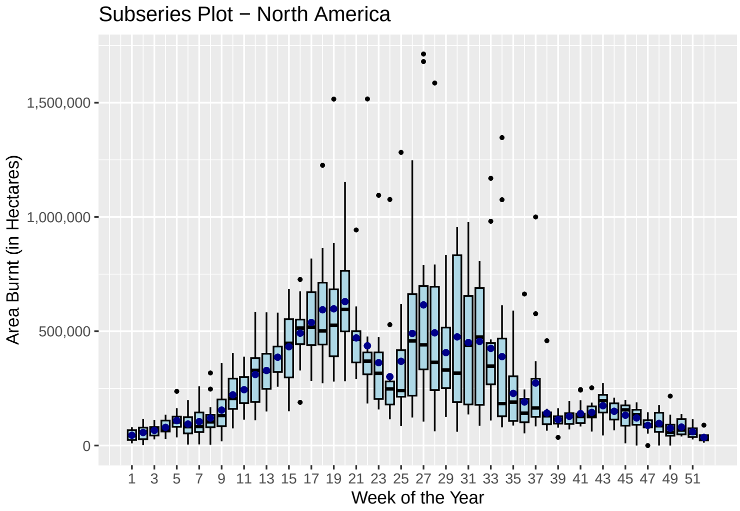
The time series plot of the aggregated weekly area burnt by wildfires across North America reveals a highly volatile structure with irregular, extreme peaks. These spikes correspond to weeks of intense fire activity, which may reflect the impact of severe weather conditions such as prolonged heatwaves or droughts, particularly common in western Canada and the western United States during summer. Despite the noise, there is a clear seasonal recurrence where elevated values tend to cluster mid-year, hinting at a cyclical component. In certain years, higher amplitudes in burnt area suggest possible structural breaks, likely linked to extraordinary climate events (e.g., the 2021 Western North America heat dome). No strong upward or downward trend is visible over the entire sample, but localized increasing patterns may signal worsening fire seasons, potentially attributable to climate change and increasing fuel loads due to forest mismanagement.

2.2 Seasonal Plot



The seasonal plot illustrates that wildfire activity consistently peaks during the summer months, especially between June and August, across most years. This is consistent with the climatic patterns of North America, where summer brings hot, dry, and windy conditions, ideal for wildfire ignition and spread. In contrast, winter months exhibit near-zero burnt areas, which can be attributed to snow cover and higher humidity. Interestingly, the amplitude of fire seasons varies significantly between years, suggesting inter-annual variability potentially tied to El Niño/La Niña cycles, changes in vegetation dryness, or land-use practices. The sharp differences between years could also reflect improvements in early detection and suppression in some years, or conversely, resource constraints or policy shifts in fire management.

2.3 Subseries Plot

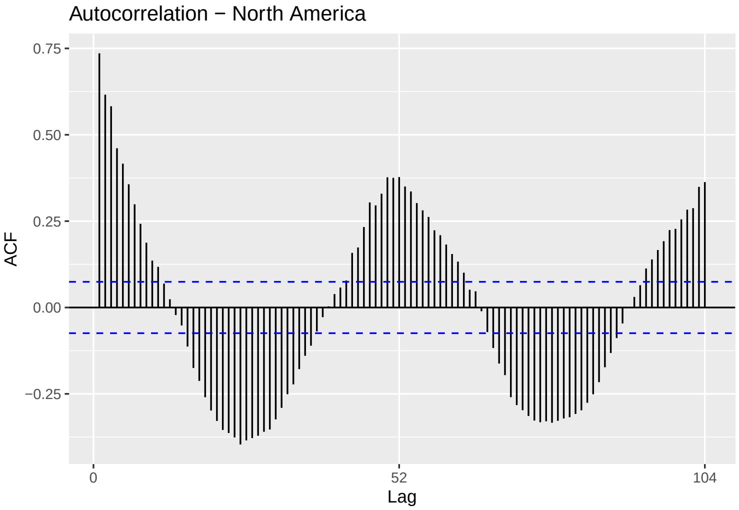


The custom subseries plot above uses weekly boxplots to visualize the distribution of area burnt by wildfires across all years, for each week of the year. Interestingly, the seasonal increase in wildfire activity is not strictly monotonic. After an initial rise in weeks 17-20, there is a noticeable dip in median and mean values between weeks 21 and 25, before activity surges again to reach a second, more intense peak between weeks 26 and 35. This pattern could reflect the averaging of asynchronous fire seasons across North America, for example, some regions may peak earlier or later due to differing climates. Alternatively, it may suggest a transitional weather phase (e.g., late spring storms) or increased suppression effectiveness during early summer. The second peak aligns more consistently with the dry mid-to-late summer, confirming it as the core fire season in the region. Again, we see a very small but surprising break in the decreasing monotony and fire activity around week 43. The double-hump pattern reinforces the importance of region-specific modeling, as aggregate trends may mask critical differences across countries.

The presence of extreme values — visible as outliers in the boxplots — is especially pronounced in peak weeks, indicating that in some years the burnt area was significantly higher than the average. The range of the interquartile spread (box height) also increases during summer, suggesting higher week-to-week volatility and unpredictability during that time. In contrast, the autumn and winter weeks (from around week 37 onward) display narrow distributions with low means and few outliers, reinforcing the idea of a highly concentrated fire season.

The blue dots representing weekly means show a smooth seasonal curve, further confirming the intra-annual regularity of wildfire dynamics. This detailed breakdown supports the inclusion of seasonal components in any modeling approach and highlights the need for special attention to peak summer periods, both for forecasting accuracy and resource planning in fire management.

2.4 Autocorrelation Plot



The ACF plot of the weekly burnt area time series shows significant positive autocorrelations in the early lags, which gradually decay over time. This indicates that fire activity in one week is statistically related to the recent past–likely reflecting the fact that wildfires often persist across several weeks, especially during the active fire season. The short-term memory structure is typical in environmental time series and suggests temporal clustering of extreme events.

Most notably, the plot reveals repeating bumps around lag 52 and beyond, indicating annual seasonality. However, the structure is not perfectly cyclical–the ACF does not exhibit crisp, consistent seasonal peaks. Instead, it displays a wavy, gradually repeating shape, likely due to year-to-year variation in the timing and intensity of fire seasons. This aligns with prior visualizations (e.g., the subseries plot), where we observed multi-modal summer peaks and irregular transitions in wildfire behavior. Together, these findings reinforce the importance of using models that capture both short-term autocorrelation and seasonal variation, such as SARIMA or Exponential Smoothing with seasonal terms.

3 Time Series Decomposition

In preparation of applying different forecasting methods on wildfire-related data, the following chapter deals with the topic of time series decomposition. The aim of this procedure is to gain a broader understanding of underlying patterns explaining the development in numbers of burnt areas and surface temperature. This includes necessary adjustments applied to the data as well as useful transformations and decomposition into trend-cycle and seasonal components using various methods in order to enhance interpretation of the results.

3.1 Data Preparation and Adjustments

The first step involves transforming the data into the correct format for time series analysis. This includes converting the datasets into tsibbles and adjusting the calendar by modifying the index.

For the wildfire-related datasets, this means aggregating the weekly data into monthly totals. The surface temperature data is already recorded at a monthly frequency, but the date column still needs to be reformatted to ensure proper monthly alignment in R.

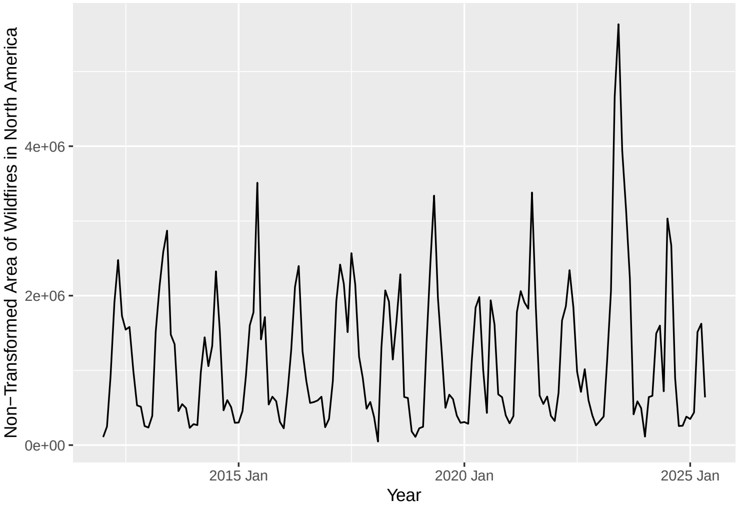
No additional adjustments for inflation, population, or other normalization are required, as the data is already relative or standardized. This analysis will focus on North American countries, including Canada, the United States, Mexico, Greenland, and others.

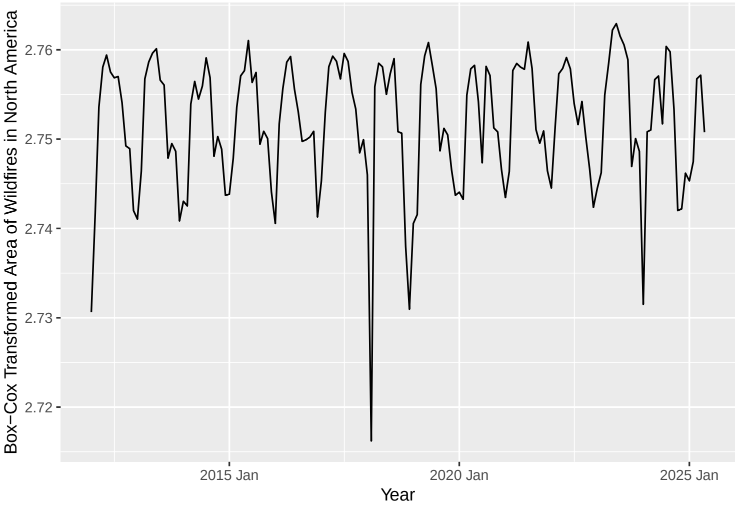
3.2 Transformations

Transformations help stabilize variance, improve normality, and sometimes linearize relationships in the data. This improves the interpretability of results in later steps such as decomposition and forecasting.

Table 1: Lambda Values for North America Countries

Country	lambda_guerrero
Bermuda	NA
Canada	-0.2483560
Greenland	0.0191811
Mexico	-0.0766638
Saint Pierre and Miquelon	NA
United States	-0.6009170





When examining North American countries individually for the dataset of area burned, several values of $\lambda \neq 0$ are observed, indicating the need for localized transformations of the wildfire-related time series. However, there is no consistent pattern across countries. For instance, Greenland shows $\lambda = -0.0022$, suggesting a log-like transformation, while the United States has $\lambda = -0.7549$, implying a different transformation to be suitable. This variability limits the interpretability of results when analyzing countries separately.

To ensure consistency, the analysis focuses on the aggregated wildfire data for North America. A Box-Cox transformation applied to the continent-wide time series yields $\lambda = -0.2514$, indicating non-constant variance and supporting a strong inverse-like transformation. After transformation, the series exhibits a more stable variance and a clearer overall structure, although the improvement is modest. In contrast, the untransformed data show increasing variance over time, likely due to more extreme weather events and prolonged droughts driven by climate change.

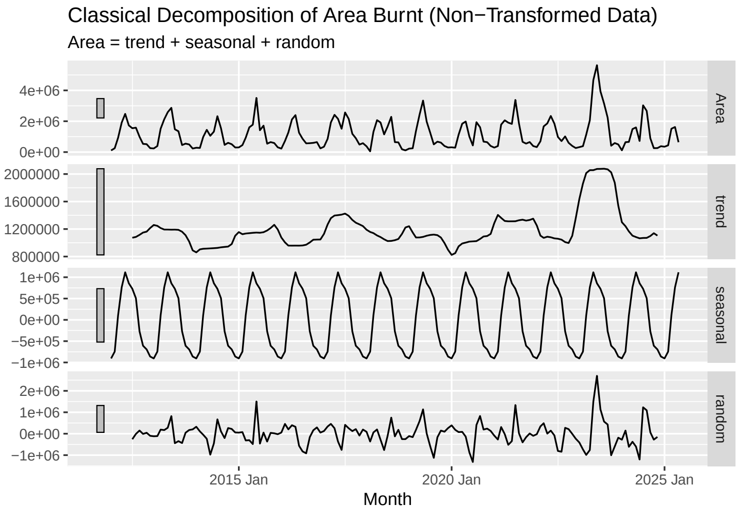
Table 2: Lambda Estimate for Surface Temperature

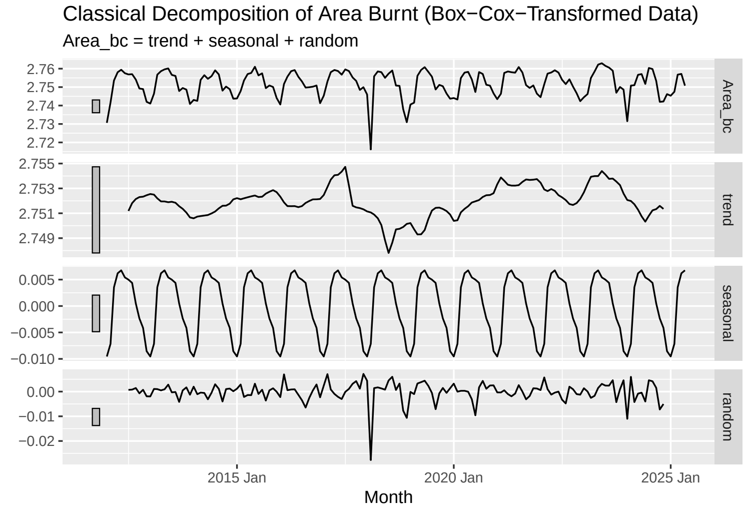
Metric	Value
Lambda (Surface Temperature, North America)	1.084134

For the surface temperature dataset, a Box-Cox analysis yields $\lambda = 1.0664$. Since this value is close to 1, it suggests that the variance is sufficiently stable, and no transformation is required for further analysis.

