Probabilistic Forecasting of Intraday Electricity Prices

In the Nordic Regions Using Deep Distributional Time Series Models

AALBORG UNIVERSITY

STUDENT REPORT

Project Report

Mathematics-Economics

Group 4.117a

Supervisor: Esben Høg

Submission date: 02/06/2023

AALBORG UNIVERSITY
STUDENT REPORT

Aalborg University
School of Engineering and Science

This document is typeset with LaTeX. Programming is performed in R.

AALBORG UNIVERSITY
STUDENT REPORT

School of Engineering and Science
Mathematics-Economics
Skjernvej 4A
9220 Aalborg Øst
http://math.aau.dk

Title:

Probabilistic Forecasting of Intraday Electricity Prices

Theme:

Probabilistic Forecasting

Project Period:

01/02 - 02/06/2023

Project Group:

4.117a

Participants:

Leonora Nielsen
Louise Neema Krog Pedersen

Supervisor:

Esben Høg

Page Numbers: 85

Submission Date: 02/06/2023

Abstract:

This study is based on the methods proposed in Klein et al., 2023, which uses Bayesian inference, echo state networks (ESNs), and copulas to tackle the problem of probabilistic forecasting of intraday electricity prices in Nordic countries. By capturing the complex dynamics of electricity pricing data, the objective is to increase the precision and dependability of forecasting models. A robust framework for probabilistic modelling is provided by Bayesian inference, which enables the inclusion of prior information and measurement of uncertainty. ESNs, like recurrent neural networks, capture temporal linkages and non-linear patterns in the data. In contrast, Copulas represent the joint distribution of several variables and take the relation between the variables into account. The investigation emphasises the need for further model development by exposing limitations in capturing extreme events and tail behaviour. Despite these difficulties, Copulas, ESNs, and Bayesian inference show promise in probabilistic forecasting; nonetheless, further study and calibration are required to raise their performance.

Preface

This project is written by the P10 group 4.117a during 01/02 − 02/06/2023. The group consists of the members; Leonora Nielsen, and Louise Neema Krog Pedersen.

There will be referred to sources with *[last name, year]*. These are listed in alphabetical order in the bibliography. If a whole chapter or section is based on the source, it will be mentioned initially. If the source is mentioned before a period, it refers to the definition or sentence immediately before.

The coding for the application has been done in RStudio.

The data utilised throughout this project is provided by Nord Pool.

The group thank our supervisor Esben Høg for his guidance throughout the completion of this project.

Aalborg University, 02/06/2023.

Leonora Nielsen
<ln18@student.aau.dk>

Louise Neema K. Pedersen
<lnkp18@student.aau.dk>

Contents

1 Motivation

This chapter is based on [Spot, 2023], [Taillon, 2023], [Segal, 2022], [Kenton, 2022], [Norway, 2023], [Stanwell, 2023], [Ørsted, 2023], [Politiken, 2021], [Bellis, 2018], [Madaleno and Pinho, 2008], and [Klein et al., 2023]

In today's world, energy in the form of electricity has become a necessary good. It not only provides light and warmth, but it also forms the foundation of all industrial endeavours. The initial historical energy sources are called primary sources, such as water wheels operating next to waterfalls. Today, most electricity is generated from so-called secondary energy sources; among these are nuclear energy and fossil fuels like coal and oil, along with renewable sources, including wind, solar, hydropower, and geothermal energy. The electricity generated from these diverse sources is transmitted using the necessary infrastructure.

The 1990s saw the deregulation of the power market in the Nordic countries Denmark, Norway, Sweden and Finland, which are the primary focus of this project. The deregulation reduced or eliminated governmental control over the sector and provided investment opportunities by allowing competitors to enter the market, encouraging innovation and potential price drops benefiting the consumers.

Wholesale and end-user sales are two categories of the power market. Power producers, power suppliers, energy businesses etc., purchase and sell large quantities on the wholesale market. In comparison, end-users are customers who buy power for personal consumption, such as homeowners, businesses, etc. In addition, individual consumers sign contracts to purchase electricity from a power supplier of their choice in the end-user market. Various markets make up the wholesale market. In these markets, bids are made and where prices are set. They comprise the intraday market, the balance market and the day-ahead market.

This project focuses on *intraday* electricity prices. In this context, intraday describes a continuous exchange of electricity on the same day it is provided. In other words, participants engage in nonstop trading throughout the day. Furthermore, due to its great degree of flexibility, intraday trading can be utilised to balance positions closer to real-time and make last-minute modifications.

The balance market is used to control consumption or production to maintain an equilibrium. A balance is established between the day-ahead and intraday markets' production and consumption of electricity. However, events, such as uneven electricity distribution, that can throw off balance within a given hour are inevitable. Consequently, the balancing market will adjust production or consumption upward or downward depending on what is required to maintain an immediate equilibrium.

