On the Suitability of Machine Learning Approaches for Long-Horizon Power Demand Forecasting in Gothenburg

Master's thesis in Engineering Mathematics and Computational Science
Master's thesis in Complex Adaptive Systems

WILLIAM KARLSSON
TERESE KÄRNELL

DEPARTMENT OF MATHEMATICAL SCIENCES
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2026
www.chalmers.se

Master's thesis 2026

On the Suitability of Machine Learning Approaches for Long-Horizon Power Demand Forecasting in Gothenburg

WILLIAM KARLSSON
TERESE KÄRNELL

Department of Mathematical sciences
Division of Analysis and Probability Theory
Chalmers University of Technology
Gothenburg, Sweden 2026

On the Suitability of Machine Learning Approaches for Long-Horizon Power Demand Forecasting in Gothenburg
WILLIAM KARLSSON
TERESE KÄRNELL

Supervisor: Peter Helgesson, Department of Mathematical Sciences
Examiner: Peter Helgesson, Department of Mathematical Sciences

Master's Thesis 2026
Department of Mathematical Sciences
Division of Anslysis and Probability Theory
Chalmers University of Technology
SE-412 96 Gothenburg
Telephone +46 31 772 1000

Typeset in LaTeX
Printed by Chalmers Reproservice
Gothenburg, Sweden 2026

On the Suitability of Machine Learning Approaches for Long-Horizon Power Demand Forecasting in Gothenburg
WILLIAM KARLSSON
TERESE KÄRNELL
Department of Mathematical Scineces
Chalmers University of Technology

Abstract

Gothenburg's electricity grid is projected to require substantial expansion to accommodate the large-scale electrification of society. Long-horizon forecasts of power peaks are critical for informing energy companies of the infrastructure investment needed to meet future demand. This thesis evaluates the capability of machine learning models to produce such forecasts. Machine learning models and traditional time series models were fitted to two datasets of differing temporal coverage and sampling frequency, incorporating target variable data provided by Göteborg Energi alongside publicly available exogenous features. Under the available data, all models projected a continually constant level of power demand, with traditional time series models demonstrating marginally higher reliability across most forecast horizons. The study was subject to several limitations, including small and sparse datasets, historically flat trends in the target variables, and exogenous features with limited explanatory power. It was concluded that the currently available data restricts the viability of machine learning-based modeling approaches for long-horizon power peak forecasting. Nevertheless, the prospects for such methods remain promising, particularly in the context of short-horizon forecasting and with access to richer, more granular and explanatory features.

Keywords: Göteborg Energi, Gothenburg, power demand, electricity consumption, forecast, machine learning, time series, Bayesian statistics, XGBoost, DeepAR.

Acknowledgements

We would first like to thank our supervisor Peter Helgesson for his expertise and valuable feedback during our weekly meetings. Our gratitude also extends to Dina Zuko and colleagues at Göteborg Energi for generously and continuously providing us with the data that made this analysis possible. We are very thankful for the domain knowledge you shared with us and for the engaging discussions held throughout the course of the project.

Finally, our sincere thanks go to our dear families and friends for their unwavering support.

William Karlsson and Terese Kärnell, Gothenburg, May 2026

List of Acronyms

Below is the list of acronyms that have been used throughout this thesis listed in alphabetical order:

ACF	Auto-Correlation Function
AR	Autoregressive
CART	Classification and Regression Tree
CPI	Consumer Price Index
CV	Cross-validation
DeepAR	Deep Autoregressive
ELIS	Elbilen i Sverige
EV	Electric Vehicle
FNN	Feedforward Neural Network
GENAB	Göteborg Energi Nät AB
HMC	Hamiltonian Monte Carlo
LSTM	Long Short-Term Memory
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
MCMC	Markov Chain Monte Carlo
ML	Machine Learning
MLE	Maximum likelihood estimates
MLP	Multi-Layer Perceptron
NLL	Negative Log-likelihood
NUTS	No-U-Turn Sampler
PACF	Partial Auto-Correlation Function
RMSE	Root Mean Squared Error
RNN	Recurrent Neural Network
SARIMA	Seasonal Auto-regressive Integrated Moving Average
SARIMAX	Seasonal Auto-regressive Integrated Moving Average with Exogenous Regressors
SCB	Statistiska Centralbyrån
SMHI	Swedish Meteorological and Hydrological Institute
XGBoost	Extreme Gradient Boosting

Nomenclature

Below is the nomenclature of indices, sets, parameters, and variables that have been used throughout this thesis.

Indices

i, j, k	Arbitrary sequence indices
t	Time index

Sets

$\{X_t\}$	Arbitrary time series
D_Y	Dataset of yearly aggregated observations
D_M	Dataset of monthly aggregated observations

Parameters

p	Parameter for the amount of autoregressive terms
d	Differencing degree
q	Parameter for the amount of moving average terms
P	Parameter for the amount of seasonal autoregressive terms
D	Seasonal differencing degree
Q	Parameter for the amount of seasonal moving average terms
s	Period of seasonality
ϕ	Coefficient of autoregression term
Φ	Coefficient of seasonal autoregression term
θ	Coefficient of moving average term
Θ	Coefficient of seasonal moving average term

σ	Standard deviation
β	Exogenous feature coefficients
h	General parameter for lag
n	Amount of observations

Variables

Z_t	White noise term
m_t	Trend component of a time series
s_t	Seasonal component of a time series
$\boldsymbol{x}_t$	Vector of exogenous variables at time t

Functions

$\gamma_X(\cdot, \cdot)$	Autocovariance function of the time series $\{X_t\}$
$\rho_X(\cdot)$	Autocorrelation function (ACF) of the time series $\{X_t\}$
$\alpha_X(\cdot)$	Partial autocorrealtion function (PACF) of the time series $\{X_t\}$
$b_t^l(X^n)$	Best linear predictor of X_t

Contents

List of Figures

List of Tables

1

Introduction

Sweden has set an ambitious target to achieve net-zero greenhouse gas emissions by 2045 via the climate law implemented January 1st 2018 [1]. To accomplish this, among other measures, large-scale electrification of society is required particularly within sectors that are currently heavily dependent on fossil fuels, such as industry and transportation [2]. This in turn places demands on the electricity grid to be able to transmit the power required, a demand that is steadily increasing [3].

The Swedish electrical grid is structured into three hierarchical levels. The *transmission grid* transports electricity from power generation facilities to the *regional grids*. The regional grids then distribute electricity to large industrial consumers and further to the *local grids*, which supply electricity to smaller industries and residential users [4]. This layered structure may give rise to transmission bottlenecks in areas where grid infrastructure is insufficiently developed or lacks the capacity required to meet increasing demand. The transmission grid is owned and operated by the Swedish state through Svenska Kraftnät, while the regional and local grids are owned by various grid operators such as Vattenfall and Ellevio. In Gothenburg, the municipal energy company Göteborg Energi (GE) owns the local grid through its subsidiary Göteborg Energi Nät AB (GENAB). The local grid receives power from both Vattenfall's and Elevio's regional grid as well as some local generation facilities.

There are ongoing plans to expand and reinforce the transmission grid in order to increase capacity and improve power transfer to the regional and local grids [5]. However, such investments are associated with long planning and implementation horizons, which may temporarily hinder the pace of electrification due to limitations in available power capacity.

One possible approach to address these challenges would be to expand the grid to ensure a capacity margin exceeding anticipated demand. However, as previously noted, grid expansions are associated with long planning and implementation timelines, as well as significant economic and environmental costs. Since GENAB is a municipally owned company, such investments are partly financed through public funds via tax revenues. Consequently, expanding the grid beyond what is necessary as a precautionary measure is difficult to justify from both an economic and societal perspective.

In addition to economic considerations, the choice of transmission technology also involves trade-offs in terms of environmental impact and operational performance. Overhead lines, which have limited applicability in densely populated areas, are relatively cost-efficient to install and maintain but are more susceptible to weather-related disturbances that can lead to outages [6]. Underground cables, on the other hand, are typically the only viable option in urban environments and are less exposed to external disruptions. However, they are more costly to install, more complex to maintain and repair, and may entail greater environmental impacts due to excavation work and the burial of metal-containing infrastructure, which can, for instance, contribute to the spread of invasive species [7].

For these reasons, GENAB aims to improve its ability to forecast future capacity needs in order to enable proactive planning and timely grid investments. In this context, capacity refers to the amount of electrical power used simultaneously within the grid. The overall objective of this project is therefore to investigate whether machine learning methods can be used to forecast power demand up until the year 2040.

Forecasting power demand poses several challenges. Electricity demand is highly time-dependent and influenced by stochastic factors such as weather conditions and economic activity. As a result, many factors that influence demand lie outside the control of GENAB. This makes it difficult to accurately predict the exact power demand at a specific hour several decades into the future. Consequently, this project focuses on developing a machine learning framework that can serve as a foundation for future forecasting work at GENAB. To evaluate the framework the project will consist of two forms of forecasts, one long-horizon forecast that will forecast total annual electricity consumption through 2040, and one shorter forecast that will forecast a monthly highest power peak through 2030.

1.1 Background

Gothenburg has historically been, and continues to be, a major industrial hub, with several large manufacturing companies such as Volvo and SKF. In combination with planned expansions, new industrial establishments, and ongoing electrification of similar large-scale industries, this development is contributing to an increasing demand for electrical capacity and grid planning. This may, for example, involve the replacement of gas-fired furnaces with electrically powered alternatives, as well as substantial electricity requirements for processes such as hydrogen production [8].

One of Göteborg Energi's strategic goals is to contribute to a sustainable energy system while supporting electrification and regional economic development [9]. To support this goal, forecasting future power demand has been an important part of GENAB's strategy since 2022. Currently, time-series combined with scaling factors such as weather data and industrial development indicators are used to estimate hourly peak loads, defined as the maximum power demand within a given hour.

While this approach represents a well-established and reliable method for forecasting electricity demand, rapid technological and societal changes are increasing the uncertainty associated with long-term planning. Historically, grid development and maintenance could be planned on long time horizons with relatively stable demand patterns. Today, however, grid operators must adapt more quickly to changing consumption patterns and emerging technologies.

In this context, new measures may be required to maintain grid stability. These include demand-side management strategies such as conditional connection agreements or other incentives designed to shift electricity consumption throughout the day. Such measures can help smooth demand peaks and improve the efficiency of grid utilization, enabling higher energy consumption without necessarily increasing peak power demand.

The Swedish electricity grid is fundamentally well-dimensioned, as it has been designed in accordance with the so-called N-1 principle [10]. This implies that the system is built to withstand individual disturbances, such as the loss of a transmission line or a generation facility. As a result, the grid is inherently robust, even though transmission bottlenecks may still occur.

Consequently, the likelihood that capacity shortages would lead to power outages is low. The limits that GENAB and this project model against are therefore, to a large extent, administrative rather than physical. Exceeding these limits primarily results in additional economic costs, rather than having a direct impact on system operation.

1.2 Purpose and Objectives

The main goal of the thesis is to investigate if machine learning can be applied to forecast power demand and the potential of machine learning to improve the current capacity forecasts used at Göteborg Energi. An important part of the project is to assess the models' ability to anticipate power peaks and their advancement until the year 2040. The results should also show which parameters and variables are the most significant to the forecast model. By the end of the project, the objective is to provide a framework that, given a set of input data and parameters, can generate and visualize forecasting models together with their respective strengths and limitations.

The purpose of the project translates into the following main objectives:

1. Identify a broad range of relevant variables and forecasting models suitable for long-term electricity capacity forecasting based on available data and literature studies.

2. Construct multiple forecasting models using different combinations of variables and modeling approaches.

3. Evaluate which models provide the most reliable and accurate predictions of future power peaks up until the year 2040.
4. Analyze the significance and influence of the variables included in the models.
5. Discuss the relevance and applicability of machine learning methods for long-term capacity forecasting.

1.3 Limitations

The thesis aims to extensively investigate how different data sources, explanatory variables, and forecasting model types affect the accuracy and confidence of electricity capacity forecasts. However, several restrictions are imposed on the variety of data, features and model types.

Göteborg Energi is primarily interested in medium- to long-term forecasts. As a result, longer time series spanning several decades with relatively coarse temporal resolution are used instead of shorter datasets with higher resolution. Furthermore, changes in measurement technologies and data collection methods over time have resulted in datasets with varying structures and resolutions. Consequently, it is difficult to consistently distinguish between different consumer segments in the historical data resulting in use of aggregated data for the entire customer base instead of segmented as would be preferable.

While this simplifies the modeling process by reducing model complexity, it also limits the ability to analyze which customer segments contribute most to peak power demand and which demand-side measures could potentially be implemented. Additionally, it restricts the set of possible explanatory variables, since some variables may only be relevant for specific customer segments.

The feature selection is therefore largely guided by domain knowledge. Specifically, this means studying established electricity demand behavior and patterns documented in literature, rather than running an exhaustive feature search. This approach both improves interpretability and reduces model complexity.

Similarly, only a limited set of model families is evaluated based on prior knowledge and relevance to the forecasting task. Hyperparameter tuning is also limited to decrease complexity. Although computational efficiency is desirable, minimizing computational time is not a primary objective beyond these constraints.

Finally, the forecasts developed in this thesis are specifically tailored to Göteborg Energi and the Gothenburg region. As a result, the findings may not be directly transferable to other regions with different climatic, demographic, or infrastructural conditions. Furthermore, long-term forecasts extending to 2040 are inherently subject to uncertainty. Future developments in policy, technology, and consumer behavior may significantly influence electricity demand and are difficult to fully capture within forecasting models.

2

Theory

This chapter introduces data and theoretical background for the used models. Figures of data and variables can be found i Appendix A along with complementary theory.

2.1 Data and Variables

An important part of any machine learning project is the selection of variables expected to contribute to model performance. For the longer forecasts the chosen variables are *consumed energy in GWh in Gothenburg*, *energy efficiency*, *population in Gothenburg*, *salary in 2025's monetary value in Sweden* and *electric vehicles per capita in Sweden*.

For the shorter forecasts the chosen variables are *maximum consumed power in Gothenburg*, *average air temperature of the five coldest days of the month*, *energy efficiency*, *population in Gothenburg*, *salary in 2025's monetary value in Sweden* and *electric vehicles per capita in Gothenburg*. Since one of the purposes of this project is to forecast future power demand the underlying assumption is that past energy consumption in combination with the other variables will capture changes in technical progress as well as raised economic standards.

In general, any kind of machine learning project requires a large amount of data in order to achieve satisfactory predictive performance. For this project, data has been collected both from Göteborg Energi's internal databases and from external sources. From the internal databases, total annual consumption between 1996 and 2025 was obtained as well as monthly maximum power peaks between 2012 and 2025. Population data for Gothenburg was retrieved from the City of Gothenburg's statistical database [11]. Air temperature data was collected from the Swedish Meteorological and Hydrological Institute (SMHI) [12].

Energy efficiency data were constructed by the authors based on results from the report *Elanvändning i Sverige 2030 och 2050* [13]. According to the report, efficiency improvements occur across all sectors and are largely "autonomous," meaning they are not directly driven by policy. Instead, they are mainly influenced by economic, technological, and structural factors, although these may be indirectly shaped by

policy measures such as taxes, regulations, and support for research and technological development. Historical energy efficiency improvements have been approximately 2–3% per year, while future improvements are projected to be in the range of 3–4% according to [13]. However, according to experts at GENAB it is more feasible to estimate a future improvement rate to be about 1%. Based on these assumptions, the authors of this thesis constructed a corresponding time series with values within this interval, both on a monthly and yearly basis. To better reflect real-world variability, stochastic noise was introduced, ensuring that the resulting efficiency trajectory did not follow a perfectly smooth trend.

Data on real salary was calculated by the authors using data on average salary and the Consumer Price Index (CPI), both obtained from Statistiska centralbyrån (SCB) [14], and the formula

$$\text{Real salary} = \text{Average salary} \times \frac{CPI_{2025}}{CPI_{year}} \tag{2.1}$$

for each year between 1996 and 2025 as well as for each month between January 2012 to December 2025. The formula calculates the average salary in 2025 monetary value.

Data on electric vehicles was obtained from Power Circle's ELIS database [15], which provides statistics on the total number of electric vehicles in Sweden from 2013 to the present day. The type of vehicles in the data set and are cars, light- and heavy duty trucks and busses. Henceforth in the report, vehicle will refer to any of the before mentioned categories. Together with population data for Sweden from SCB, the authors calculated the number of electric vehicles per capita using the formula

$$\text{Electric Vehicles per Capita} = \frac{\text{Number of Electric Vehicles}}{\text{Population}} \tag{2.2}$$

for the years between 2013 and 2025 as well as each month between January 2012 to December 2025. All data have been collected in a table, one for yearly data and one for monthly data, for convenient access during modeling.

Forecasted values for each variable have also been compiled in a table. Data on consumption and power peaks was obtained from Göteborg Energi's internal database, which includes projections for future industrial investments and electrification. Population forecasts were collected from SCB, while the forecasted temperature data were generated by the author based on historical observations.

The forecast for real salary was produced by the authors using linear regression based on the historical real salary described above. Since salary formation in Sweden is typically based on negotiations conducted by the major labor unions, annual salary increases are similar and common, resulting in a roughly linear upward trend. Therefore, it is reasonable to model real salary according to this trend, even though it does not strictly follow changes in monetary value. This is because the CPI, which is required to convert average salary into real salary, is highly dependent on inflation fluctuations and other macroeconomic factors that are beyond the scope of this project to fully analyze and model.

For the number of electric vehicles, the authors applied a Gompertz growth curve, which is a special case of a sigmoidal growth model. The formula for the Gompertz curve is

$$f(x) = a \cdot e^{-b \cdot e^{-c \cdot x}} \tag{2.3}$$

where a represents the carrying capacity, b represents displacement along the x-axis and c represents the growth rate. This model is an asymmetric S-curve that assumes slower growth in the beginning and end of the curve than the standard logistic curve, meaning that initial growth is slower followed by rapid growth that tapers off again closer to saturation. Choosing a logistic curve to model the number of electric vehicles is reasonable since the total demand for vehicles is unlikely to increase substantially. Instead, the transition to electric mobility is expected to occur gradually, as existing internal combustion engine vehicles are replaced with electric vehicles once they reach the end of their lifespan [16]. Furthermore, a Gompertz curve is especially fitting to model technological advancements such as electric vehicles since initial costs of new technology often limits the initial growth before the technology becomes more widely accessible. The prediction is based on the histori-

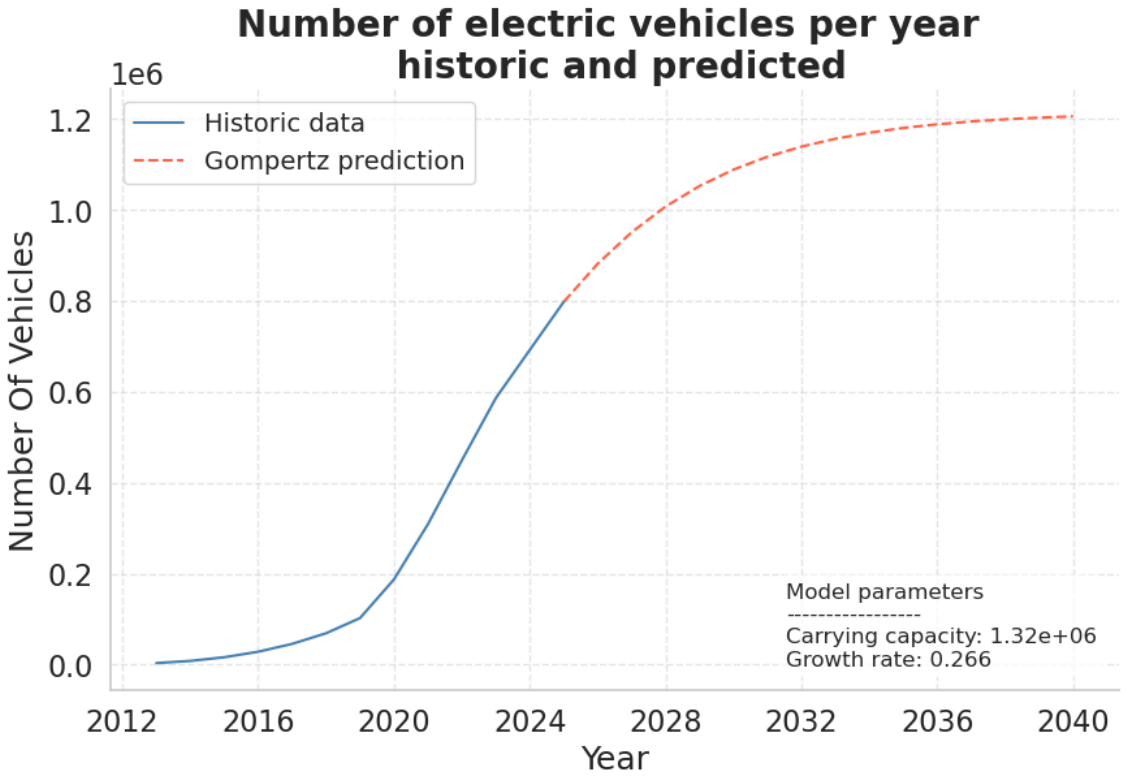

Figure 2.1: Figure of number of electric vehicles. The prediction shows a saturation at about 1.3 million vehicles.

cal data described above and is most likely conservative. According to Trafikanalys, which is a government agency, there are approximately 5.8 million vehicles, including both electrical and non-electric vehicles, in use in Sweden in 2025 [17]. As can be seen in figure 2.1 the forecast suggests that the number of vehicles saturates at $\approx$ 1.3 million vehicles. In comparison to the number of total vehicles in use today it is

clear that if a complete transition to electric vehicles is to take place, the prediction made based on the historical data is conservative.

It is important to note that the growth rate of electric vehicles is strongly linked to political decisions and policy measures, making the forecast highly uncertain. The projection was then combined with population forecasts from SCB to calculate electric vehicles per capita, as described above.

2.2 Introduction to Bayesian statistics

Bayesian statistics is introduced here as it forms the basis of one of the models considered in this study.

Bayesian statistics originates from work by Thomas Bayes and was later formalized by Pierre-Simon Laplace [18]. Although the theoretical framework dates back to the 18th century, its practical application expanded significantly during the second half of the 20th century, largely due to advances in computational methods. Today, Bayesian approaches are widely used in modern data analysis and machine learning [19].

In contrast to the frequentist framework, where probability is interpreted as the long-run frequency of repeated experiments and parameters are treated as fixed but unknown constants, Bayesian statistics interprets probability as a measure of uncertainty or belief. This distinction allows prior knowledge to be incorporated directly into the analysis. Consequently, parameters are modeled as stochastic variables, and inference is based on probability distributions rather than single point estimates [20].

A central objective in statistical inference is to learn about one or more unknown parameters, denoted by θ. In the Bayesian framework, prior knowledge about these parameters is represented by a prior distribution, $f(\theta)$. This information may be derived from previous studies, expert assessments, or other relevant sources.

Observed data, denoted by x, provide additional information about the parameters through the likelihood function, $f(x \mid \theta)$, which describes the probability of the data given the parameters. By combining the prior distribution and the likelihood using Bayes' theorem, the posterior distribution is obtained:

$$f(\theta \mid x) = \frac{f(x \mid \theta)\, f(\theta)}{f(x)}. \tag{2.4}$$

The posterior distribution represents the updated beliefs about the parameters after accounting for the observed data and forms the basis for inference, prediction, and uncertainty quantification.

In Bayes' theorem, $f(x)$ is known as the marginal likelihood (or evidence) and represents the probability of the observed data. It acts as a normalizing constant ensuring that the posterior distribution is a valid probability distribution. Since it

does not depend on the parameters, it is often omitted in practice and 2.4 is written as $f(\theta \mid x) \propto f(x \mid \theta) f(\theta)$ [20].

A key advantage of the Bayesian framework is that it provides a natural representation of uncertainty through posterior distributions rather than point estimates. This allows for more informative interpretations, such as credible intervals and probabilistic statements about parameters. However, the approach also introduces sensitivity to the choice of prior. Strongly informative priors may improve estimation in cases of limited data but can also introduce bias if not carefully justified. The extent of this risk depends on the level of informativeness assumed in the prior [19].

2.3 Overview of Time Series

This section about the fundamentals of time series is largely inspired by the book *Time Series: Theory and Methods*, written by Peter J. Brockwell and Richard A. Davis [21].

A time series is a set of chronological sequence of real-valued observations $\{X_t\}$, recorded at time t. A discrete time series $\{X_t\},\ t \in \mathbb{T}$, has a countable index set $\mathbb{T} \subset \mathbb{R}$. The concept of covariance is for time series extended to handle the self-dependency on previous observations.

Definition 1. For a process $\{X_t\}$ with $\mathrm{Var}(X_t) < \infty$ for all $t \in \mathbb{Z}$, the *autocovariance function* $\gamma_X(\cdot, \cdot)$ is defined by

$$\gamma_X(r, s) = Cov(X_r, X_s) = E[(X_r - E[X_r])(X_s - E[X_s])], \quad \forall r, s \in \mathbb{Z}.$$

Using the definition of autocovariance, it is also possible to define weak stationarity for time series.

Definition 2. The time series $\{X_t\},\ t \in \mathbb{Z}$ is *weakly stationary* if

(i) $Var(X_t) < \infty, \quad \forall t \in \mathbb{Z}$,

(ii) $E[X_t] = m, \quad \forall t \in \mathbb{Z},$

(iii) $\gamma_X(r, s) = \gamma_X(r + t, s + t), \quad \forall r, s, t \in \mathbb{Z}$

Stationarity is an important property in the context of time series, as there is extensive theory around modelling, analysis and predictions for such series. However, many real-life time series contain trends and seasonality that disrupt stationarity. Classically, a time series $\{X_t\}$ is modeled using the decomposition.

$$X_t = m_t + s_t + Y_t, \tag{2.5}$$

where m_t is the trend component, s_t is the seasonal component and Y_t is a stationary time series with mean zero. For forecasting, the trend and seasonality components are eliminated through differencing. The *difference operator* ∇ is defined as

$$\nabla X_t := X_t - X_{t-1} = (1 - B)X_t, \quad t \geq 2,$$

where B is the so called *backward shift operator* such that

$$BX_t := X_{t-1}. \tag{2.6}$$

Powers of B are defined as

$$B^j X_t := X_{t-j}, \quad t > j \tag{2.7}$$

or similarly for the difference operator [21]

$$\nabla^j X_t = \nabla \nabla^{j-1} X_t.$$

Assuming the trend component m_t from Equation (2.5) is expressed as the polynomial

$$m_t := \sum_{j=0}^{q} a_j t^j, \quad q < n,$$

for $n \in \mathbb{N}$, one can show that

$$\nabla^q m_t = q!\, a_q$$

If $s_t = 0$, Equation (2.5) can be rewritten as

$$\nabla^q X_t = q!\, a_q + \nabla^q Y_t.$$

Since Y_t is stationary, it is also true that $\nabla^q Y_t$ is stationary. Then $\nabla^q X_t$ is a mean $q!\, a_q$ process without a trend component, making it stationary.

Similarly, the seasonal component s_t with period δ is dealt with by introducing the *lag-δ differencing operator* ∇_δ defined by

$$\nabla_\delta X_t := X_t - X_{t-\delta} = (1 - B^\delta)X_t$$

Applying the operator on the decomposition model, together with the seasonality condition $s_t - s_{t-\delta} = 0$, this yields

$$\nabla_\delta X_t = \nabla_\delta m_t + \nabla_\delta Y_t.$$

Since Y_t is stationary, it follows that $\nabla_\delta Y_t$ is stationary. The term $\nabla_\delta m_t$ is eliminated using the regular difference operator as previously mentioned [21].

Assume that both the trend and the seasonal component have been eliminated using ∇^N and ∇_δ^M respectively. Then there are coefficients b_k, or so called best linear predictors, and a mean zero stationary process $\tilde{Y}_t$ such that

$$\tilde{Y}_t = \nabla^N \nabla_\delta^M X_t = (1 - B)^N (1 - B^\delta)^M X_t = \sum_{k=0}^{N+M\delta} b_k B^k X_t = \sum_{k=0}^{N+M\delta} b_k X_{t-k}, \quad \forall t \in \mathbb{Z},$$

where $b_0 = 1$. Specifically this implies that

$$X_{t+h} = \tilde{Y}_{t+h} - \sum_{k=1}^{N+M\delta} b_k X_{t+h-k}, \quad h \in \mathbb{Z}$$

Best linear predictors are defined as follows.

Definition 3. For a time series with $\mathrm{Var}(X_t) < \infty$, $\forall t \in \mathbb{Z}$ and a collection X^n of random variables of the time series at time n different times such that $X^n := (X_{t_1}, \ldots, X_{t_n})$, the function $b_t^l(X^n)$ is called a *best linear predictor* of X_t, $t \in \mathbb{Z}$ if it minimizes the mean squared error

$$b_t^l(X^n) := \underset{g(X^n)}{\arg\min}\, \mathrm{MSE}(g(X^n), X_t) = \underset{g(X^n)}{\arg\min}\, \mathbb{E}((g(X^n) - X_t)^2),$$

over all affine functions $g(X^n) := a_0 + a_1 X_{t_n} + a_2 X_{t_{n-1}} + \cdots + a_n X_{t_1}, \quad a_0, \ldots, a_n \in \mathbb{R}$. The function is a linear operator and it therefore also has the property

$$b_t^l(cX^n + dY^n) = cb_t^l(X^n) + db_t^l(Y^n), \tag{2.8}$$

for two arbitrary collections of random variables X^n and Y^n.

For a general time series, forecasts are made by recursively calculating the best linear predictors forward in time. Suppose that X is observed at times $-N - M\delta + 1, \ldots, n, \quad n \in \mathbb{N}$ and that $\tilde{Y}$ therefore is observed at times $1, \ldots, n$, such that $X^{n+N+M\delta} = (X_{-N-M\delta+1}, \ldots, X_n)$ and $\tilde{Y}^n = (\tilde{Y}_1, \ldots, \tilde{Y}_n)$. The assumption that the initial observations $X_{N-M\delta+1}, \ldots, X_0$ are uncorrelated with $\tilde{Y}^n$, yields that

$$b_{n+h}^l(X^{n+N+M\delta}) = b_{n+h}^l(\tilde{Y}^n) - \sum_{k=1}^{N+M\delta} b_k b_{n+h-k}^l(X^{n+N+M\delta}). \tag{2.9}$$

Linear predictors of stationary time series such as $b_{n+h}^l(\tilde{Y}^n)$ are easily computed. What remains is a formula that allows for recursive calculations of X^{n+h} by computing forward $b_{n+1}^l(X^{n+N+M\delta})$, then $b_{n+2}^l(X^{n+N+M\delta})$ and so forth. For entries with $h \leq k$, it holds that $b_{n+h-k}^l(X^{n+N+M\delta}) = X_{n+h-k}$ [21] [22].

2.4 Time Series Models for Forecasting

The history of time series forecasting and its development is approximately a century long. Starting in the early 1920s, Yule and Walker established the framework for autoregressive models, from which the traditional linear family of models was developed into the 1970s. Some extensions to the linear model family were made to also capture seasonality, trends and exogenous variables [23][24]. Combining the theory with Bayesian methods allows for the inclusion of prior knowledge and yields prediction distributions rather than point estimates, which better quantifies uncertainty [25].

Machine learning methods emerged in the 1990s but were at that time computationally limited. While traditional forecasting models excel in finding linear patterns within small datasets, machine learning and deep learning approaches tend to perform better on larger, complex and non-linear datasets. This also explains why machine learning methods have not been strong forecasting contenders until recently, as the available amount of data has been insufficient. Despite the growing digitalization and data collection, the traditional forecasting methods remain relevant to this day. These models have strong mathematical foundations as well as high interpretability in the sense that the resulting regression coefficients have direct statistical meaning [26]. Some of the most recent models therefore incorporate both methodologies in hybrid approaches [27].

2.4.1 Classical Time Series Models

In this section, the ARIMA-family of classical time series models is discussed. They are all expressed as linear combinations of previous values and error terms, and get increasingly intricate as trends, seasonality and exogenous variables are included. The book *Time Series: Theory and Methods* by P. J. Brockwell and R. A. Davis [21], serves as the main reference throughout this section.

The two most elemental classical models are the autoregression and moving average processes, and are subsequently defined as the linear combination of previous values or previous error terms in the following way.

Definition 4. A *p-order autoregression* or AR(p) *process* is a stationary time series $\{X_t\}$ satisfying the equations

$$X_t = Z_t + \sum_{i=1}^{p} \phi_i X_{t-i}, \quad \forall t \in \mathbb{Z} \tag{2.10}$$

where $\{Z_t\} \sim \mathrm{WN}(0, \sigma^2)$, and Z_t is uncorrelated with X_s for all $s < t$.

Definition 5. A *q-order moving average* or MA(q) *process* is a stationary time series $\{X_t\}$ satisfying

$$X_t = Z_t + \sum_{j=1}^{q} \theta_j Z_{t-j}, \quad \forall t \in \mathbb{Z}, \tag{2.11}$$

with $\{Z_t\} \sim \mathrm{WN}(0, \sigma^2)$, and Z_t is uncorrelated with X_s for all $s < t$.

Combining the AR(p) and MA(q) processes yields the following model.

Definition 6. The process $\{X_t\}$ is an *autoregressive-moving average* (ARMA(p, q)) process if it satisfies

$$X_t - \sum_{i=1}^{p} \phi_i X_{t-i} = Z_t + \sum_{j=1}^{q} \theta_j Z_{t-j}, \quad \forall t \in Z, \tag{2.12}$$

where $\{Z_t\} \sim \mathrm{WN}(0, \sigma^2)$.