3.3 Classical Decomposition

In time-series decomposition, data is broken down into its underlying components to better understand trend-cycle, seasonal, and irregular effects. To ensure robust and representative results, multiple decomposition methods are applied and analyzed in the following section. These include classical decomposition, STL decomposition, and official statistical methods such as X-11 and X-13-ARIMA.

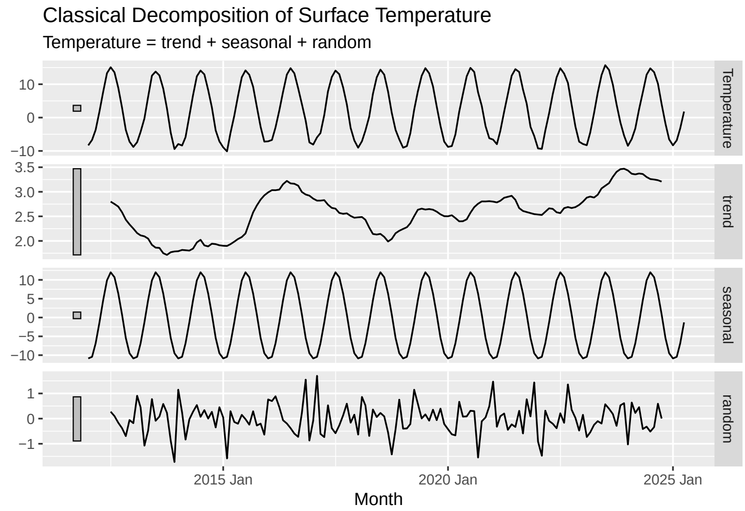




The classical decomposition of the time series of area burned reveals a strong seasonal pattern in both the transformed and non-transformed data, confirming the high seasonality of wildfires. Peaks consistently occur during the summer months (May to August), reflecting the increased likelihood of wildfires during hot and dry conditions. In contrast, troughs appear in the winter months (November to February), likely due to cooler temperatures and higher humidity reducing fire risk.

In the trend component of the non-transformed data, the area burned remains relatively stable with moderate fluctuations between 2014 and 2020. A clear upward shift begins in 2021, culminating in a sharp peak in 2023. For the Box-Cox-transformed series, the trend peaks in both 2017 and 2023, with a steady rise beginning in 2018. These trends may signal worsening wildfire conditions in recent years, potentially driven by climate change, increased forest fuel loads, changes in land management policies, or diminished fire suppression capacity.

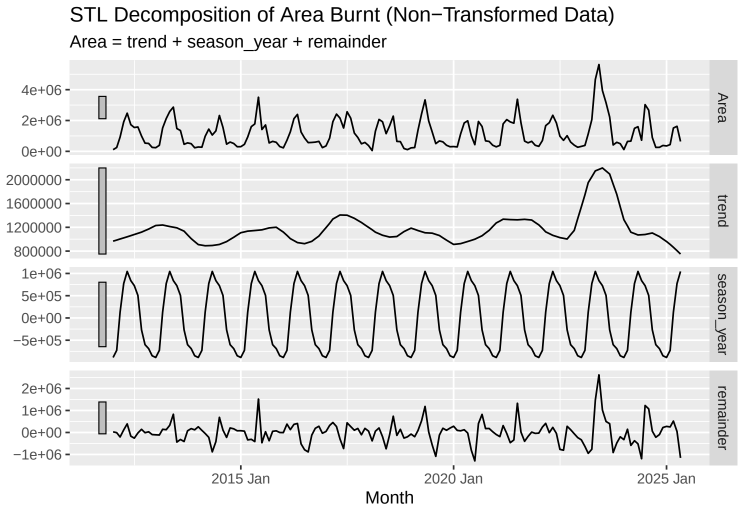
The remainder component also shows significant spikes in 2023 in the non-transformed data, possibly reflecting extreme and unpredictable weather events such as heatwaves or prolonged droughts.

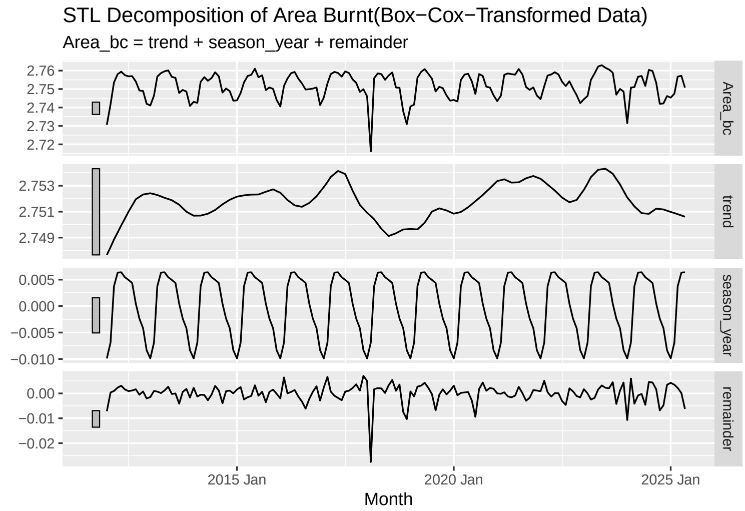


The classical decomposition of average surface temperature in North American countries reveals a generally rising trend beginning around 2018. This supports the hypothesis that increasing temperatures in recent years may be contributing to a rise in wildfire activity.

A well-defined and consistent seasonal component is also evident, reflecting the stable and recurring seasonal patterns typical of the region. The residual component remains relatively stable overall, though some outliers are present, potentially corresponding to short-term anomalies or extreme weather events.

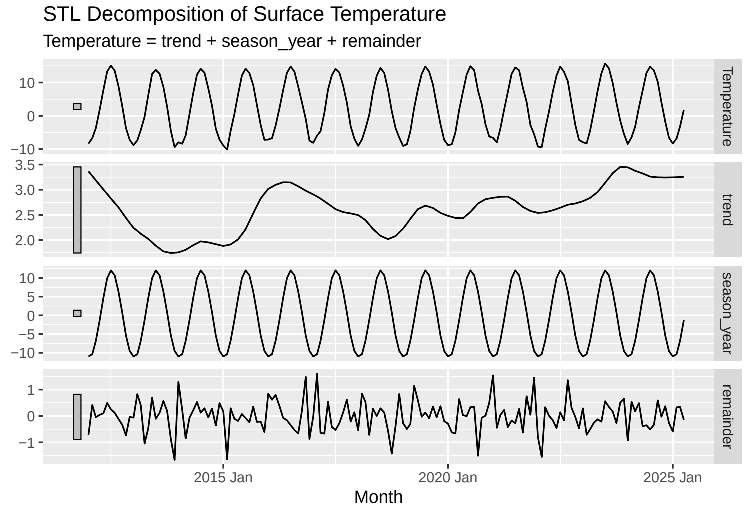
3.4 STL Decomposition





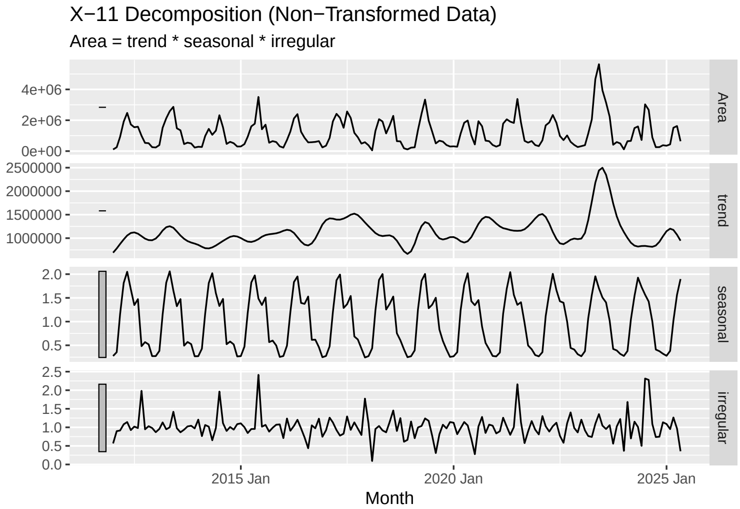
STL decomposition, known for its flexibility and robustness to changing seasonality, was applied to the dataset of burned area. Compared to classical decomposition, STL produces a smoother trend component in both the non-transformed and Box-Cox-transformed series. Despite the smoother appearance, the overall structure, particularly the presence of seasonal peaks and an upward trend in the non-transformed data, remains largely consistent with the classical method.

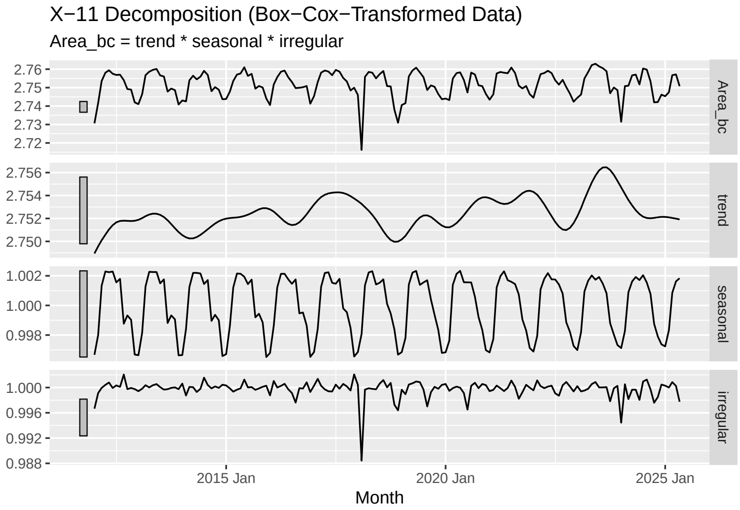
The seasonal component shows a nearly identical pattern across both decomposition approaches, with clear peaks during the summer months and troughs during the winter, highlighting the strong seasonality of wildfire activity. Similarly, the remainder component closely matches that of the classical decomposition, including a pronounced outlier in 2018 visible in the Box-Cox-transformed series.



For the surface temperature dataset, STL decomposition reveals patterns that closely mirror those found using classical decomposition. The trend component displays smoother variations, offering a clearer view of long-term temperature changes. Additionally, due to the use of a periodic seasonal window, no variation in seasonal patterns over time is detected, reinforcing the consistency of seasonal temperature fluctuations across the observed period.

3.5 X-11 Decomposition

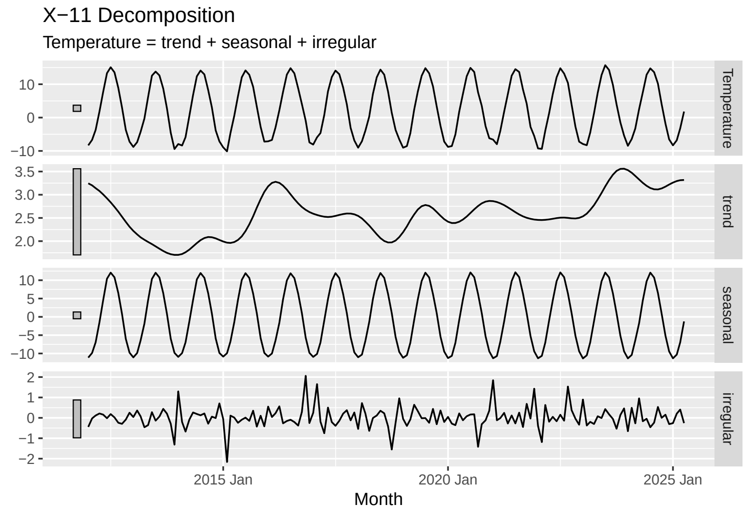




Applying an X-11 decomposition to the area burned time series yields a trend component that closely resembles the results from both the non-transformed and transformed versions, as previously observed with STL decomposition. However, the seasonal components behave differently under this method.