In the day-ahead market, agreements are formed for deliveries between buyers and sellers the next day. In other words, it describes the purchase or sale of electricity before its actual production or delivery.

Returning to the intraday market, a theoretical price called the *system price*, which is based on the presumption that no grid congestion is determined daily. Power producers place bids outlining the volume of energy they are willing to produce for a given price. This price is directly related to the cost of operating a power plant and reflects the value producers place on their output. Power suppliers place bids stating how much they are willing to buy at various prices. As a result, the system price is established at the equilibrium where supply and demand are equal. It should be noted that electricity cannot be stored, at least not directly; as a result, it is produced and used instantaneously. Supply and demand must, therefore, consistently be balanced in real-time. This balance is created using an *order book*, which is an electronic list of buy and sell orders. Those orders represent the current supply and demand conditions in a particular market area. The energy price will hereafter be determined by matching supply and demand orders using data from the order book. Since trades are the product of intense competition among order exchange participants in an open and transparent environment, they always represent the most up-to-date information.

In addition to system prices, *area prices*, or the cost of electricity in a particular area, also need to be considered. Since this project primarily focuses on the Nordic countries of Denmark, Norway, Sweden, and Finland, the regions or "bidding zones" considered are DK1 and DK2 in Denmark, NO1-NO5 in Norway, Sweden having four SE1-SE4, and Finland having one FI.

These area prices differ, and in the event of congestion brought on by, for example, changes in demand, insufficiently planned transactions, etc., some places may experience a deficit of energy, which is when supply is insufficient compared to demand. In contrast, others may experience a surplus of energy, that is, when supply is greater than demand. As a result, power is moved from locations with a surplus to those with a shortfall. As a result, prices are often higher in locations with a power shortage than in those with a surplus. Nonetheless, local pricing will match the system price throughout all areas if there are no restrictions on the electricity grid.

A phenomenon of *negative prices* exists, exclusively observed in the wholesale market. It occurs when the supply offered at negative prices is higher than the demand, a power surplus. Negative price scenarios are typically observed in the middle of the day due to competition between generators trying to dispatch their energy. This phenomenon indicates that supply must be constrained or demand must be raised. How costly and fast energy production can stop and start depends on the energy source. Sources like solar and wind energy can stop and start relatively quickly, meaning that these types of energy sources can avoid negative price periods. Nuclear power and coal-fired generators are two examples of energy sources that require hours to restart and incur significant costs when they are stopped and started. These energy sources continue to produce energy during the negative price phase because it is the most cost-effective. Also, this guarantees that the energy demand is still satisfied in the evening when, for example, solar energy is no longer available.

Some stylised facts about electrical market pricing, in general, are provided below.

- Seasonality, or the varying supply and/or demand of power, can, e.g., be caused by business operations. When intraday trading is taken into account, seasonality is connected to the structure of the business day-weekend structure. In other words, prices start to rise as the workday officially begins, then subsequently decline when it finishes and demand shifts predominantly towards domestic usage. The weather also regulates the seasonal effect. That is, shifting climate factors like temperature and the length of daylight directly affect prices.

- The mean-reverting or anti-persistent nature of electricity prices. In other words, data shocks are temporary and revert to the previous price level.

- Unexpected jumps or spikes in cost are the most noticeable aspect of power prices. That is to say, system costs might rise significantly quickly and then drop back to their previous level. Such jumps are particularly infamous during, for instance, peak consumption seasons like the winter. These spikes are typically a result of supply changes brought on by severe weather conditions and/or power outages.

- Leptokurtosis. Electricity prices commonly exhibit leptokurtosis, meaning that both minor and significant price changes occur more frequently than they would under a normal distribution. To put it another way, the distribution of electricity prices shows fat tails.

- Volatility. Electricity prices are generally very volatile because of transmission and storage issues and the market's need to establish equilibrium pricing immediately. Due to the difficulty of resolving supply and demand imbalances in

the short term, the price changes in the electricity market are more dramatic than those in other financial or commodity markets.

Some other factors affecting the electricity price are how various factors influence the different types of energy sources. Examples include the absence of wind, which would reduce energy output and supply. CO2-quotas, or the right to emit CO2, are another consideration. Each CO2 quota entitles the holder to one ton of CO2 emission. These CO2 quotas are intended to be a tool for reducing CO2 emissions. The quotas are distributed among others to businesses producing energy and emitting CO2. Thus, a smaller supply of quotas would result in an increase in their price, which would have an impact on energy prices. War and politics are a couple of other things that affect energy prices. Here, sanctions can directly affect energy prices, which was observed when Russia cut back on and eventually stopped exporting gas to Europe due to European sanctions against Russia.