Equation (2.12) is often simplified as

$$\phi(B)X_t = \theta(B)Z_t, \quad \forall t \in Z, \tag{2.13}$$

with polynomials

$$\begin{aligned} \phi(z) &= 1 - \phi_1 z - \cdots - \phi_p z^p, \\ \theta(z) &= 1 + \theta_1 z + \cdots + \theta_q z^q, \end{aligned} \tag{2.14}$$

and the backward shift operator B defined in Equations (2.6, 2.7). The polynomials $\phi(z)$ and $\theta(z)$ do not share any roots in an ARMA(p, q) process.

As previously stated in section 2.3, elimination of trend and seasonality from a non-stationary time series can be done through differencing. The difference operator $\nabla = 1 - B$ is applied one or more times to remove the trend component. Similarly, the seasonal component is deleted through applying the lag-s differencing operator $\nabla_s = 1 - B^s$. Assuming that the underlying differenced stationary series is an ARMA(p, q) process, the approach can be formalized as ARIMA and SARIMA processes where trend and seasonality are considered.

Definition 7. An ARMA(p, q) process defined by $\phi(B)X_t = \theta(B)Z_t$ is *causal* if there exists a sequence of constants $\{\psi_j\}$ such that

(i) $X_t = \sum_{j=0}^{\infty}, \quad \forall t \in \mathbb{Z},$

(ii) $\sum_{j=0}^{\infty} |\psi_j| < \infty$

Definition 8. For a non-negative integer d, $\{X_t\}$ is an *autoregressive-integrated moving average* (ARIMA(p, d, q)) *process* if

$$\phi(B)(1 - B)^d X_t = \theta(B)Z_t, \quad \{Z_t\} \sim \text{WN}(0, \sigma^2), \quad \forall t \in \mathbb{Z}. \tag{2.15}$$

or equivalently that the process $Y_t := (1 - B)^d X_t = \nabla^d X_t$ is a causal ARMA(p, q) process.

Definition 9. If d and D are non-negative integers, then $\{X_t\}$ is said to be a SARIMA$((p, d, q) \times (P, D, Q)_s)$ *process* with period s if the already differenced process $Y_t := (1 - B)^d (1 - B^s)^D X_t$ is a causal ARMA process

$$\phi(B)\Phi(B^s)Y_t = \theta(B)\Theta(B^s)Z_t, \quad \{Z_t\} \sim \text{WN}(0, \sigma^2), \quad \forall t \in \mathbb{Z}, \tag{2.16}$$

with $\phi(z)$ and $\theta(z)$ as defined in Equation (2.14), as well as the polynomials $\Phi(B^s) = 1 - \Phi_1 z - \cdots - \Phi_P z^P$, $\quad \Theta(B^s) = 1 + \Theta_1 z + \cdots + \Theta_Q z^Q$

[21].

Lastly, exogenous variables are added in the SARIMAX model.

Definition 10. Assuming all conditions in Definition 9 hold and that $\mathbf{x_t^T}$ is a vector of exogenous variables at time t with corresponding coefficients $\boldsymbol{\beta}$, the following time series $\{Y_t\}$ is a SARIMAX-*process*

$$\begin{gathered} Y_t = \boldsymbol{\beta}\mathbf{x_t^T} + u_t, \\ \phi(B)\Phi(B^s)(1-B)^d(1-B^s)^D u_t = \theta(B)\Theta(B^s)Z_t, \quad \{Z_t\} \sim \mathrm{WN}(0,\sigma^2), \quad \forall t \in \mathbb{Z}. \end{gathered} \tag{2.17}$$

Note that the goal variable and the exogenous features are both measured at time t. This is not a problem when fitting the model in-sample, but introduces additional uncertainty in forecasts since predicting Y_{t+1} requires the knowledge of x_{t+1}. It is therefore advantageous to add delays to variables that are not clearly contemporaneous, to avoid unnecessary uncertainty [28].

2.4.2 Bayesian Estimates of SARIMAX Coefficients

When estimating the parameters of a SARIMAX model (see Section 2.4.1), the coefficients $\boldsymbol{\beta}$, ϕ, Φ, θ, Θ and the variance σ^2 are traditionally obtained using maximum likelihood estimation (MLE). In this framework, parameters are treated as fixed but unknown, and estimation is performed by maximizing the likelihood function.

In the Bayesian setting, the same parameters are instead treated as random variables with specified prior distributions. Inference is based on the posterior distribution, which combines the likelihood implied by the SARIMAX model with prior beliefs about the parameters, as introduced in Section 2.2.

This approach requires the specification of prior distributions for all model parameters. These priors may reflect domain knowledge, previous studies, or be chosen to be weakly informative. The resulting posterior distribution does not generally admit a closed-form solution, as the marginal likelihood is analytically intractable for SARIMAX models.

Consequently, posterior inference is typically performed using sampling-based methods. In this study, the No-U-Turn Sampler (NUTS), an extension of Hamiltonian Monte Carlo (HMC), is employed. These methods leverage gradient information to efficiently explore the posterior distribution and avoid the random-walk behavior associated with traditional Markov Chain Monte Carlo (MCMC) methods. NUTS further improves efficiency by adaptively selecting the trajectory length [29]. Under mild conditions, the resulting Markov chain is ergodic, ensuring convergence to the true posterior distribution as the number of samples increases [30].

A key advantage of Bayesian estimation in this context is that uncertainty in the parameters is naturally quantified through posterior distributions rather than point estimates. This is particularly beneficial in settings with limited or noisy data, where prior information can help stabilize inference. However, the results may be

sensitive to the choice of prior, and overly informative priors can introduce bias if not carefully justified.

2.5 Machine Learning Models for Forecasting

Research around optimal machine learning approaches for forecasting is vast and ongoing, especially in the context of economics and portfolio optimization. Consequently, there exists a wide variety of machine learning models with increasing complexity and specific characteristics [23]. Since the focus of this report is on investigating the suitability of machine learning methods in general, the scope of models is narrowed down. In this section, two forecasting approaches are explained in depth: XGBoost and DeepAR. While the tree-based XGBoost model works as a strong baseline [26], the deep learning model DeepAR represents a more sophisticated approach designed for probabilistic forecasting [31]. Together, they enable a meaningful comparison between two fundamentally different paradigms.

2.5.1 Tree-based Machine Learning and XGBoost

The regression in Extreme Gradient Boosting, often called XGBoost, is based on an ensemble of Classification and Regression Trees (CARTs) that sort data points into leaves based on their feature values $\boldsymbol{x}_i,\ i = 1, \ldots, n$, and give scores accordingly. Predicted data points $\hat{y}_i$ are estimated as the sum over all trees $k = 1, \ldots, K$

$$\hat{y}_i = \sum_{k=1}^{K} f_k(\boldsymbol{x}_i), \quad f_k \in \mathcal{F},$$

where f_k is the function corresponding to the structure of the k-th tree and its score, and $\mathcal{F}$ is the set of all possible decision trees. For a specific tree f_k with leaf count T_k, we can reformulate f_k in terms of function $q_k : \mathbb{R}^p \to \{1, \ldots, T_k\}$ mapping each data point to a leaf in the k-th tree, and vector $w_k \in \mathbb{R}^{T_k}$ of the tree's leaf weights so that $f_k(\boldsymbol{x}_i) = w_{k,q_k(\boldsymbol{x}_i)}$. In comparison to classical time series models where coefficients of lagged terms and exogenous features are optimally estimated through linear regression, the objective of XGBoost is to find structures q_k and w_k that minimize

$$\text{obj} = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k), \tag{2.18}$$

with loss function $l(y_i, \hat{y}_i)$ and tree complexity $\Omega(f_k)$. In most situations it is intractable to learn all trees simultaneously. Predictions are estimated progressively with additive training instead

$$\hat{y}_i^{(t)} = \sum_{k=1}^{t} f_k(\boldsymbol{x}_i) = \hat{y}_i^{(t-1)} + f_t(\boldsymbol{x}_i), \tag{2.19}$$

where the next tree selected for iteration t best optimizes the objective function (2.18) up to $k = t$. This reduces the problem to finding a single tree f_t for the current iteration, rather than optimizing all K trees jointly [32].

Loss functions vary in complexity. In general, the Taylor expansion of the loss function up to the second order is utilized. Removing training errors from previous iterations, the objective for finding tree f_t becomes

$$\text{obj}(f_t) = \sum_{i=1}^{n} \left[g_i f_t(\boldsymbol{x}_i) + \frac{1}{2} h_i f_t^2(\boldsymbol{x}_i) \right] + \Omega(f_t), \tag{2.20}$$

$$g_i = \partial_{\hat{y}_i^{(t-1)}} \, l(y_i, \hat{y}_i^{(t-1)}),$$
$$h_i = \partial^2_{\hat{y}_i^{(t-1)}} \, l(y_i, \hat{y}_i^{(t-1)}).$$

For XGBoost, the complexity term is defined as

$$\Omega(f_k) = \gamma T_k + \frac{1}{2} \lambda \sum_{j=1}^{T_k} w_{kj}^2,$$

with regularization hyperparameters γ, λ. Equation (2.20) together with the additive training in Equation (2.19) introduces gradient boosting and makes XGBoost a machine learning method. If adding a new tree f_t does not sufficiently reduce the loss to negate its regularization cost, the objective (2.20) does not improve and training stops early.

Using the set of indices $I_j = \{i | q(x_i) = j\}$ of data points assigned to the j-th leaf, objective (2.20) can be compressed to

$$\text{obj}(f_t) = \sum_{j=1}^{T} \left(G_j w_j + \frac{1}{2} (H_j + \lambda) w_j^2 \right) + \gamma T,$$
$$G_j = \sum_{i \in I_j} g_i, \quad H_j = \sum_{i \in I_j} h_i.$$

Also, the loss function gain attributed to a feature x is defined as

$$\text{Gain(x)} = \frac{1}{2} \left[\frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{(G_L + G_R)^2}{H_L + H_R + \lambda} \right] - \gamma,$$

with G_L, G_R, H_L, H_R representing respective sums of indices assigned to the left and right leaves from j. Averaging this gain across all nodes creates a measure of feature importance [32].

Note that XGBoost is not inherently a time series model. The time series structure can be arranged by manually adding lagged features to the model and forecasting iteratively.

2.5.2 Deep Learning Fundamentals and DeepAR

The Deep Autoregressive (DeepAR) model was developed in the late 2010s by Amazon Corporation and is an extended Long-Short Term Memory (LSTM) model utilizing the structure of time series and autoregression. DeepAR stands out among other ML-based forecasting approaches for its ability to capture global patterns among several individual time series, and for its property of providing probabilistic prediction intervals instead of single values [31]. Before discussing the specifics of DeepAR and its implications, an introduction to Multi-Layer Perceptrons (MLP) and Recurrent Neural Networks (RNN) is needed.

MLPs, also called a Feedforward Neural Network (FNN), count as the most fundamental deep learning models. They essentially define a mapping $y = f(x; \boldsymbol{\theta})$ and learn the parameter values θ that yield the best function approximation. The map in its most basic form consists of a directed acyclic graph of functions creating a network of layers. For example, the chain $f(x) = f^{(L)}(f^{(L-1)}(\cdots f^{(1)}(x)))$ builds a network of depth L, where $f^{(1)}$ is considered the first layer and $f^{(L)}$ the output layer. Between the first and the last layer, the network's objective is to learn how to best use all layers to reproduce the output y. Since these layers do not show the desired output, they are called hidden layers. All hidden layers consist of vectors of neurons that work in parallel. The number of units per layer constitutes the network's width. Each neuron is assigned a weight transformation depending on the input importance, which in turn is propagated forward in the network through a typically non-linear activation function. A FNN does not contain recurrent feedback connections. Instead, networks with that property are called RNNs [33].

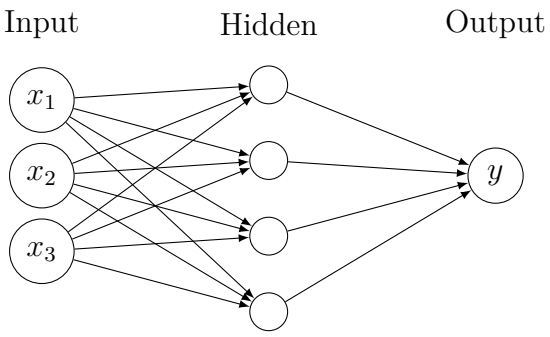

Figure 2.2: Illustration of a feedforward network with depth $L = 3$.

Unlike MLPs, which define static functions without internal state dynamics, RNNs can be viewed as dynamical systems with an internal state that evolves over time [34]. As with MLPs, a RNN consists of an input layer and one or more hidden layers; however, the hidden layers are recursively connected across time steps. In a strict sense, a RNN can be formulated without an explicit output layer, although an output layer is required in practical applications to produce predictions.

A key distinction between MLPs and RNNs is the presence of recurrent connections in the latter, whereby the hidden state at one time step influences the subsequent

state. This can be expressed as $h^{(t)} = f(h^{(t-1)}, x^{(t)}; \theta)$, where $x^{(t)}$ denotes the input at time step t, $h^{(t)}$ the corresponding hidden state, and θ the parameters of the function f [31]. This structure implies that the same parameters are reused across time, which is particularly suitable for modeling sequential data. However, the recursive nature of the network may give rise to vanishing or exploding gradients when trained using gradient-based methods. To mitigate these issues, more advanced recurrent architectures, such as LSTM networks, have been developed. These introduce gating mechanisms that regulate the flow of information and enable the retention of relevant information over longer time horizons.

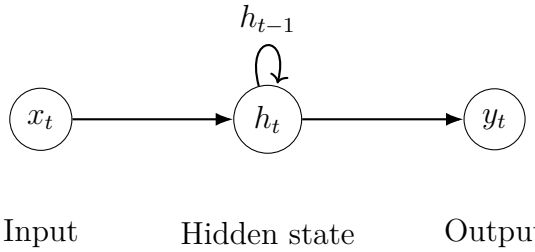

Figure 2.3: Illustration of a recurrent neural network

Given a time series $Z_{i,t}$ for item i and time step t the objective of DeepAR is to model the conditional probability distribution

$$P(\mathbf{Z}_{i,t_0:T} \mid \mathbf{Z}_{i,1:t_0-1}, \mathbf{X}_{i,1:T}),$$

where $\mathbf{Z}_{i,1:t_0-1}$ denotes the observed conditioning range, $\mathbf{Z}_{i,t_0:T}$ the prediction range and $\mathbf{X}_{i,1:T}$ are covariates known for all time steps. Using the hidden states of the recurrent network $\mathbf{h}_{i,t} = h\left(\mathbf{h}_{i,t-1}, Z_{i,t-1}, \mathbf{X}_{i,t}\right)$ the parameters of the predictive distribution are computed as $\theta_t = \theta\left(\mathbf{h}_{i,t}, \Theta\right)$ where Θ denotes the model parameters.

The conditional distribution is factorized autoregressively as

$$\begin{aligned} P\left(\mathbf{Z}_{i,t_0:T}\right) &= \prod_{t=t_0}^{T} P\left(Z_{i,t} \mid \mathbf{Z}_{i,1:t-1}, \mathbf{X}_{i,1:T}\right) \\ &= \prod_{t=t_0}^{T} P\left(Z_{i,t} \mid \theta\left(\mathbf{h}_{i,t}, \Theta\right)\right) \\ &= \prod_{t=t_0}^{T} P\left(Z_{i,t} \mid \theta_t\right). \end{aligned}$$

The model is trained by maximizing the likelihood, i.e.

$$\mathcal{L} = \sum_{i=1}^{N} \sum_{t=t_0}^{T} \log\left(P\left(Z_{i,t} \mid \theta_t\right)\right).$$

Forecasting is performed autoregressively. Historical observations are first fed into the network to initialize the hidden state. Future values are then generated sequentially by sampling from the predicted probability distribution at each time step and

feeding the sampled value back into the network as input for the next prediction step. Repeating this process multiple times yields Monte Carlo sample trajectories that approximate the full predictive distribution.

A major advantage of DeepAR is its ability to jointly learn temporal patterns across many related time series. This reduces the amount of manual feature engineering compared to traditional forecasting approaches and enables the model to generalize to products or entities with limited historical data. Furthermore, the probabilistic formulation provides uncertainty estimates through prediction intervals and quantiles, which are valuable in applications such as inventory management and demand forecasting [31].

2.6 Evaluations

This section addresses the theoretical basis of selected assessment metrics, including autocorrelation, Granger causality, and negative log-likelihood estimates. A summary of more fundamental evaluation metrics is provided in Appendix A.

2.6.1 ACF and PACF

If the series $\{X_t\}$ is stationary, it holds for the autocovariance from Definition 1 that $\gamma_X(r,s) = \gamma_X(r-s,0), \quad \forall r,s \in \mathbb{Z}$. The autocovariance of a stationary series can therefore be redefined as the function of only one variable such that

$$\gamma_x(h) \equiv \gamma_X(h,0) = \mathrm{Cov}(X_{t+h}, X_t), \quad \forall t,h \in \mathbb{Z}.$$

Analogously, the following definitions are made.

Definition 11. For a stationary process $\{X_t\}$, the *autocorrelation function (ACF)* $\rho_X(h)$ is defined by

$$\rho_X(h) \equiv \frac{\gamma_X(h)}{\gamma_X(0)} = \mathrm{Corr}(X_{t+h}, X_t), \quad \forall t,h \in \mathbb{Z}.$$

Definition 12. The *partial autocorrelation function (PACF)* $\alpha_X(h)$ of a stationary time series is defined as

$$\begin{aligned} &\alpha_X(0) = 1, \\ &\alpha_X(1) = \mathrm{Corr}(X_2, X_1) = \rho_X(1), \\ &\alpha_X(h) = \delta_{hh}, \quad \forall h \geq 1, \end{aligned}$$

where δ_{hh} is the last component of

$$\delta_h = ((\gamma_X(i-j))_{i,j=1}^h)^{-1}(\gamma_X(1), \gamma_X(2), \ldots, \gamma_X(3))'$$

[21][35]. In time series modeling, ACF and PACF estimates are computed up to a point h in order to determine the ARMA parameters p, q. The PACF values suggest an appropriate parameter p for an AR(p) process, and the ACF estimates propose a value for q in a MA(q) series. In fact, it can be proven that the ACF and PACF values are associated to the mentioned parameters in accordance with Table 2.1. For a full proof, the reader is referred to the book *Time Series Analysis and Its Applications: With R Examples*, by R. H. Shumway and D. S. Stoffer [36]. It is also exemplified in Figure 2.4, where ACF and PACF values are estimated for the AR(3) process given by equation,

$$X_t = \frac{9}{10}X_{t-1} - \frac{1}{2}X_{t-2} + \frac{1}{5}X_{t-3} + Z_t, \tag{2.21}$$

and the MA(3) process given by,

$$X_t = Z_t + \frac{9}{10}Z_{t-1} - \frac{1}{2}Z_{t-2} + \frac{1}{5}Z_{t-3}, \tag{2.22}$$

for time series X_t and white noise term Z_t. For the MA(3) process there is a clear ACF cut-off after $h = 3$ in figure 2.4(a), while the ACF value of the AR(3) slowly decays. The contrary holds for the PACF values in Figure 2.4(b) [22].

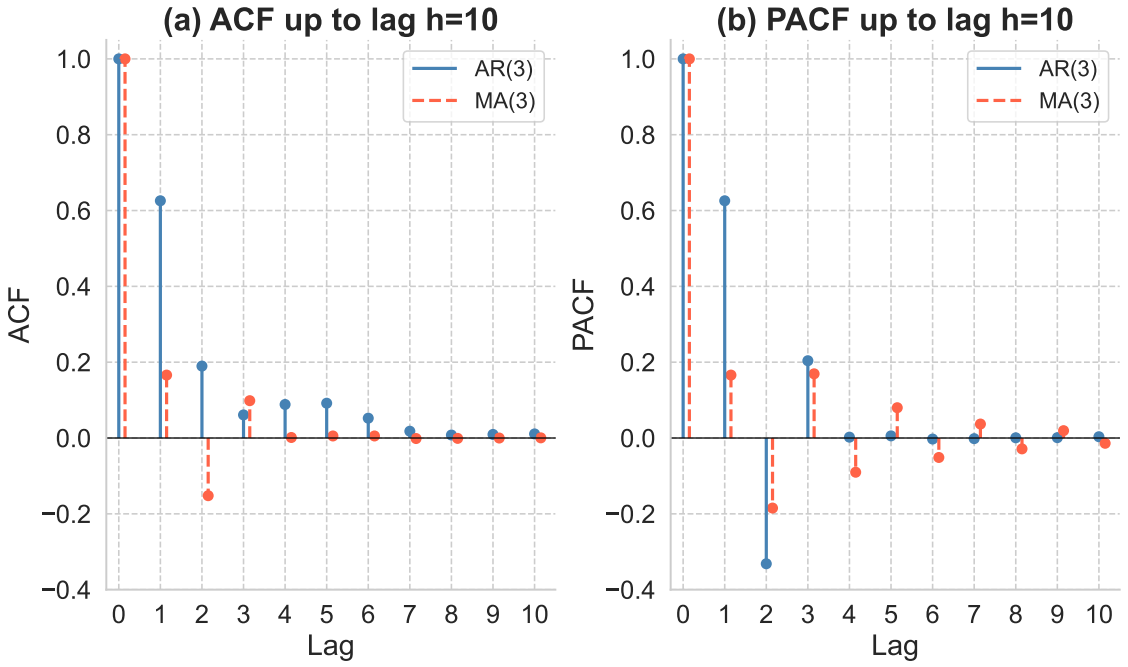

Figure 2.4: ACF and PACF values for processes given by Equations 2.21 and 2.22 up to lag $h = 10$.

	AR(p)	MA(q)	ARMA(p,q)
ACF	Tails off	Close to 0 after lag q	Tails off
PACF	Close to 0 after lag p	Tails off	Tails off

Table 2.1: Behavior of ACF and PACF estimates for AR(p), MA(q) and ARMA(p, q) models. The table is taken from the book *Time Series Analysis and Its Applications: With R Examples*, by R. H. Shumway and D. S. Stoffer [36].

2.6.2 Granger Causality

Determining whether one time series $\{X_t\}$ offers useful information in forecasting another time series $\{Y_t\}$ involves utilizing statistical hypothesis tests such as the Granger Causality test. It is said that X_t Granger causes Y_t if

$$\sigma^2(Y_t|Y_0, \ldots, Y_{t-1}) > \sigma^2(Y_t|Y_0, \ldots, Y_{t-1}, X_0, \ldots, X_{t-1}),$$

with $\sigma^2(Y|\mathcal{I})$ representing the variance of the optimal linear forecast of Y_t, given the information set $\mathcal{I}$. Feedback is said to occur if the causality holds in both directions such that Y_t Granger causes X_t and vice versa. In practice, Granger Causality is tested using a Vector Autoregression (VAR). Considering a bivariate system of X_t, Y_t, the VAR is formulated as

$$X_t = \sum_{h=1}^{H} a_h X_{t-h} + \sum_{h=1}^{H} b_h Y_{t-h} + \epsilon'_t,$$
$$Y_t = \sum_{h=1}^{H} c_h X_{t-h} + \sum_{h=1}^{H} d_h Y_{t-h} + \epsilon''_t,$$

with maximum lag H and white noise error terms $\epsilon'_t, \epsilon''_t$. Then X_t Granger causes Y_t if and only if not all $c_h = 0$. This is tested with a standard F-Test or Wald test on the null hypothesis $H_0 : c_1 = c_2 = \cdots = c_H = 0$. The tests assume that only stationary series are involved. In the non-stationary case, the variance $\sigma(Y_t|\mathcal{I})$ is time-dependent and the causality may alter over time [37].

2.6.3 Negative Log-Likelihood Estimates

Given a set of n data points $\mathbb{X} = \{x^{(1)}, \ldots, x^{(n)}\}$ drawn from the true but unknown generating distribution $p_{\text{data}}(\boldsymbol{x})$, the modeling principle in MLE is to maximize the likelihood of $p_{\text{model}}(\boldsymbol{x}; \boldsymbol{\theta})$ being the underlying distribution. In the context of machine

learning and software in general, MLE is often reformulated as finding the optimal model parameters $\boldsymbol{\theta}$ to the loss minimization problem

$$\boldsymbol{\theta}_{NLL} = \arg\min_{\boldsymbol{\theta}} \left(-\sum_{i=1}^{n} \log p_{\text{model}}(\boldsymbol{x}^{(i)}; \boldsymbol{\theta}) \right),$$

where the negative sum is called the *Negative Log-likelihood* (NLL). Although raw NLL-values do not carry any universal meaning, they are used as a comparative metric between models and losses within models [33].

3

Methods

The data workflow of the thesis is discussed in detail throughout this chapter, serving as the link between the theoretical framework established earlier and the empirical results presented in subsequent chapters. It outlines a complete methodology, covering the pre-processing and preparatory analysis of the data, the approach to parameter tuning and model validation, as well as the metrics selected for visual evaluation. A dedicated section addresses the choice of models, including their library configurations and parameter settings. It should be noted that the entire data pipeline was implemented using Python 3.14.3 and the latest releases of all relevant model packages.

3.1 Data Pipeline

As described in Section 2.1, forecasts were developed for two datasets: D_Y with yearly observations from 1996-2025, and D_M, with monthly observations from 2012-2025. A summary of the goal variable and exogenous features for each dataset is given in Table 3.1. The forecasting horizons differed accordingly, with 3-15 years ahead for D_Y and 12-60 months for D_M. Therefore, exogenous features were also projected to 2040 and 2030 respectively, as outlined in Section 2.1. Finally, previous GENAB forecasts of the goal variables were stored in both datasets for future comparisons. The two datasets were included to investigate how the chosen models respond to differences in data frequency and forecasting horizon.

It is important to note that the forecasts produced in this thesis, both for the target variables and the exogenous variables, are entirely data-driven and based on historical observations. Consequently, the generated forecasts primarily reflect the behavior and structures observed within the historical datasets used during model training.

Some preprocessing of the dataset was required in order to address missing observations and ensure consistency across variables. In addition to the calculations of EVs per capita and real salaries described in Section 2.1, several datasets contained incomplete historical observations that required estimation. For example, the monthly salary dataset contained observations between 1995 and 2019. Given the relatively

Dataset D_Y	Yearly data, 1996-2025 (30 datapoints)
Features	Aggregated total network consumption [GWh]
	Population in Gothenburg
	Average real salary in Sweden [kr]
	Energy efficiency
	Electric vehicles per capita in Sweden

(a) Frequency, time horizon and features of Dataset D_Y.

Dataset D_M	Monthly data, 2012-2025 (168 datapoints)
Features	Highest network power usage [MW]
	Average coldest five air temperatures [$^\circ C$]
	Population in Gothenburg
	Average real salary in Sweden [kr]
	Energy efficiency
	Electric vehicles per capita in Gothenburg

(b) Frequency, time horizon and features of Dataset D_M.

Table 3.1: Summary of datasets, with goal variables highlighted in gray.

stable and near-linear salary growth pattern in Sweden, linear regression was used to estimate average monthly salaries for 2020-2025. Since the EV data was recorded on a monthly basis, the number of EVs registered in December of each year was used to represent the yearly value in the annual dataset. The EV observations were first recorded in 2013. Missing observations prior to 2013 were set to zero for D_Y, and was completed through linear regression in D_M due to its shorter interval of missing data. The maximum hourly power peak dataset, which is the base for the monthly maximum power, was largely complete. It was missing a few observations that were filled in using adjacent averages and the maximum value was selected to represent the monthly power peak. Data of efficiency rate was constructed both on a monthly and a yearly basis following the description in Section 2.1, and missing temperature observations were supplemented using nearby stations or interpolation between adjacent values.

In order to validate the forecasts, the historic data was split into training and validation data. A common rule of thumb for forecast validation is to match the length of the validation split to the forecast horizon. For instance, if the goal is to forecast 5 years ahead, the validation set should span 5 years. The process of splitting the data consequently differed depending on the forecast horizon and dataset. A simple hold-out split was used for longer forecasts, due to lack of older data. For shorter forecasts, a rolling cross-validation (CV) split was utilized instead. Figure 3.1 visualizes this principle in the case where D_Y is used to forecast 15 years and 3 years ahead respectively. In the context of this thesis, the forecast horizons of D_Y were $T_Y^{\text{forecast}} \in [3, 5, 10, 15]$ years, with rolling CV split being used for $T_Y^{\text{forecast}} = 3$

and hold-out split being utilized for the remaining prediction windows. Similarly, for the selected forecast horizons $T_M^{\text{forecast}} \in [12, 24, 36, 60]$ of D_M, rolling CV splits were applied for the smallest forecast period $T_M^{\text{forecast}} = 12$ and a hold-out split was implemented for the rest.

Due to time series' temporal ordering, model training and validation must preserve chronological structure, ensuring that information from future observations is never used to inform past estimates. This requirement motivates the use of sequential validation schemes, such as rolling cross-validation, illustrated in Figure 3.1. Consequently, whereas non-temporal datasets with independent observations allow for a wide range of train–test partitions, data splitting strategies for time series is limited by temporal dependency.

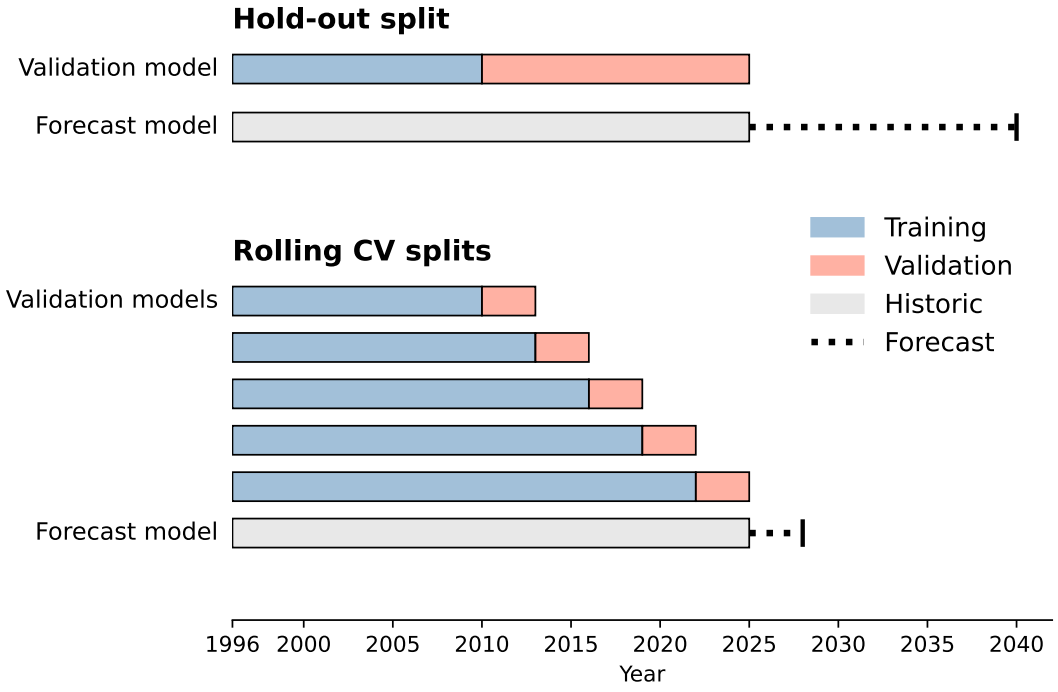

Figure 3.1: Example of appropriate situations for hold-out and rolling CV splits.

Preparatory analysis was conducted on the training part of both datasets. Pearson correlation coefficients were calculated between the model features to identify potential multicollinearity issues. Granger causality tests were conducted between each exogenous feature and the goal variable to assess their predictive relevance. Differencing as well as ACF and PACF computations were also performed on the goal variables, to guide the selection of SARIMA-parameters. ACF and PACF estimates are shown as part of the method, while Pearson correlations and Granger causality tests are displayed in the first part of the results in Chapter 4.

Finally, all selected models were estimated on both the yearly and the monthly data across their respective forecast periods, directly addressing objectives 2 and 3 from

Section 1.2. Point predictions over each forecast window were stored and visualized, and for probabilistic models, full predictive distributions were presented instead. To support model evaluation, fitted values on the training and validation sets were retained, from which residuals as well as summary error metrics RMSE and MAE were computed. A brief theoretical treatment of these metrics is provided in Appendix A. In cases where rolling cross-validation was used, errors were aggregated across splits to characterize the stability and variance of model performance. Lastly, to approach objectives 4 and 5, the contribution of exogenous features was summarized for each model, offering insight into which variables drove the forecasts and how well machine learning methods lend themselves to long-term capacity forecasting.

3.2 Details of Model Selection

In total, six strategically diverse forecasting models were implemented in this thesis. They allowed for a broad methodological comparison while maintaining a manageable scope. Four models were selected from the traditional time series family of methods: two baseline AR and SARIMA models along with two SARIMAX models, where one optimized coefficient values with MLE and the other through Bayesian estimates. The AR baseline model is used for the yearly aggregated dataset D_Y, while the seasonal D_M utilizes SARIMA as a baseline. To include both the tree-based and the deep learning paradigms of machine learning modelling, XGBoost and DeepAR were added. A brief overview of model packages and parameters is provided in Table 3.2. What follows describes the technical specifics of each model beyond what is summarized in the table.

SARIMA-family parameters are often decided by tuning a large amount of combinations of $(p, d, q,) \times (P, D, Q)_s$ down to one, through AIC optimization. This method rapidly gets complex and time-consuming to compute, particularly in models that in addition rely heavily on sampling, such as the Bayesian SARIMAX model. All SARIMA-family parameters were therefore instead guided by differencing as well as ACF and PACF values in this thesis. Figure 3.2 presents the ACF and PACF values for the goal variables of both datasets. It shows, in subfigures 3.2(a,c), that the PACF of the total yearly consumtion cuts off after lag $h = 1$, while the ACF gradually decreases. This motivated the choice of AR(1) for the yearly dataset. In subfigures 3.2(b,d), the lag $h = [1, 11, 12]$ peaks are slightly more pronounced for the ACF than those of the PACF, which suggested a SARIMA$(0, 1, 1) \times (0, 1, 1)_{12}$ model.