In both wildfire time series, including area burned, there appears to be a gradual smoothing of seasonal patterns over time. Unlike the classical and STL decompositions, the X-11 method reveals a discontinuous seasonal pattern early in the observation period. Specifically, it suggests a secondary wildfire peak in August, which is not evident in other methods.

The non-monotonic increase in the seasonal component showing two distinct peaks during the warmer season confirms the observations of Chapter 2.3. The occurence of asynchronous fire seasons in North America within a year could be justified by the fact that climate and wildfire favoring weather conditions differ across the continent.



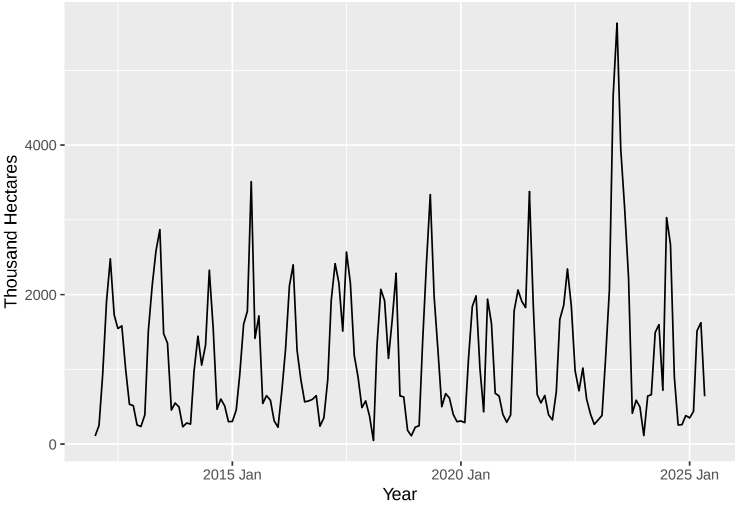
When applying X-11 decomposition to the surface temperature time series, the trend component exhibits a development similar to that observed with STL decomposition, though with slightly more fluctuations. The seasonal component remains consistent with previous models, showing no significant deviations.

Notably, the remainder component appears more stable, with fewer outliers compared to other decomposition methods. This increased stability may reflect the robustness of the X-11 approach when applied to the surface temperature data.

4 The Forecaster's Toolbox

This chapter introduces a practical approach to selecting and evaluating simple forecasting methods using both intuition and formal accuracy measures. The analysis begins with graphical exploration of each time series to guide the choice of an appropriate forecasting method, as outlined in the textbook. To assess model performance, the data is split into training and test sets, and both traditional and cross-validation techniques are used to compare forecasting accuracy. The method with the best predictive performance is identified and examined further. Finally, residual diagnostics are conducted to determine whether the forecast errors resemble white noise, ensuring the model captures the underlying patterns in the data effectively.

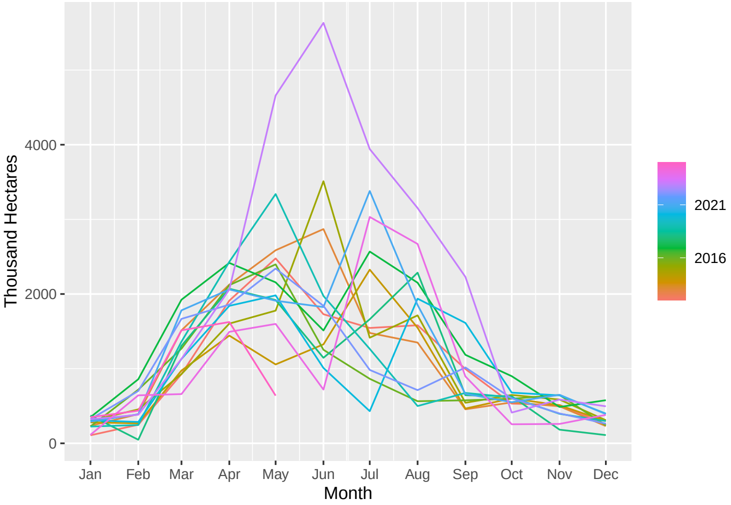
4.1 Comparison of Forecasts and Actuals



The time series chart of monthly area burnt in North America (in thousand hectares) displays clear seasonal patterns, with regular spikes occurring approximately once a year, typically during the warmer months. These seasonal peaks indicate a strong annual cyclicality in wildfire activity. Additionally, while the amplitude of peaks varies significantly across years, particularly a noticeable surge around mid-2023, there does not appear to be a consistent long-term trend.

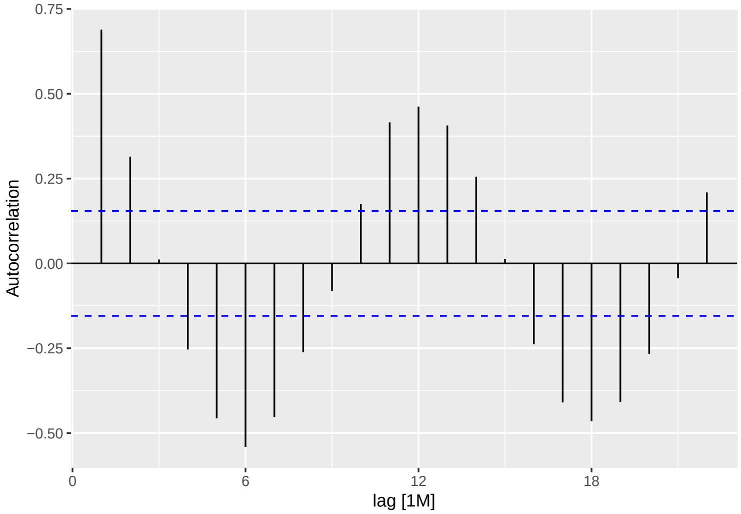
Given these characteristics, seasonality is the dominant component of this series. Among the simple forecasting methods discussed in the textbook, the seasonal naive method is the most appropriate choice for this data. This method forecasts each period using the observed value from the same period in the previous year. Because of the recurring seasonal spikes and relative absence of trend, this method would capture the yearly wildfire cycle more effectively than other simple models without overcomplicating the model.

4.2 Seasonal Pattern: Mid-Year Peak in Area Burned



As already seen in Chapter 2.2, a strong seasonal structure showing consistent mid-year peaks in burnt area can be identified across most years. While there is some variability in the magnitude of the peaks, the seasonal timing of fire activity remains relatively stable. Such a clear and repeated seasonal pattern supports the previous recommendation of using the seasonal naive method for forecasting.

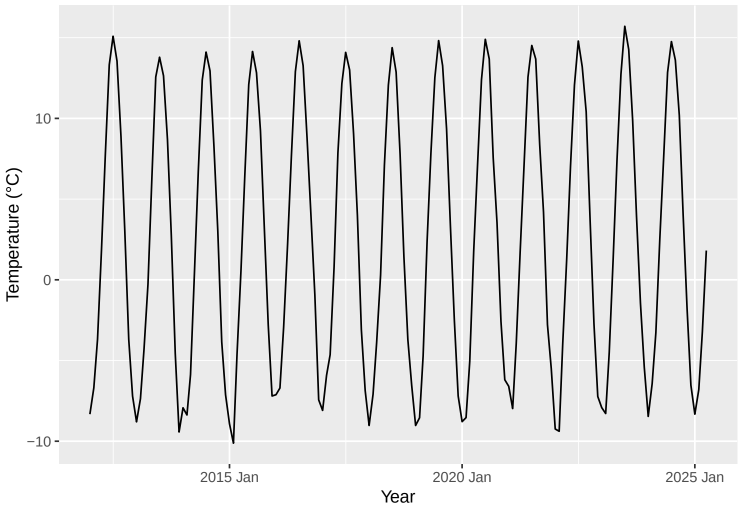
4.3 Autocorrelation Function of Area Burned



The autocorrelation function (ACF) plot of the area burnt series shows a significant spike at lag 12, along with smaller peaks at multiples of 12. This confirms a strong annual seasonality, as the autocorrelation is high for data points one year apart. There are also smaller positive autocorrelations at lags close to 1 and 2, but they decline quickly, suggesting that short-term dependencies are weaker.

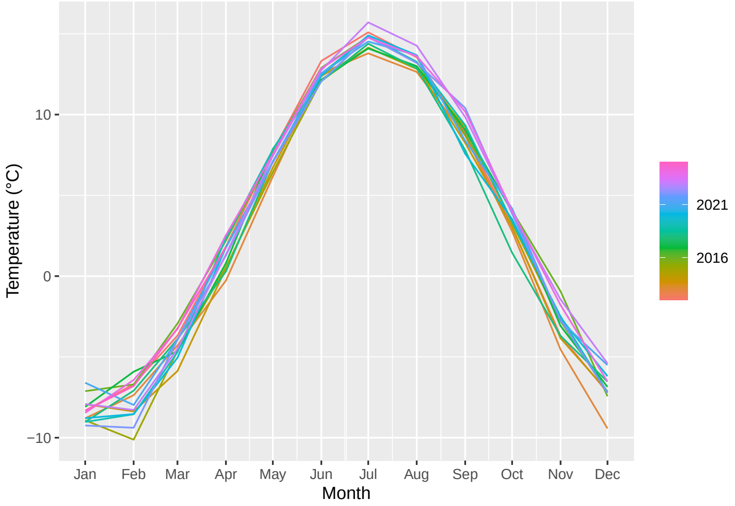
The cyclical pattern in the ACF, alternating between negative and positive correlations, also indicates a regular seasonal structure, likely driven by the annual fire cycle in North America. These observations strongly validate the earlier graphical conclusions and reinforce the choice of the seasonal naive method as an effective forecasting approach.

4.4 Monthly Surface Temperature Trends in North America



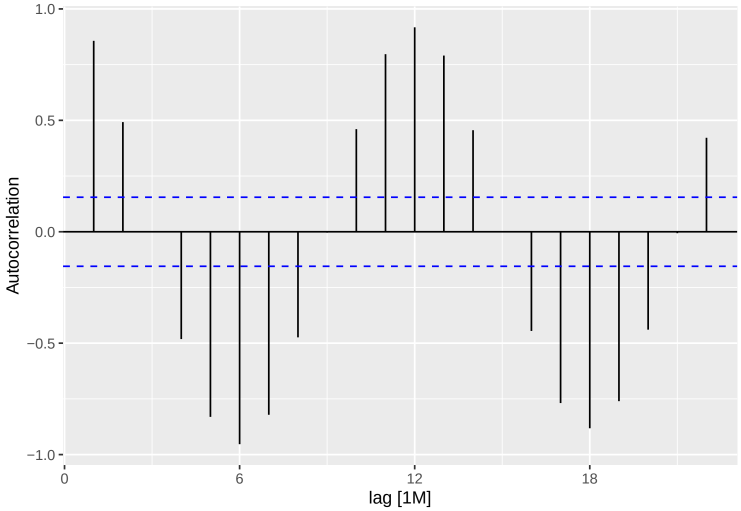
The time series of monthly average surface temperature exhibits clear seasonality with a strong and consistent annual pattern. The temperature rises and falls in a regular cycle each year, peaking around mid-year and dipping in the early months of each year. While there are minor variations from year to year, there is no obvious long-term upward or downward trend over the time span displayed, so Seasonal Naive Method is recommended for forecasting the monthly average surface temperature series due to its regular and strong seasonal cycle.