This project's primary goal is to apply some of the methods mentioned in the recently published paper [Klein et al., 2023] since their approaches are innovative in generating probabilistic price forecasts for energy prices using data from the Australian market. However, as mentioned, this project focuses on the Nordic region's energy market. For the market to run smoothly and profitably for participants, accurate intraday price predictions are essential. Probabilistic forecasts are particularly interesting because it is a stylized fact that electricity prices frequently exhibit leptokurtosis, and hence it is not just the mean and variance that matters. The paper [Klein et al., 2023] introduces statistical time series models based on *echo state networks*, a type of recurrent neural network. Here, the output layer coefficients of the echo state network are estimated by Bayesian techniques. In addition, an output layer coefficient shrinkage prior is included to provide regulation or control. Additionally, a different strategy is used, employing implicit copulas of time series derived from an echo state network. The copula model is coupled with a marginal distribution of the data to represent the serial dependence in the time series accurately. After the models are in place, a probabilistic forecast is created using predictive distributions over K different weight configurations.

The project is outlined as follows:

- Chapter 2: Briefly introduces the later utilised theory from Bayesian statistics to obtain the predictive distribution for forecasting intraday electricity prices in the Nordic region.
- Chapter 3: Provides information on neural networks, focusing on recurrent neural networks and echo state networks, which form the basis of the final forecasting models.

- Chapter 4: Outlines the theory of copulas, which is incorporated into the second forecasting model.
- Chapter 5: Describes the Gaussian probabilistic and Copula models, the two specific models used for making predictions.
- Chapter 6: Applies the models to actual data provided by Nord Pool. The chapter begins with an introduction and preliminary examination of the data, followed by obtaining probabilistic forecasts using the models.
- Chapter 7: Discusses the different approaches used in the two models and some of the decisions made during the project.
- Chapter 8: Presents the conclusion of the project.
- Chapter 9: Discusses additional viewpoints or methods that could have been incorporated into the models and the project.
- Appendix: Contains supplementary figures related to the application in Chapter 6 and formulas used throughout the project.

Problem Statement

The primary objective is to investigate the application of neural networks and the incorporation of copulas to develop two distinct models capable of generating probabilistic forecasts for intraday electricity prices in the Nordic regions encompassing Denmark, Norway, Sweden, and Finland.

A Brief Introduction to Bayes Inference 2

This chapter is based on [Marin and Robert, 2014], [Lee, 2004], [Hoff, 2009a], [Taboga, 2023a], [Taboga, 2023b], and [Hoff, 2009b]

Throughout this project *Bayesian statistics* will be used, a theory in the field of statistics based on the *Bayesian interpretation* of probability. Therefore, the probability is understood as a reasonable expectation indicating a level of knowledge or a measurement of a personal conviction rather than the frequency or propensity of any occurrence. One of the primary purposes of introducing Bayesian inference is to get access to the so-called predictive distribution used to construct probabilistic forecasts. The distinctive feature of Bayesian statistics is the application of Bayes' theorem in a broader range of circumstances than in classical statistics. In particular, Bayesian statisticians are always willing to talk about the probability of a hypothesis, both unconditionally (its *prior probability*) and given some evidence (its *posterior probability*). In contrast, other statisticians will only talk about the probability of a hypothesis in restricted circumstances. In broad outline, prior beliefs are assumed about various possible hypotheses, and these prior beliefs are modified in the light of more relevant data collected to arrive at posterior beliefs.

For unknown parameters, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$ for $k \in \mathbb{N}$, a prior belief about their value can be expressed in terms of their pdf,

$$\textbf{Prior}: \quad p(\boldsymbol{\theta}). \tag{2.1}$$

Furthermore, for n observations of relevant data on their values, $\mathbf{X} = (X_1, \ldots, X_n)$, for $n \in \mathbb{N}$, have a probability distribution that depends on the k unknown quantities as parameters so that the pdf (continuous or discrete) of the vector $\mathbf{X}$ depends on the vector $\boldsymbol{\theta}$ in a known way. Typically the components of $\boldsymbol{\theta}$ and $\mathbf{X}$ will be integers or real numbers so that components of $\mathbf{X}$ are random variables, and hence the dependence of $\mathbf{X}$ can be expressed in terms of a pdf,

$$\textbf{Likelihood}: \quad p(\mathbf{X} \mid \boldsymbol{\theta}). \tag{2.2}$$

This pdf, considered as a function of $\mathbf{X}$ for a fixed $\boldsymbol{\theta}$, is a density. However, often the pdf is thought of as a function of $\boldsymbol{\theta}$ for fixed $\mathbf{X}$. If this is the case, it does not have quite the same properties, e.g., it is not necessary to sum (or integrate) to unity. Thus, in the extreme case where $\pi(\mathbf{X} \mid \boldsymbol{\theta})$ does not depend on $\boldsymbol{\theta}$, it is possible for it to sum (or integrate) to ∞. Thus, when considering $\pi(\mathbf{X} \mid \boldsymbol{\theta})$ as a function of $\boldsymbol{\theta}$, it is referred to as the likelihood function, as stated in (2.2). Thus,

$$l(\boldsymbol{\theta} \mid \boldsymbol{X}) = p(\boldsymbol{X} \mid \boldsymbol{\theta}).$$