Model	Package	Parameters
AR	`statsmodels: AutoReg`	D_Y: $p = 1$
SARIMA	`statsmodels: SARIMAX`	D_M: $(p,d,q) \times (P,D,Q)_s =$ $(0,1,1) \times (0,1,1)_{12}$
SARIMAX (MLE)	`statsmodels: SARIMAX`	D_Y: $(1,1,0) \times (0,0,0)_1$ D_M: $(0,1,1) \times (0,1,1)_{12}$ $\alpha \in [0.5, 5]$
SARIMAX (Bayesian)	`pymc_extras: BayesianSARIMAX`	D_Y: $(1,1,0) \times (0,0,0)_1$ D_M: $(0,1,1) \times (0,1,1)_{12}$ Priors*: β, σ, ϕ, θ, Φ, Θ
	`pymc: sample`	Settings*: `draws`, `tune`, `target_accept`
XGBoost	`xgboost: XGBRegressor` `sklearn: TimeSeriesSplit, RandomizedSearchCV`	`max_depth` $\in [1, 5]$, `learning_rate` $\in [0.005, 0.05]$, `n_estimators` $\in [25, 200]$, `min_child_weight` $\in [1, 3]$, `subsample` $\in [0.5, 0.9]$, `colsample_bytree` $\in [0.5, 0.9]$

DeepAR	`pytorch_forecasting: DeepAR, TimeSeriesDataSet`
	`lightning.pytorch`
	`optuna`
		`learning_rate` $\in [0.001, 0.01, 0.05, 0.1]$,
		`hidden_size` $\in [16, 32, 64, 128]$,
		`rnn_layers` $\in [1, 2, 3]$,
		`dropout` $\in [0.1, 0.2, 0.3]$

Table 3.2: Overview of implemented forecasting models and their parameters. Highlighted parameters are tuned. *Priors and settings for the Bayesian SARIMAX model are listed in detail in Table 3.3.

The selection of priors for the Bayesian SARIMAX coefficients were based on previous domain knowledge as well as ACF and PACF estimations on training data, displayed in Figure 3.2. They are condensed, together with choices of sampling settings, into the comprehensive Table 3.3. For example, it is known that the goal variable and most features in dataset D_Y have increased over longer periods of time, motivating a slightly positive prior for β in the table. Similarly, the prior distribution σ for the monthly data was assigned a higher standard deviation than that of the yearly data, because monthly power peaks are known to vary more in comparison to yearly aggregated total electricity consumption. However, the selection of the priors $(\phi, \theta, \Phi, \Theta)$ were directly inspired by the ACF and PACF Figure 3.2. It

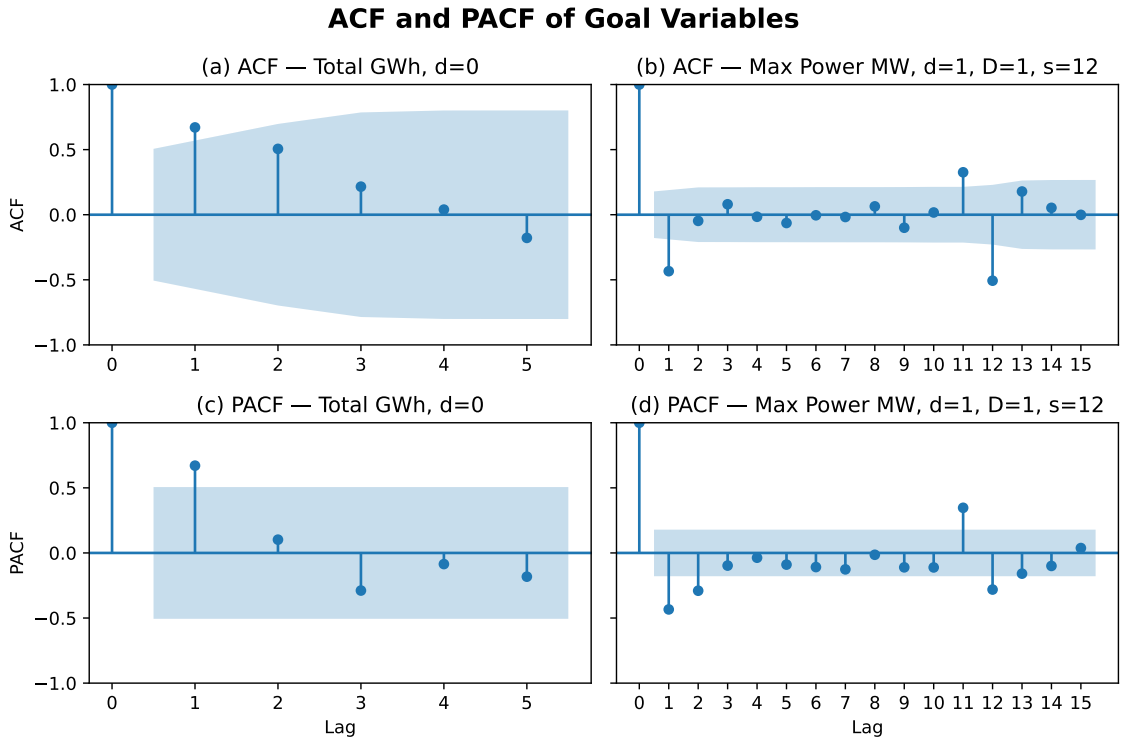

Figure 3.2: ACF and PACF values of the goal variables in D_Y and D_M up to lags of $h_Y = 5$ and $h_M = 15$ respectively. The monthly trend and seasonal components are removed through differencing.

should be noted that the prior specifications employed in this thesis are heuristic in nature and should not be interpreted as exact or authoritative. The SARIMAX model utilizes L1-regularization with $\alpha \in [0.5, 5]$ as seen in Table 3.2, in order to penalize the inclusion of insignificant features. Bayesian SARIMAX models instead use priors as a form of regularization. The amount of draws mentioned in Table 3.3 correspond to the number of samples kept for the posterior distribution, while the tunes calibrate the NUTS sampler and are discarded before sampling [38]. It is optimal to use high values of both draws and tunes to more accurately estimate posterior distributions. They were however both decreased for the SARIMAX model based on D_M, due to its inherently high complexity with added seasonality terms.

As described in subsection 2.5.1, the XGBoost model is not a native time series model. Therefore, two adaptations were made to impose temporal structure. First, lagged columns of the target variables were added as features, serving as AR-terms. Secondly, forecasts were generated forward iteratively, so that each prediction could feed into subsequent steps as lagged inputs. All XGBoost parameters stated in Table 3.2 were tuned on training data using the function `RandomizedSearchCV` from scikit-learn, which compared 100 parameter combinations on 3 time series CV splits. The parameters `max_depth`, `min_child_weight`, `subsample` and `colsample_bytree` regularize the model to avoid overfitting. In addition, the default L_2 regularization implemented in XGBoost was retained throughout the training process. Lastly, squared errors were used as the minimization objective of XGBoost regressions.

Priors and Sampler Settings	
D_Y	D_M
$\beta \sim \mathcal{N}(0.1,\ 0.25^2)$	$\beta \sim \mathcal{N}(0,\ 0.4^2)$
$\sigma \sim \mathcal{N}^+(0.05)$	$\sigma \sim \mathcal{N}^+(0.1)$
$\phi \sim \mathcal{N}(0.6,\ 0.25^2)$	$\phi \sim \mathcal{N}(0,\ 0.25^2)$
$\theta \sim \mathcal{N}(0,\ 0.25^2)$	$\theta \sim \mathcal{N}(0.75,\ 0.25^2)$
	$\Phi \sim \mathcal{N}(0,\ 0.25^2)$
	$\Theta \sim \mathcal{N}(0.75,\ 0.25^2)$
`draws = 500`	`draws = 200`
`tune = 1000`	`tune = 400`
`target_accept = 0.9`	`target_accept = 0.9`
`nuts_sampler = "blackjax"`	`nuts_sampler = "blackjax"`

Table 3.3: Priors and sampler settings for the yearly dataset D_Y and monthly dataset D_M. The priors $(\phi, \theta, \Phi, \Theta)$ correspond to the coefficents of SARIMA-parameters (p, q, P, Q) as seen in subsection 2.3. Priors for exogenous feature coefficients β are also stated, along with the prior for the standard deviation σ.

DeepAR hyperparameters were tuned using Optuna over 20 trials, with `dropout` functioning as a regularization parameter. A normal distribution loss is used for regression and NLL estimates are stored as results in place of training and testing errors. For more details on hyperparameters and their purpose, see Appendix B.

4

Results

This section will present the results and short discussions. More results and figures are found in Appendix C.

4.1 Results of the Preparatory Analysis

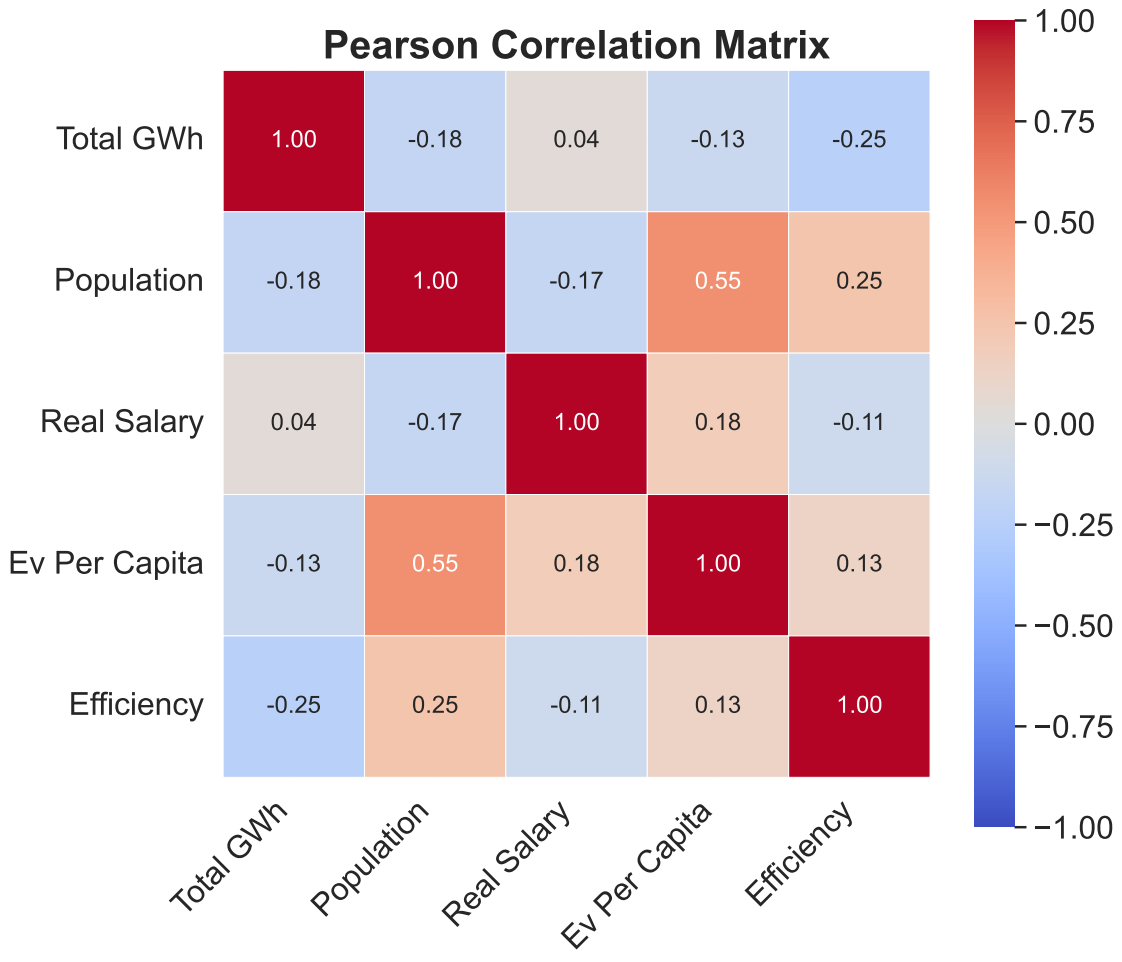

Figure 4.1: Correlation matrix of D_Y.

Feature	F-statistic	p-value	Significant
Efficiency	0.2539	0.6212	No
Population	0.169	0.6865	No
Real Salary	0.0107	0.9188	No

Table 4.1: Granger causality tests for each selected exogenous feature D_Y against its target variable. EV per capita is not included in the test since it is equal to zero for most of the training data. However, it is still included while modeling.

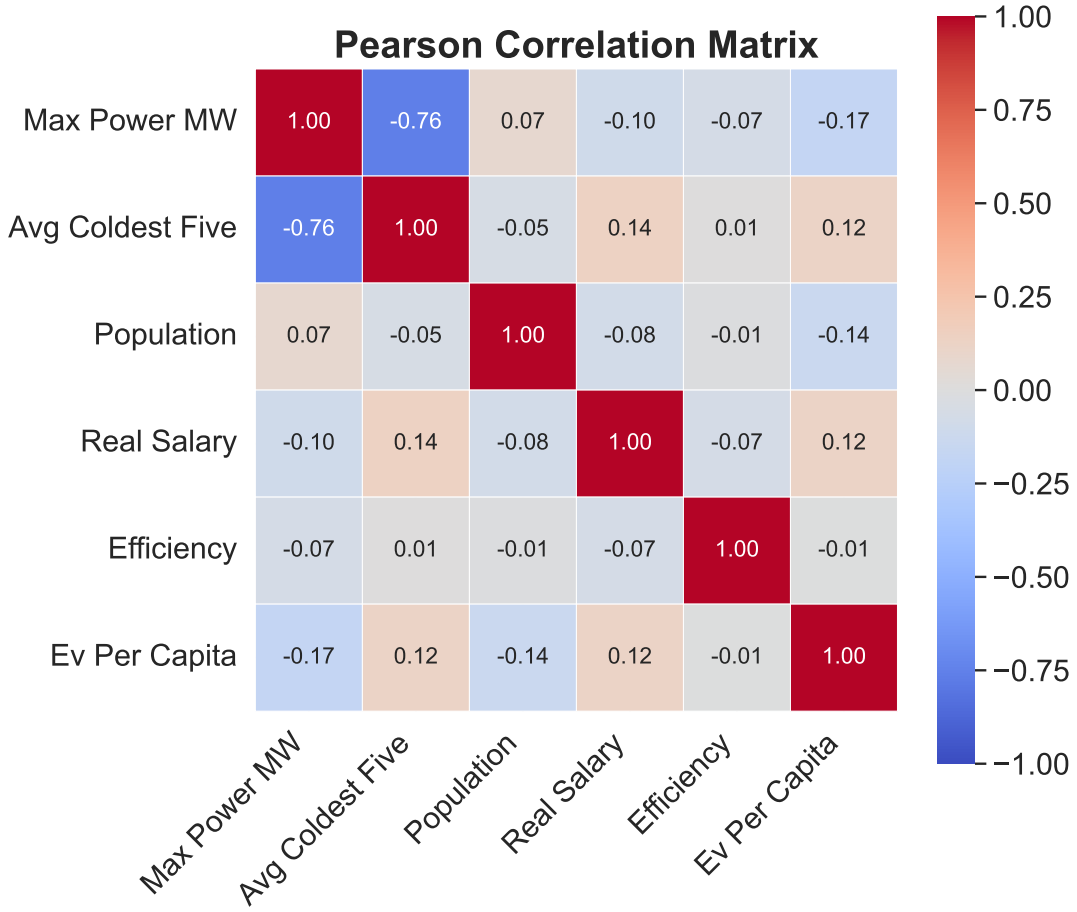

Figure 4.2: Correlation matrix of D_M.

Feature	F-statistic	p-value	Significant
Population	2.0151	0.0308	Yes
Real Salary	1.8367	0.0527	No
Avg Coldest Five	1.2457	0.2641	No
Ev Per Capita	0.7211	0.7276	No
Efficiency	0.5425	0.8816	No

Table 4.2: Separate bivariate Granger causality tests for each selected exogenous feature in dataset D_M against its target variable.

Granger causality tests and correlation matrices were produced after the data had been stationarised as described in Section 2.3.

For the annual dataset (see Figure 4.1), the variables were differenced in order to remove long-term trends before the correlation analysis was performed. After differencing, most correlations between the variables were relatively weak, suggesting that short-term changes in the variables do not move closely together. This indicates that there is no clear evidence of multicollinearity among the independent variables. The only relationship that stands out to some extent is the correlation between population and EVs per capita. This may partly be explained by the large number of zero values in the EV per capita variable during much of the training period. It may also reflect a natural dependency between EV adoption and population size.

For the monthly dataset (see Figure 4.2), the data were differenced to remove both long-term trends and seasonal effects prior to analysis. Similar to the annual dataset, most correlations remained weak after differencing, indicating that short-term fluctuations in the variables are relatively independent. No significant multicollinearity was identified among the independent variables. However, temperature showed a strong negative correlation with maximum power demand, which is reasonable given the increased need for heating during colder periods. This suggests that short-term variations in maximum power demand are primarily explained by temperature changes rather than by economic or demographic factors.

The significance of the exogenous features for each dataset is presented in Tables 4.1 and 4.2. The Granger causality test evaluates whether lagged values of an exogenous variable provide additional predictive information for a target time series. More specifically, the test examines whether one or more lagged observations of an exogenous feature contribute significantly to predicting the current value of the target variable. Although a clear relationship exists between power demand peaks and temperature in D_M, the results of the Granger causality test suggest that past temperature observations do not provide significant additional explanatory power for predicting the current power peak.

4.2 Forecast Results with Yearly Aggregated Data

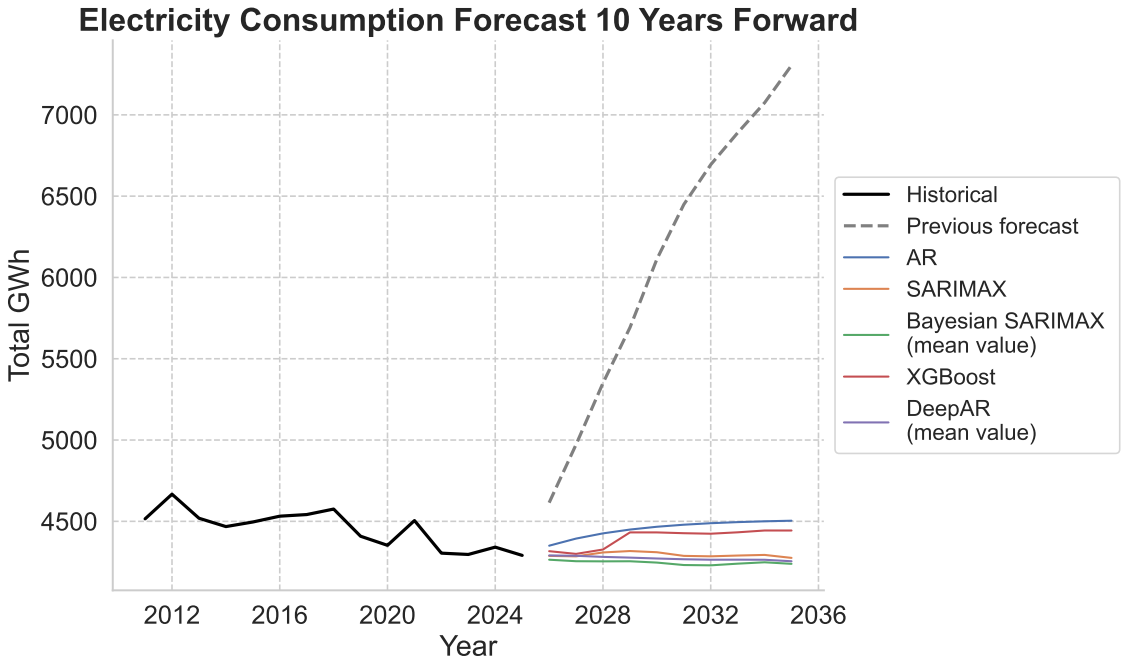

Figure 4.3: Forecast 10 years forward.

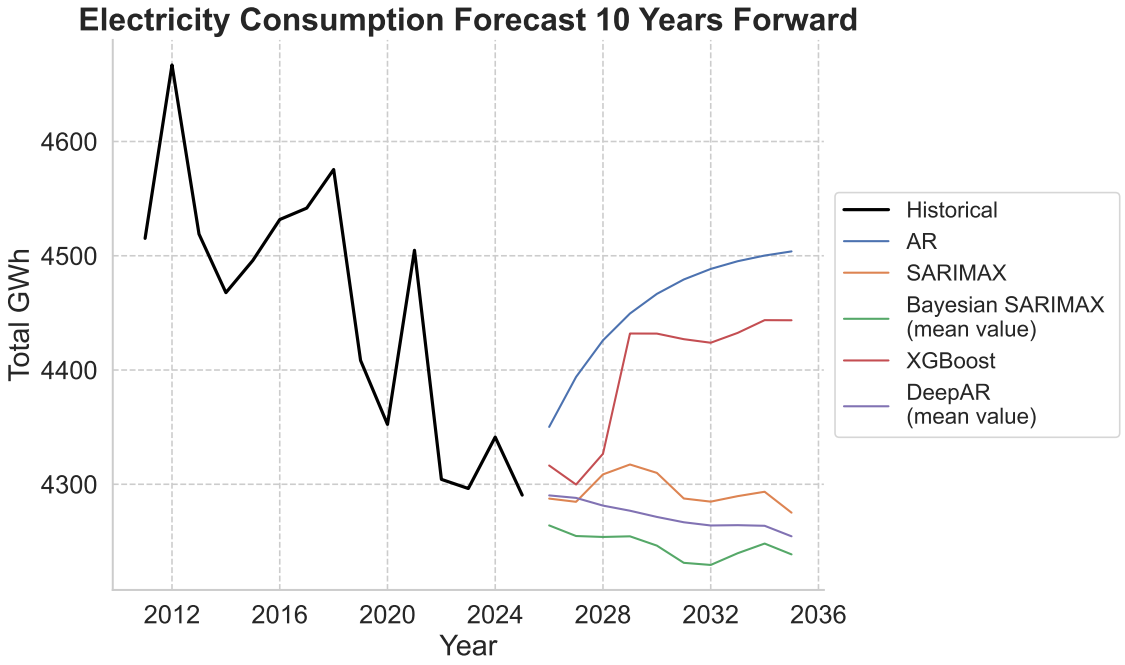

Figure 4.4: Forecast 10 years forward without GENABs' previous forecast. This makes it easier to see the forecasts of the models.

Baseline Errors, D_Y		
Forecast Length (Years)	**Training RMSE**	**Validation RMSE**
3	106.84	136.73
5	99.93	182.61
10	100.65	200.26
15	93.33	242.61

SARIMAX Errors, D_Y		
Forecast Length (Years)	**Training RMSE**	**Validation RMSE**
3	88.47	74.01
5	88.135	101.635
10	82.60	96.09
15	72.44	619.50

B-SARIMAX Errors, D_Y		
Forecast Length (Years)	**Training RMSE**	**Validation RMSE**
3	91.14	49.27
5	84.73	440.83
10	84.04	2220.58
15	83.29	482.23

XGBoost Errors, D_Y		
Forecast Length (Years)	**Training RMSE**	**Validation RMSE**
3	43.90	164.92
5	5.11	170.73
10	4.40	151.20
15	0.95	220.09

DeepAR Errors, D_Y		
Forecast Length (Years)	**Training RMSE**	**Validation RMSE**
3	—	44.08
5	—	119.89
10	—	71.01
15	—	—

Table 4.3: Training and validation RMSE for all five models on dataset D_Y, across forecast horizons of 3, 5, 10, and 15 years. DeepAR training RMSE is not reported as the model does not yield an equivalent in-sample metric. The dash for DeepAR at the 15-year validation error indicate that results were unavailable for that horizon.

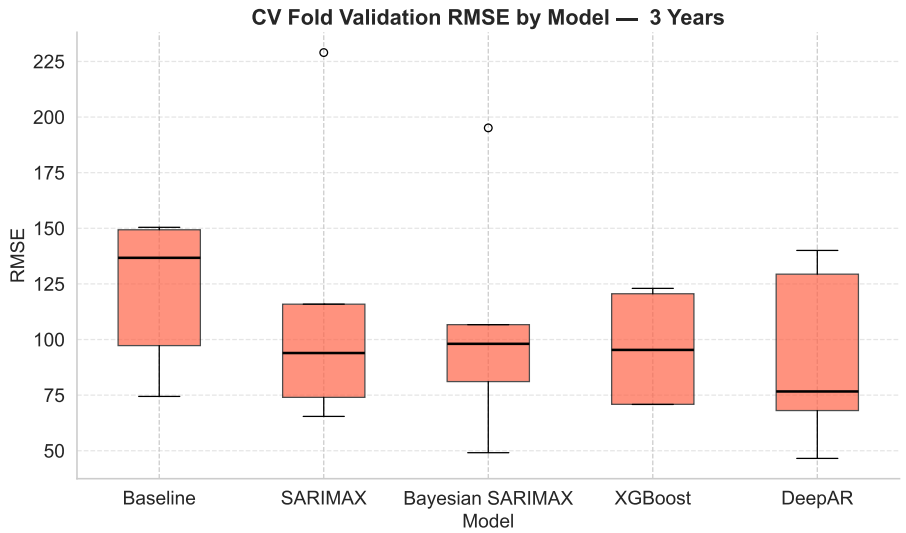

Figure 4.5: Spread of rolling CV split validation errors for all models. The models forecast 3 years forward on D_Y with 5 splits.

According to Figures 4.3 and 4.4, the forecasts generated by the models indicate a continuation of the current consumption levels rather than a substantial long-term upward trend. The dotted line representing the previous forecast in Figure 4.3 corresponds to GENAB's current forecast, which is based on estimated future changes in electricity consumption driven by new large-scale industrial establishments as well as the electrification of existing industries and society. While GENAB forecasts that total electricity consumption will almost double before 2040, most of the evaluated models instead predict a slight decrease in consumption, with the exception of the AR and XGBoost models. This result is consistent with the downward trend observed in the data between 2012 and 2025.

As shown in Table 4.3, all models exhibit indications of overfitting to the training data, particularly for longer forecasting horizons. The largest discrepancies between training and validation RMSE values are observed for the XGBoost model. This behavior may be explained by the limited size of the dataset, as machine learning models are generally more prone to overfitting when trained on small datasets.

Figure 4.5 presents a boxplot illustrating the distribution of validation errors obtained using the rolling cross-validation splits described in Section 3.1. The height of each box represents the interquartile range (IQR), while the horizontal black line inside the box indicates the median value. The whiskers illustrate the spread of the data outside the IQR, and the circles represent outliers. Together, these elements convey the stability and consistency of model performance across different training and validation splits. Nevertheless, the boxplot alone does not constitute a comprehensive assessment of forecasting quality, and its interpretation should be considered alongside the additional performance metrics presented in this section.

In general, models with lower median RMSE values demonstrate better predictive performance. Overall, all models exhibit relatively high RMSE values. The AR model shows both a high median RMSE and a large spread of errors, indicating unstable performance across the validation splits. In contrast, the SARIMAX and Bayesian SARIMAX models achieve lower RMSE values and display less variation between splits, suggesting improved generalisation capability. The XGBoost model demonstrates a comparatively low median RMSE together with a smaller overall spread, indicating stable predictive performance. Although the DeepAR model achieves the lowest median RMSE, it also exhibits the largest spread of errors, suggesting greater variability across the validation splits. These results indicate that the SARIMAX, Bayesian SARIMAX, and XGBoost models provide the most stable performance for this forecasting horizon.

The residual analysis indicates that only a limited number of models exhibit residual distributions that resemble a normal distribution, although with a shifted mean value. For the majority of the models, the residuals deviate from normality. In multiple cases, the residuals also display visible trends over time rather than fluctuating randomly around zero indicating non-random behavior. An example is presented in Figure 4.9 with additional figures in Appendix C

For all forecasting scenarios, the limitations associated with the small dataset are

evident. This issue is particularly pronounced for the 15-year forecasting horizon, where the training split primarily contains observations indicating an upward trend, while the validation split of the observed data instead reflects a downward trend. Consequently, the models are trained on patterns that differ substantially from those present in the validation period, which negatively affects their ability to generalize and produce reliable forecasts.

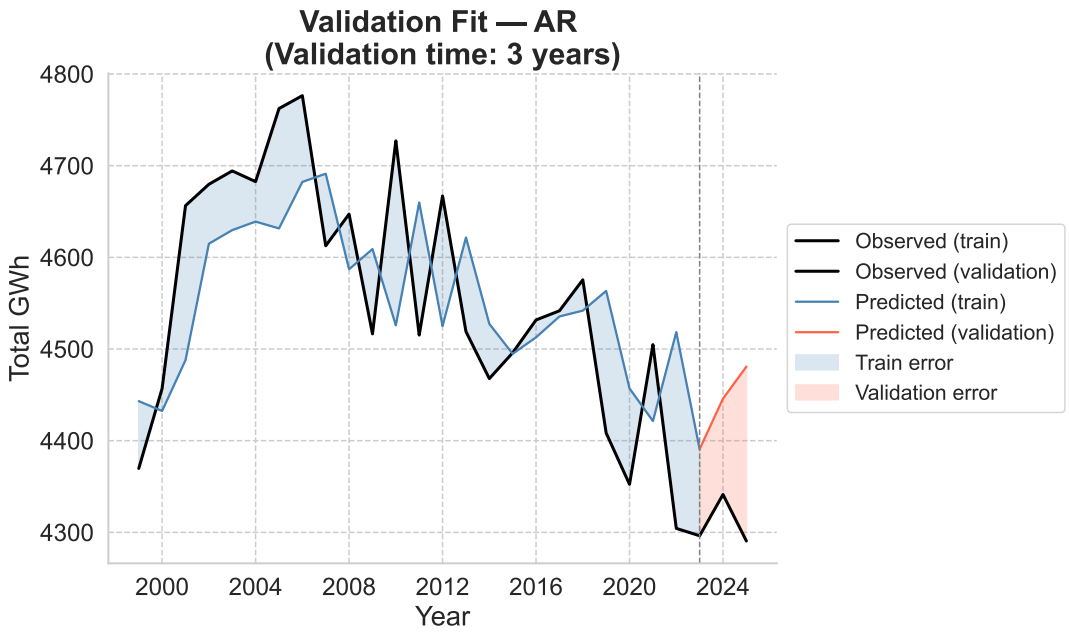

Figure 4.6: Model diagnostics of AR with validation time 3 years.

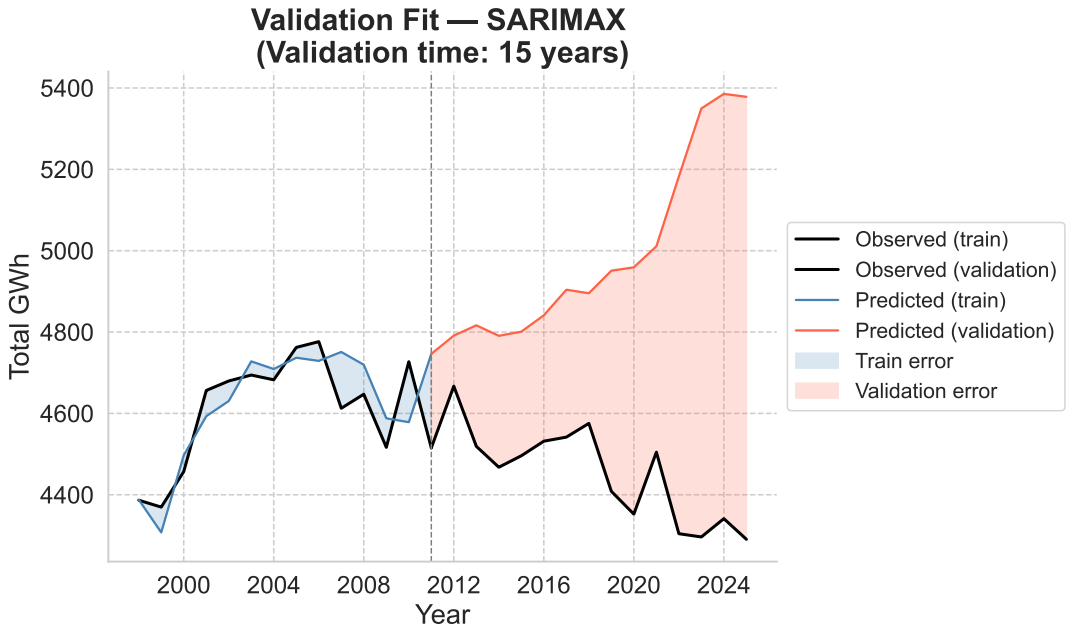

Figure 4.7: Model diagnostics of SARIMAX with validation time 15 years.

SARIMAX Validation Fit Summary (15 Years)				
Feature	**Coefficient**	**Std. Error**	**z-value**	**p-value**
Population	0.00469	0.00836	0.561	0.574
Real Salary	-0.051	0.0643	-0.793	0.428
Efficiency	-1910	814	-2.34	0.0193
ϕ: $p = 1$	-0.502	0.517	-0.971	0.332
Variance	5930	3570	1.66	0.0967

SARIMAX Forecast Fit Summary (15 Years)				
Feature	**Coefficient**	**Std. Error**	**z-value**	**p-value**
Population	-0.00399	0.00231	-1.73	0.0836
Real Salary	0.012	0.0223	0.535	0.592
Efficiency	-793	451	-1.76	0.0786
ϕ: $p = 1$	-0.562	0.162	-3.46	0.000536
Variance	7040	2440	2.89	0.00388

Table 4.4: SARIMAX coefficient estimates, standard errors, z-values, and p-values for the 15-year forecast horizon, fitted on the validation period (top) and the full forecast period (bottom). Due to regularization, EV per capita was removed as a feature from both the validation and forecast fit.

The AR model demonstrates relatively stable training errors, especially for the shorter forecasting horizons. This is displayed in Figure 4.6. However, there is a considerable gap between the training and validation errors, indicating limited generalization performance and suggesting that the model struggles to accurately predict unseen data despite its stable training behavior.