4.5 Seasonal Pattern in Surface Temperature



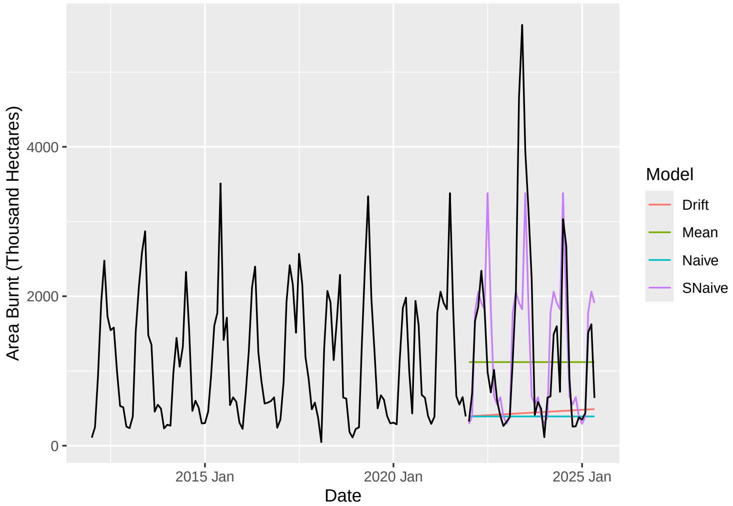
This plot illustrates the monthly average surface temperature for multiple years (2016-2021), allowing for a direct comparison of seasonal behavior across years. The shape of each year's line follows a consistent pattern, confirming the presence of a strong, regular seasonal cycle. Temperatures reliably increase from January to July, peak during the summer months (June-August), and then decrease back toward December. Although there are slight variations between years, such as marginally warmer summers or colder winters, the overall seasonal structure remains highly stable. This reinforces the earlier conclusions.

4.6 Autocorrelation Function of Surface Temperature



The ACF plot of the monthly surface temperature series further confirms the strong seasonal behavior observed in earlier analyses.

4.7 Preparing Training and Test Sets



This plot presents a comparison between the actual area burnt (in thousands of hectares) and forecasts generated by four simple models–Mean, Naive, Drift, and Seasonal Naive (SNaive)–on the test set.

The Mean and Drift models perform poorly, producing flat or gradually changing forecasts that completely miss the recurring seasonal spikes.

The Naive model, which uses the most recent value as the forecast, fails to anticipate the seasonal upswings and tends to underestimate peak values.

The Seasonal Naive (SNaive) model, which reuses values from the same time in the previous year, captures the general timing of the seasonal spikes more effectively than the others. However, it still struggles with the varying magnitude of those spikes, particularly during extreme fire seasons.

To conclude, the models highlight the importance of capturing seasonal dynamics in the data despite their simplicity. The SNaive model outperforms the others, thanks to its incorporation of annual seasonality. Nonetheless, its inability to adapt to changes in peak intensity suggests that more advanced forecasting techniques would be necessary for more accurate and responsive forecasts.

Table 3: Forecast Accuracy on Test Set

.model	.type	ME	RMSE	MAE	MPE	MAPE	MASE	RMSSE	ACF1
Drift	Test	899.87964	1547.2533	971.2236	26.88839	59.71039	NaN	NaN	0.7371062
Mean	Test	222.74830	1275.3414	920.1762	-83.84049	122.21148	NaN	NaN	0.7356760
Naive	Test	949.99359	1574.6005	992.5964	35.65792	57.91073	NaN	NaN	0.7356760
SNaive	Test	32.40761	994.1431	588.2636	-28.62712	54.14953	NaN	NaN	0.4724649

The table above compares the accuracy of four simple forecasting models–Drift, Mean, Naive, and Seasonal Naive (SNaive)–based on several performance metrics evaluated on the test set.

- SNaive (Seasonal Naive) performs the best overall, with the lowest RMSE (1014) and lowest MAE (599), indicating that it has the smallest average and squared forecast errors. It also has the lowest

autocorrelation in residuals (ACF1 = 0.459), suggesting fewer systematic patterns left in the residuals–a sign of a better-fitting model.
- Naive and Drift models perform similarly, but worse than SNaive. Naive has slightly higher RMSE (1582) and MAE (983), with a high residual autocorrelation (ACF1 = 0.735), indicating persistent errors and missed seasonality.
- Mean model performs particularly poorly. While its RMSE (1290) and MAE (933) are slightly better than Drift and Naive, its very high MAPE (125%) and extreme negative MPE (-87.7%) indicate systematic underprediction. This makes sense since the Mean model doesn't account for seasonality or recent trends, leading to biased forecasts, especially during high-activity months.
- Drift model also struggles, with a high RMSE (1556) and MAE (964). Its high positive MPE (25.4%) shows a tendency to overestimate, and the ACF1 (0.737) indicates strong autocorrelation in residuals.

4.8 Cross-Validation

Table 4: Cross-Validation Accuracy Summary (1-Step Forecasts)

.model	MAE	RMSE
Naive	582.0700	797.2101
Drift	585.6283	800.8733
SNaive	643.3569	1074.1986
Mean	812.2045	1040.8367

Key insights:

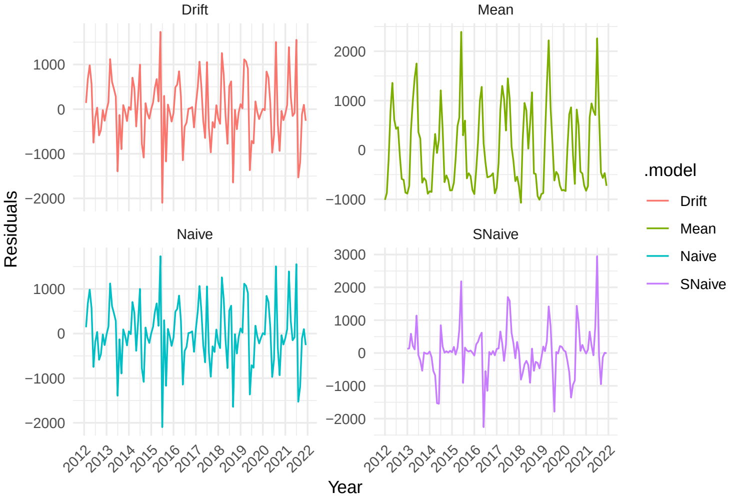
- Naive model performs best overall, with the lowest MAE (587) and lowest RMSE (801). This suggests that simply using the previous month's value is surprisingly effective for short-term wildfire area forecasting.

- Drift model is a close second, with slightly higher error metrics than the Naive model. It accounts for trends by projecting forward based on historical change, but that seems to offer minimal benefit here.

- SNaive model, which performed best in the test set, ranks lower in this cross-validation. This might be because seasonal structure is more valuable at longer horizons (e.g., 12 months ahead) than in a 1-step-ahead setting, where recent values dominate.

- Mean model performs the worst, with the highest MAE (817) and highest RMSE (1046). Averaging past values fails to capture either the seasonal or short-term dynamics of wildfire activity.

4.9 Comparison of Model Residuals



This figure shows the residuals (forecast errors) from the Drift, Mean, Naive, and Seasonal Naive models. Residuals are important for diagnosing model performance — ideally, they should resemble white noise: randomly scattered around zero, with no clear pattern.

The Drift model produces residuals that fluctuate randomly but show large variability over time. Some periods exhibit clusters of large positive or negative errors, which may indicate non-stationarity or structural changes in the data. This suggests that the Drift model struggles to capture both the level and variability in the series.

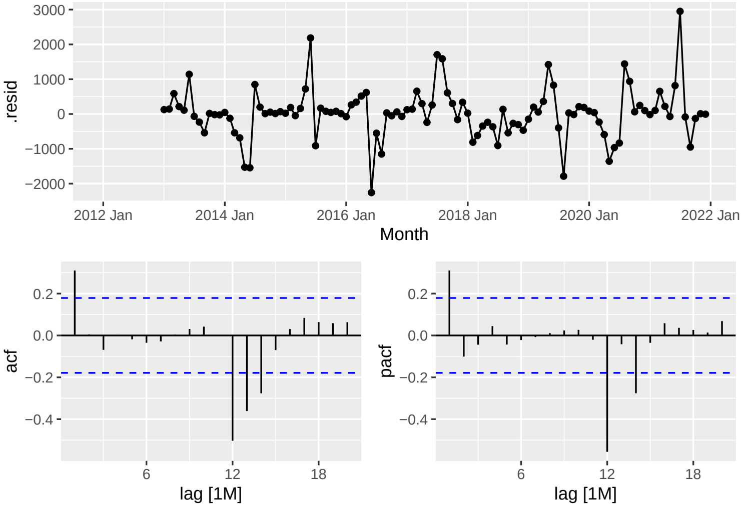
The Mean model displays clear periodic patterns in its residuals, particularly during high-burn seasons. This indicates that the model systematically underpredicts during high-activity periods and overpredicts during low-activity periods. It fails to capture seasonality and sudden spikes in wildfire activity.

The Naive model shows more randomness than the Mean model but still features noticeable swings, especially during periods of intense burning. The residuals remain relatively persistent, reflecting that last-period values are often poor predictors of sharp increases or decreases. While it performs better than the Mean model, the Naive model still lacks sensitivity to sudden changes.

The Seasonal Naive (SNaive) model has residuals that are more centered and less structured than those of the other models. Although some spikes remain—likely caused by unusually high wildfire activity exceeding typical seasonal levels—SNaive overall exhibits the most balanced residuals, with fewer extreme errors and no apparent long-term trends.

In conclusion, none of the models completely eliminate patterns in the residuals, indicating room for improvement. However, the SNaive model outperforms the others by producing residuals that are closer to white noise.

4.10 Residual Analysis: Seasonal Naive Model



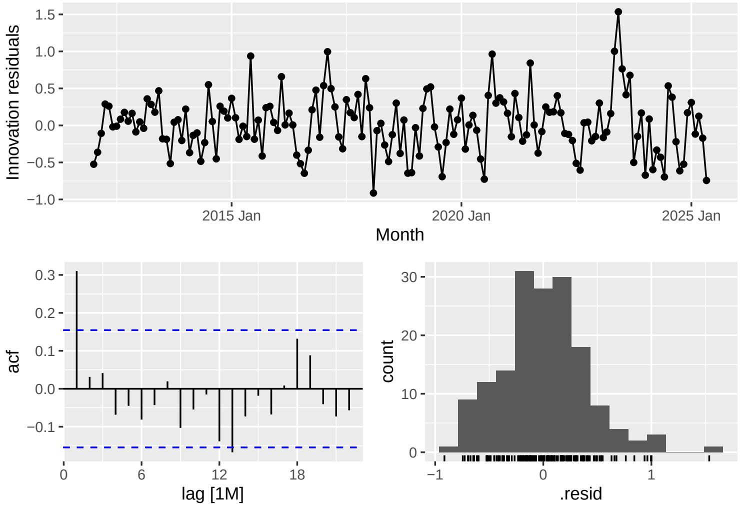
The residuals mostly hover around zero, with some moderate to large spikes (both positive and negative). Notably, the largest residuals seem to coincide with outlier events, likely unusually severe or mild wildfire seasons that deviate from normal seasonal patterns. There is no strong trend in the residuals over time. The SNaive model is mostly unbiased, but still misses some of the larger seasonal anomalies.

In the autocorrelation function (ACF) there is a significant spike at lag 12, which matches the seasonal period (12 months). The presence of autocorrelation at lag 12 suggests that seasonal structure still remains in the residuals. A few other lags are near the threshold but mostly within bounds, indicating that short-term autocorrelation is not a big issue. The SNaive model doesn't fully eliminate the seasonal structure. This might be due to seasonal variation over years or nonstationary seasonality (e.g., increasing wildfire intensity).

In the partial autocorrelation function (PACF), there is a strong negative partial autocorrelation at lag 12, consistent with what we saw in the ACF. PACF drops off quickly after lag 1–2, which suggests that residuals don't exhibit strong dependencies beyond seasonality.

5 Exponential Smoothing

5.1 Model Training, Visualization, and Residual Analysis



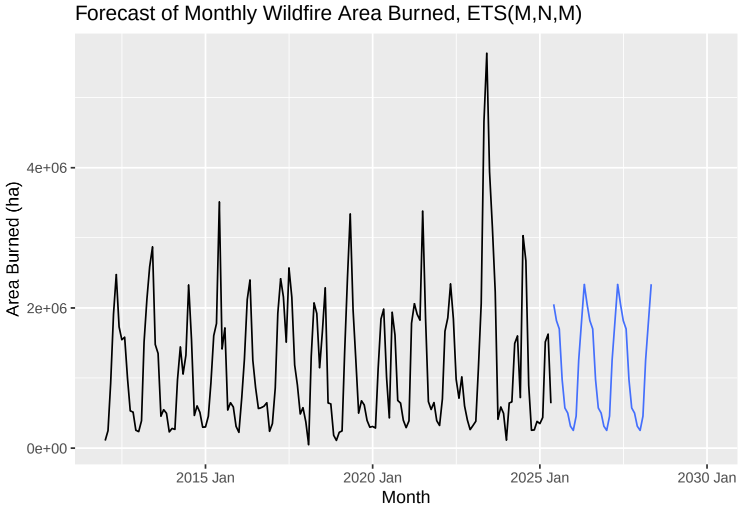


Table 5: Comparison of ETS vs. Simple Models

.model	.type	ME	RMSE	MAE	MPE	MAPE	MASE	RMSSE	ACF1
ETS	Training	2363.825	618135.4	384542.2	-24.587	46.286	0.678	0.641	0.442
Naive	Training	3320.550	743786.2	537054.7	-25.554	60.671	0.948	0.771	0.111
Drift	Training	0.000	743778.8	537054.7	-26.131	60.838	0.948	0.771	0.111
SNaive	Training	-296.564	964835.7	566804.9	-32.281	65.807	1.000	1.000	0.522

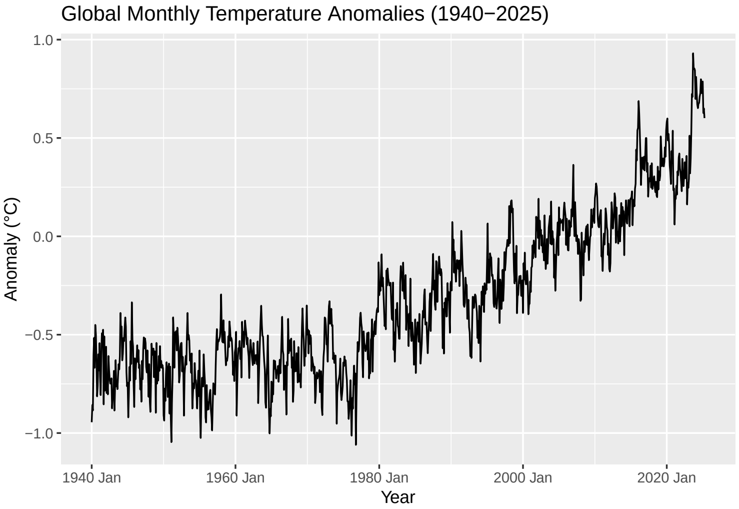
We applied Exponential Smoothing (ETS) to model the monthly area burned by wildfires in North America from 2012 to 2025. Based on the data's strong seasonal structure and variable scale, the automatically selected model, ETS(M, N, M) with multiplicative error and multiplicative seasonality, was found to be the most appropriate.

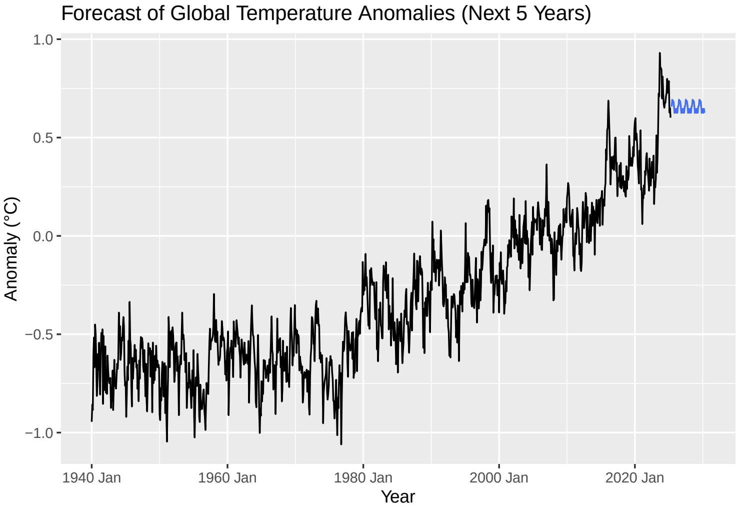
The model estimated a smoothing level $\alpha = 0.0950$ and a seasonal smoothing $\gamma \approx 0.0001$, indicating high inertia and stable seasonal effects. Initial states included a level of 1,055,610 hectares and strong seasonal multipliers ranging from 0.195 to 2.075. The model achieved an AIC of 4926.64 and residual variance $\sigma^2 = 0.1644$, confirming a good fit.

We generated a 36-month forecast, capturing sharp seasonal peaks consistent with historical trends. Residual diagnostics confirmed adequacy, with minimal autocorrelation and white-noise behavior. Compared to simple benchmarks (Naive, Drift, Seasonal Naive), ETS achieved the best performance, with RMSE = 610293.9 and MAE = 396879.6, highlighting its effectiveness for forecasting seasonal wildfire behavior.

These results suggest that ETS is a suitable and reliable method for forecasting seasonal wildfire behavior in this context.

5.2 Analysis of Global Temperature Anomalies





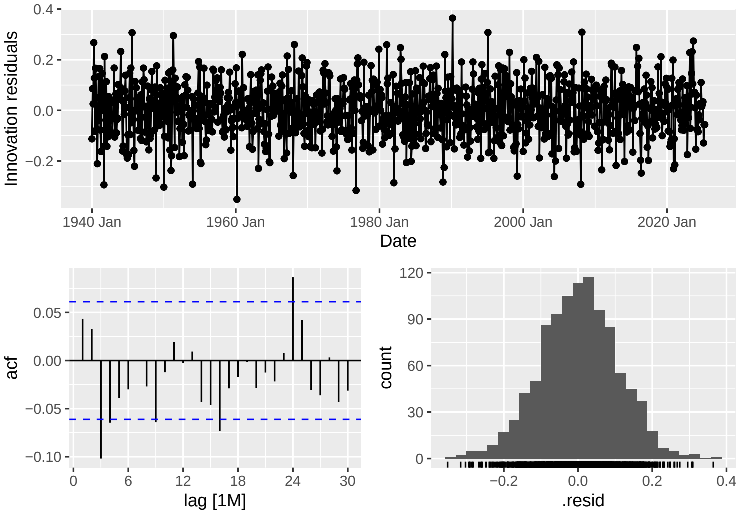


Table 6: Comparison of ETS vs. Simple Models

.model	.type	ME	RMSE	MAE	MPE	MAPE	MASE	RMSSE	ACF1
Drift	Test	0.135	0.244	0.159	13.335	37.695	NaN	NaN	0.790
ETS	Test	0.031	0.217	0.156	-22.554	48.426	NaN	NaN	0.795
Naive	Test	0.157	0.265	0.177	19.346	41.264	NaN	NaN	0.800
SNaive	Test	-0.040	0.278	0.222	-59.996	86.203	NaN	NaN	0.635

We applied Exponential Smoothing (ETS) to model global monthly temperature anomalies from 1940 to 2025. The automatically selected model was ETS(A, N, A), featuring additive errors and additive seasonality but no trend component. The smoothing parameters were estimated as $\alpha = 0.5331$ for the level and $\gamma = 0.0001$ for the seasonal component, indicating moderate sensitivity to recent changes and a stable annual seasonal pattern. The model's initial level was -0.886, and seasonal multipliers ranged between -0.04 and $+0.05$, reflecting subtle but consistent month-to-month variation.

Using this model, we produced a 60-month forecast that projects a continuation of the current warming trajectory, embedded within slight seasonal fluctuations. The model achieved a residual variance of $\sigma^2 = 0.0111$, with an AIC of 2498.46, confirming a strong statistical fit.

We compared the ETS model against simple baselines including Naive, Seasonal Naive, and Drift models. ETS yielded the lowest RMSE (0.217) and MAE (0.156) on a 36-month test set, slightly outperforming Drift and significantly surpassing SNaive. Residual diagnostics confirmed the adequacy of the model, though mild autocorrelation remained present (ACF1 $\approx$ 0.795), suggesting room for structural model improvement.

These results suggest that ETS(A,N,A) is a solid baseline for forecasting temperature anomalies, especially when focusing on short- to medium-term climate signals. For long-term structural modeling, more flexible approaches like ARIMA may provide complementary insights.

6 ARIMA

ARIMA, which stands for AutoRegressive Integrated Moving Average, is one of the most widely used and versatile approaches for time series forecasting. ARIMA models are particularly well-suited for univariate time series that exhibit trends, cycles, or autocorrelated patterns. The structure of an ARIMA model is defined by three key components: the autoregressive (AR) term, which incorporates dependence on past values; the integrated (I) term, which represents the number of times the series needs to be differenced to achieve stationarity; and the moving average (MA) term, which accounts for past forecast errors.

To build a suitable ARIMA model, we follow the Box-Jenkins methodology, a systematic approach for identifying, estimating, and diagnosing time series models. The process begins with checking stationarity and applying differencing if necessary. Next, we analyze the sample autocorrelation (SACF) and sample partial autocorrelation (SPACF) plots to determine appropriate values for the AR and MA parameters. Once candidate models are estimated, we perform diagnostic checks on residuals to ensure they resemble white noise, indicating a good fit. Finally, we use information criteria such as AIC or BIC to compare and select the best-performing model.

This chapter applies the Box-Jenkins methodology to two climate-related time series for North America. The first dataset, Area Burnt, measures the total area burned by wildfires each month, aggregated from weekly records across Canada, the United States, Mexico, Greenland, Bermuda, and Saint Pierre and Miquelon. The second dataset, Surface Temperature, captures the average monthly surface temperature across the same countries, using daily observations averaged by month. By modeling these two series, we aim to evaluate their respective trends, seasonal patterns, and forecasting performance using ARIMA models.

6.1 Box-Cox Transformations

Table 7: Estimated Box-Cox Lambda Values

Dataset	Lambda
Area Burnt	-0.227
Surface Temperature	1.066

After calculating lambda, a Box-Cox transformation will be applied to the Area Burnt dataset. The estimated lambda value ($\lambda \approx -0.23$) is far from one, indicating non-constant variance. To stabilize the variance and meet the assumptions needed for reliable modeling, a log-like transformation will be used.

In contrast, the Surface Temperature dataset does not require transformation. Its estimated lambda value ($\lambda \approx 1.07$) is close to one, suggesting that the variance is already stable and the data is on an appropriate scale for analysis.

6.2 Testing for Stationarity

The Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test is used to assess stationarity in a time series. Its null hypothesis states that the data is stationary. If the p-value is less than 0.05, the null hypothesis is rejected, indicating that the data is non-stationary.

Table 8: KPSS Test Results

Dataset	kpss_stat	kpss_pvalue
Area Burnt	0.0290136	0.1
Surface Temperature	0.0115452	0.1

The KPSS test results for the Area Burnt dataset do not provide strong evidence against stationarity. With a low test statistic (≈ -0.027) and a high p-value ($p = 0.1$), the null hypothesis of stationarity is not rejected. This indicates no clear signs of trend non-stationarity. However, failing to reject the null does not confirm the series is stationary, it only suggests that the data does not obviously violate stationarity assumptions. As a result, differencing may not be necessary at this stage, though further analysis will follow.

A similar outcome is observed for the Surface Temperature dataset. The test statistic is also low (≈ -0.012), and the p-value remains high ($p = 0.1$), leading again to a failure to reject the null hypothesis. This points to no strong evidence of a deterministic trend in the data. Nonetheless, to ensure a more reliable assessment, additional stationarity tests will be conducted to support or refine these findings.

The Phillips-Perron test evaluates whether a time series is non-stationary. Its null hypothesis assumes non-stationarity. If the p-value is less than 0.05, the null is rejected, indicating that the data is stationary.

Table 9: Phillips-Perron Test Results

Dataset	pp_stat	pp_pvalue
Area Burnt	-6.850843	0.01
Surface Temperature	-5.445114	0.01

The Phillips-Perron test results are consistent with those from the KPSS test. For both the Area Burnt and Surface Temperature datasets, the Phillips-Perron test yields low p-values ($p = 0.01$), leading to the rejection of the null hypothesis of non-stationarity. This indicates that both series can be considered stationary.

Together, the KPSS and Phillips-Perron tests provide converging evidence. The KPSS test fails to reject the null hypothesis of stationarity, while the Phillips-Perron test rejects the null of non-stationarity. This agreement reinforces the conclusion that both datasets are stationary and suitable for modeling without additional non-seasonal differencing.

6.3 Differencing

The ndiffs() function estimates the number of non-seasonal differences required to achieve stationarity in a time series. It synthesizes results from multiple unit root tests, such as KPSS and Phillips-Perron, to identify the minimum degree of differencing needed. A result of zero indicates that the series is already stationary and does not require further differencing.

Table 10: Estimated Number of Non-Seasonal Differences

Dataset	ndiffs
Area Burnt	0
Surface Temperature	0

Both datasets returned an ndiffs() value of 0, indicating that no non-seasonal differencing is required to achieve stationarity. For the Area Burnt dataset, this suggests that the series does not exhibit a strong long-term trend. The statistical properties, such as the mean and variance, appear stable over time, supporting the assumption of stationarity.

The Surface Temperature dataset shows a similar result. With no differencing needed, the series can be considered stationary, implying that its variability is not driven by a persistent trend. This supports the use of the data in its current form for time series modeling.