Table 4.3 shows that the training errors for the 3-, 5-, and 10-year forecasting horizons are relatively consistent for the SARIMAX model, whereas the results for the 15-year horizon differ more substantially. The validation fit of the 15-year SARIMAX forecast is illustrated in Figure 4.7. Among the validation results presented in Table 4.3, only the 3-year forecasting horizon demonstrates satisfactory performance, while the validation errors for the remaining horizons are considerably higher. This indicates that the forecasting ability of the model deteriorates as the forecasting horizon increases, suggesting a greater tendency to overfit the exogenous variables in longer-term predictions.

This behavior is also reflected in Table 4.4, where the efficiency variable obtains a substantially lower coefficient and statistical significance in the validation fit compared to the forecast fit. At the same time, the autoregressive coefficient, ϕ, is less significant in the validation fit, allowing the exogenous variables to exert a greater influence on the upward trend of the forecast.

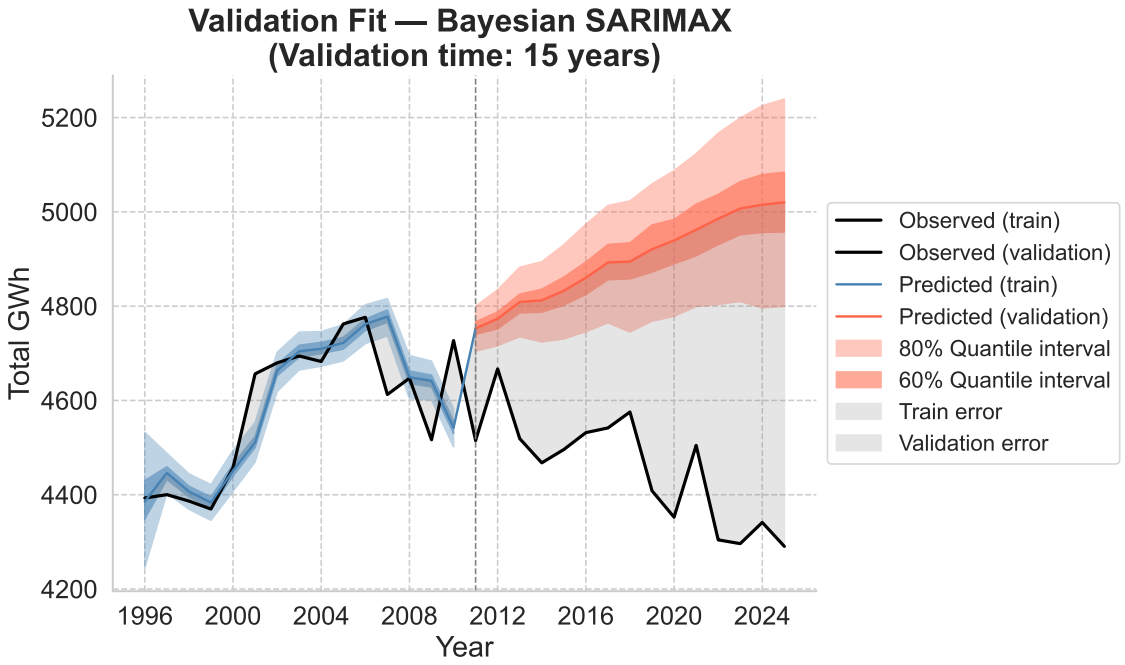

Figure 4.8: Model diagnostics of Bayesian SARIMAX with validation time 15 years.

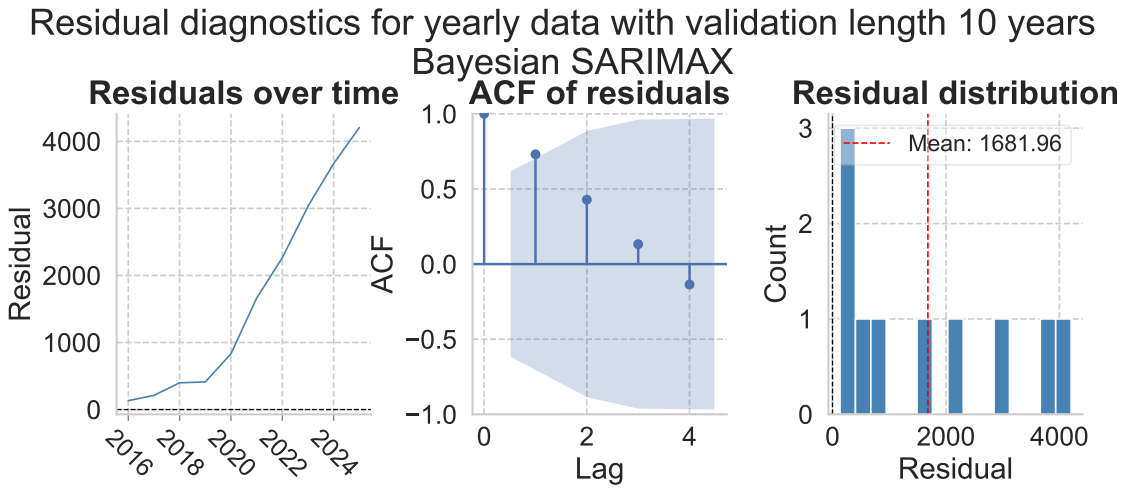

Figure 4.9: Figures of residual diagnostics of the Bayesian SARIMAX model for 10 years of validation.

The Bayesian SARIMAX model shows stable training errors across all forecasting horizons in Table 4.3. However, the overall behavior is similar to that of the standard SARIMAX model. While the shorter forecasting horizon produces comparatively reasonable validation performance, the longer forecasting horizons result in substantially larger validation errors, indicating reduced generalization ability and signs of overfitting.

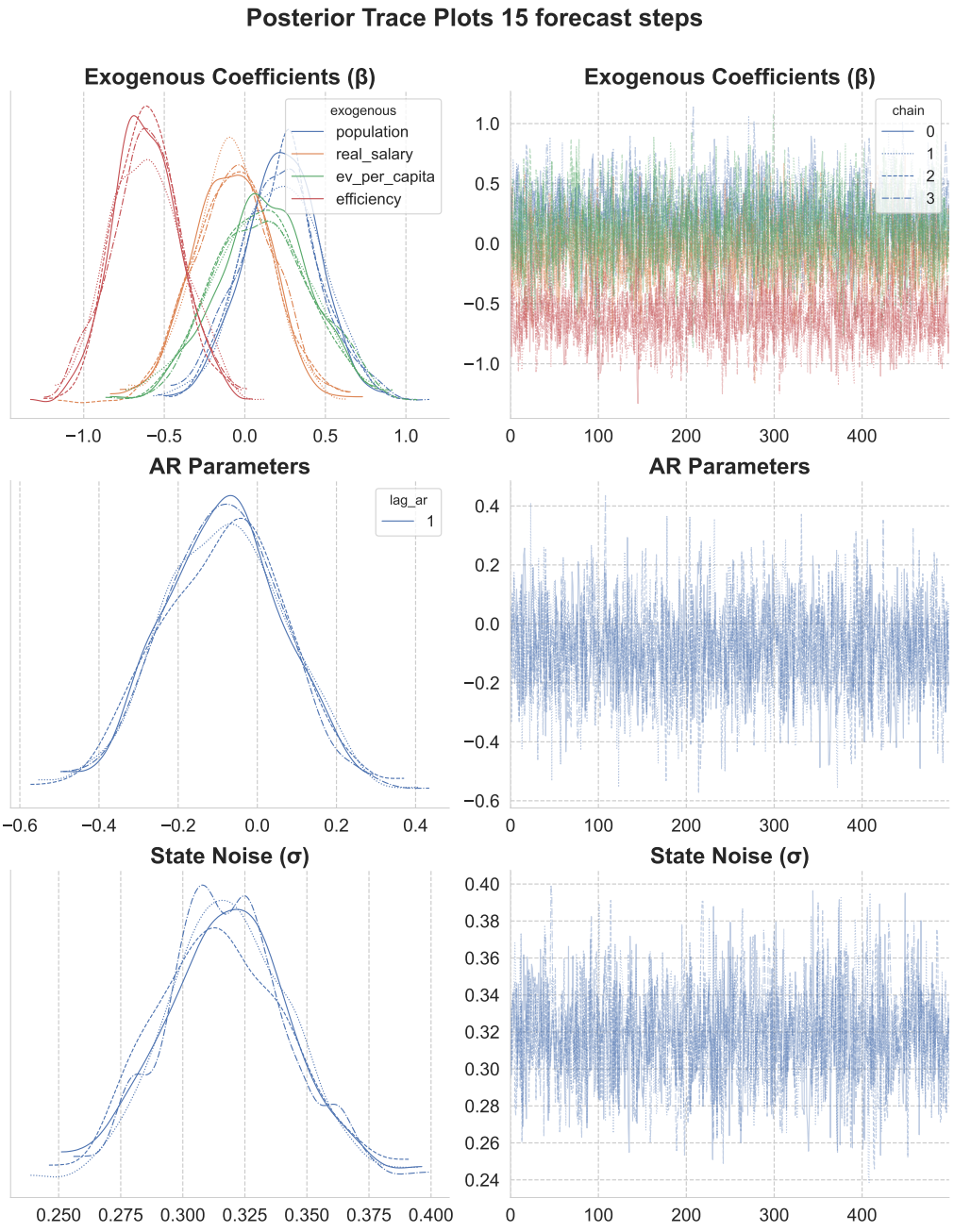

Figure 4.10: Traceplot over the posterior distributions of Bayesian coefficients for the 15 year forecast horizon.

Trace plots were examined to assess the convergence and mixing properties of the MCMC chains. The plots display the posterior distributions of the coefficients associated with the exogenous variables, as well as the autoregressive (AR) and moving average (MA) parameters. For the annual data models, 500 posterior draws, 1,000 tuning iterations and 4 chains were used. Figure 4.10 present the posterior distribution for the 15 years forecast horizon, more traceplots can be found in Appendix C.

Overall, the trace plots indicate satisfactory convergence for most models. The chains appear to fluctuate around stable mean values without exhibiting noticeable trends or systematic drifts over time. Furthermore, no abrupt jumps between different regions of the parameter space were observed, suggesting that the chains mix well and are sampling from a common posterior distribution.

Figure 4.10 also reveal that several of the highest density intervals (HDIs) for the exogenous variables include zero. This indicates limited evidence that these variables have a substantial effect on the target variable within the respective models. For the annual data set, the energy efficiency variable appears to have the strongest influence, as its posterior distribution suggests a more pronounced and consistent effect compared to the other explanatory variables.

XGBoost Validation Fit Summary (10 Years)	
Feature	**Gain**
Population	0.456
Real Salary	0.153
Total GWh, Lag 1	0.107
Efficiency	0.0355

XGBoost Forecast Fit Summary (10 Years)	
Feature	**Gain**
Total GWh, Lag 1	0.512
Population	0.229
EV Per Capita	0.143
Real Salary	0.0428
Efficiency	0.0195

Table 4.5: XGBoost feature importance (gain) for the 10-year forecast horizon, fitted on the validation period (left) and the full forecast period (right). Note that EV per capita is not included as a validation feature due to lack of training data.

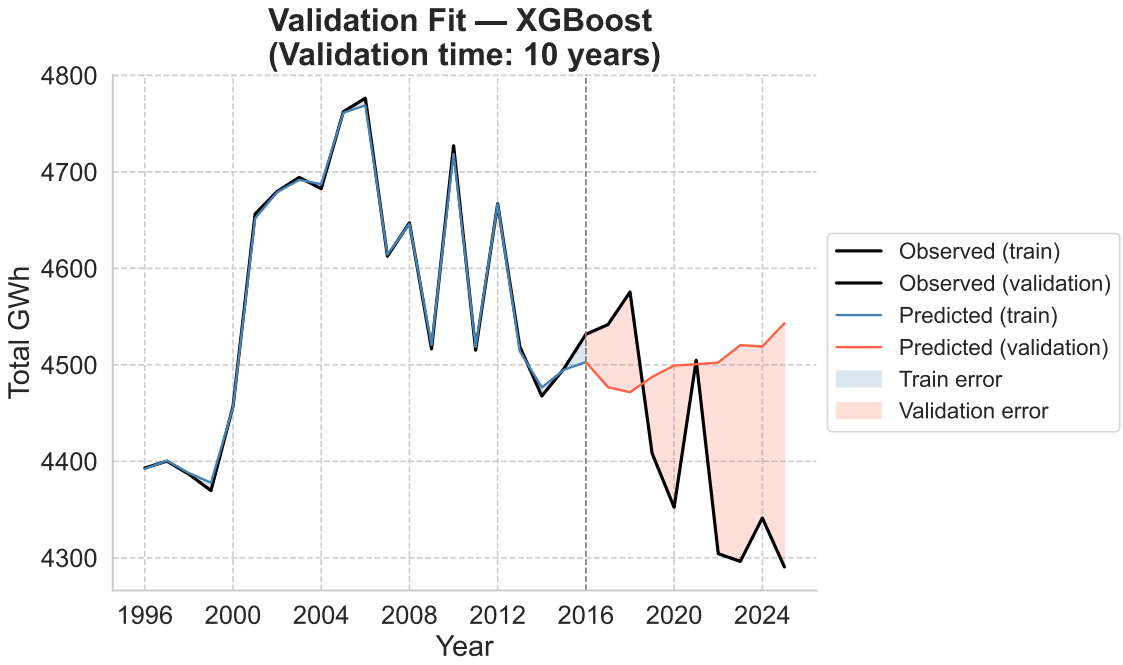

Figure 4.11: Model diagnostics of XGBoost with validation time 10 years.

For the XGBoost model, the training errors vary considerably across the different forecasting scenarios, indicating less stable model behavior across forecasting horizons. In addition, the validation errors are substantially larger than the corresponding training errors in all cases, strongly suggesting that the model overfits the training data and generalises poorly to unseen observations.

As shown in Table 4.5, population is the most influential exogenous feature for both the validation and forecast fits, contributing to an increase in the predicted electricity consumption in both cases. In contrast, efficiency and real salary appear to have a comparatively smaller influence on the XGBoost forecasts. The limited contribution from the efficiency variable may be explained by its exponentially decreasing trend, which differs from the behavior of the other exogenous variables and the total electricity consumption.

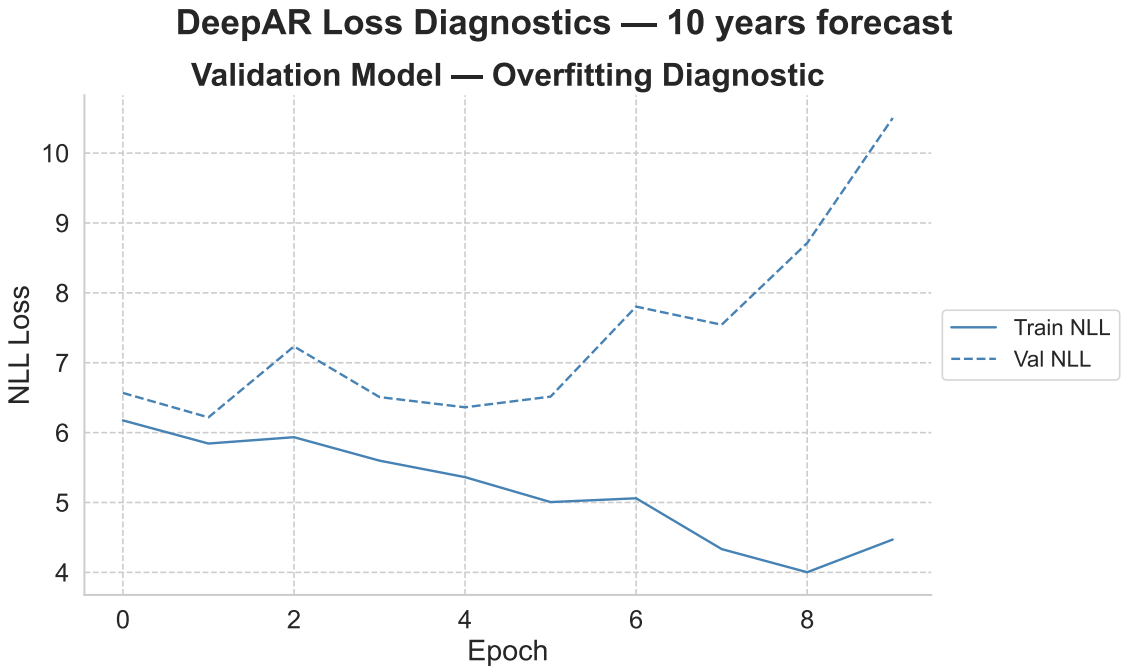

Figure 4.12: NLL for DeepAR with 10 year forecast. One can clearly see that the validation error is significantly higher than the training error indicating overfitting.

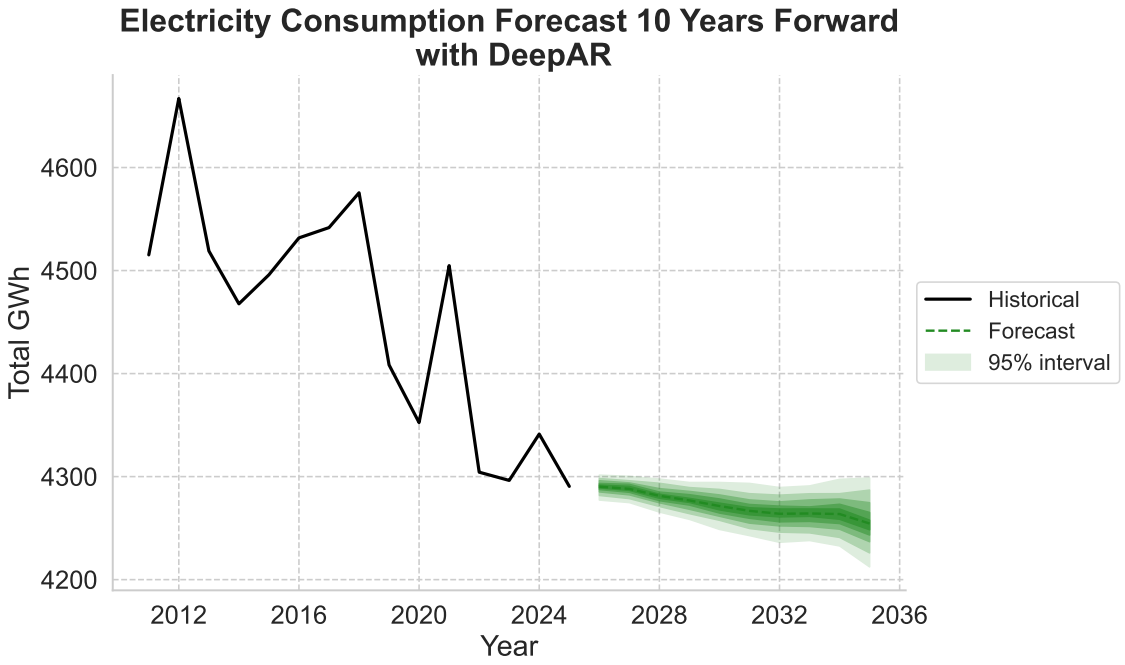

Figure 4.13: Forecast of energy consumption 10 years forward with DeepAR. Note that the forecast is shown as a distribution of likely values.

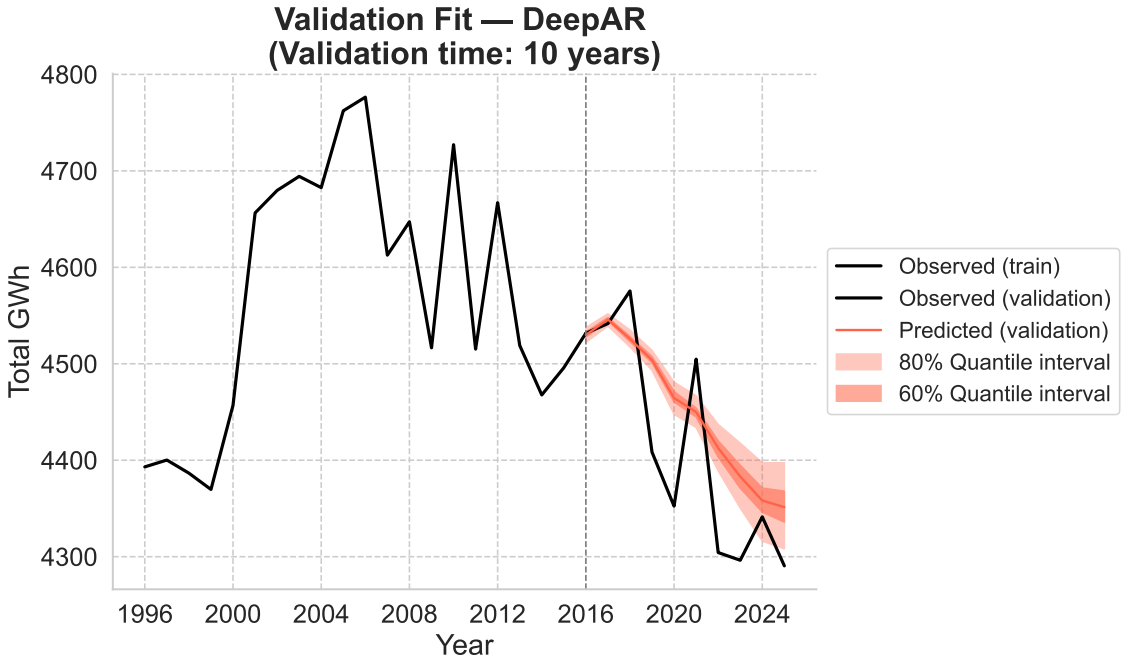

Figure 4.14: Model diagnostics of DeepAR with validation time 10 years.

For the DeepAR model, forecasts were generated for the 3-, 5-, and 10-year forecasting horizons due to the limited size of the dataset. Unlike the previous models, model performance was evaluated using the Negative Log-Likelihood (NLL) loss as

the primary diagnostic metric. The training NLL-loss decreases consistently across all forecasting horizons, indicating that the model successfully captures patterns in the training data. However, the validation loss initially decreases before subsequently increasing, which is a characteristic indication of overfitting. An example of this behavior for the 10-year forecasting horizon is presented in Figure 4.12 and Figures C.3 and C.5 in Appendix C.

These results suggest that, although the model improves its fit to the training data over time, its ability to generalise to unseen data does not improve correspondingly. Nevertheless, the validation forecast for the 10-year horizon shown in Figure 4.14 still follows the observed values reasonably well.

4.3 Forecast Results with Monthly Aggregated Data

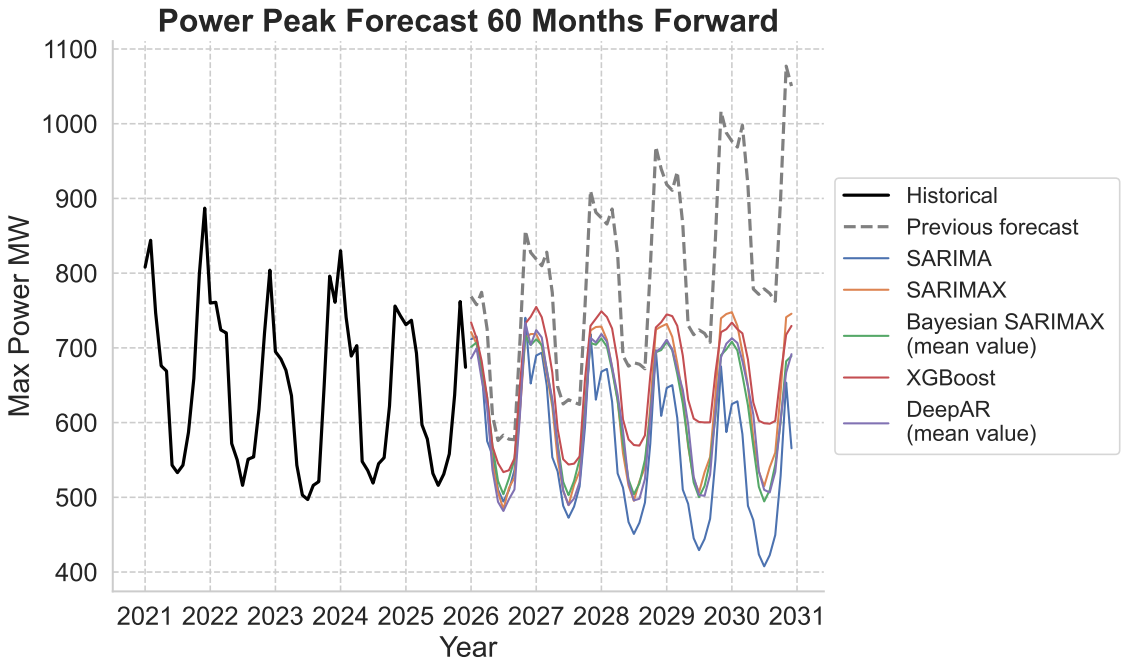

Figure 4.15: Max power forecasts 60 months forward for all selected models.

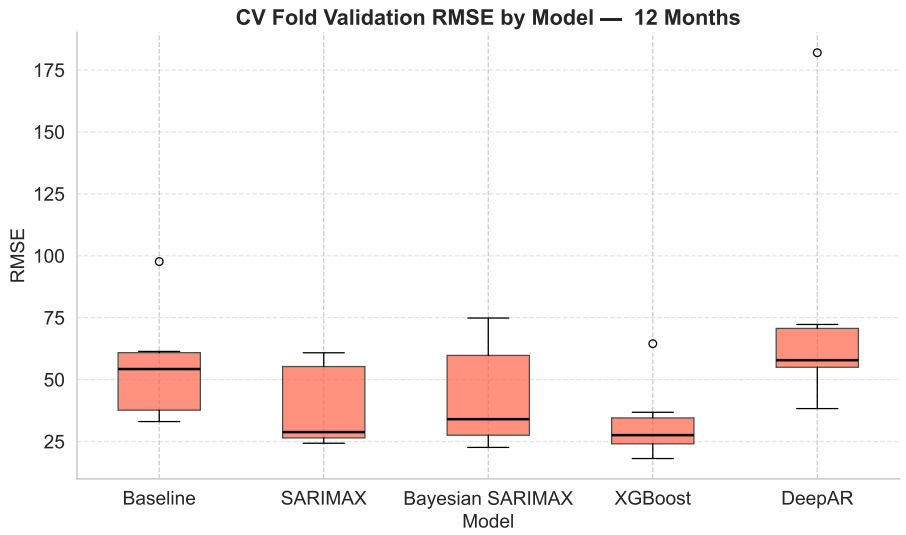

Figure 4.16: Spread of rolling CV split validation errors for all models. The models forecast 12 months forward on D_M with 7 splits.

Baseline Errors, D_M		
Forecast Length (Months)	**Training RMSE**	**Validation RMSE**
12	46.38	44.20
24	45.04	53.03
36	45.31	45.43
60	42.28	44.41

SARIMAX Errors, D_M		
Forecast Length (Months)	**Training RMSE**	**Validation RMSE**
12	23.51	28.55
24	23.87	24.73
36	23.95	23.33
60	24.02	25.70

B-SARIMAX Errors, D_M		
Forecast Length (Months)	**Training RMSE**	**Validation RMSE**
12	24.74	27.26
24	24.61	28.18
36	25.04	26.88
60	25.27	59.63

XGBoost Errors, D_M		
Forecast Length (Months)	**Training RMSE**	**Validation RMSE**
12	3.44	25.33
24	15.41	20.12
36	14.04	46.75
60	15.13	42.13

DeepAR Errors, D_M		
Forecast Length (Months)	**Training RMSE**	**Validation RMSE**
12	—	28.73
24	—	32.10
36	—	61.05
60	—	43.50

Table 4.6: Training and validation RMSE for all five models on dataset D_M, across forecast horizons of 12, 24, 36, and 60 months. DeepAR training RMSE is not reported as the model does not yield an equivalent in-sample metric.

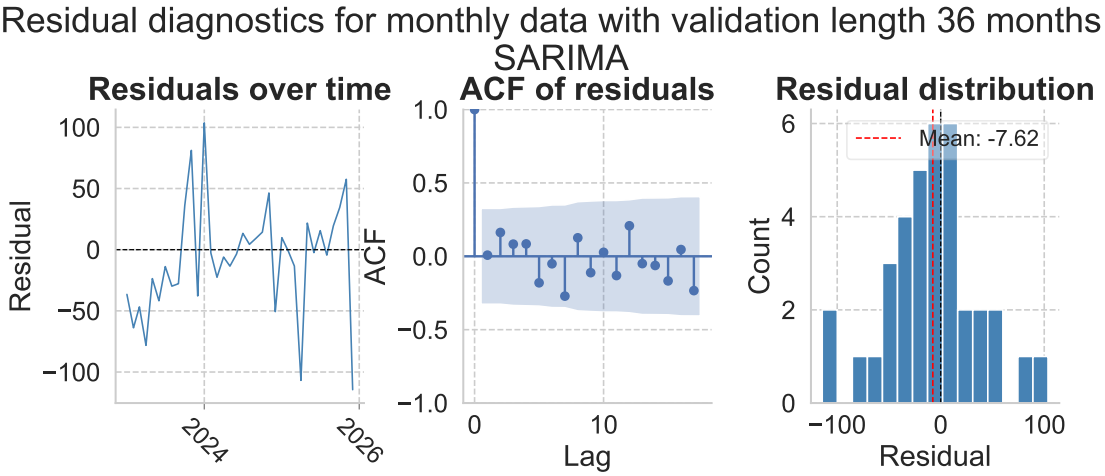

Figure 4.17: Residuals for the SARIMA model with 36 months validation time.

Overall, the statistical time series models demonstrate relatively stable generalization performance, whereas the machine learning–based models exhibit more pronounced signs of overfitting, likely due to the limited size of the available dataset. In addition, all models generate forecasts characterised by an overall downward trend. In contrast, GENAB's previous forecast of monthly power peaks, illustrated by the dotted line in Figure 4.15, indicates an increasing trend.

As shown in Figure 4.16, all models exhibit relatively low RMSE values across the validation splits. The SARIMA model demonstrates one of the highest median RMSE values, although with a relatively small overall spread, indicating consistent but comparatively weaker predictive performance. In contrast, the SARIMAX and Bayesian SARIMAX models achieve some of the lowest median RMSE values, albeit with a larger spread across the splits, suggesting increased variability in performance. The XGBoost model exhibits both a low median RMSE and a limited spread, indicating stable and robust predictive performance across the validation splits. Conversely, the DeepAR model shows the highest RMSE values and the highest median error, indicating comparatively weaker forecasting performance for this prediction horizon.

The residual analysis shows a consistent pattern across both D_M and D_Y, suggesting similar underlying model behaviour in both datasets.

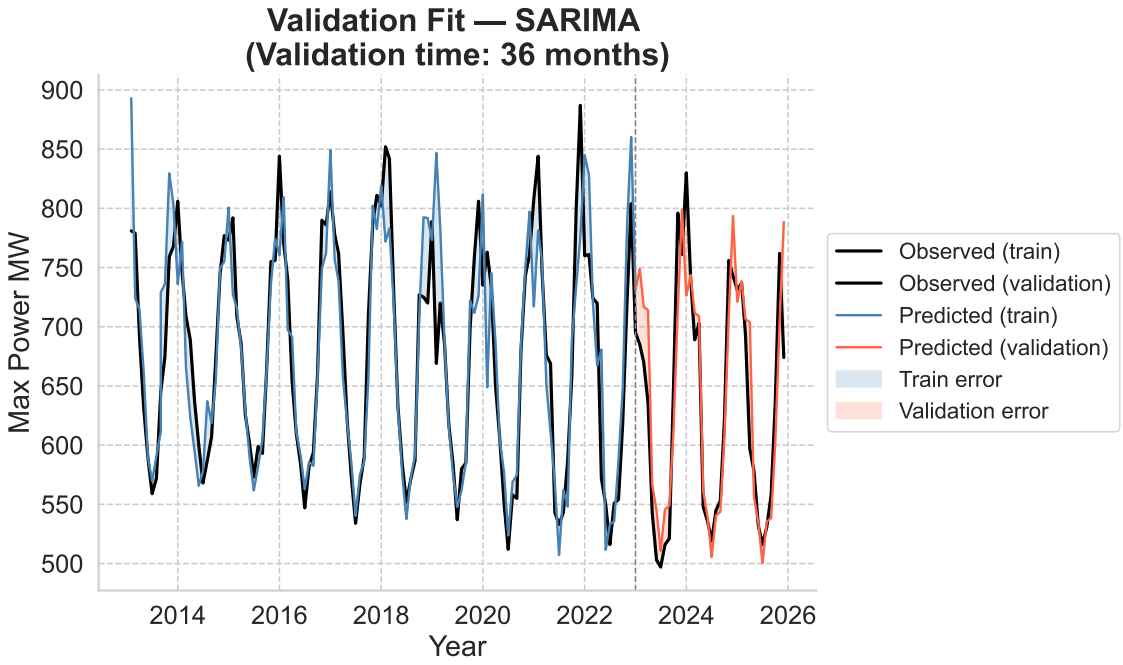

Figure 4.18: Model diagnostics of SARIMA with validation time 36 months.

The SARIMA model serves as the baseline model for comparison with the other methods. Results indicate that the model generalizes the data relatively well, as the training error remains stable across all horizons in Table 4.6. The validation error is also stable and small relative to the training error, with the exception of the 24-month scenario where the difference is somewhat larger. However, this difference is considered relatively minor in relation to the overall magnitude of the errors and is therefore of limited significance. At the same time, the SARIMA model exhibits the largest errors among all models, indicating limited forecasting accuracy despite its stability.