6.4 Seasonal Differencing

The nsdiffs() function estimates the number of seasonal differences needed to make a time series stationary. It identifies whether repeating seasonal patterns in the data cause non-stationarity. A result greater than 0 suggests that seasonal differencing is required to remove these patterns, while a result of 0 indicates that the data is already seasonally stationary.

Table 11: Estimated Number of Seasonal Differences

Dataset	nsdiffs
Area Burnt	1
Surface Temperature	1

For the Area Burnt dataset, the nsdiffs() test returned 1, indicating seasonal non-stationarity. This suggests a repeating seasonal pattern that must be removed before modeling. Applying one seasonal difference should address this and help achieve stationarity.

The Surface Temperature dataset showed the same result, with nsdiffs() also equal to 1. This indicates the presence of non-stationary seasonal variation. One seasonal difference will be applied to stabilize the series for accurate time series modeling.

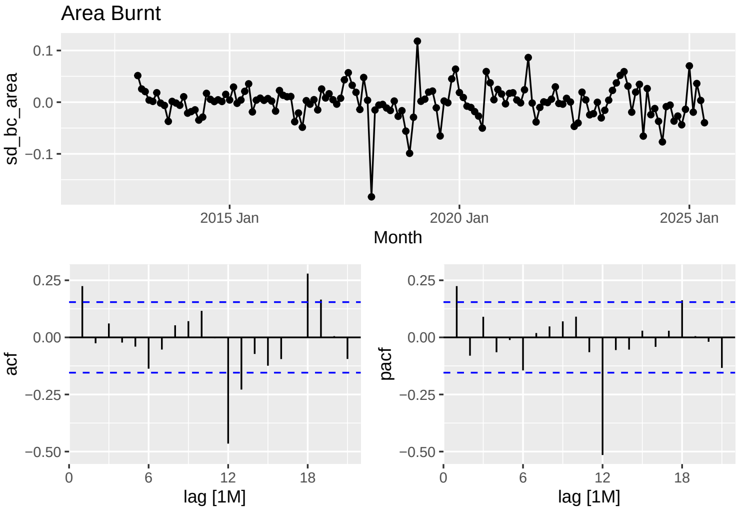
Table 12: Estimated Remaining Seasonal Differences After One Seasonal Difference

Dataset	nsdiffs
Area Burnt	0
Surface Temperature	0

After applying a seasonal difference with a lag of 12 to both datasets, the nsdiffs() function returns zero, indicating that the seasonal non-stationarity has been resolved. The seasonal patterns have been effectively removed, and both time series are now seasonally stationary. No additional seasonal differencing is required, and the datasets are ready for modeling.

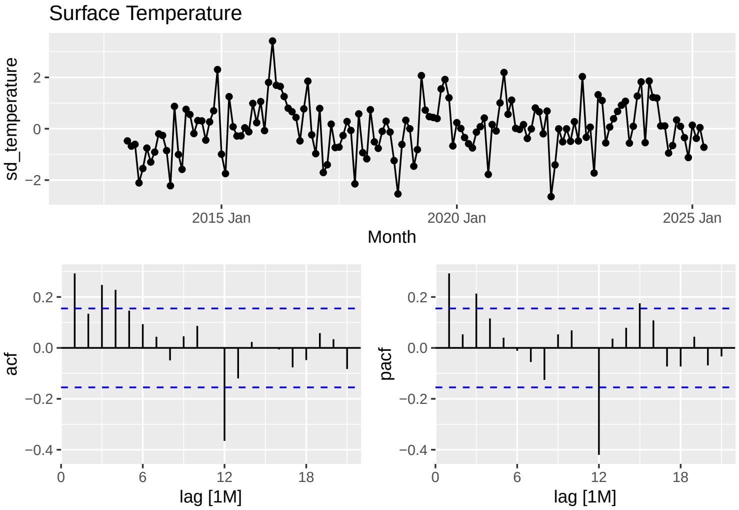
6.5 Autocorrelation and Partial Autocorrelation Analysis

The Sample Autocorrelation Function (SACF) measures the correlation between a time series and its past values at different lags, helping to identify patterns such as trend or seasonality. The Sample Partial Autocorrelation Function (SPACF) shows the correlation between a time series and its lags after removing the effects of shorter lags. Together, SACF and SPACF are used to identify appropriate ARIMA model components by highlighting potential autoregressive (AR) and moving average (MA) terms.



For the Area Burnt dataset, the ACF plot shows a strong spike at lag 1 followed by a rapid decline, suggesting a moving average component of order 1 (MA(1), q = 1). Similarly, the PACF has a noticeable spike at lag 1 before tapering off, indicating an autoregressive component of order 1 (AR(1), p = 1). Since the data has already been seasonally differenced and no additional non-seasonal differencing is needed, we set d = 0. Based on these observations, possible non-seasonal model configurations include (1, 0, 1), (1, 0, 0), or (0, 0, 1).

Seasonal patterns are also visible in the Area Burnt data. Strong spikes at lag 12 in both the ACF and PACF indicate a yearly seasonal effect. The PACF cuts off at lag 12, suggesting a seasonal autoregressive term of order 1 (P = 1), while the ACF shows prominent spikes at seasonal lags, pointing to potential seasonal moving average terms (Q = 1 or Q = 2). With seasonal differencing already applied, D is set to 1, leading to potential seasonal configurations such as (1, 1, 1) or (1, 1, 0).



In the Surface Temperature dataset, the ACF tapers off gradually with several small but meaningful lags, which is characteristic of an autoregressive process. This pattern suggests the presence of an AR(1) component (p = 1). While the moving average component is less pronounced, a small MA term (q = 0 or q = 1) may still be appropriate to test. As seasonal differencing has already been applied and no further non-seasonal differencing is needed, we set d = 0. Based on this, initial non-seasonal model options include (1, 0, 0) or (1, 0, 1).

Seasonal effects are also evident in the Surface Temperature data. Clear spikes at lag 12 appear in both the ACF and PACF. The PACF spike suggests a seasonal autoregressive term (P = 1), while the ACF spike indicates a seasonal moving average term (Q = 1). With one seasonal difference already applied, D is set to 1. Suggested seasonal configurations include (1, 1, 1) or (1, 1, 0).

6.6 ARIMA Model Estimation

To compare time series models, we use several performance metrics. Lower residual variance (σ^2) indicates better model accuracy. A higher log-likelihood (log_lik) means a better fit to the data. For model selection, lower AIC, AICc, and BIC values are preferred, as they balance goodness of fit with model complexity to reduce overfitting.

Table 13: Comparison of Candidate ARIMA Models for Area Burnt

.model	sigma2	log_lik	AIC	AICc	BIC
A1	0.0005411	336.1312	-660.2625	-659.6709	-642.2388
A2	0.0008157	318.8870	-627.7740	-627.3544	-612.7542
A3	0.0005436	335.1613	-660.3227	-659.9031	-645.3029
A4	0.0008228	317.7260	-627.4519	-627.1741	-615.4361
A5	0.0008178	318.1920	-628.3839	-628.1061	-616.3681
A6	0.0005426	335.3603	-660.7205	-660.3009	-645.7008

Based on the model comparison results for the Area Burnt dataset, the ARIMA(0, 0, 1)(1, 1, 1)[12] model (A6) and ARIMA(1, 0, 1)(1, 1, 1)[12] model (A1) performed the best overall. A6 shows the most favorable information criteria values, while A1 remains a strong and competitive alternative.

Both models produce very low residual variance, with A1 performing slightly better in this regard. This suggests that both models fit the Box-Cox-transformed area burnt data well and effectively capture its variance and seasonal structure.

Model A6 likely benefits from a more complex seasonal moving average component, contributing to its strong performance in terms of information criteria. In contrast, A1 offers a simpler structure while achieving nearly equivalent results. Given its slightly lower residual variance and more streamlined design, A1 may be preferable in practice, depending on further diagnostic checks and forecasting accuracy.

Table 14: Comparison of Candidate ARIMA Models for Surface Temperature

.model	sigma2	log_lik	AIC	AICc	BIC
A1	0.5253596	-174.9705	359.9410	360.3636	374.9271
A2	0.8126734	-194.3194	396.6388	396.9185	408.6277
A3	0.5062717	-171.7440	355.4881	356.0838	373.4713
A4	0.7623761	-189.3057	388.6115	389.0340	403.5976

Based on the model comparison results for the Surface Temperature dataset, the ARIMA(1, 0, 1)(1, 1, 1)[12] model (A3) and ARIMA(1, 0, 0)(1, 1, 1)[12] model (A1) demonstrated the best performance among the tested configurations for modeling surface temperature.

Model A3 achieved the most favorable information criteria values, indicating the best overall fit. Model A1 followed closely, making it a strong alternative. Both models also exhibited relatively low error variances, with A3 performing slightly better in this regard.

A3 includes both autoregressive and moving average components in both the seasonal and non-seasonal parts of the model, allowing it to capture more complexity in the temperature data. A1, on the other hand, uses a simpler structure with only a first-order autoregressive term in the non-seasonal part. While A1 offers a more interpretable model, A3 provides a better statistical fit and lower residual variance, making it the preferred choice for forecasting surface temperature.

6.7 Diagnostic Testing

Diagnostic testing in time series analysis checks whether a model's residuals meet key assumptions, such as independence and constant variance. It helps determine if the model adequately captures the structure in the data. One common method is the Ljung-Box test, which assesses whether residuals are autocorrelated. If the residuals pass diagnostic tests, the model is considered a good fit for the data.

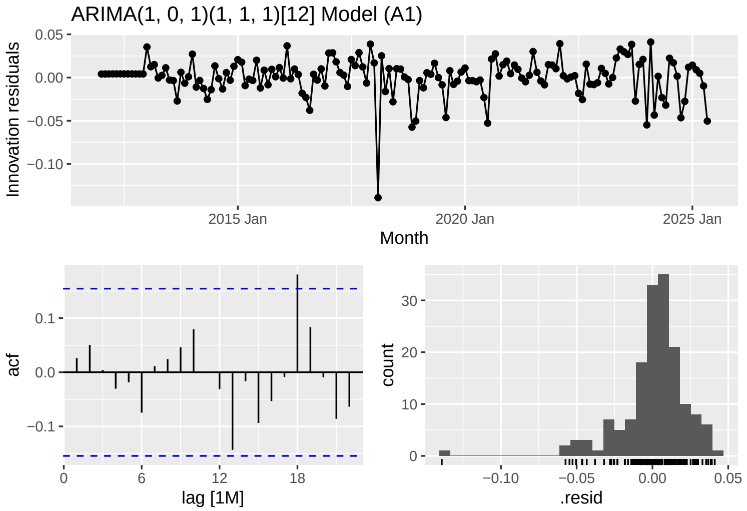
Table 15: Ljung-Box Test for ARIMA(1, 0, 1)(1, 1, 1)[12] Model (A1)

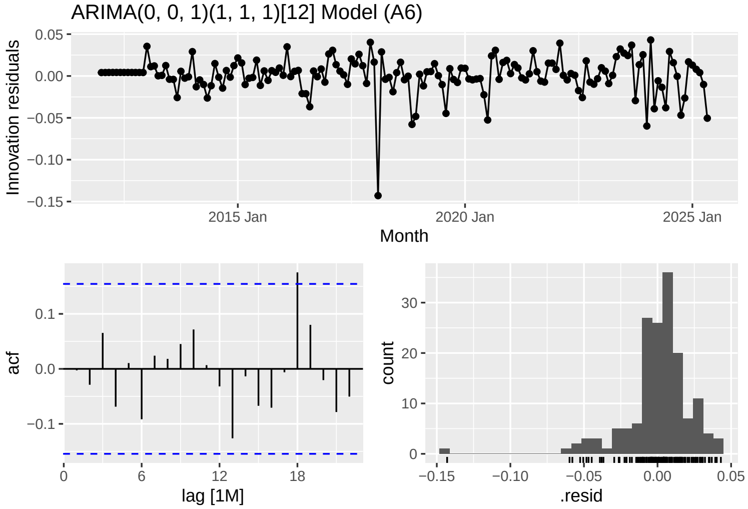
.model	lb_stat	lb_pvalue
A1	3.437874	0.9039574

Table 16: Ljung-Box Test for ARIMA(0, 0, 1)(1, 1, 1)[12] Model (A6)

.model	lb_stat	lb_pvalue
A6	4.665132	0.8624685

Diagnostic testing was conducted on the residuals of the ARIMA(1,0,1)(1,1,1)[12] model (A1) and ARIMA(0,0,1)(1,1,1)[12] model (A6), fitted to the Area Burnt time series. Both models passed the Ljung-Box test at lag 12, indicating no significant autocorrelation in the residuals and suggesting they can be treated as white noise. The p-value for A1 was slightly higher than for AR6, implying that A1 produced more random residuals.





Visual inspection of the residual plots reinforced this finding. The A1 model showed residuals that were more tightly centered around zero and appeared to have slightly lower variance. While A1 includes one additional non-seasonal AR term compared to A6, this added complexity is justified by the improvement in residual behavior. Overall, diagnostic results support A1 as the better model for capturing the structure of the transformed time series.

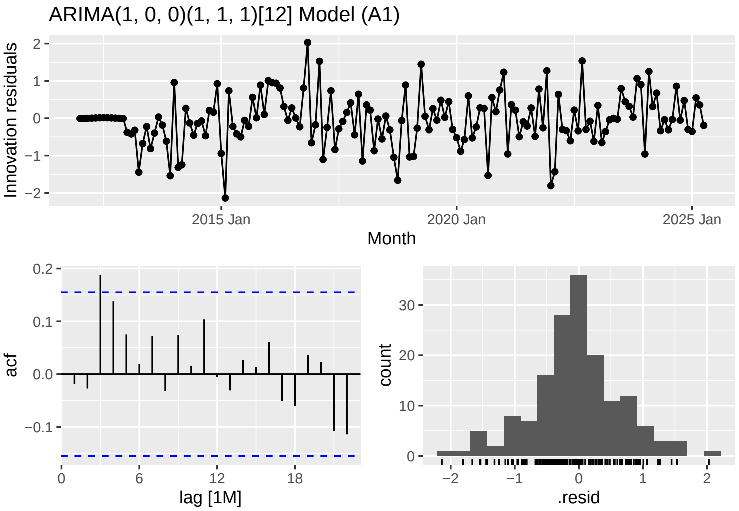
Table 17: Ljung-Box Test for ARIMA(1, 0, 0)(1, 1, 1)[12] Model (A1)

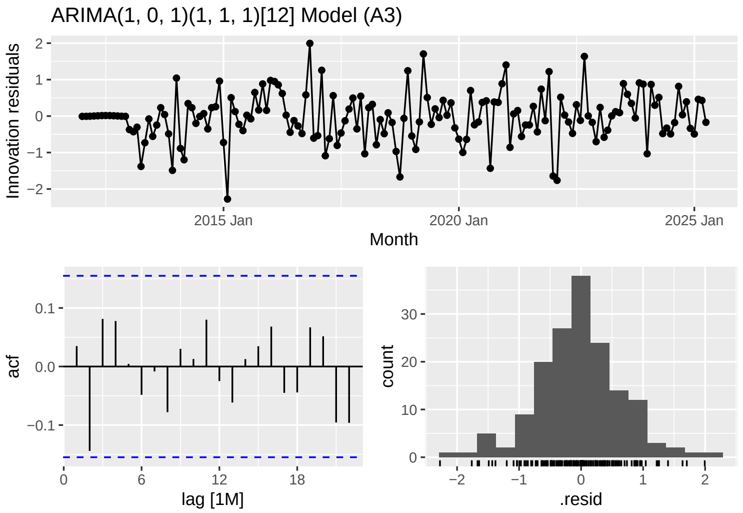
.model	lb_stat	lb_pvalue
A1	14.11338	0.1183489

Table 18: Ljung-Box Test for ARIMA(1, 0, 1)(1, 1, 1)[12] Model (A3)

.model	lb_stat	lb_pvalue
A3	8.550196	0.3816481

Diagnostic testing was conducted on the residuals of the ARIMA(1,0,0)(1,1,1)[12] model (A1) and ARIMA(1,0,1)(1,1,1)[12] model (A3), fitted to the Surface Temperature time series. Both models passed the Ljung-Box test at lag 12, indicating no significant autocorrelation in the residuals. However, A3 had a higher p-value compared to A1, suggesting that A3 produced more random and less autocorrelated residuals, indicating a better statistical fit.





Visual inspection of the residual plots supports these results. Both models show residuals centered around zero with similar distribution shapes, but the ACF plot for A3 displays smaller spikes, all within the confidence bounds. In contrast, A1 has a few values closer to the threshold. Although A3 includes an additional moving average term, the improved residual behavior justifies the extra complexity. Based on these diagnostics, A3 is the stronger choice for modeling the surface temperature time series.

6.8 ARIMA Model Selection

For the Area Burnt dataset, the information criteria slightly favor A6, which has a lower AIC and BIC compared to A1. These values suggest that A6 offers a marginally better fit with a simpler structure. However, residual diagnostics previously indicated that A1 had better-behaved residuals. Given the small differences in AIC and BIC and the improved residual characteristics of A1, the added complexity is justifiable. As a result, ARIMA(1, 0, 1)(1, 1, 1)[12] model A1 remains the preferred model due to its stronger diagnostic performance.

For the Surface Temperature dataset, both AIC and BIC clearly support A3 over A1. These improvements align with earlier diagnostic results showing that A3 produced more random, less autocorrelated residuals. Since A3 enhances both model fit and residual quality with only one additional parameter, the slight increase in complexity is warranted. Therefore, ARIMA(1, 0, 1)(1, 1, 1)[12] model A3 is the preferred model for capturing the structure of the surface temperature series.

6.9 Auto-ARIMA Implementation

For Area Burnt, the selected model is ARIMA(0,0,1)(2,1,0)[12], applied to the Box-Cox transformed data. It includes one non-seasonal moving average (MA) term and two seasonal autoregressive terms. A seasonal difference of order 1 at lag 12 captures annual seasonality. This configuration effectively models short-term shocks and recurring seasonal patterns in the monthly area burnt data.

For Surface Temperature, the chosen model is ARIMA(1,0,0)(2,1,1)[12], applied to the original, untransformed temperature series. It includes one non-seasonal autoregressive (AR) term, two seasonal AR terms, and one seasonal MA term. With a seasonal difference of order 1 at lag 12, the model accounts for annual cycles and captures both persistence and seasonal variation in the surface temperature data.

Table 19: Selected ARIMA Model vs. Auto ARIMA Model for Area Burnt

.model	sigma2	log_lik	AIC	AICc	BIC
A1	0.0005411	336.1312	-660.2625	-659.6709	-642.2388
Auto	0.0007180	326.3985	-644.7971	-644.5193	-632.7813

For the Area Burnt dataset, the ARIMA(1,0,1)(1,1,1)[12] model (A1) outperforms the auto ARIMA model (Auto) on all evaluation metrics. A1 has a higher log-likelihood and lower values for AIC, AICc, BIC, and residual variance, indicating a better overall fit for modeling the burned area data.

Table 20: Selected ARIMA Model vs. Auto ARIMA Model for Surface Temperature

.model	sigma2	log_lik	AIC	AICc	BIC
A3	0.5062717	-171.7440	355.4881	356.0838	373.4713
Auto	0.5516450	-173.3845	360.7691	361.5691	381.7495

For the Surface Temperature dataset, the ARIMA(1,0,1)(1,1,1)[12] model (A3) also outperforms the auto ARIMA model (Auto). It achieves lower AIC, AICc, BIC, and residual variance, along with a higher log-likelihood. These results suggest that A3 more effectively captures the structure of the temperature time series than the automated model.

6.10 Forecast Accuracy Comparison

Table 21: Forecast Accuracy Comparison for Area Burnt Models

.model	.type	RMSE	MAE	MAPE	MASE	RMSSE
A1	Test	0.0247187	0.0195641	0.4657212	0.9119472	0.7073347
Auto	Test	0.0245604	0.0191416	0.4558044	0.8922564	0.7028032
ETS	Test	0.0242740	0.0189861	0.4520514	0.8850081	0.6946093
Naive	Test	0.0655133	0.0554322	1.3115219	2.5838791	1.8746842
SNaive	Test	0.0298409	0.0237571	0.5646594	1.1073965	0.8539059

Among the models tested on the Area Burnt dataset, the ETS model delivers the best overall performance. It achieves the lowest values across all evaluated error metrics, including RMSE, MAE, MAPE, MASE, and RMSSE. The auto ARIMA and ARIMA(1, 0, 1)(1, 1, 1)[12] models (Auto and A1) also perform well, with only slightly higher error values, though the differences are minimal. The seasonal naive model performs noticeably worse, and the naive model performs the worst by a wide margin. Overall, ETS stands out as the most accurate and reliable model for forecasting Area Burnt.

Table 22: TS-CV Forecast Accuracy Comparison for Area Burnt Models

.model	.type	RMSE	MAE	MAPE	MASE	RMSSE
A1	Test	0.0282565	0.0189188	0.4526012	0.8406229	0.8456315
Auto	Test	0.0282733	0.0193120	0.4616260	0.8580925	0.8461340
ETS	Test	0.0293359	0.0200138	0.4783591	0.8892766	0.8779349

.model	.type	RMSE	MAE	MAPE	MASE	RMSSE
Naive	Test	0.0581702	0.0451952	1.0783043	2.0081677	1.7408600
SNaive	Test	0.0348892	0.0236375	0.5650431	1.0502874	1.0441287

The time series cross-validation results support the earlier findings, with the ARIMA models (A1 and Auto) and the ETS model continuing to outperform the naive and seasonal naive baselines. However, the rankings differ slightly compared to the holdout test evaluation.

In this analysis, the ARIMA(1,0,1)(1,1,1)[12] model (A1) performs best overall. It records the lowest values for MAE, MAPE, and MASE. The auto ARIMA model (Auto) closely follows, with nearly identical performance across all metrics. The ETS model, which ranked highest in the holdout test, now ranks third, with slightly higher errors across all metrics.

As before, the naive and seasonal naive models perform significantly worse. While the seasonal naive model shows some seasonal awareness, its error metrics are notably higher than those of the top three models.

Overall, the cross-validation results confirm the strength of the ARIMA and ETS models. However, they suggest that ARIMA models may offer more consistent performance across time, whereas ETS may be more sensitive to recent patterns.

Table 23: Forecast Accuracy Comparison for Surface Temperature Models

.model	.type	RMSE	MAE	MAPE	MASE	RMSSE
A3	Test	0.7827370	0.6017210	12.15088	0.6802303	0.7009151
Auto	Test	0.8665138	0.6998976	14.60201	0.7912164	0.7759345
ETS	Test	0.7890160	0.6011184	11.57367	0.6795490	0.7065378
Naive	Test	13.0368061	10.4337618	168.76927	11.7951020	11.6740294
SNaive	Test	0.8302519	0.6409263	14.47744	0.7245509	0.7434632

Among the models tested on the Surface Temperature dataset, the ETS model again performs the best overall. It achieves the lowest values for several key error metrics, including MAE, MAPE, and MASE, and ranks second in RMSE and RMSSE, closely behind the ARIMA(1, 0, 1)(1, 1, 1)[12] model (A3). While A3 slightly outperforms ETS in those two metrics, the differences are minimal. In contrast, the auto ARIMA and seasonal naive models perform moderately worse across all metrics. The naive model performs significantly worse than all others, with error values that are substantially higher. Overall, ETS provides the most balanced and accurate forecasts across the evaluated metrics.

Table 24: TS-CV Forecast Accuracy Comparison for Surface Temperature Models

.model	.type	RMSE	MAE	MAPE	MASE	RMSSE
A3	Test	0.9213497	0.7019581	19.56348	0.8926533	0.8939067
Auto	Test	0.9901017	0.7735505	21.11625	0.9836946	0.9606108
ETS	Test	0.8516884	0.6630540	17.80972	0.8431804	0.8263203
Naive	Test	11.6757091	9.4851190	208.75240	12.0618626	11.3279399
SNaive	Test	1.0203462	0.7679146	19.74484	0.9765275	0.9899545

The time series cross-validation results reinforce earlier conclusions, with the ETS model again performing best overall. It achieves the lowest values across all evaluated error metrics, including RMSE, MAE, MAPE, MASE, and RMSSE.

The ARIMA(1, 0, 1)(1, 1, 1)[12] model (A3) ranks second, followed closely by the seasonal naive model (SNaive). The auto ARIMA model performs slightly worse, particularly in MAE, MAPE, and MASE. As in previous analyses, the naïve model is the poorest performer, with error values significantly higher than those of the other models.

While the performance gaps among ETS, A3, and SNaive are narrower in cross-validation than in the holdout test, ETS maintains its lead, indicating strong generalization across different time windows and forecast origins. These results confirm that ETS remains the most accurate and stable model for forecasting surface temperature in this dataset.

7 Other Forecasting Methods

7.1 Prophet Model Evaluation

The Prophet model is a forecasting tool developed by Facebook for time series data that exhibits strong seasonal effects and trends. It uses an additive model where the time series is expressed as a combination of trend, seasonality, and holiday effects. Prophet automatically detects and handles missing data, outliers, and changes in trend, making it robust and easy to use with minimal tuning. It supports daily, weekly, and yearly seasonality and is well-suited for business and economic forecasting tasks with irregular observations or multiple seasonal patterns.