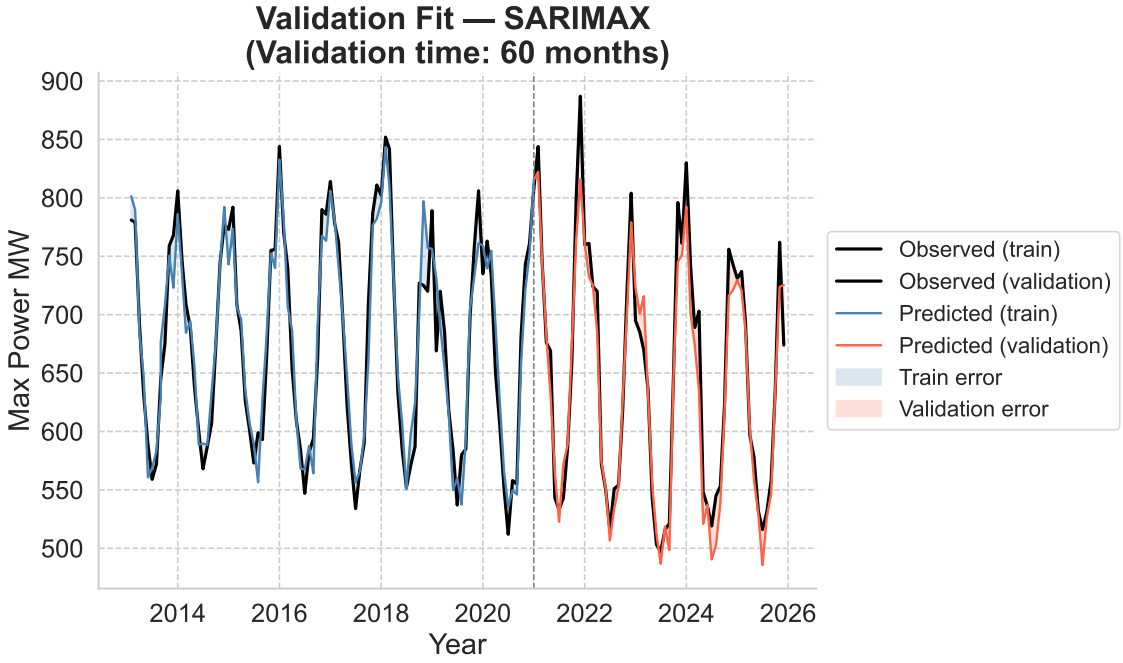

Figure 4.19: Model diagnostics of SARIMAX with validation time 60 months.

SARIMAX Validation Fit Summary (60 Months)				
Features	**Coefficient**	**Std. Error**	**z-value**	**p-value**
Avg Coldest Five	-9	0.853	-10.6	4.78e-25
Population	-0.00139	0.00733	-0.185	0.853
Real Salary	0.00938	0.016	0.591	0.554
θ: $(q = 1)$	-0.545	0.0898	-6.11	9.88e-10
Θ: $(Q = 1)$	-0.712	0.122	-5.85	3.17e-09
Variance	537	88.6	6.06	1.35e-09

SARIMAX Forecast Fit Summary (60 Months)				
Feature	**Coefficient**	**Std. Error**	**z-value**	**p-value**
Avg. Coldest Five	-10.2	0.647	-15.9	8.01e-57
Population	-0.000421	0.00358	-0.12	0.905
Real Salary	0.00675	0.0059	1.14	0.253
θ: $(q = 1)$	-0.626	0.0645	-9.65	4.88e-22
Θ: $(Q = 1)$	-0.977	0.877	-1.13	0.026
Variance	467	399	1.17	0.241

Table 4.7: SARIMAX coefficient estimates, standard errors, z-values, and p-values for the 60-month forecast horizon, fitted on the validation period (top) and the full forecast period (bottom). Due to regularization, both efficiency and EV per capita were removed as features from the validation and forecast fits.

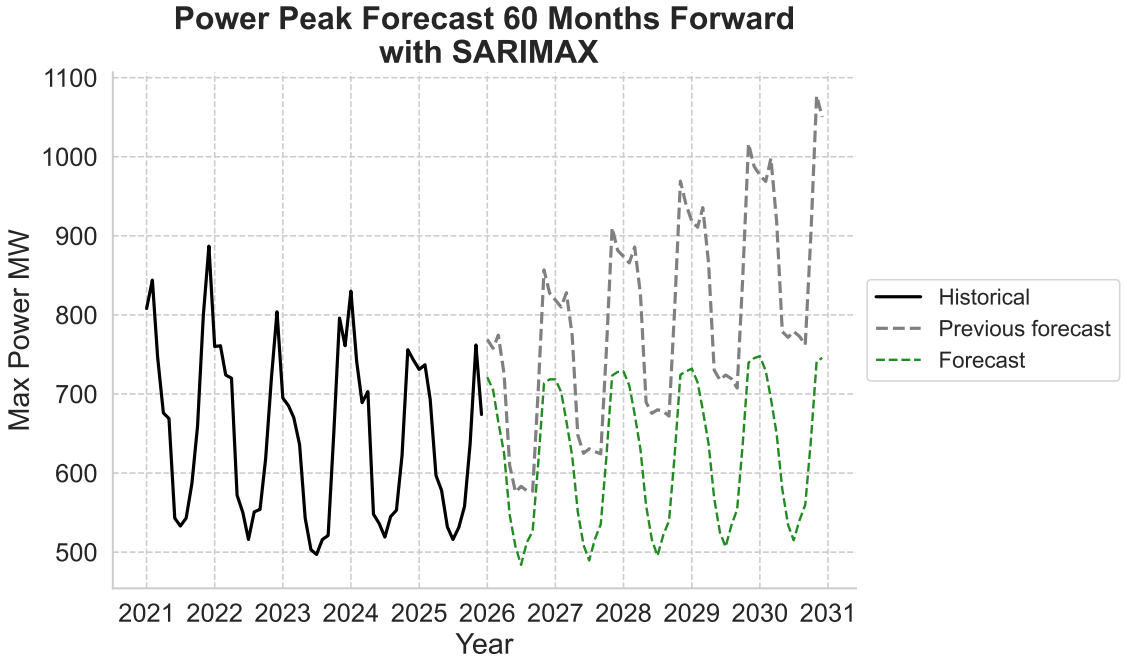

Figure 4.20: Forecast of power peaks 60 months forward with SARIMAX.

For the SARIMAX models, the training errors in Table 4.6 remain relatively consistent across all evaluated scenarios, indicating that the models are able to capture the underlying patterns in the training data effectively. Similarly, the validation errors are stable across the different forecasting scenarios and closely correspond to the training errors, suggesting good generalization performance and a limited degree of overfitting.

Among all evaluated models, SARIMAX achieved the best overall performance for the monthly dataset D_M. This result is expected, as the model incorporates exogenous variables while simultaneously applying regularization through the SARIMAX framework, thereby reducing the risk of overfitting to the explanatory variables. In addition, the monthly dataset exhibits a relatively consistent trend throughout the observed time period, with only minor variations between the training and validation segments. Consequently, the statistical properties of the training and validation data are more homogeneous for D_M compared to D_Y, allowing the model to generalize more effectively across forecasting scenarios.

The Bayesian SARIMAX model also demonstrates consistent training errors across the different scenarios in Table 4.6. The validation errors for the 12-, 24-, and 36-month horizons are relatively stable and small in relation to the training errors, indicating good generalization performance for these forecasting horizons. This is exemplified in Figure 4.22. However, the validation error is more than twice as large as the training error in the 60-month scenario, suggesting that this model is in need of further regularization through more informative priors.

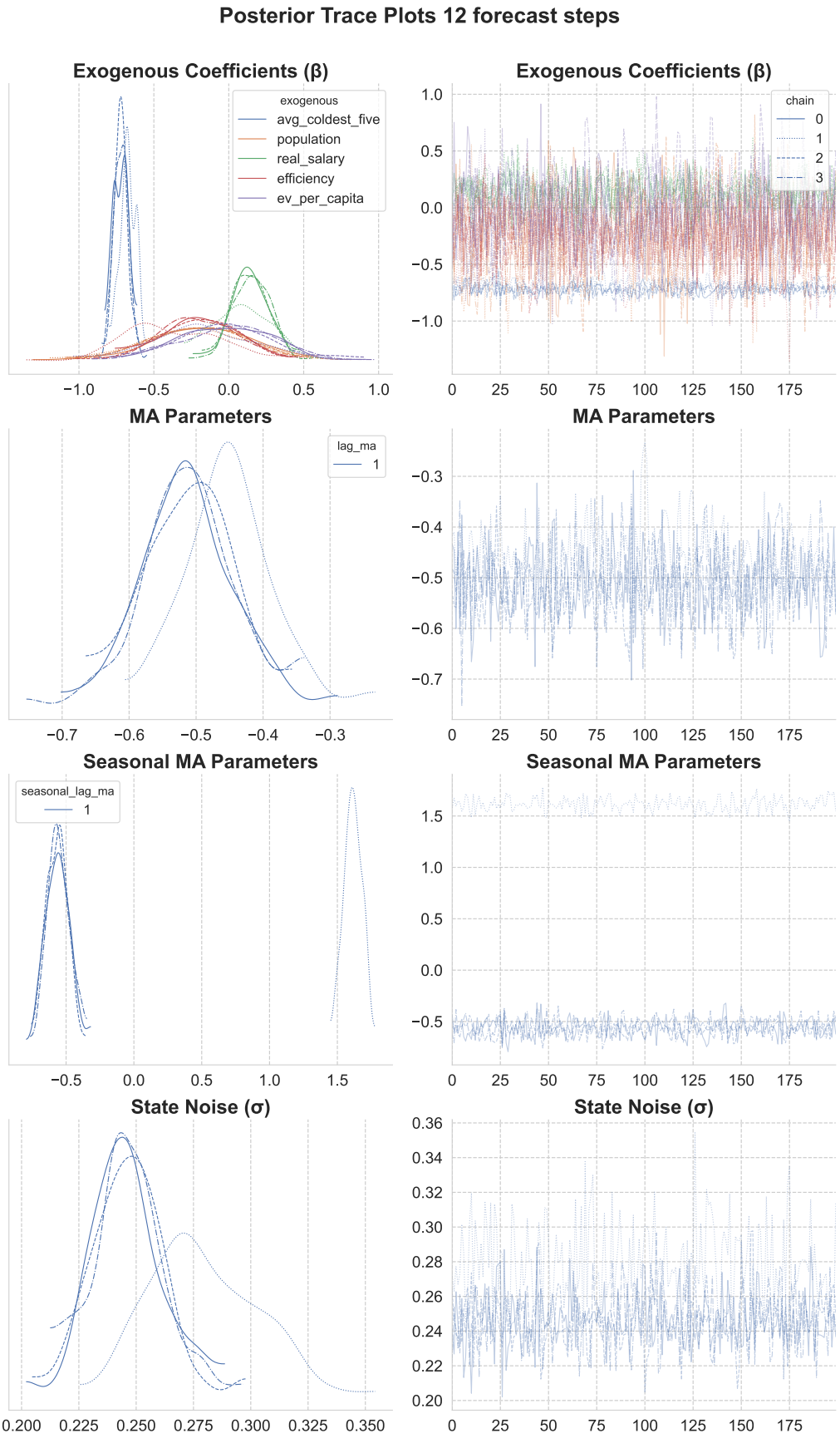

Figure 4.21: Traceplot over the posterior distributions of Bayesian coefficients for the 12 months forecast horizon.

Figure 4.21 show that the model have some convergence issues for the 12-month forecasting horizon. In these models, 200 posterior draws, 400 tuning iterations and 4 chains were employed. The trace plots suggest that the MCMC chains experience greater difficulty in fully exploring the posterior distribution, leading to less stable parameter estimates. Increasing the number of draws could mitigate this convergence problem. Nevertheless, the models corresponding to longer forecasting horizons demonstrate improved convergence behavior and more reliable posterior estimates.

Among the exogenous variables included in the monthly data set, temperature appears to have the most substantial impact on model performance. This is reflected in the posterior distributions, where the estimated effect of temperature is more clearly distinguished from zero than those of the remaining explanatory variables.

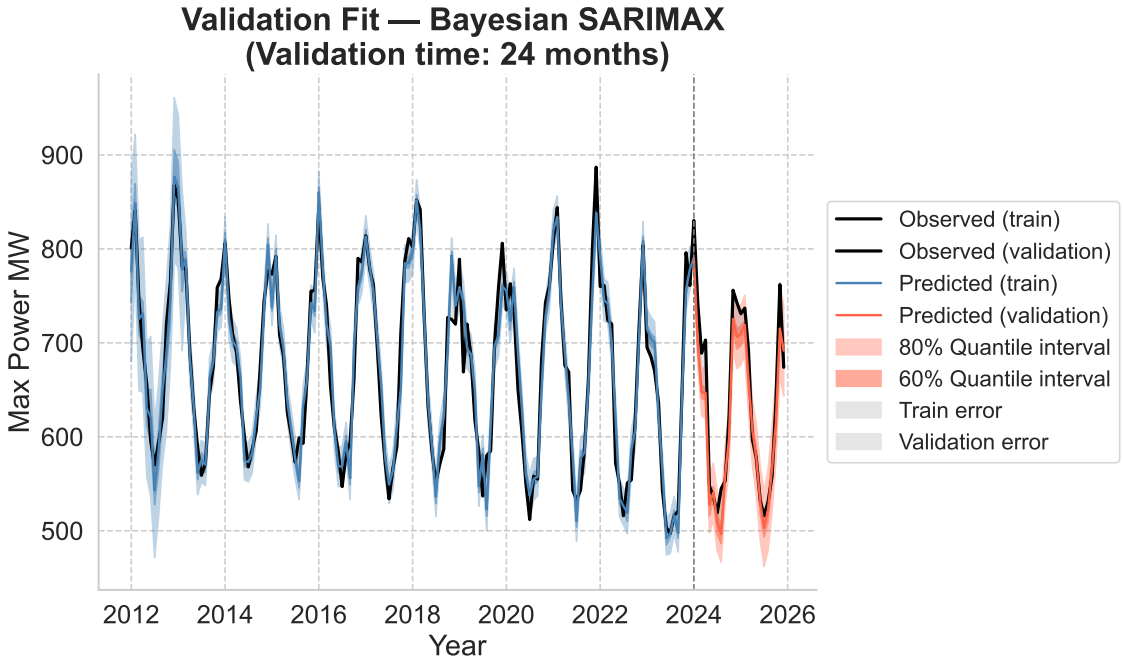

Figure 4.22: Model diagnostics of Bayesian SARIMAX with validation time 24 months.

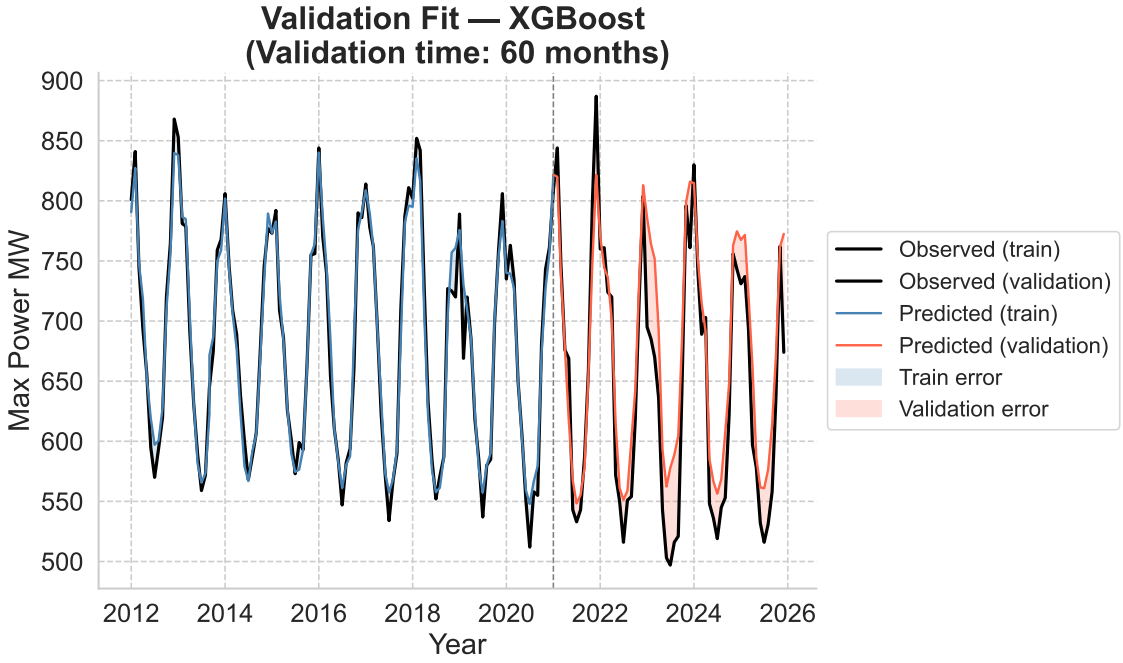

Figure 4.23: Model diagnostics of XGBoost with validation time 60 months.

XGBoost Validation Fit Summary (60 Months)	
Feature	**Gain**
Max Power MW, Lag 1	0.291
Max Power MW, Lag 12	0.22
Avg. Coldest Five	0.144
Efficiency	0.0147
Real Salary	0.0143
Population	0.0138
EV Per Capita	0.0134

XGBoost Forecast Fit Summary (60 Months)	
Feature	**Gain**
Max Power MW, Lag 12	0.242
Max Power MW, Lag 1	0.196
Avg. Coldest Five	0.171
Population	0.0224
EV Per Capita	0.0179
Efficiency	0.0114
Real Salary	0.00675

Table 4.8: XGBoost feature importance (gain) for the 60-month forecast horizon, fitted on the validation period (left) and the full forecast period (right). Features are ranked in descending order of gain.

For the XGBoost model, the training errors exhibit greater variation across forecasting horizons compared to the statistical time series models presented in Table 4.6. Overall, the training errors remain relatively low, indicating that the model captures patterns in the training data effectively. However, the substantial gap between training and validation errors across all scenarios suggests that the model suffers from overfitting and has limited ability to generalize to unseen observations.

In comparison with the traditional time series models, XGBoost nevertheless maintains validation errors at a relatively reasonable level, even for longer forecasting horizons. This behavior is illustrated in Figure 4.23. Furthermore, Table 4.8 indicates that the average of the five coldest temperatures is the most important

exogenous feature. This result is expected, as the variable is more closely related to the target variable than the remaining features. The consistently low importance assigned to the other exogenous variables across both model fits further supports the need for additional feature regularization or feature selection.

Figure 4.24 shows that the training NLL loss generally decreases throughout the 60-month forecasting scenario for the DeepAR model, indicating that the model progressively learns patterns in the training data. However, the validation loss decreases only during the initial epochs before either increasing or plateauing, depending on the forecasting horizon. This suggests that the model's generalization ability does not improve at the same rate as its training performance and indicates the presence of overfitting across all scenarios. The widening gap between training and validation loss further supports this interpretation.

This behavior is consistently observed across all forecasting scenarios, as illustrated in Figures C.6, C.7, and C.8 in Appendix C.

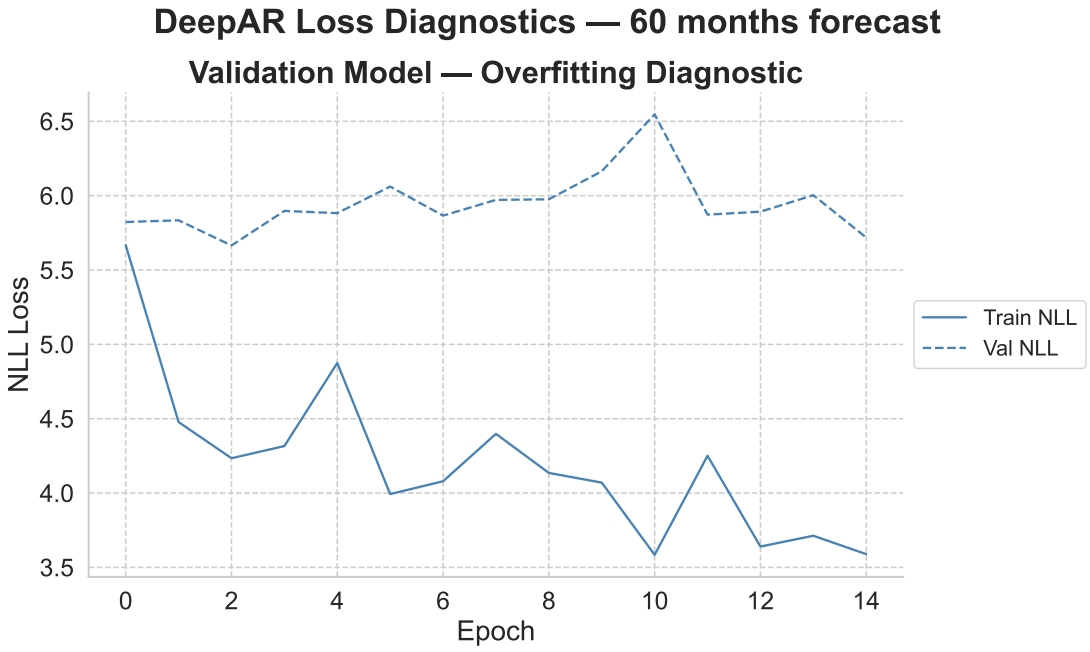

Figure 4.24: NLL for DeepAR with 60 months forecast. One can clearly see that the validation error is significantly higher than the training error indicating overfitting.

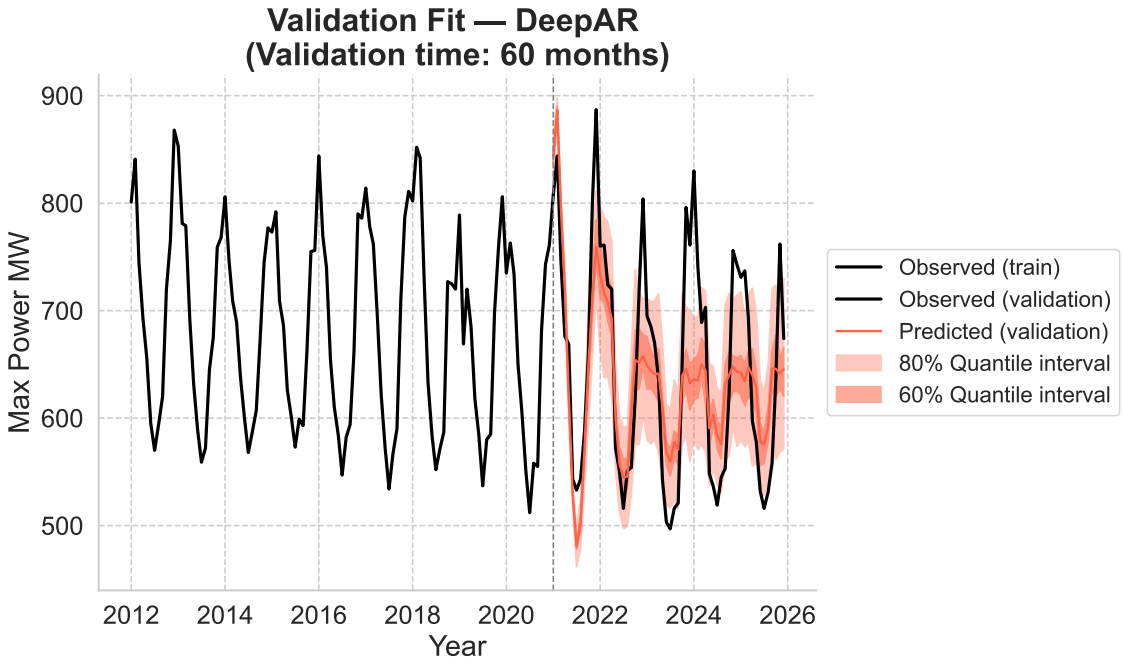

Figure 4.25: Model diagnostics of DeepAR with validation time 60 months.

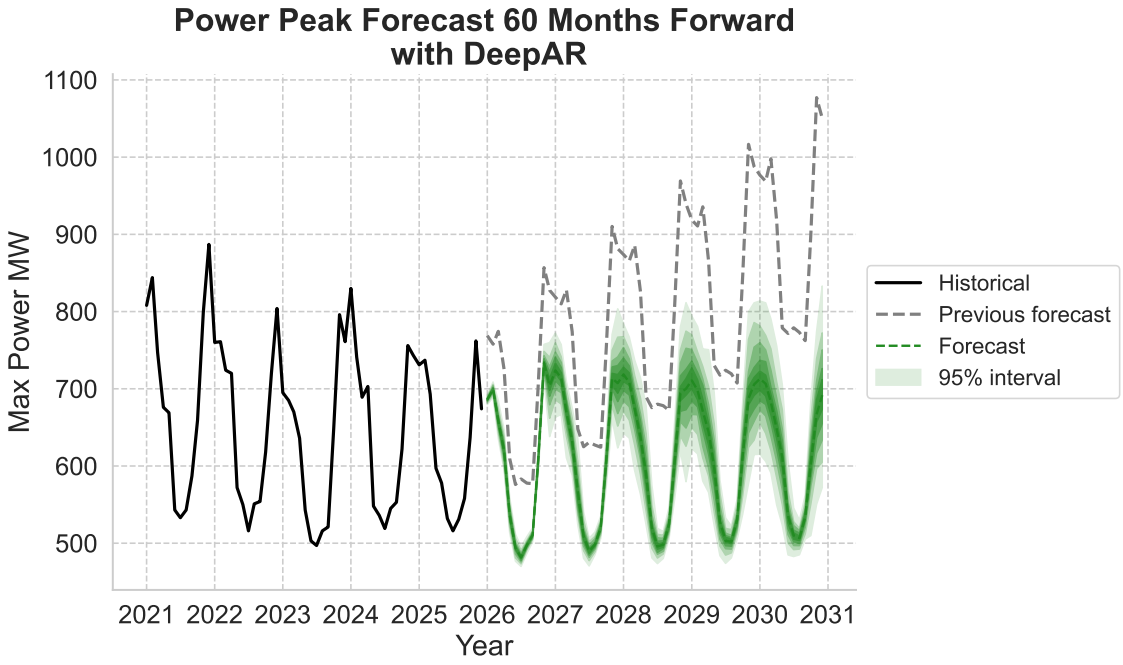

Figure 4.26: Forecast of power peaks 60 months forward with DeepAR. Note that the forecast is shown as a distribution of likely values.

4.4 Discussion of Results

The residual behavior suggests that the evaluated models were unable to fully capture important temporal dynamics and structural relationships in the data. The presence of systematic and non-random residual patterns indicates potential model misspecification, implying that relevant information may still remain unexplained. This suggests that additional explanatory variables, alternative model structures, or further adjustments to the time-series specification may be necessary to improve model performance, robustness, and forecasting reliability. Overall, these residual patterns indicate that the forecasting performance is affected by several underlying limitations related to the available data, model assumptions, and explanatory variables. The results can primarily be attributed to three main challenges.

First, the datasets are too small to fully leverage the strengths of machine learning methods. The annual dataset contains only 30 historical observations in total, and since the models must be validated on the same number of observations as the forecasting horizon, only 15 observations remain available for training in the 15-year forecasting scenario. Such a limited amount of training data restricts the models' ability to learn robust and generalizable patterns while also reducing the number of possible validation splits, thereby increasing the risk of biased evaluations. Although the monthly dataset contains a larger number of observations, it is still considered relatively small in the context of machine learning applications, particularly for deep learning–based forecasting models. The results therefore only serve as an exhaustive investigation of model capabilities given the data available for this specific thesis, instead of an assessment of long-term ML-forecasts in general.

Second, while future increases in electricity consumption and power demand are expected, these developments are not reflected in the historical data. The time series does not exhibit a sufficiently clear upward trend that would allow the models to extrapolate long-term growth patterns. As a result, the models primarily learn from the existing historical behavior of the data, leading to forecasts that remain closely aligned with past observations. For the annual dataset, the historical data instead suggests a declining trend, while the monthly dataset also exhibits a slight downward tendency. Consequently, the generated forecasts tend to reflect stable or decreasing consumption levels rather than the anticipated future increase.

The tendency of the models to reproduce historical patterns rather than anticipated future growth also highlights an important methodological difference between the forecasting approach applied in this study and the forecasts produced by GENAB. The models developed in this thesis are entirely data-driven and rely on historical relationships within the available datasets. As a result, the forecasts are constrained by the information and patterns present in the historical data. In contrast, GENAB's forecasting methodology is based on existing consumption levels combined with expected future developments and adjustments informed by domain expertise and industry knowledge. Such an approach allows for the incorporation of anticipated structural changes, policy developments, technological transitions, and planned ex-

pansions that are not yet visible in the historical time series.

Therefore, the differences between the forecasts should not necessarily be interpreted as contradictions, but rather as consequences of fundamentally different forecasting methodologies and assumptions. The forecasts presented in this study reflect the behavior implied by the historical data that was available at the time of analysis, whereas GENAB's forecasts incorporate expert-based expectations regarding future developments in electricity consumption and power demand.

Third, the selected target variables with the exogenous variables do not sufficiently explain the variation in the data and therefore exhibit limited statistical significance. Conversely, certain exogenous features were found to decrease forecast performance, indicating that further regularization may be warranted. As a consequence, the models struggle to capture meaningful relationships and underlying drivers of electricity consumption and power demand, ultimately limiting forecasting accuracy across the evaluated scenarios.

Overall, the findings demonstrate the challenges of applying machine learning methods to long-term electricity demand forecasting when historical datasets are limited and future structural changes are not yet reflected in the available data.

5

Conclusion

In conclusion, the first two objectives presented in Section 1.2 were achieved through the implementation of relevant variables and forecasting models derived from existing literature. Based on the validation RMSE values, the machine learning models marginally outperformed the traditional statistical models for both D_Y and D_M. However, the reliability of these results was constrained by bias introduced during the validation process. Overall, the forecasting models struggled to capture signals of increased electricity consumption and power demand from the exogenous variables, instead producing predictions that remained relatively stable over time. This finding addresses the third objective of the study.

With respect to the fourth objective, the results indicate that several of the selected exogenous variables should be retained in future forecasting models. In particular, population levels appear to constitute an important explanatory factor, as population growth is naturally associated with increased electricity consumption. Temperature was also identified as a highly relevant variable, especially in forecasting peak power demand, since colder weather conditions are strongly associated with increased electricity usage due to higher heating demand. Several other variables showed low predictive value and contributed to increased forecast uncertainty and reduced accuracy.

Regarding the fifth objective, the results demonstrate that machine learning approaches can be successfully applied to electricity consumption and power demand forecasting. However, larger and more comprehensive datasets are required in order to fully utilize the strengths of these approaches. The limited amount of available historical data restricts the models' ability to learn robust and generalizable patterns, especially for longer forecasting horizons. Although outside the scope of this thesis, an alternative approach would be to shorten the forecasting horizon to days, weeks, or months in order to utilize currently available high-frequency data. This represents a promising direction for future research at GENAB, particularly given the larger availability of exogenous variables and the extensive academic literature in this area.

A notable limitation throughout this thesis was the scarcity of exogenous variables with historical records extending back to 1996 for D_Y, which constrained the feature selection process. Several potentially relevant variables were excluded either because they fell outside the scope of the thesis or because sufficiently long his-

torical records were unavailable. Nevertheless, such variables may improve future forecasting performance and enhance the models' ability to capture the underlying drivers of electricity demand.

One potentially important aspect is load aggregation, which describes how electricity consumption from different sectors is distributed over time. For example, electric buses charging during nighttime hours may aggregate efficiently with industrial production that primarily occurs during daytime hours. In such cases, electricity demand becomes more evenly distributed throughout the day, thereby reducing peak power demand even if total electricity consumption increases. The management of power peaks through load aggregation has become increasingly relevant for energy companies in long-term planning. Incorporating such data could therefore provide a more nuanced understanding of future power demand.

The transport sector is also expected to become increasingly important, as the electrification of transportation will likely contribute substantially to both electricity consumption and peak power demand. Variables related to electric vehicle adoption, charging infrastructure, and transport electrification could therefore provide valuable explanatory information for future forecasting models.

Industrial activity likewise constitutes an important explanatory factor. This includes both the current level of electricity consumption within the industrial sector and broader indicators of industrial development over time. Since industrial expansion and electrification are expected to contribute significantly to increased electricity consumption in the coming years, variables capturing industrial growth may improve the explanatory power of the forecasting models. Such variables would, however, most likely need to be represented as categorical variables or stepwise functions rather than as conventional continuous time series.

Furthermore, political decisions may have a considerable influence on electricity demand. Policy measures such as subsidies for electric vehicles, industrial electrification initiatives, and energy transition regulations can substantially affect long-term consumption patterns. Similar to industrial activity variables, these factors are often best represented categorically, particularly when historical observations are limited or when policy effects emerge gradually over time. In this context, the DeepAR model may be particularly suitable, as it is designed to handle categorical variables and multiple time series efficiently.

A significant improvement for future studies would therefore be access to more detailed historical data from the industrial and transport sectors. Such data could provide a clearer understanding of the underlying drivers of increasing electricity consumption and power demand, thereby improving both model interpretability and forecasting accuracy.

References

[1] Sveriges riksdag, *Klimatlag (2017:720)*, Stockholm, Jun. 22, 2017. Accessed: Mar. 16, 2026. [Online]. Available: https://www.riksdagen.se/sv/dokument-och-lagar/dokument/svensk-forfattningssamling/klimatlag-2017720_sfs-2017-720/.

[2] Naturvårdsverket, "Underlag till regeringens kommande klimathandlingsplan och klimatredovisning," Naturvårdsverket, Stockholm, Apr. 13, 2023. Accessed: Mar. 16, 2026. [Online]. Available: https://www.naturvardsverket.se/contentassets/4c414b0778e9409fb2836fc4d3dc6259/underlag-till-regeringens-kommande-klimathandlingsplan-och-klimatredovisning-2023-04-13.pdf.