Table 25: Forecast Accuracy Comparison for Area Burnt Models

.model	.type	RMSE	MAE	MAPE	MASE	RMSSE
A1	Test	0.0247187	0.0195641	0.4657212	0.9119472	0.7073347
Auto	Test	0.0245604	0.0191416	0.4558044	0.8922564	0.7028032
ETS	Test	0.0242740	0.0189861	0.4520514	0.8850081	0.6946093
Prophet	Test	0.0372904	0.0319285	0.7581801	1.4882947	1.0670768

For the Area Burnt dataset, the Prophet model performs noticeably worse than the ARIMA and ETS models. It records the highest error values across all metrics. In contrast, the ETS model achieves the lowest errors overall, closely followed by the auto ARIMA and ARIMA(1,0,1)(1,1,1)[12] (A1) models.

While Prophet still produces reasonable forecasts, its accuracy lags behind the other models, particularly in MAE and MASE, where its values are nearly 70% higher than those of ETS. This suggests that Prophet may not capture the seasonal and structural nuances of the transformed area burnt data as effectively as the time-series-specific models provided by the fable framework.

Overall, ETS remains the best-performing model, with ARIMA models close behind, while Prophet is the least accurate in this comparison.

Table 26: Forecast Accuracy Comparison for Surface Temperature Models

.model	.type	RMSE	MAE	MAPE	MASE	RMSSE
A3	Test	0.7827370	0.6017210	12.15088	0.6802303	0.7009151
Auto	Test	0.8665138	0.6998976	14.60201	0.7912164	0.7759345
ETS	Test	0.7890160	0.6011184	11.57367	0.6795490	0.7065378
Prophet	Test	0.7196889	0.5542652	10.95407	0.6265827	0.6444577

On the Surface Temperature dataset, the Prophet model performs best overall. It achieves the lowest error values across all metrics. These results indicate strong accuracy and reliability in capturing the seasonal and trend components of the data.

The ETS model and ARIMA(1,0,1)(1,1,1)[12] (A3) follow closely, with slightly higher but still competitive error values. The auto ARIMA model performs slightly worse across all metrics but remains within a reasonable range.

Overall, the results suggest that Prophet provides the most accurate forecasts for Surface Temperature in this comparison. Its use of additive seasonality appears well-suited to the structure of the data, where seasonal patterns remain consistent over time.

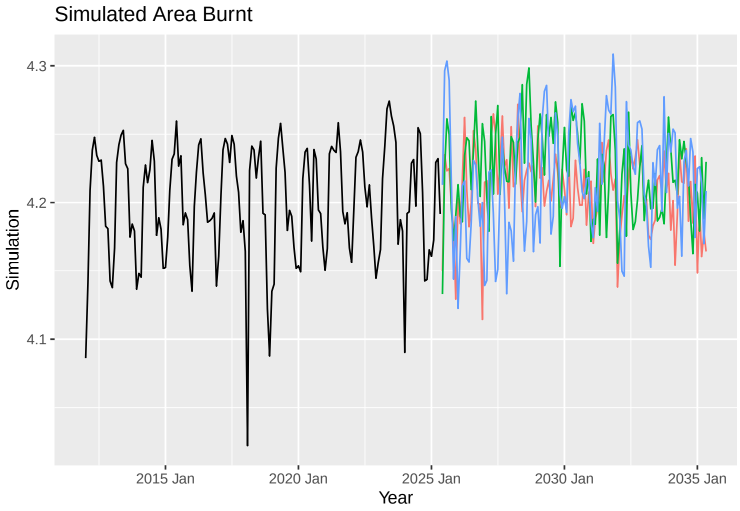
7.2 Neural Network Autoregression Model

Neural network models are machine learning algorithms inspired by the structure of the human brain. They consist of layers of interconnected nodes (or "neurons") that process data by applying weighted transformations and nonlinear activation functions. In time series forecasting, neural networks can model complex, non-linear relationships and capture patterns that traditional statistical models might miss. They are particularly useful when the data has intricate interactions or lacks clear seasonal or trend structures, but they often require more data and tuning to perform well.

Table 27: Forecast Accuracy Comparison for Area Burnt Models

.model	.type	RMSE	MAE	MAPE	MASE	RMSSE
A1	Test	0.0247187	0.0195641	0.4657212	0.9119472	0.7073347
Auto	Test	0.0245604	0.0191416	0.4558044	0.8922564	0.7028032
ETS	Test	0.0242740	0.0189861	0.4520514	0.8850081	0.6946093
NN	Test	0.0332213	0.0263161	0.6241304	1.2266797	0.9506385
Prophet	Test	0.0373156	0.0318955	0.7573675	1.4867582	1.0677984

On the Area Burnt dataset, the neural network model performs moderately well but does not outperform the ARIMA or ETS models. It records higher error values across all metrics, including RMSE, MAE, MAPE, MASE, and RMSSE. However, it still performs better than the Prophet model, which has the highest errors overall. This suggests that while the neural network can capture some non-linear patterns, it may not be as effective as traditional time series models for this particular dataset.



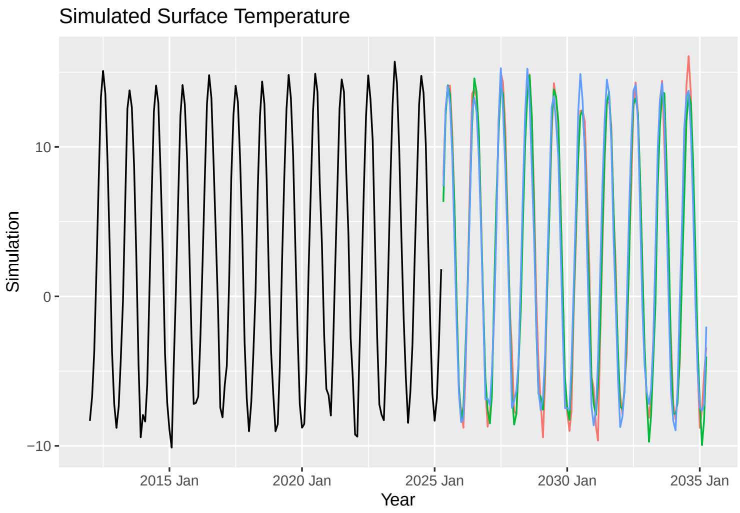
The simulation paths show moderate variability in both the magnitude and timing of seasonal peaks and troughs. This spread reflects the forecast uncertainty captured by the neural network model, particularly as the horizon extends further into the future. While the simulations diverge, they remain within a reasonable range, indicating that the model captures underlying seasonality and short-term dynamics without overfitting to noise.

All simulated paths maintain a stable mean level and preserve the regular seasonal rhythm seen in the historical data. This suggests the model does not anticipate any significant structural changes or long-term trends in the transformed area burnt values. Instead, it expects future behavior to resemble past patterns, with variability driven by natural seasonal fluctuations rather than large-scale shifts.

Table 28: Forecast Accuracy Comparison for Surface Temperature Models

.model	.type	RMSE	MAE	MAPE	MASE	RMSSE
A3	Test	0.7827370	0.6017210	12.15088	0.6802303	0.7009151
Auto	Test	0.8665138	0.6998976	14.60201	0.7912164	0.7759345
ETS	Test	0.7890160	0.6011184	11.57367	0.6795490	0.7065378
NN	Test	1.0711009	0.8382981	19.08544	0.9476746	0.9591355
Prophet	Test	0.7193100	0.5535378	10.94516	0.6257604	0.6441184

On the Surface Temperature dataset, the neural network model performs the worst among the models tested. It records the highest error values across all metrics. These results indicate that the model struggles to capture the structure of the series as effectively as the others. While the neural network model may still capture some non-linear behavior, it appears less suitable for this dataset's seasonal and trend components.



The simulation paths display low to moderate variability, with slight differences in amplitude and timing of seasonal peaks. While all paths preserve the clear annual seasonal pattern, the divergence among them increases slightly over the 10-year horizon, reflecting forecast uncertainty in the neural network model.

Despite these differences, the simulations remain close in structure, indicating the model expects consistent seasonal cycles with relatively stable temperature ranges. There are no signs of large-scale trend shifts or abrupt structural changes, suggesting the model anticipates future surface temperatures to follow historical patterns with predictable seasonal variation.

8 Conclusion

In conclusion, this project aimed to analyze and forecast wildfire activity and surface temperature using various statistical and machine learning techniques. The analysis of wildfire time series data in North America (primarily from 2012 to 2024) revealed a strong seasonality with annual peaks in the summer months and significant inter-annual variability. Although a strong long-term trend was not visible across the entire period, higher amplitudes in burned area in certain years suggested potential structural breaks, likely linked to extraordinary climate events. Decomposition analysis confirmed the distinct seasonality and indicated a possible upward trend in burned area starting around 2021, with a notable peak in 2023 in the non-transformed data. The surface temperature data showed a generally rising trend starting around 2018, along with a consistent annual seasonality.

In terms of forecasting, several methods were applied, including simple benchmark models, Exponential Smoothing (ETS), ARIMA models, Prophet, and Neural Network Autoregression (NNAR). For forecasting burned wildfire area, ETS and ARIMA generally proved to be the most effective models, with the seasonal naive method serving as a useful simple comparison. The Prophet and NNAR models showed lower accuracy in this application. The superior performance of the seasonal naive model over simpler non-seasonal models highlighted the importance of considering seasonal dynamics.

For forecasting surface temperature, the Prophet model performed the best, followed by ETS and ARIMA, which also delivered competitive results. The NNAR model exhibited the poorest performance here.

The findings of this analysis emphasize the significant influence of seasonal patterns on both wildfire activity and surface temperatures. The identified trends, particularly the recent potential increase in burned area and the clear rise in surface temperatures, underscore the potential impacts of climate change. The developed and compared forecasting models provide valuable tools for environmental planning, resource allocation, and public policy by offering insights into future patterns of wildfire activity and climate trends. Furthermore, the identification of the strengths and weaknesses of different forecasting methods contributes to a more informed approach in selecting appropriate tools for various types of wildfire-related data.

9 References

Bowman, D.M., Moreira-Muñoz, A., Kolden, C.A., Chávez, R.O., Muñoz, A.A., Salinas, F. and Johnston, F.H., 2020. Human–environmental drivers and impacts of the globally extreme 2017 Chilean fires. Ambio, 49(2), pp.350–362. Available at: https://doi.org/10.1007/s13280-019-01129-7 [Accessed 28 Apr. 2025].

Box, G.E.P., Jenkins, G.M., Reinsel, G.C. and Ljung, G.M., 2015. Time series analysis: Forecasting and control. 5th ed. Hoboken, NJ: Wiley.

Hyndman, R.J. and Athanasopoulos, G., 2021. Forecasting: Principles and practice. 3rd ed. Melbourne: OTexts. Available at: https://otexts.com/fpp3/ [Accessed 28 Apr. 2025].

Intergovernmental Panel on Climate Change (IPCC), 2021. Climate Change 2021: The Physical Science Basis. Cambridge: Cambridge University Press. Available at: https://www.ipcc.ch/report/ar6/wg1/ [Accessed 28 Apr. 2025].

Johnston, F.H., Henderson, S.B., Chen, Y., Randerson, J.T., Marlier, M., DeFries, R.S. and Brauer, M., 2012. Estimated global mortality attributable to smoke from landscape fires. Environmental Health Perspectives, 120(5), pp.695–701. Available at: https://doi.org/10.1289/ehp.1104422 [Accessed 28 Apr. 2025].

Our World in Data, n.d.-a. Wildfires. [online] Available at: https://ourworldindata.org/wildfires [Accessed 28 Apr. 2025].

Our World in Data, n.d.-b. CO_2 and Greenhouse Gas Emissions. [online] Available at: https://ourworldindata.org/co2-and-greenhouse-gas-emissions [Accessed 28 Apr. 2025].

Taylor, S.J. and Letham, B., 2018. Forecasting at scale. The American Statistician, 72(1), pp.37–45. Available at: https://doi.org/10.1080/00031305.2017.1380080 [Accessed 28 Apr. 2025].

van der Werf, G.R., Randerson, J.T., Giglio, L., van Leeuwen, T.T., Chen, Y., Rogers, B.M. and Kasibhatla, P.S., 2017. Global fire emissions estimates during 1997–2016. Earth System Science Data, 9(2), pp.697–720. Available at: https://doi.org/10.5194/essd-9-697-2017 [Accessed 28 Apr. 2025].