[3] Naturvårdsverket, "Underlag till klimatredovisning 2022," Naturvårdsverket, Stockholm, Mar. 30, 2022. Accessed: Mar. 16, 2026. [Online]. Available: https://www.naturvardsverket.se/contentassets/caf14fb0008a41d29b9d51228f874fcb/underlag-klimatredovisning-2022.pdf.

[4] Vattenfall Eldistribution AB. "Elnätets uppbyggnad," Vattenfall Eldistribution, Accessed: Mar. 16, 2026. [Online]. Available: https://www.vattenfalleldistribution.se/var-verksamhet/om-elnatet/elnatets-uppbyggnad/.

[5] Svenska kraftnät, "Nätutvecklingsplan 2026–2035," Svenska kraftnät, Stockholm, 2025. Accessed: Mar. 16, 2026. [Online]. Available: https://www.svk.se/4af4e9/siteassets/om-oss/rapporter/natutvecklingsplanen-2026-2035/svk_natutveckling_nup_2026-2035.pdf.

[6] Svenska kraftnät, *Technology*, 2025. Accessed: Mar. 18, 2026. [Online]. Available: https://www.svk.se/en/grid-development/the-construction-process/technology/.

[7] J.-O. Helldin and M. Kågström, "Miljöeffekter av elnät: Förstudie," Naturvårdsverket, Mar. 2023. Accessed: Mar. 18, 2026. [Online]. Available: https://www.naturvardsverket.se/4acd13/contentassets/f703a57284f14d6faa0765c79018bea5/miljoeffekter-av-elnat-rapport-2023-03-03.pdf.

[8] Göteborgs Stad, "Vätgasens betydelse i energi- och klimatomställningen," Göteborgs Stad, N800 R 2023:10, 2023. Accessed: May 25, 2026. [Online]. Available: https://goteborg.se/wps/wcm/connect/a418c2b2-bb64-45ee-880c-54581ee5a6e2/N800_R_2023_10_V%C3%A4tgasens+betydelse+i+energi-+och+klimatomst%C3%A4llningen.pdf?MOD=AJPERES.

[9] Göteborg Energi, *Vad vi gör*, 2026. Accessed: Mar. 18, 2026. [Online]. Available: https://www.goteborgenergi.se/om-oss/vad-vi-gor.

[10] Svenska kraftnät, "Driftsäkerhet i kraftsystemet: Framtidens behov och förmågor," Svenska kraftnät, Aug. 2025. [Online]. Available: https://www.svk.se/4a5f32/siteassets/om-oss/rapporter/2025/rapport_driftsakerhet_i_kraftsystemet_augusti_2025.pdf.

[11] Göteborg Stad. "Statistikdatabas," Accessed: Jan. 30, 2026. [Online]. Available: https://statistikdatabas.goteborg.se/pxweb/sv/.

[12] SMHI - Sveriges meteorologiska och hydrologiska institut. "Ladda ner väderobservationer: Lufttemperatur (instant) – station 71420," Accessed: May 4, 2026. [Online]. Available: https://www.smhi.se/data/hitta-data-for-en-plats/ladda-ner-vaderobservationer/airtemperatureInstant/71420.

[13] NEPP - North European Power Perspectives, "Elanvändningen i sverige 2030 och 2050," NEPP, Rapport, 2015. Accessed: Mar. 19, 2026. [Online]. Available: https://www.nepp.se/etapp1/pdf/Elanv%C3%A4ndningen_i_Sverige_2030_2050.pdf.

[14] Statistiska centralbyrån (SCB). "Statistikdatabasen," Accessed: Mar. 19, 2026. [Online]. Available: https://www.statistikdatabasen.scb.se/pxweb/sv/ssd/.

[15] Power Circle. "Elbilsstatistik," Accessed: Mar. 19, 2026. [Online]. Available: https://powercircle.org/elbilsstatistik/.

[16] A. L. Porter, *Forecasting and management of technology*. John Wiley & Sons, 1991, vol. 18.

[17] Trafikanalys, "Fordon 2025," Trafikanalys, 2026. Accessed: Apr. 23, 2026. [Online]. Available: https://www.trafa.se/globalassets/statistik/vagtrafik/fordon/2026/fordon-2025.pdf.

[18] S. B. Mcgrayne, *The Theory That Would Not Die: How Bayes' Rule Cracked the Enigma Code, Hunted Down Russian Submarines, and Emerged Triumphant from Two Centuries of Controversy*. Yale University Press, 2011, ISBN: 9780300169690. Accessed: May 6, 2026. [Online]. Available: http://www.jstor.org/stable/j.ctt1np76s.

[19] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin, *Bayesian Data Analysis*, 3rd ed. CRC Press, 2013.

[20] D. Ríos Insua, F. Ruggeri, and M. P. Wiper, *Bayesian Analysis of Stochastic Process Models*. Hoboken, NJ: John Wiley & Sons, 2012, ISBN: 9780470744536. DOI: 10.1002/9780470975916. [Online]. Available: https://ebookcentral.proquest.com/lib/chalmers/detail.action?docID=877779.

[21] P. J. Brockwell and R. A. Davis, *Time Series: Theory and Methods* (Springer Series in Statistics), 2nd ed. New York, NY: Springer, 1991, ISBN: 978-0-387-97429-3. DOI: 10.1007/978-1-4419-0320-4.

[22] P. A. Lang. A, *Financial time series*, 2022.

[23] J. G. De Gooijer and R. J. Hyndman, "25 years of time series forecasting," *International Journal of Forecasting*, vol. 22, no. 3, pp. 443–473, 2006. DOI: 10.1016/j.ijforecast.2006.01.001.

[24] C. H. Weiß, "Time series modelling," *Entropy*, vol. 23, no. 9, p. 1163, 2021. DOI: 10.3390/e23091163.

[25] H. Tyralis and G. Papacharalampous, "A review of predictive uncertainty estimation with machine learning," *Artificial Intelligence Review*, vol. 57, no. 4, p. 94, 2024. DOI: 10.1007/s10462-023-10698-8.

[26] J. Kim, H. Kim, et al., "A comprehensive survey of deep learning for time series forecasting: Architectural diversity and open challenges," *Artificial Intelligence Review*, 2025. DOI: 10.1007/s10462-025-11223-9.

[27] Z. Liu, Z. Zhang, and W. Zhang, "A hybrid framework integrating traditional models and deep learning for multi-scale time series forecasting," *Entropy*, vol. 27, no. 7, p. 695, 2025. DOI: 10.3390/e27070695.

[28] statsmodels developers, *Sarimax: Introduction*, https://www.statsmodels.org/stable/examples/notebooks/generated/statespace_sarimax_stata.html, statsmodels 0.14.6, accessed April 2026, 2024.

[29] M. D. Hoffman and A. Gelman, "The No-U-Turn Sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo," *Journal of Machine Learning Research*, vol. 15, pp. 1593–1623, 2014. [Online]. Available: https://jmlr.org/papers/v15/hoffman14a.html.

[30] A. Durmus, S. Gruffaz, M. Kailas, E. Saksman, and M. Vihola, *On the convergence of dynamic implementations of Hamiltonian Monte Carlo and No-U-Turn Samplers*, Jul. 2023. [Online]. Available: https://arxiv.org/abs/2307.03460.

[31] D. Salinas, V. Flunkert, J. Gasthaus, and T. Januschowski, "Deepar: Probabilistic forecasting with autoregressive recurrent networks," *International Journal of Forecasting*, vol. 36, no. 3, pp. 1181–1191, 2020. DOI: 10.1016/j.ijforecast.2019.07.001.

[32] XGBoost developers, *Introduction to boosted trees*, https://xgboost.readthedocs.io/en/release_3.2.0/tutorials/model.html, XGBoost 3.2.1 documentation, accessed April 2026, 2025.

[33] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016, http://www.deeplearningbook.org.

[34] B. Mehlig, *Machine Learning with Neural Networks: An Introduction for Scientists and Engineers*. Cambridge University Press, 2021.

[35] P. J. Brockwell and R. A. Davis, *Introduction to Time Series and Forecasting* (Springer Texts in Statistics), 3rd ed. Cham, Switzerland: Springer, 2016, ISBN: 978-3-319-29852-8. DOI: 10.1007/978-3-319-29854-2.

[36] R. H. Shumway and D. S. Stoffer, *Time Series Analysis and Its Applications: With R Examples* (Springer Texts in Statistics), 4th. Cham: Springer, 2017, ISBN: 978-3-319-52451-1. DOI: 10.1007/978-3-319-52452-8.

[37] C. W. J. Granger, "Investigating causal relations by econometric models and cross-spectral methods," *Econometrica*, vol. 37, no. 3, pp. 424–438, 1969. [Online]. Available: https://www.jstor.org/stable/1912791.

[38] PyMC Development Team, *pymc.sample — PyMC 6.0.0 documentation*, Accessed: May 17, 2026, 2025. [Online]. Available: https://www.pymc.io/projects/docs/en/stable/api/generated/pymc.sample.html.

[39] D. C. Montgomery and G. C. Runger, *Applied Statistics and Probability for Engineers*, 6th. Hoboken, NJ: Wiley, 2014.

[40] R. J. Hyndman and G. Athanasopoulos, *Forecasting: Principles and Practice*, 3rd. Melbourne, Australia: OTexts, 2021, Accessed: May 16, 2026. [Online]. Available: https://otexts.com/fpp3.

A

Appendix - Theory

A.1 Fundamental Definitions

What follows in this chapter are a few fundamental definitions left out of the theory chapter.

Definition 13. The process $\{Z_t\}$ is considered to be *white noise* with mean 0 and variance σ^2, also written as

$$\{Z_t\} \sim WN(0, \sigma^2),$$

if and only if $\{Z_t\}$ has zero mean and autocovariance function

$$\gamma(h) = \begin{cases} \sigma^2 & \text{if } h = 0, \\ 0 & \text{if } h \neq 0. \end{cases}$$

[21]

Definition 14. The Pearson's correlation coefficient ρ_{X_t,Y_t} between two time series X_t and Y_t is expressed as

$$\rho_{X_t,Y_t} = \frac{\mathrm{Cov}(X_t, Y_t)}{\sigma_{X_t}\sigma_{Y_t}},$$

where σ_{X_t} and σ_{Y_t} are standard deviations of the two time series, and $\mathrm{Cov}(\cdot)$ is the covariance between them.

[39]

Definition 15. A *forecast error* e_{T+h} at a future time point h is the difference

$$e_{T+h} = y_{T+h} - \hat{y}_{T+h|T},$$

between an observed value y_{T+h} and its forecast $\hat{y}_{T+h|T}$. The training data is given by $\{y_1, \ldots, y_T\}$ and the test data is given by $\{y_{T+1}, y_{T+2}, \ldots\}$.

Note that residuals are given by the same equation as in Definition 9, but calculated on the training set instead. The measures MAE, RMSE and MAPE are defined below using the concepts from Definition 15.

Definition 16. The *mean absolute error (MAE)* of a forecast with horizon h is expressed as

$$MAE = \frac{1}{h}\sum_{t=T+1}^{T+h}|e_t|.$$

Definition 17. The *root mean squared error (RMSE)* of a forecast with horizon h is defined by

$$RMSE = \sqrt{\frac{1}{h}\sum_{t=T+1}^{T+h}e_t^2}$$

Definition 18. The *mean absolute percentage error (MAPE)* is given by

$$MAPE = \frac{1}{h}\sum_{t=T+1}^{T+h}100\frac{e_t}{y_t},$$

for a forecast with horizon h.

[40]

- Härledning av Bayes sats?

A.2 Visualizations of Data

A.2.1 Visualization of D_M

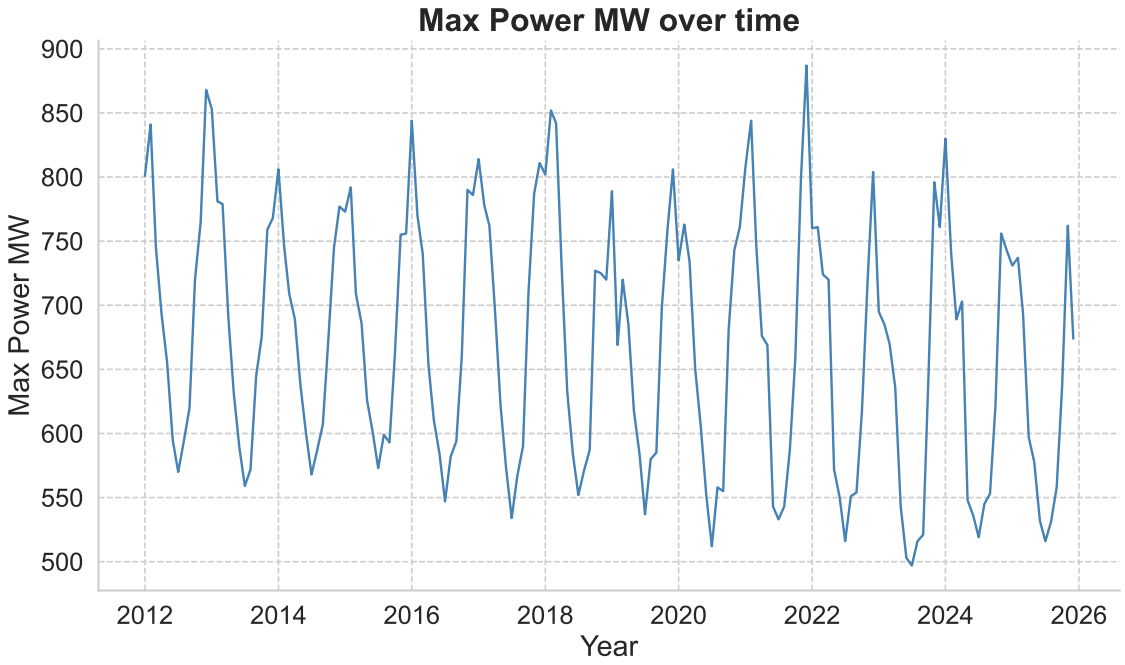

Figure A.1: Figure of power peak data.

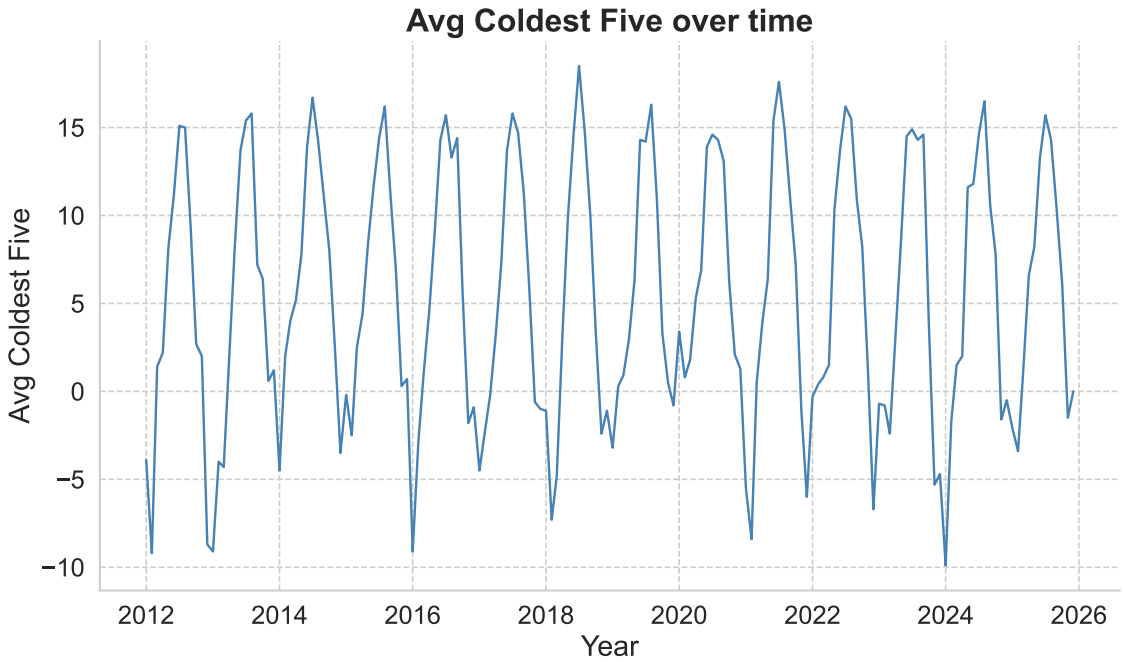

Figure A.2: Figure of data for average coldest five variable.

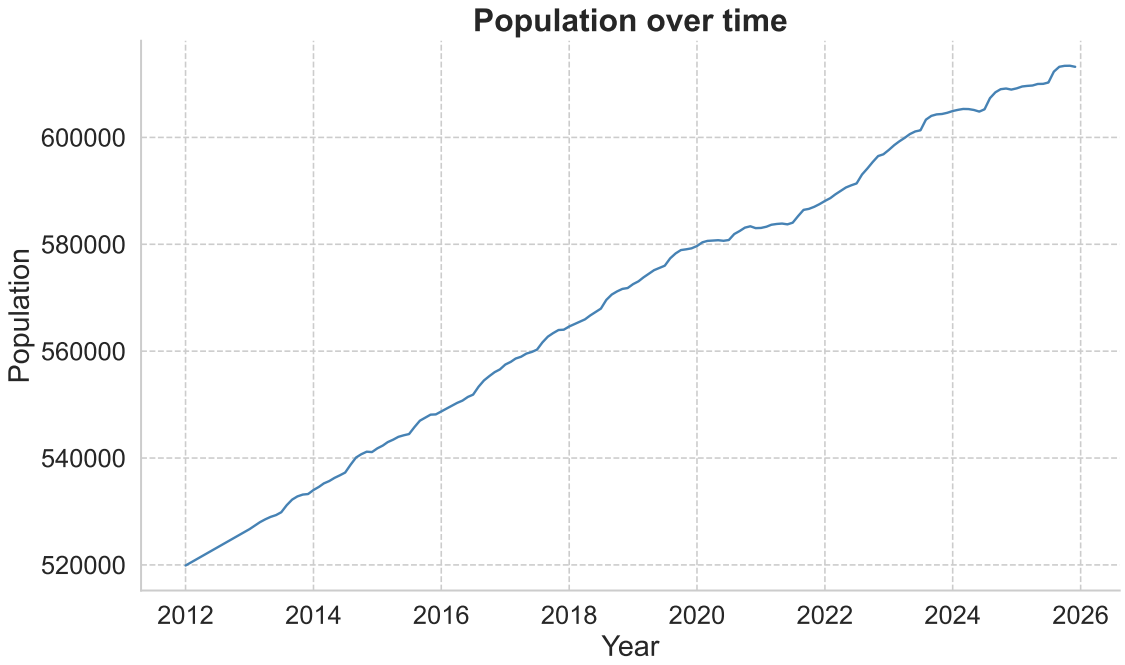

Figure A.3: Figure of data for population variable.

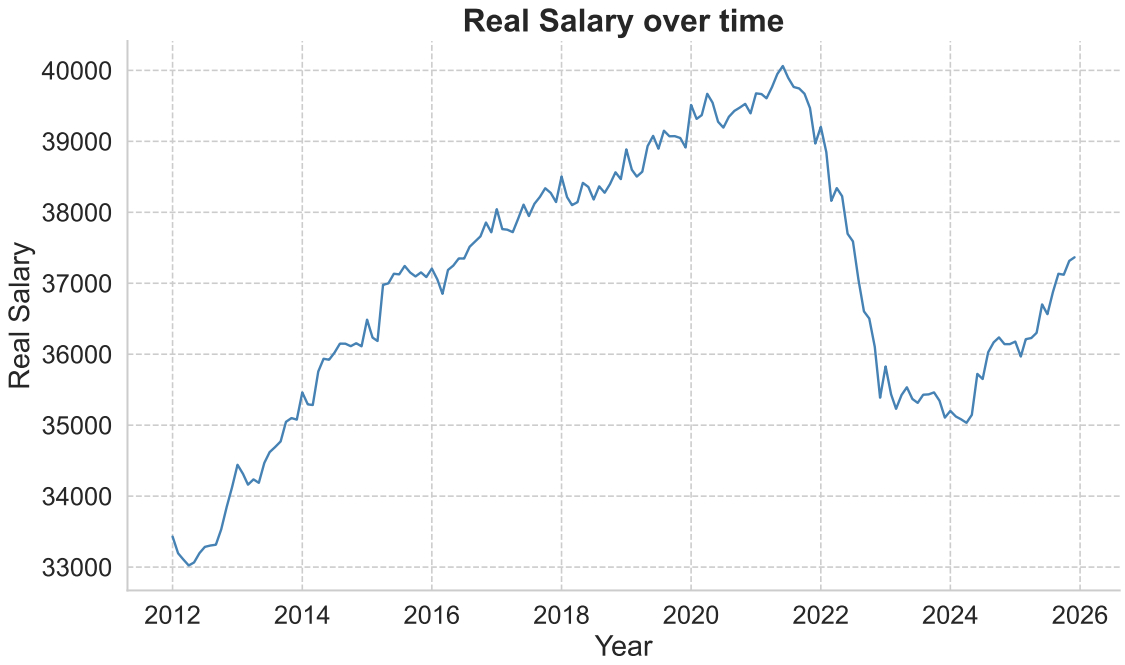

Figure A.4: Figure of data for real salary variable.

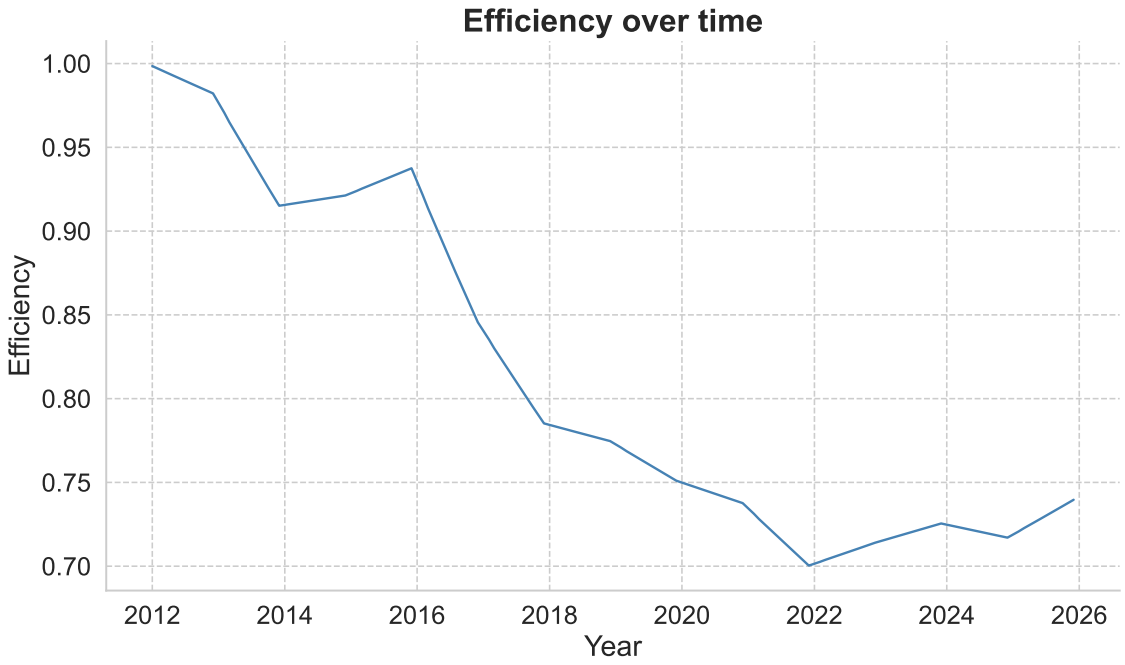

Figure A.5: Figure of data for efficiency variable.

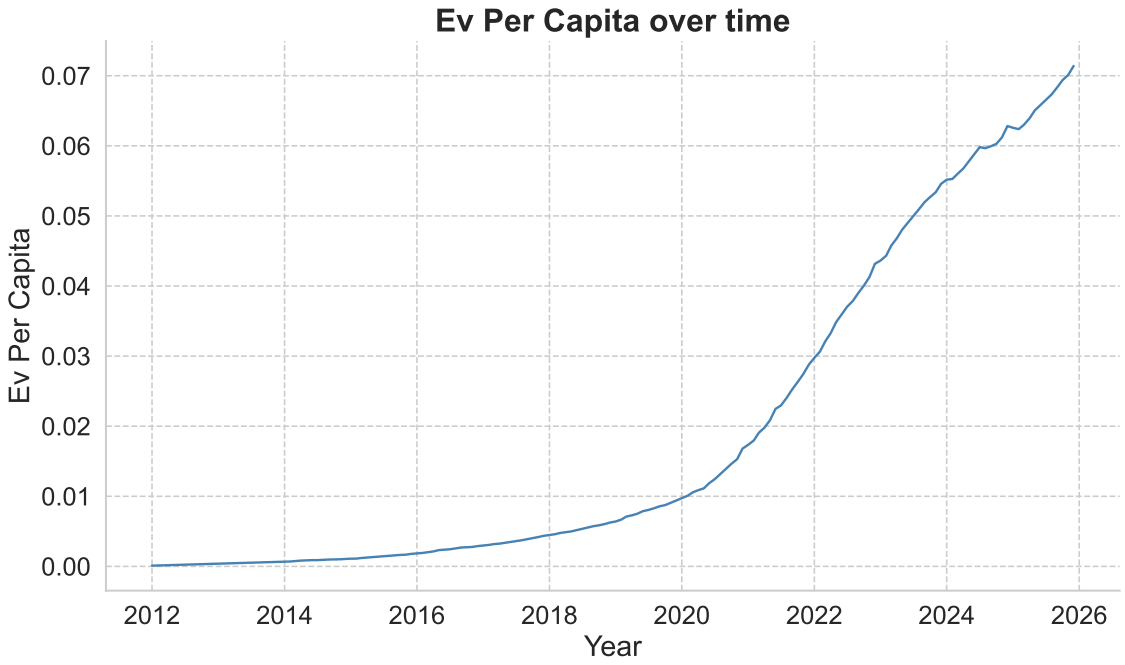

Figure A.6: Figure of data for EV per capita variable. Note that for the D_M data set the EV per capita is calculated for Gothenburg.

A.2.2 Visualization of D_Y

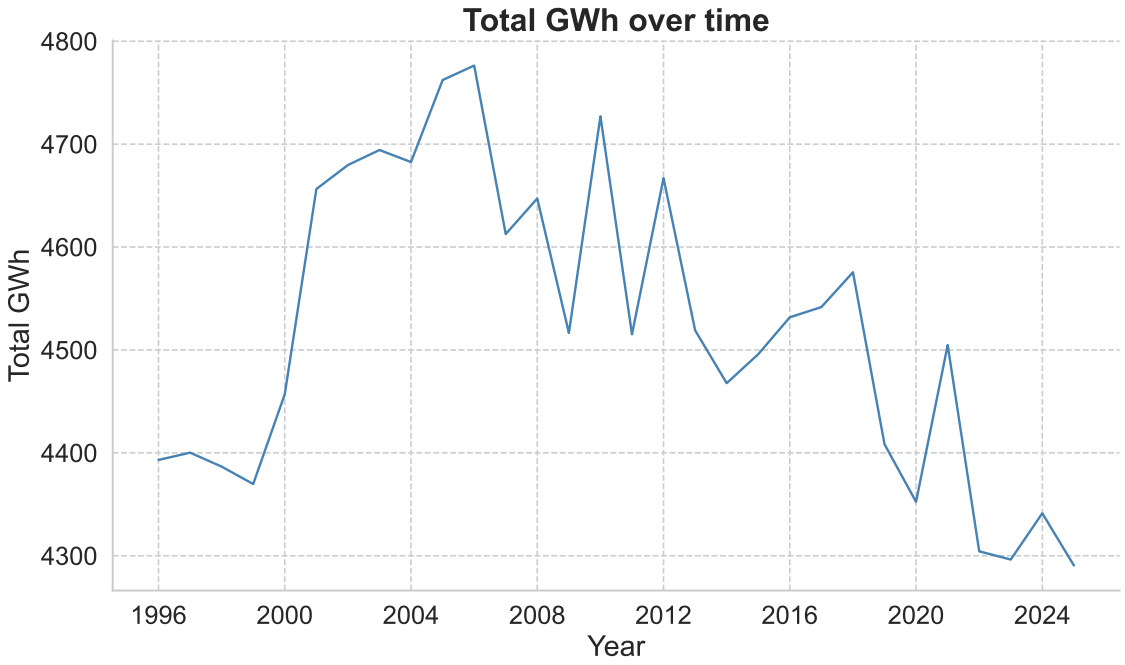

Figure A.7: Figure of energy consumption data.

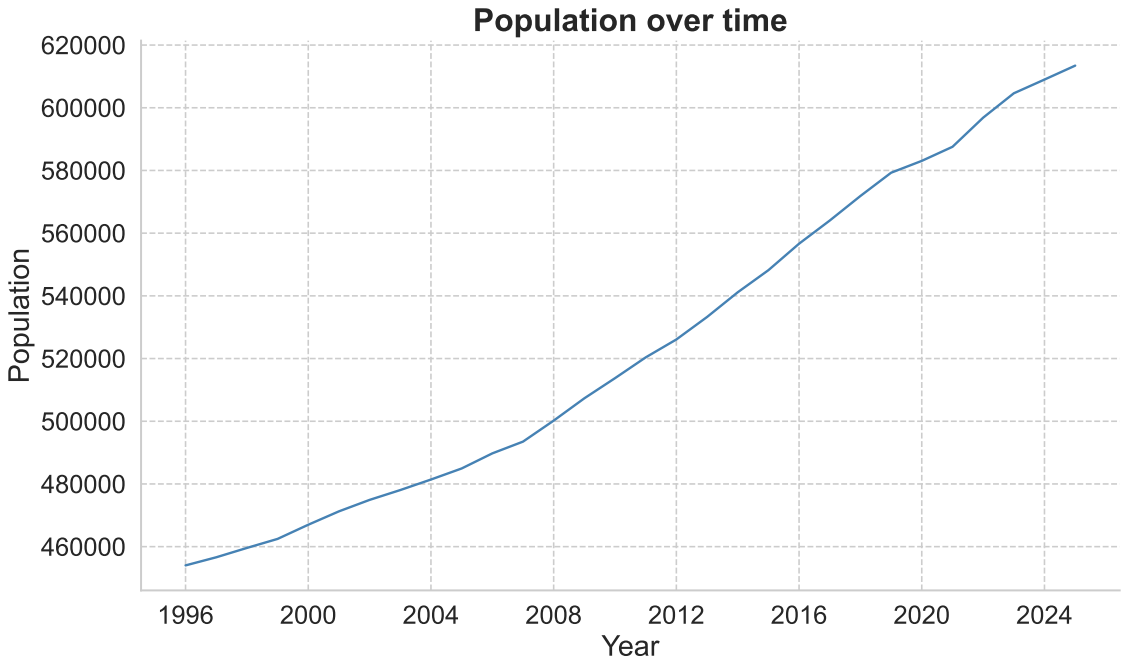

Figure A.8: Figure of data for population variable.

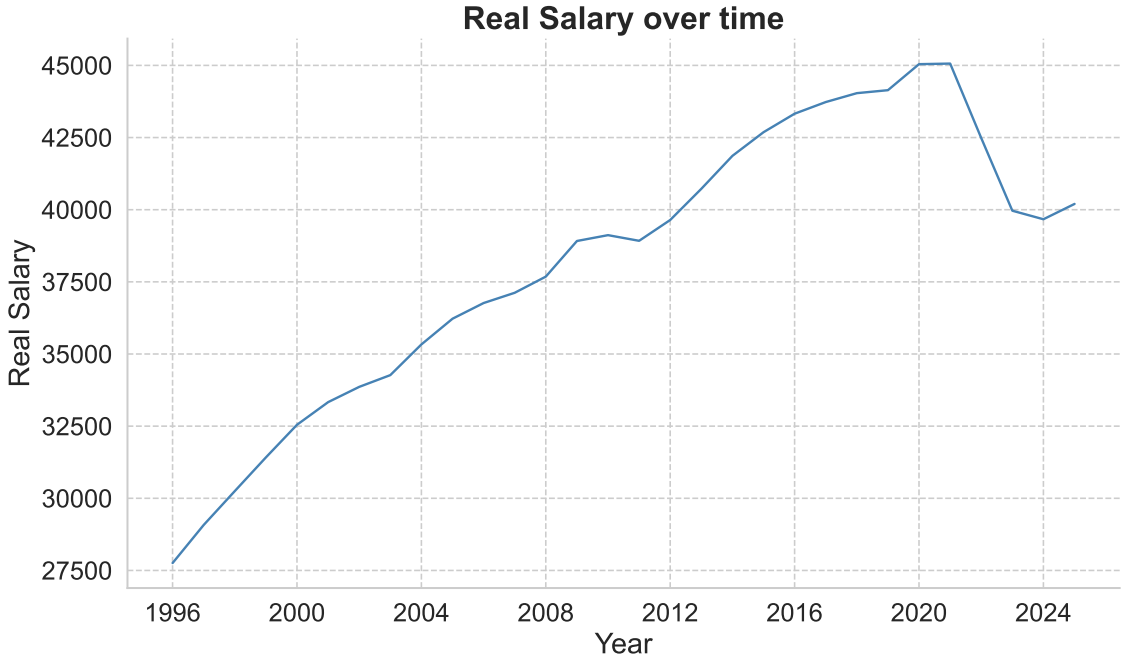

Figure A.9: Figure of data for real salary variable. Real salary have been calculated with a CPI which is an average over the year.

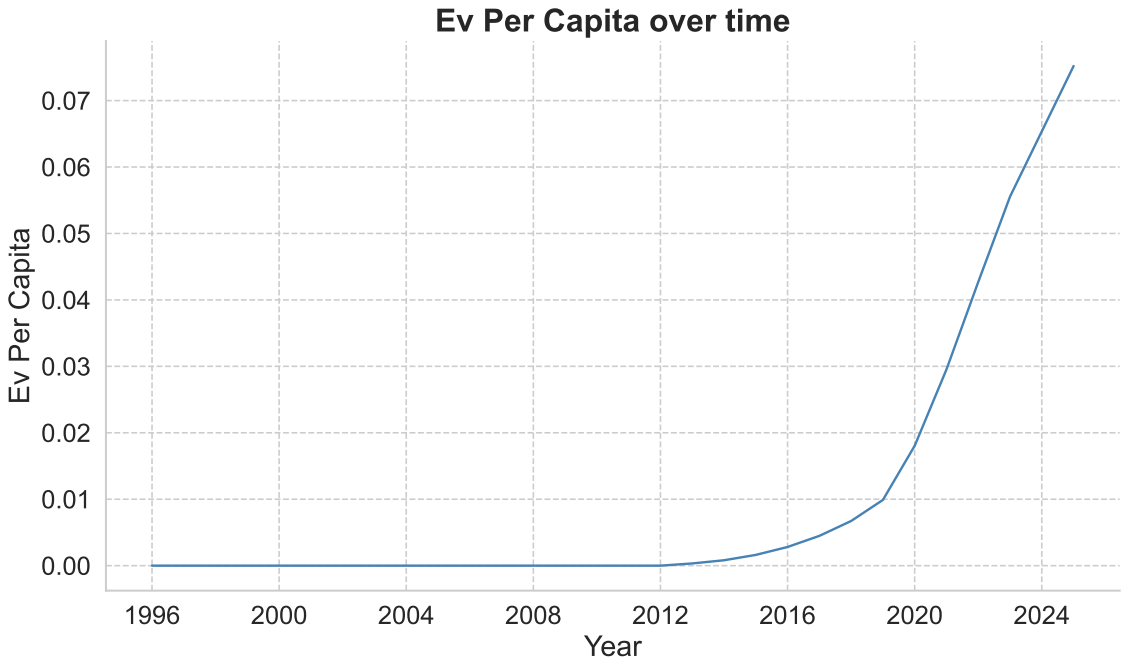

Figure A.10: Figure of data for EV per capita variable. Note that for the D_Y data set the EV per capita is calculated on a national level.

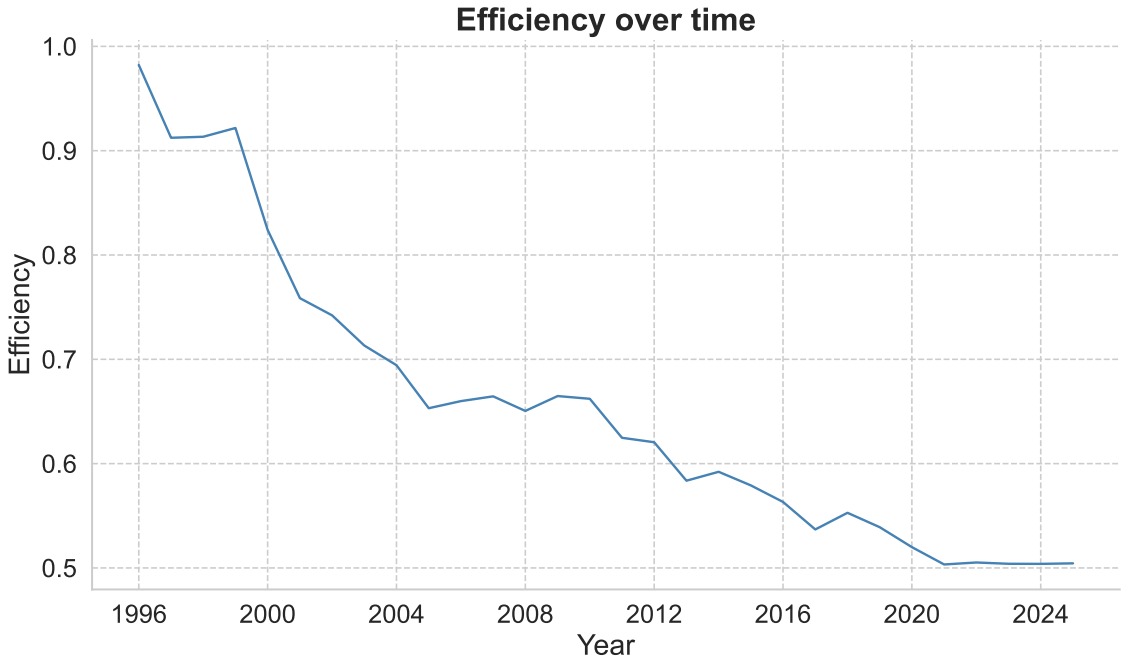

Figure A.11: Figure of data for efficiency variable.

B

Appendix - Method

This Appendix contains short explanations of the hyperparameters of XGBoost and DeepAr models.

XGBoost

The XGBoost models were tuned using several key hyperparameters that control model complexity, learning behavior, and the risk of overfitting.

The `learning rate` (often denoted in text as η) determines the step size used during the boosting process. In XGBoost, the parameter η reduces the contribution of each individual tree, making the learning process more conservative. Lower learning rates generally improve model robustness and reduce the risk of overfitting, although they often require a larger number of trees to achieve good predictive performance. The parameter is constrained to values within the interval $[0, 1]$.

The `max_depth` parameter specifies the maximum depth of each decision tree. Lower values produce shallower trees with reduced model complexity, while higher values allow the model to learn more complex relationships within the data. However, deeper trees increase the risk of overfitting, particularly when the dataset is limited in size. The parameter takes integer values within the range $[0, \infty)$.

The `n_estimators` parameter defines the total number of trees included in the ensemble. Increasing the number of trees generally increases the model's ability to capture patterns in the data, but it also leads to a more complex model and may contribute to overfitting if not properly regularized.

The `min_child_weight` parameter controls the minimum sum of instance weights required in a child node before further partitioning is allowed. Larger values make the model more conservative by preventing the creation of nodes containing only a small number of observations. If the sum of weights in a potential child node falls below the specified threshold, the partitioning process stops and the node remains a leaf node. This parameter takes values within the interval $[0, \infty)$.

The `subsample` parameter specifies the proportion of the training dataset randomly

sampled before growing each tree. For example, a value of 0.5 means that half of the training observations are randomly selected for each boosting iteration. Subsampling introduces randomness into the training process, which can improve generalization performance and reduce overfitting. The parameter must take values within the interval $(0, 1]$.

Finally, the `colsample_bytree` parameter determines the proportion of features randomly sampled for each tree. Similar to row subsampling, feature subsampling reduces the correlation between trees and can improve model robustness and generalization performance.

DeepAR

The DeepAR models were configured using several important hyperparameters that influence the training process, model stability, and generalization performance.

The `learning rate` controls how much the model weights are adjusted during each optimization step in the training process. A higher learning rate allows the model to learn more quickly but may lead to unstable training or convergence issues, while a lower learning rate generally results in a more stable optimization process at the cost of longer training times. Selecting an appropriate learning rate is therefore important for balancing convergence speed and model performance.

The `dropout` parameter specifies the proportion of neurons that are randomly deactivated in the hidden layers during training. Dropout acts as a regularization technique by preventing the model from relying too heavily on specific neurons or learned patterns. This reduces the risk of overfitting and improves the model's ability to generalize to unseen data, which is particularly important when training deep learning models on relatively small datasets. Higher dropout values increase regularization strength, whereas lower values allow the model to retain more of its learned complexity.

C

Appendix - Results

This Appendix consists of additional figures of the result along with small analysis of residuals.

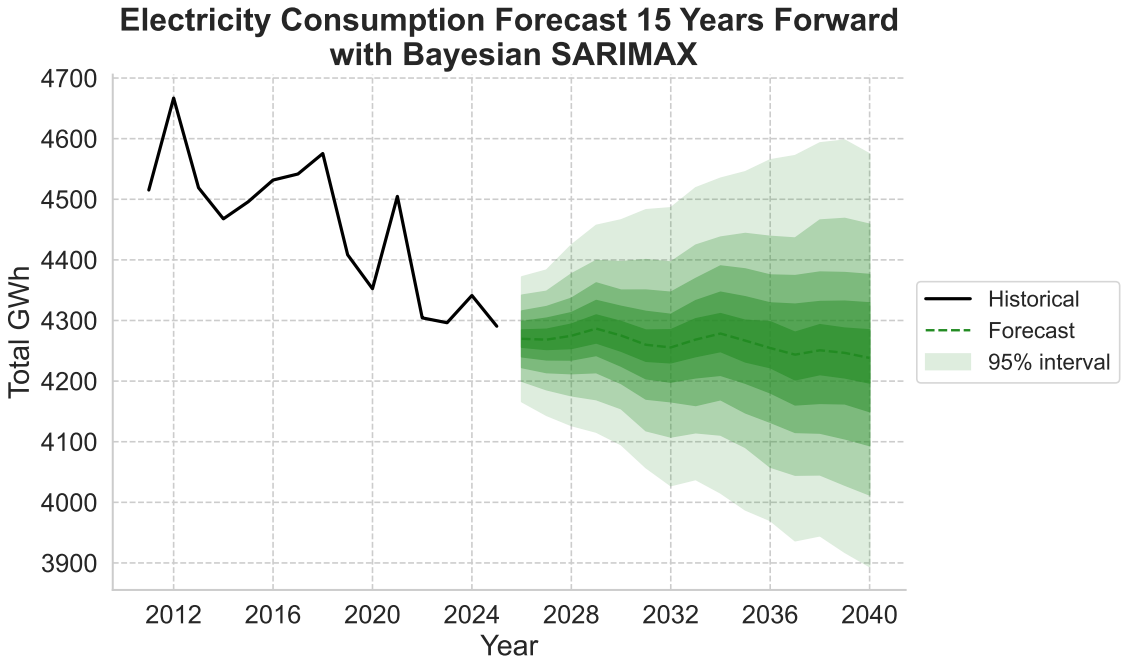

Figure C.1: Forecast of energy consumption for 15 years with Bayesian SARIMAX. Note that the forecast is shown as a distribution of likely values.

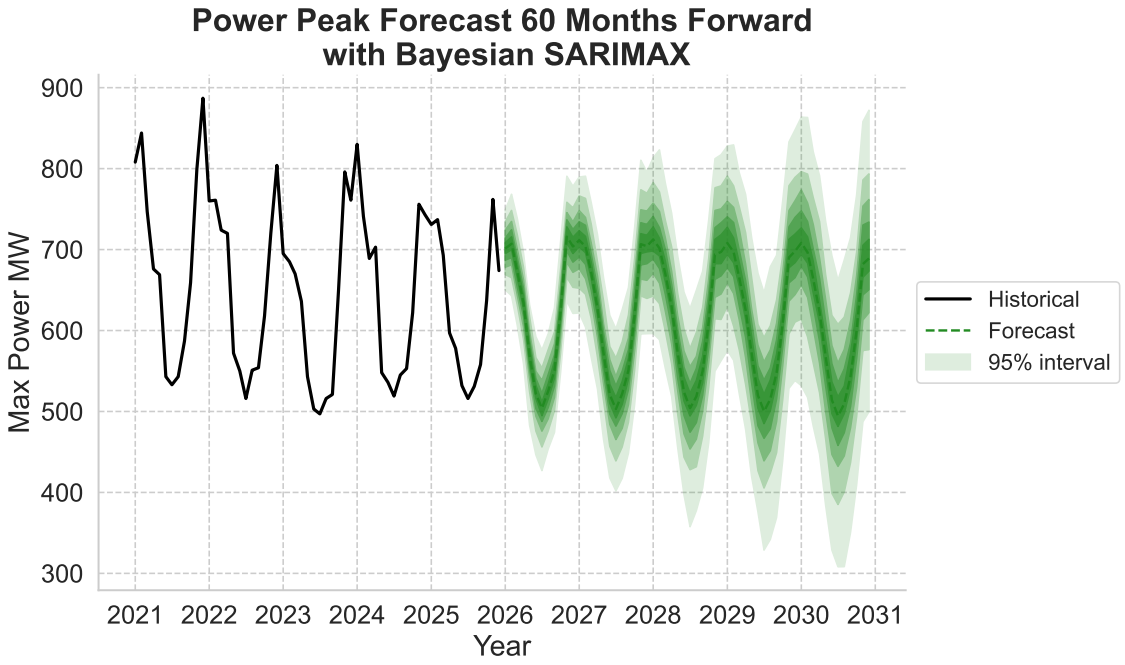

Figure C.2: Forecast of power peaks for 60 months with Bayesian SARIMAX. Note that the forecast is shown as a distribution of likely values.

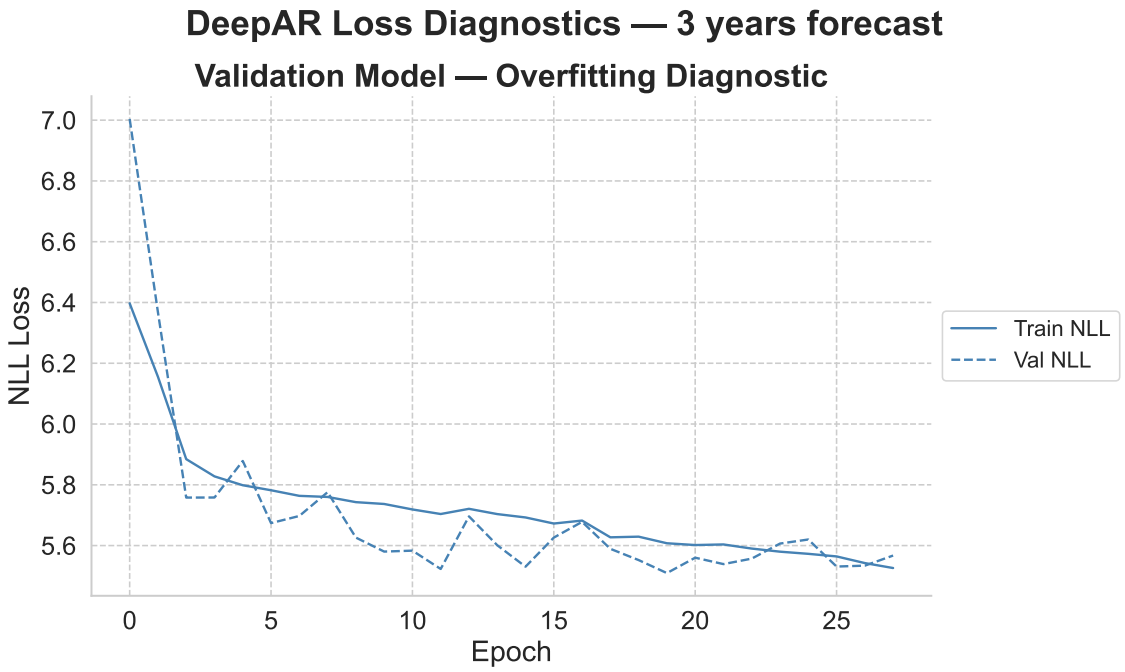

Figure C.3: NLL loss for DeepAR with 3 years forecast. The NLL loss for the DeepAR model with a 3-year forecasting horizon indicates comparatively better performance than the remaining DeepAR configurations. In this scenario, the validation loss follows the training loss more closely and exhibits less pronounced divergence, suggesting a lower degree of overfitting and improved generalization ability relative to the 5- and 10-year forecasting horizons.

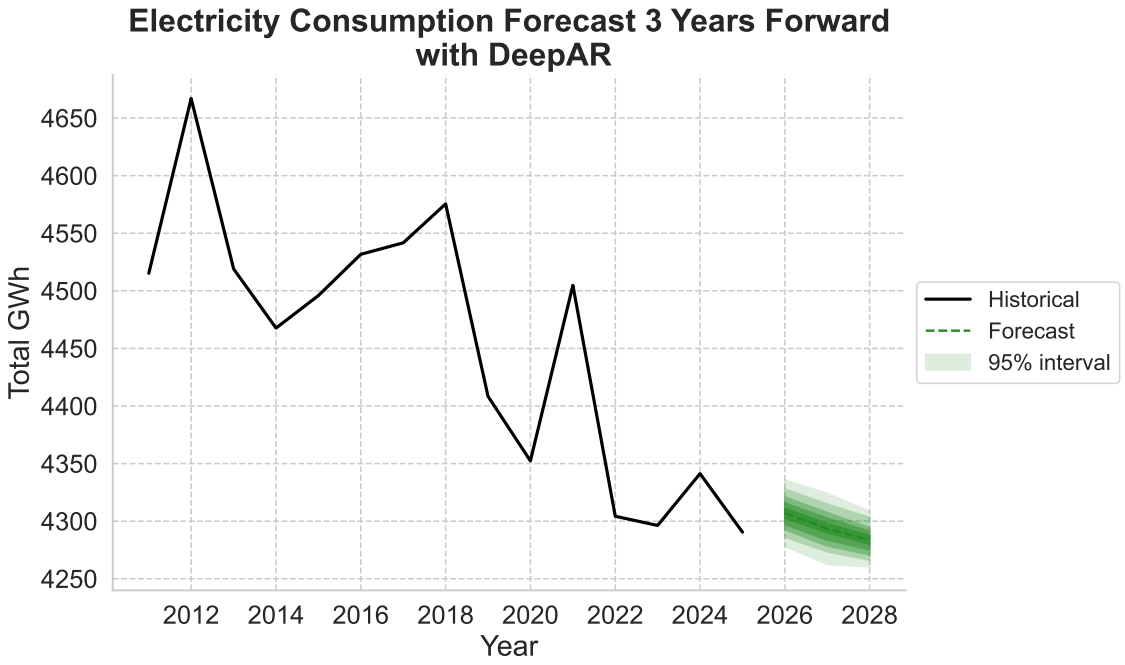

Figure C.4: Forecast of energy consumption for 3 years with DeepAR. Note that the forecast is shown as a distribution of likely values.

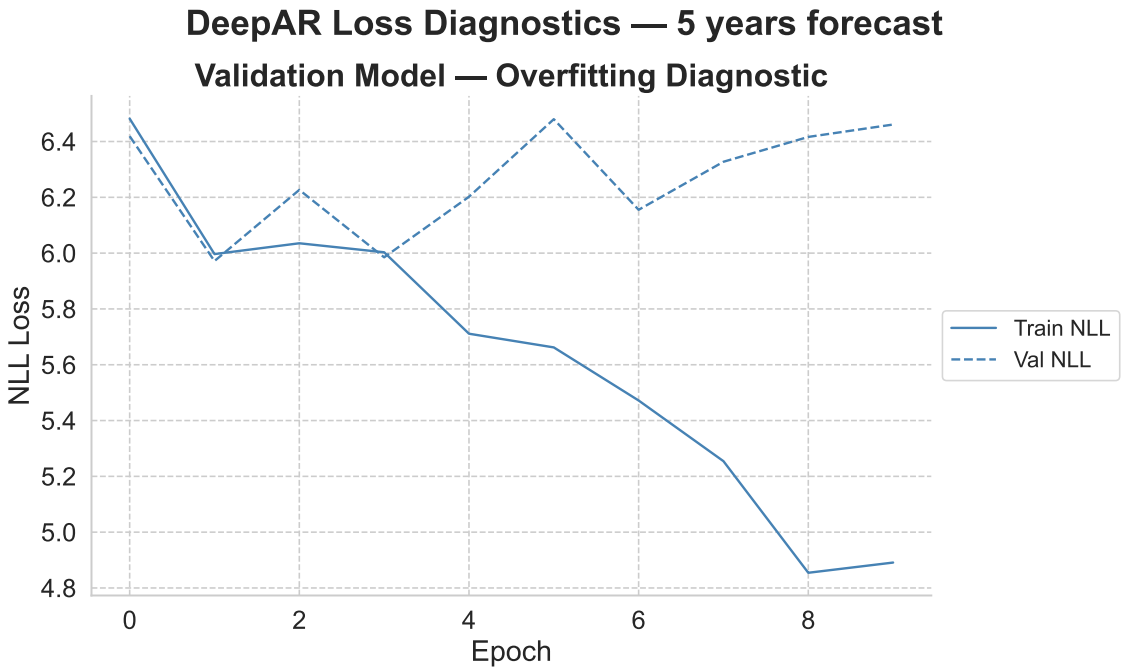

Figure C.5: NLL loss for DeepAR with 5 years forecast.

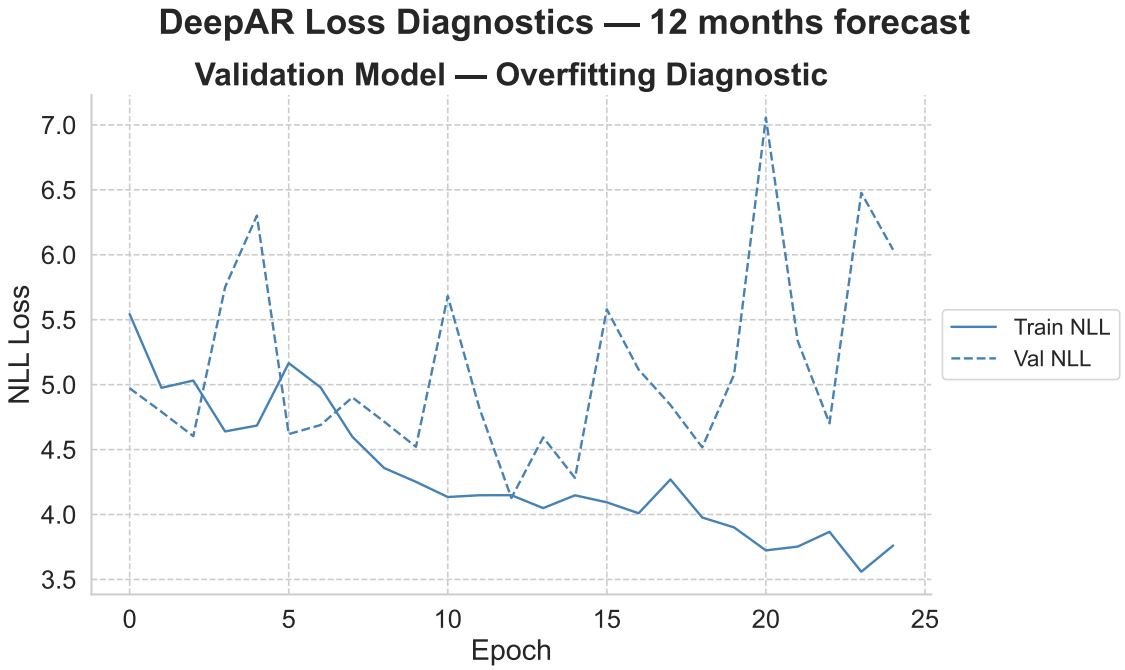

Figure C.6: NLL loss for DeepAR with 12 months forecast.

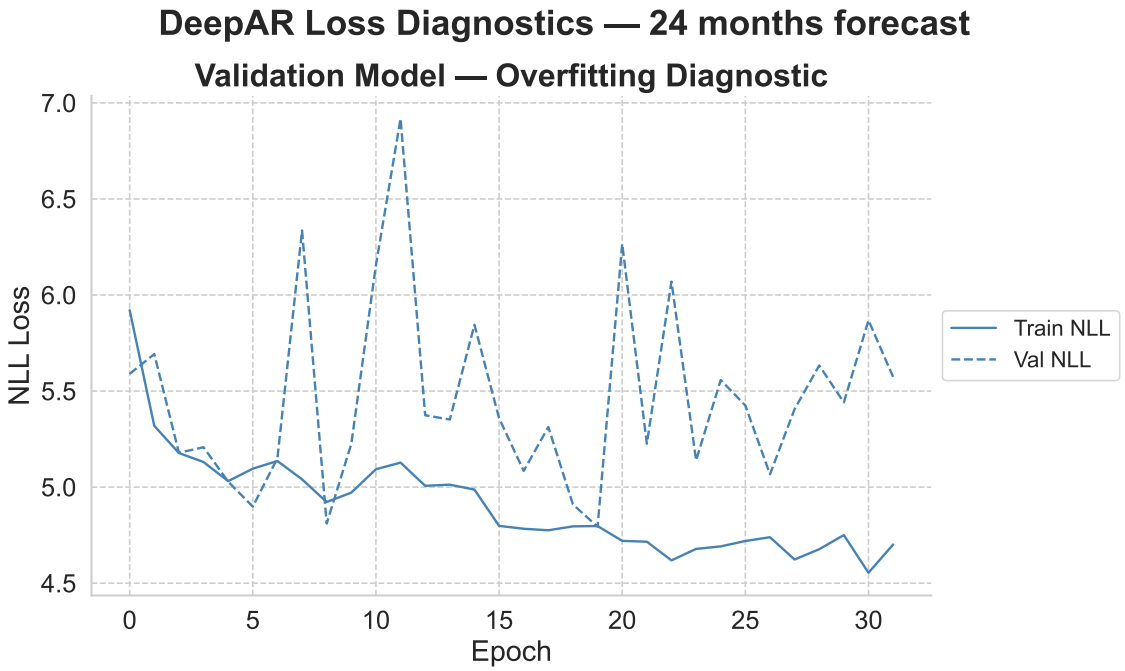

Figure C.7: NLL loss for DeepAR with 24 months forecast.

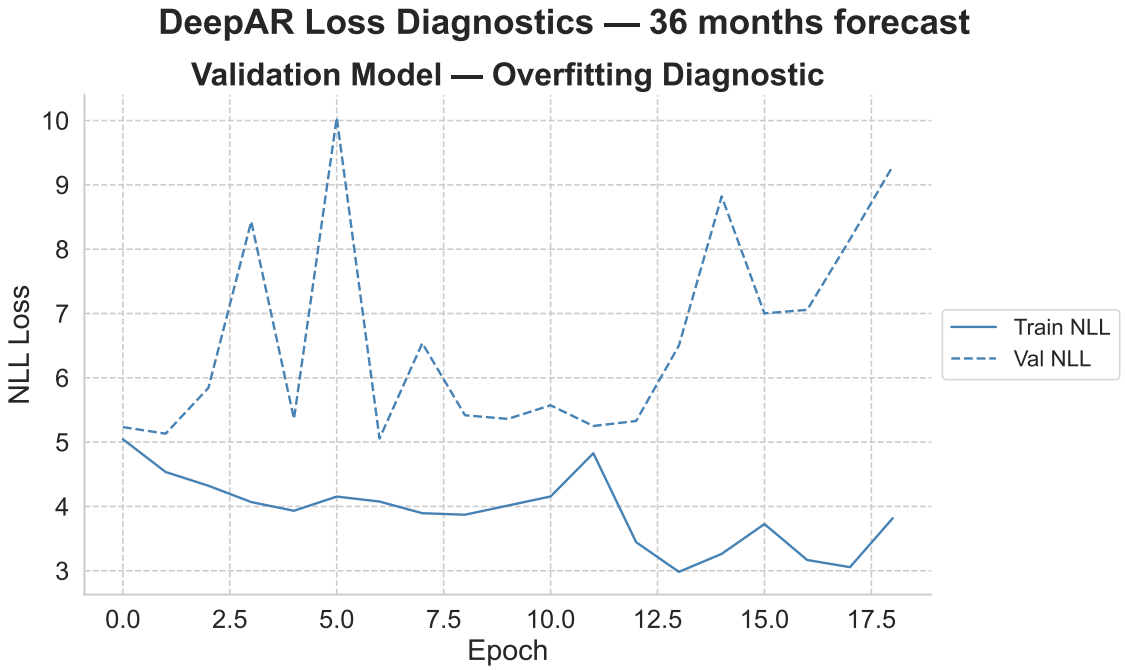

Figure C.8: NLL loss for DeepAR with 36 months forecast.

Forecast length (Months)	Type of error	Train error	Validation error
12	RMSE	23.51	28.55
12	MAE	19.11	18.91
24	RMSE	23.44	22.05
24	MAE	19.20	18.39
36	RMSE	23.36	31.60
36	MAE	18.64	26.45
60	RMSE	24.03	338.21
60	MAE	19.45	303.03

Table C.1: Table of errors for the SARIMAX model with D_M.

Forecast length (Months)	Type of error	Train error	Validation error
12	RMSE	46.38	44.20
12	MAE	34.85	31.74
24	RMSE	23.87	24.73
24	MAE	19.41	16.54
36	RMSE	23.95	23.33
36	MAE	19.29	16.58
60	RMSE	24.02	25.70
60	MAE	19.70	19.21

Table C.2: Table of errors for the SARIMA model with D_M.

Forecast length (Months)	Type of error	Train error	Validation error
12	RMSE	24.74	27.26
12	MAE	20.05	23.67
24	RMSE	24.61	28.18
24	MAE	19.57	23.42
36	RMSE	25.04	26.88
36	MAE	19.92	21.85
60	RMSE	25.27	59.63
60	MAE	19.70	50.25

Table C.3: Table of errors for the Bayesian SARIMAX model with D_M.

Forecast length (Months)	Type of error	Train error	Validation error
12	RMSE	3.44	25.33
12	MAE	2.48	14.92
24	RMSE	15.41	20.12
24	MAE	11.59	15.56
36	RMSE	14.04	46.75
36	MAE	10.73	38.52
60	RMSE	15.13	42.13
60	MAE	11.11	33.88

Table C.4: Table of errors for the XGBoost model with D_M.

Forecast length (Months)	Type of error	Validation error
12	RMSE	28.73
12	MAE	19.16
24	RMSE	32.10
24	MAE	25.93
36	RMSE	61.05
36	MAE	47.14
60	RMSE	43.50
60	MAE	34.48

Table C.5: Table of errors for the DeepAR model with D_M, only validation error is implemented with RMSE and MAE due to model configurations.

Forecast length (Years)	Type of error	Train error	Validation error
3	RMSE	106.84	136.73
3	MAE	90.13	129.78
5	RMSE	99.93	182.61
5	MAE	83.64	174.56
10	RMSE	100.65	200.26
10	MAE	87.85	161.58
15	RMSE	93.33	242.61
15	MAE	72.84	219.43

Table C.6: Table of errors for the AR model with D_Y.

Forecast length (Years)	Type of error	Train error	Validation error
3	RMSE	88.47	74.01
3	MAE	74.25	60.10
5	RMSE	88.135	101.35
5	MAE	74.43	91.51
10	RMSE	82.60	96.09
10	MAE	66.01	81.83
15	RMSE	72.44	619.50
15	MAE	59.98	532.73

Table C.7: Table of errors for the SARIMAX model with D_Y.

Forecast length (Years)	Type of error	Train error	Validation error
3	RMSE	91.14	49.27
3	MAE	73.37	43.74
5	RMSE	84.73	440.83
5	MAE	67.25	416.88
10	RMSE	84.04	2220.58
10	MAE	67.44	1681.96
15	RMSE	83.29	482.23
15	MAE	54.95	444.34

Table C.8: Table of errors for the Bayesian SARIMAX model with D_Y.

Forecast length (Years)	Type of error	Train error	Validation error
3	RMSE	43.90	164.92
3	MAE	33.80	158.69
5	RMSE	5.11	170.73
5	MAE	3.64	163.91
10	RMSE	4.40	151.20
10	MAE	3.34	128.04
15	RMSE	0.95	220.09
15	MAE	0.68	191.59

Table C.9: Table of errors for the XGBoost model with D_Y.

Forecast length (Years)	Type of error	Validation error
3	RMSE	44.08
3	MAE	34.24
5	RMSE	119.89
5	MAE	102.89
10	RMSE	71.01
10	MAE	59.10

Table C.10: Table of errors for the DeepAR model with D_Y, only validation error is implemented with RMSE and MAE due to model configurations.

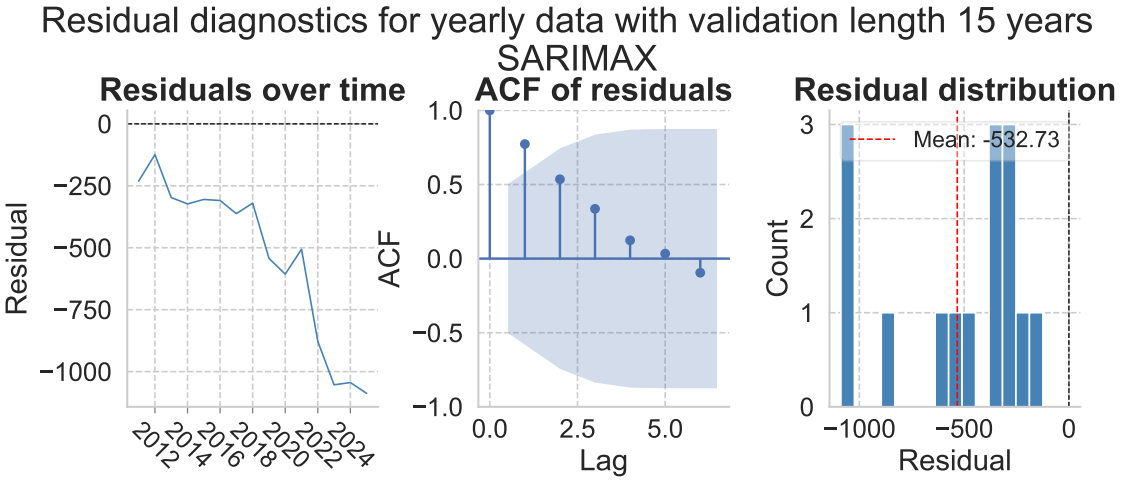

Figure C.9: Figures of residual diagnostics of the SARIMAX model for 15 years of validation.

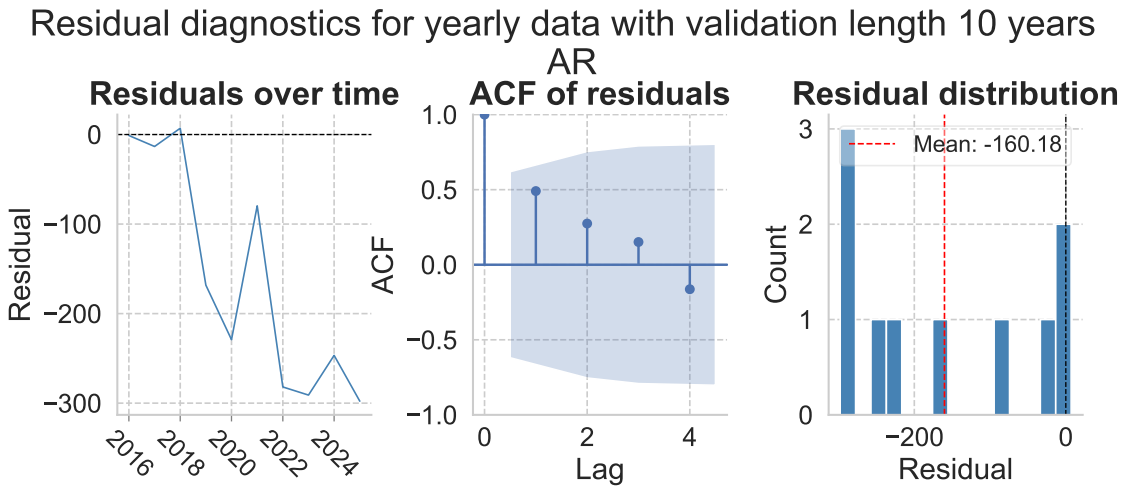

Figure C.10: Figures of residual diagnostics of the AR model for 10 years of validation.

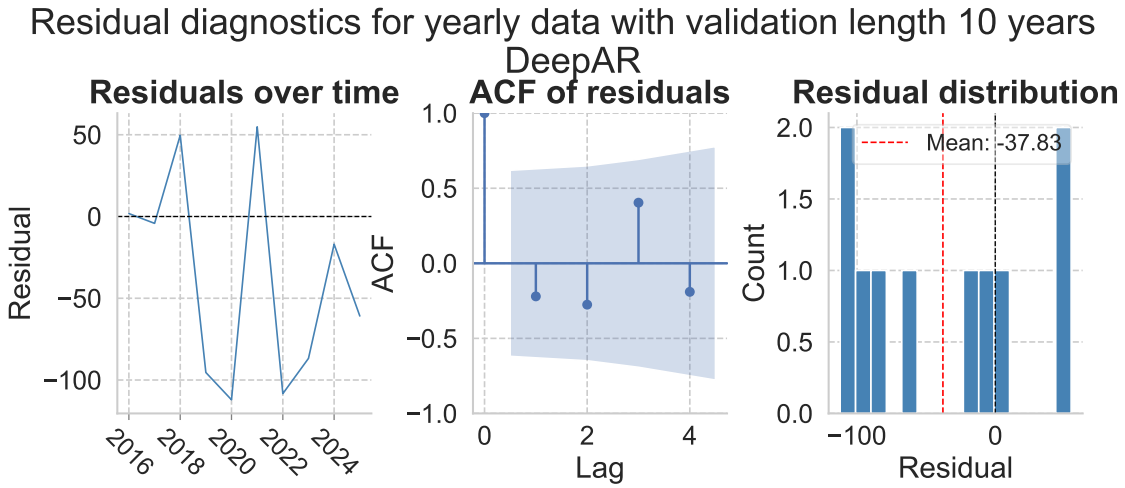

Figure C.11: Figures of residual diagnostics of the DeepAR model for 10 years of validation.

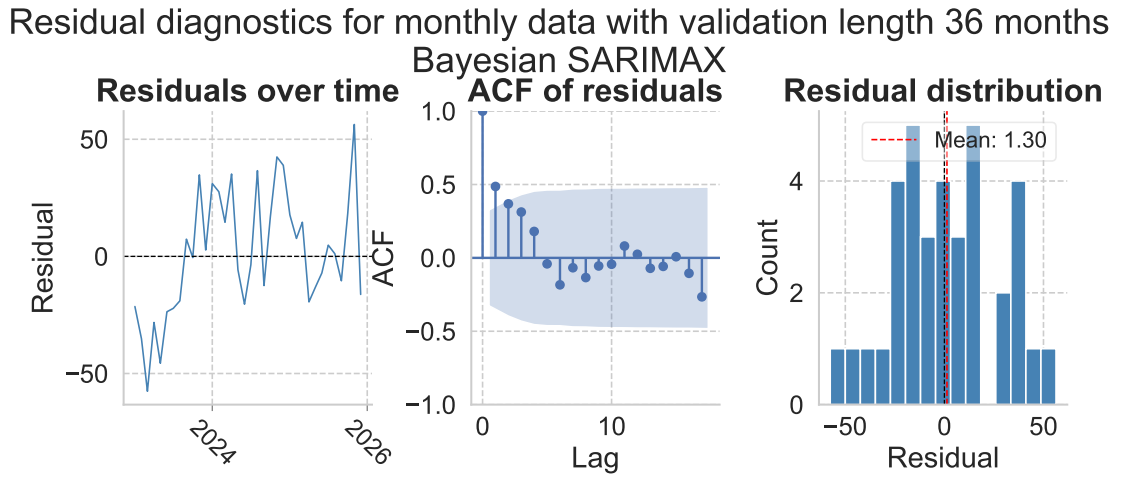

Figure C.12: Figures of residual diagnostics of the baseline SARIMA model for 36 months of validation.

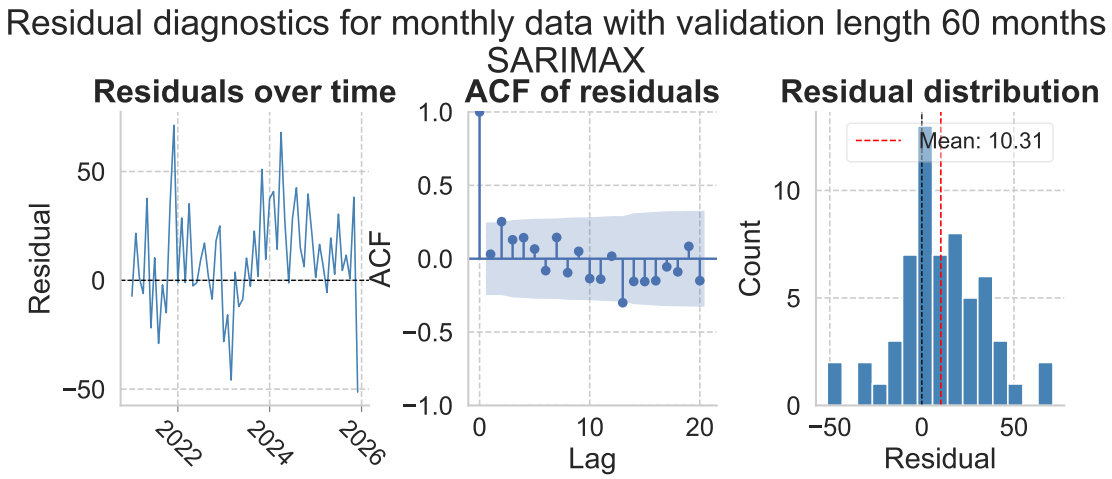

Figure C.13: Figures of residual diagnostics of the SARIMAX model for 60 months of validation.

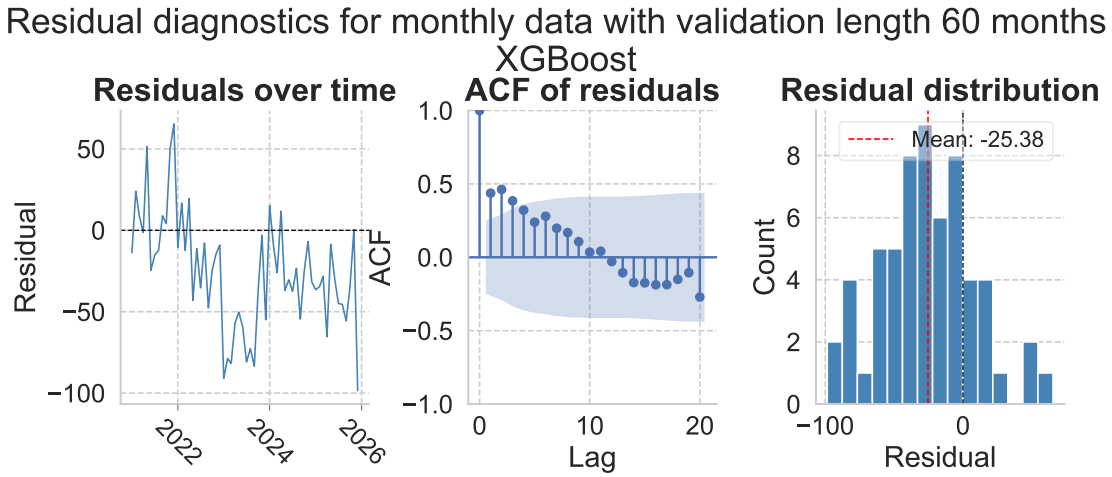

Figure C.14: Figures of residual diagnostics of the XGBoost model for 60 months of validation.

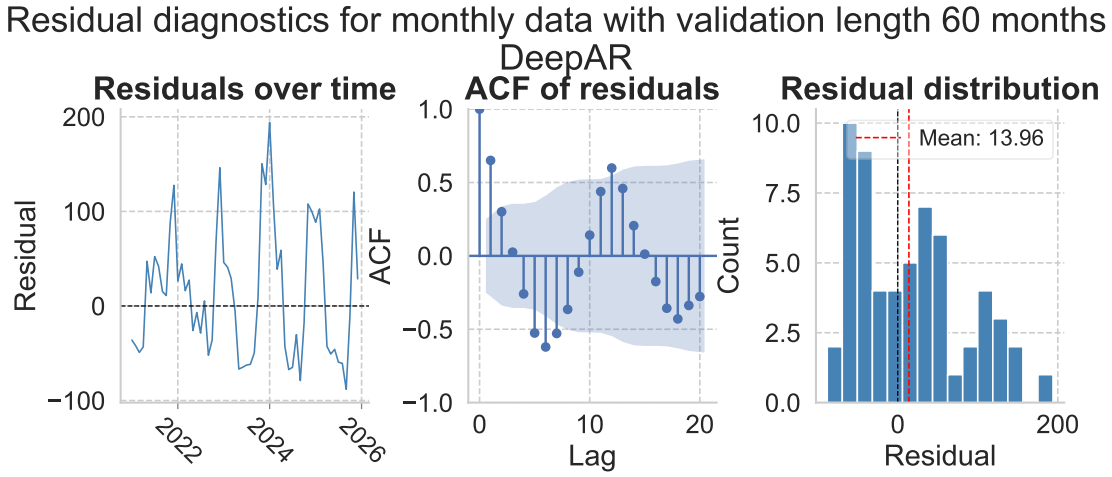

Figure C.15: Figures of residual diagnostics of the DeepAR model for 60 months of validation.

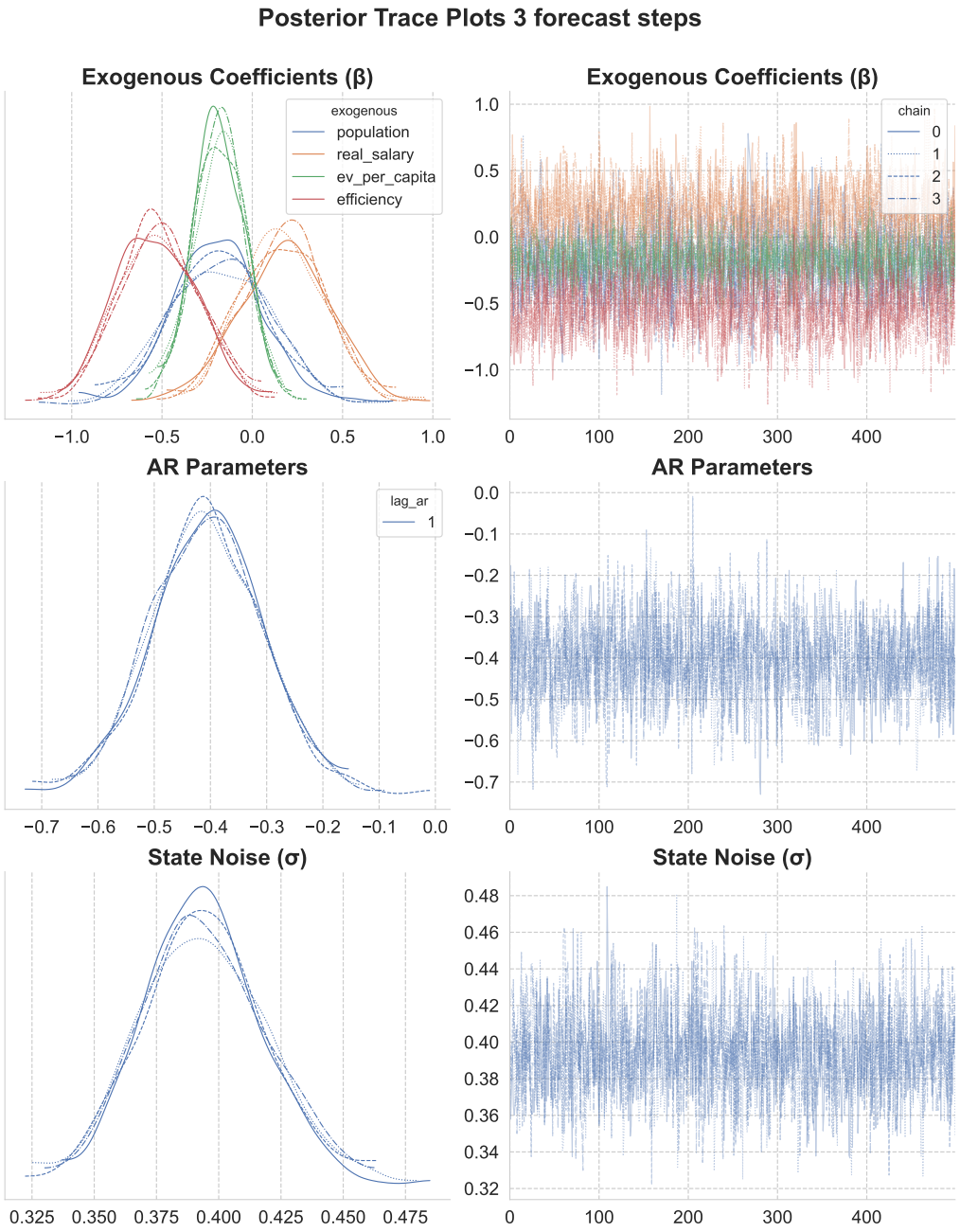

Figure C.16: Traceplot over the posterior distributions of Bayesian coefficients for the 3 year forecast horizon.

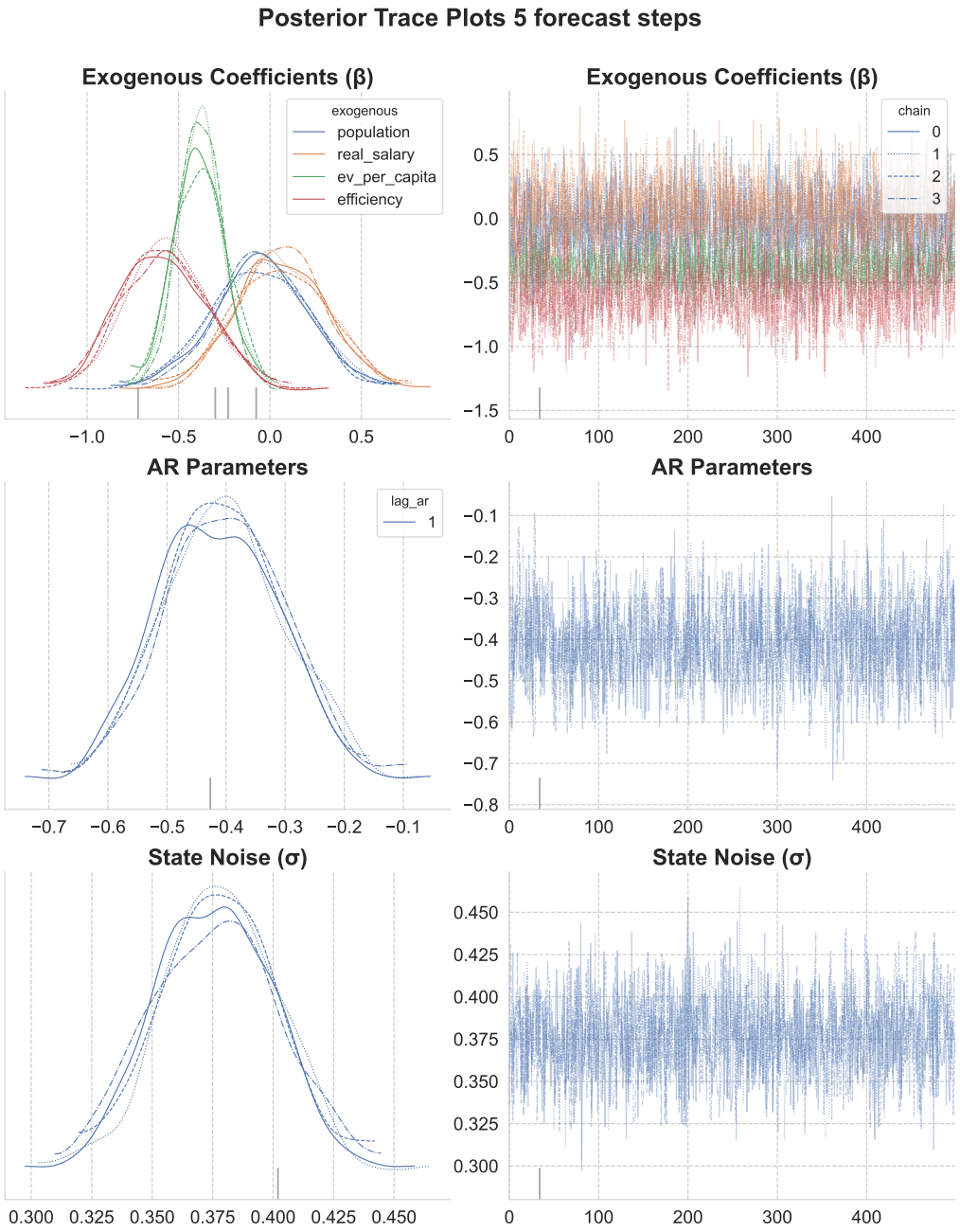

Figure C.17: Traceplot over the posterior distributions of Bayesian coefficients for the 5 year forecast horizon.

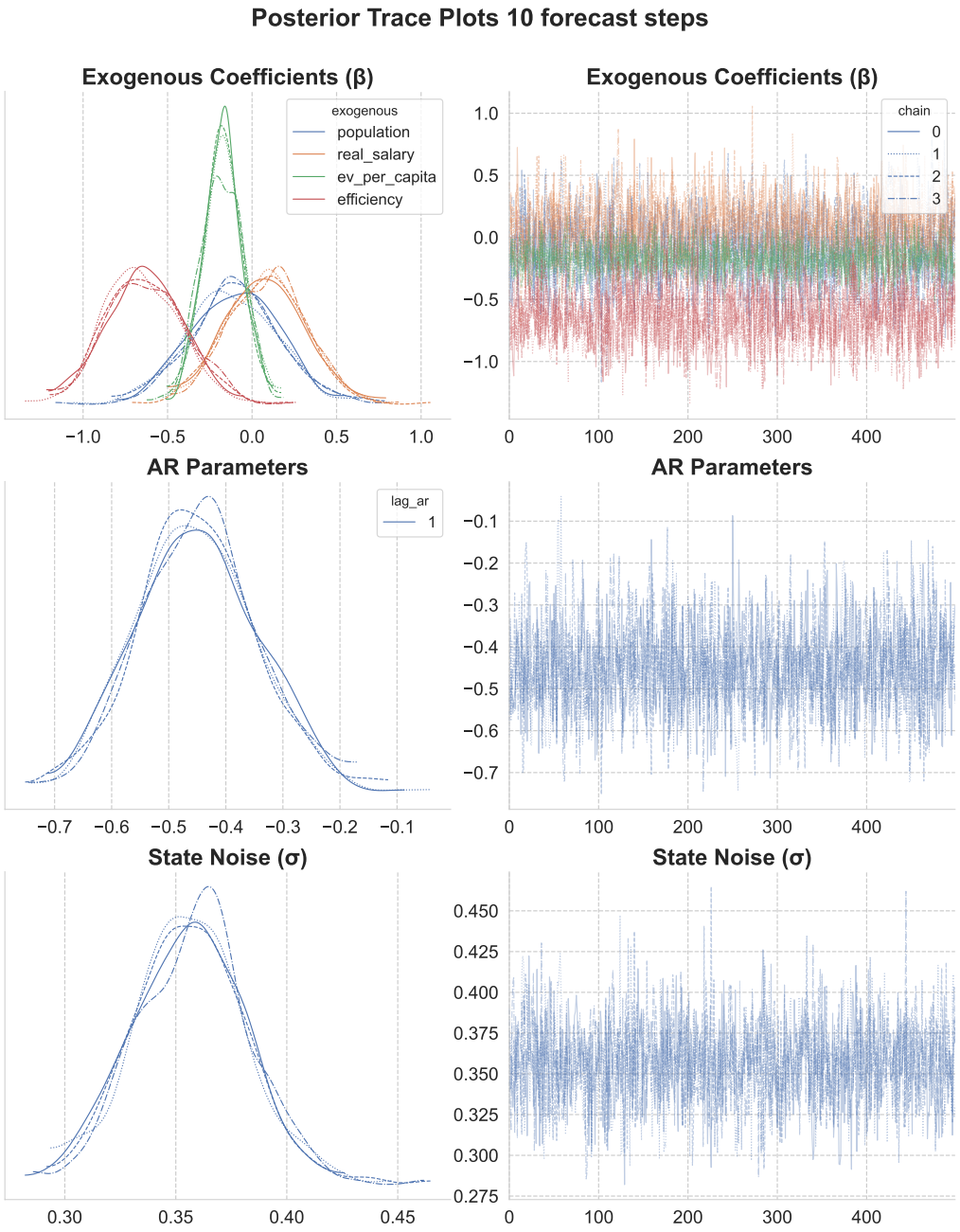

Figure C.18: Traceplot over the posterior distributions of Bayesian coefficients for the 10 year forecast horizon.

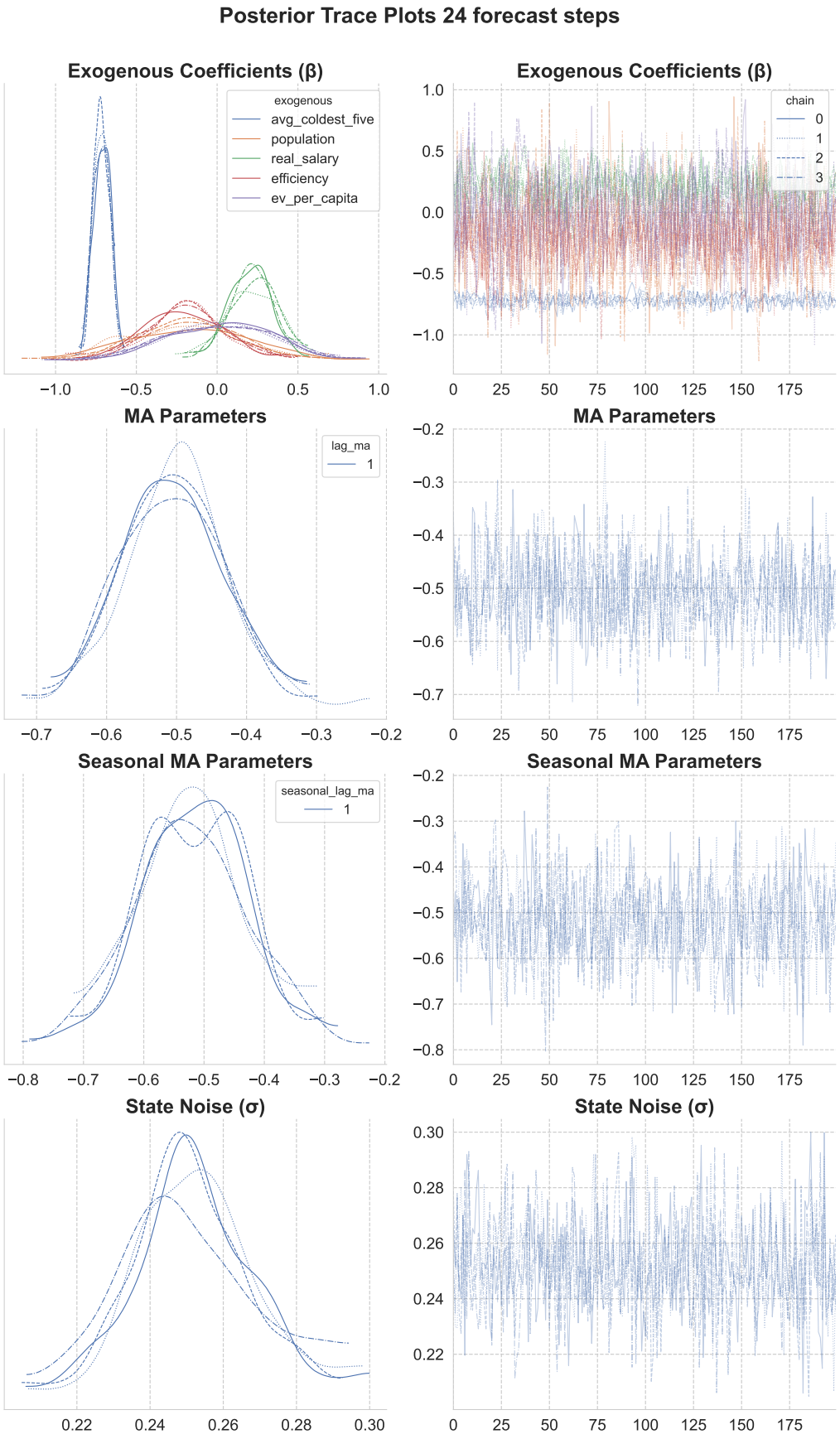

Figure C.19: Traceplot over the posterior distributions of Bayesian coefficients for the 24 months forecast horizon.

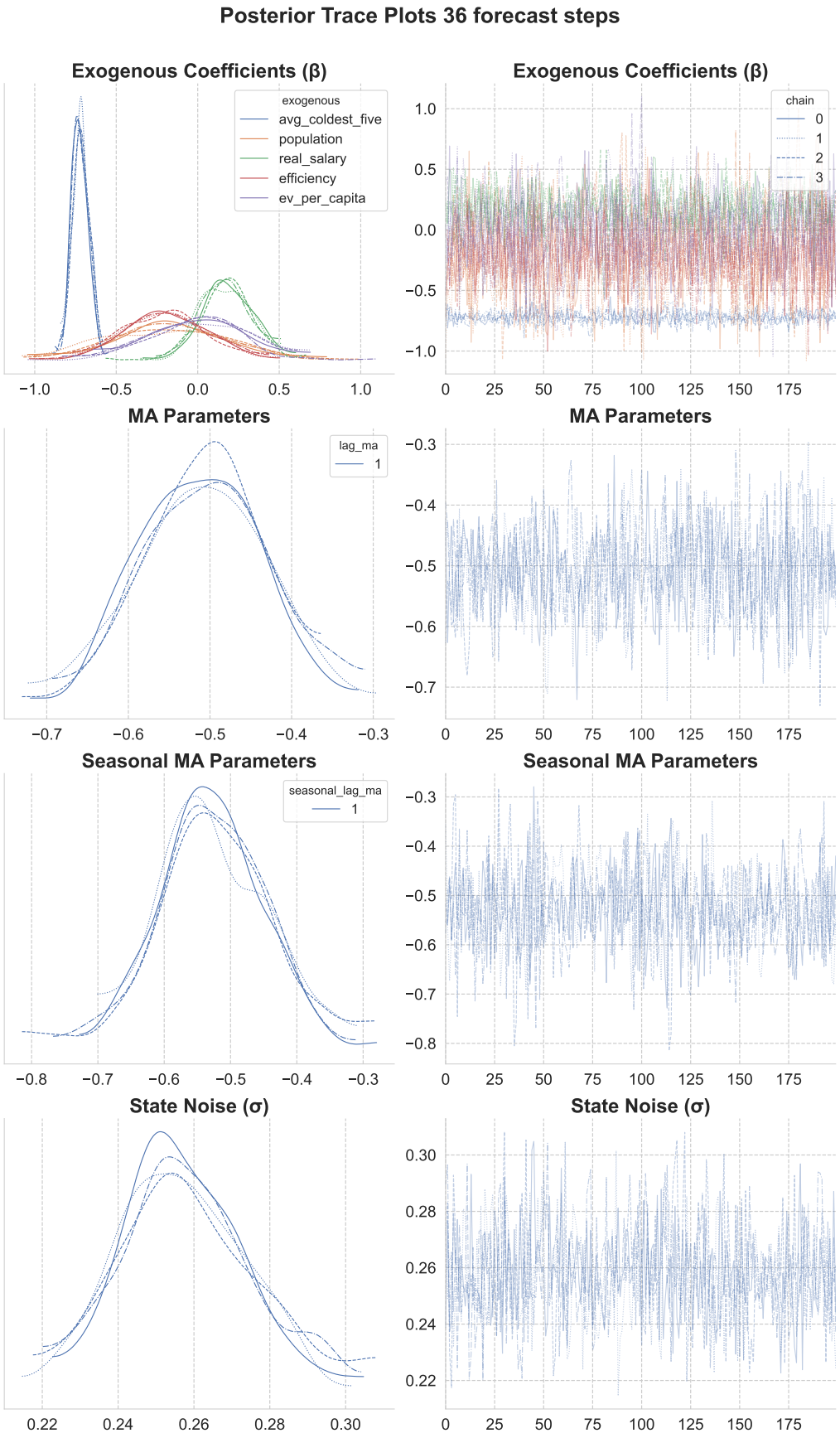

Figure C.20: Traceplot over the posterior distributions of Bayesian coefficients for the 36 months forecast horizon.

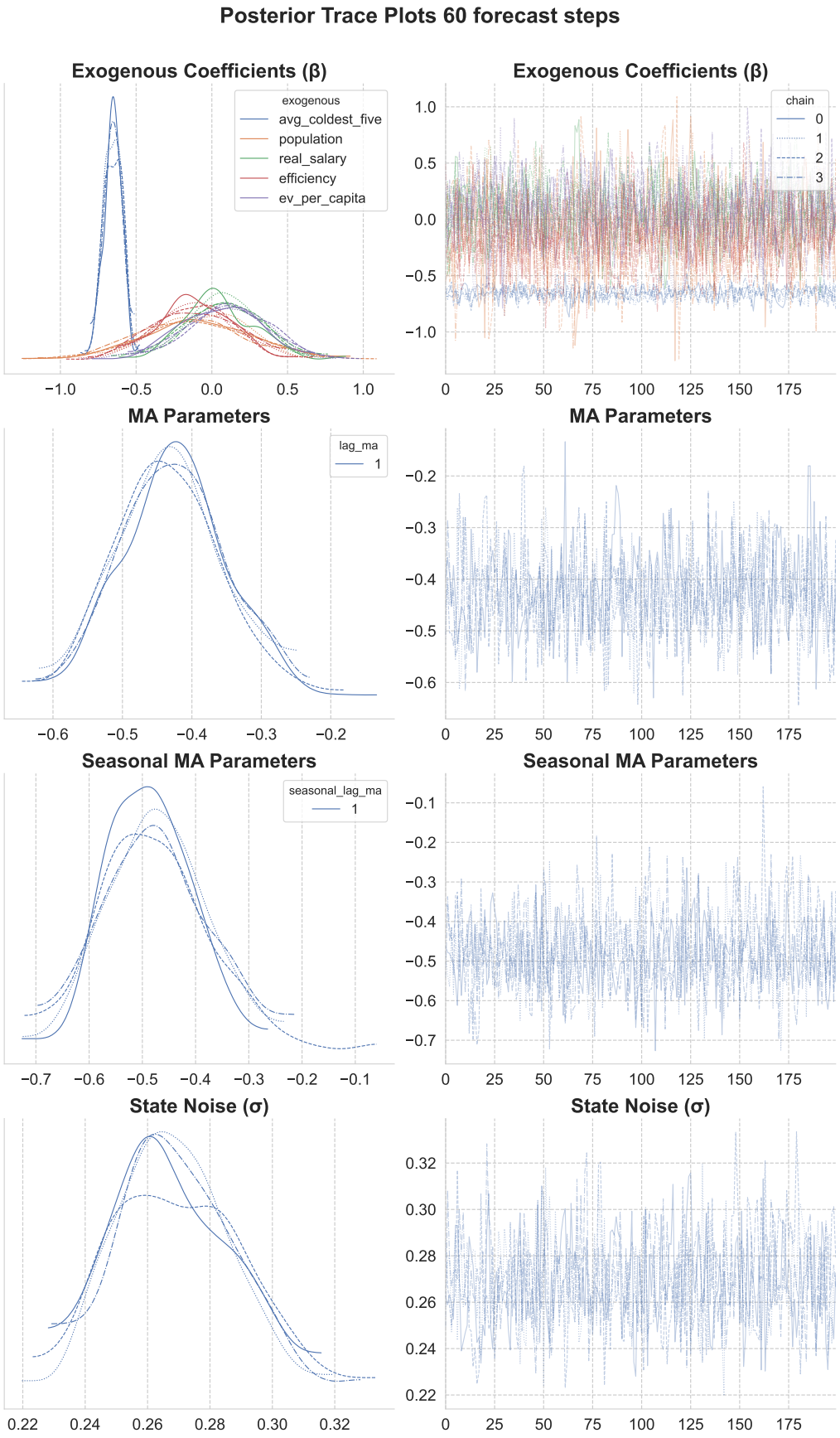

Figure C.21: Traceplot over the posterior distributions of Bayesian coefficients for the 60 months forecast horizon.

DEPARTMENT OF MATHEMATICAL SCIENCES
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden
www.chalmers.se

CHALMERS
UNIVERSITY OF TECHNOLOGY

UNIVERSITY OF GOTHENBURG