Hence the posterior belief is given by Bayes' Theorem for random variables, generalised to deal with random vectors, where it is known that,

$$\textbf{Posterior}: \quad p(\boldsymbol{\theta} \mid \boldsymbol{X}) \propto p(\boldsymbol{\theta}) p(\boldsymbol{X} \mid \boldsymbol{\theta}). \tag{2.3}$$

It should be noted that differing priors result in varying posterior beliefs; however, with enough collected data, the posterior beliefs will usually become very close.

Markov Chain Monte Carlo (MCMC) techniques are frequently employed when the posterior distribution cannot be determined analytically. Monte Carlo methods approximate a feature of the probability distribution of a random variable Y, such as the expected value.

In MCMC, the draws produced by the computer, $y_1, \cdots, y_n$, are serially correlated rather than independent, which gives it a unique spin on the traditional Monte Carlo approach. The draws are more specifically realisations of the n random variables $Y_1, \cdots, Y_n$ that make up a Markov Chain. In particular, if a random sequence Y_n meets the *Markov Property*, it is referred to as a Markov Chain.

Proposition 2.1 (Markov Property).
Regardless of the chain's historical trajectory, the probability distribution of its future values solely depends on its current values Y_n.

$$\mathbb{P}(Y_{n+t} = y \mid Y_n, Y_{n-1}, \cdots, Y_{n-k}) = \mathbb{P}(Y_{n+t} = y \mid Y_n).$$

The proof of this proposition is omitted from the project.

The chains generated by MCMC have the property of asymptotic independence. In other words, two variables Y_n and Y_{n+t} are not independent, but they approach independence as $n \to \infty$. This specifically indicates that as n becomes larger $f(y_{n+t} \mid y_n)$ converge to $f(y_{n+t})$. This property is significant because it shows that as n increases, the initial value of the chain, which is often selected randomly, has a decreasing

impact on the distribution.

There is often a difference between the distribution of the first terms of the chain and the *target distribution*, which is the distribution from which samples are extracted at the end. In the case of this project, the target distribution is the posterior distribution. As a result of this difference, an MCMC sample's initial draws are frequently discarded, called the *burn-in sample*. Hence, draws from the burn-in sample are eliminated, that is, draws that are distant from the target distribution, while draws closer to the target distribution are retained.

After the burn-in sample is eliminated, a sample of draws from a distribution that closely resembles the target distribution is acquired; the draws, however, are not independent. Here the idea of the *effective sample size* can be employed, meaning that a smaller number of independent observations is equivalent to n dependent observations. The effective sample size decreases, and the accuracy of the MCMC approximation generally declines as the correlation between adjacent observations increases. As a result, most work in MCMC samplers is focused on minimising correlation.

Specifically, in the application, Chapter 6, the *Metropolis-Hasting* algorithm will be employed. The Metropolis-Hasting algorithm is stated as follows:

Algorithm 1 Metropolis-Hasting

1: Choose a proposal kernel $q(x, y)$.
2: Define the Hasting ratio as follows:

$$H(x, y) = \frac{\pi(y)q(y, x)}{\pi(x)q(x, y)},$$

with $H(x, y) = \infty$ if $\pi(x)q(x, y) = 0$.
3: The acceptance probability is given by

$$a(x, y) = \min\left\{1, H(x, y)\right\}.$$

Another feature of Bayesian statistics is the existence of a *predictive distribution*. A predictive distribution of a random variable $\mathbf{Y}$ is a probability distribution such that known quantities are conditioned on, and unknown quantities are integrated out.

Suppose that a new data point y is acquired after the data $\mathbf{X}$ have been observed and the posterior distribution described in (2.3) has been determined. Assume further that the distribution of y is independent of $\mathbf{X}$ conditional on $\boldsymbol{\theta}$, but dependent on $\boldsymbol{\theta}$. That is,

$$p(y \mid \boldsymbol{\theta}, \mathbf{X}) = p(y \mid \boldsymbol{\theta}).$$

The distribution of y given $\mathbf{X}$ is then given by

$$\begin{aligned} p(y \mid \mathbf{X}) &= \int_{\theta} p(y, \boldsymbol{\theta} \mid \mathbf{X}) d\boldsymbol{\theta} \\ &= \int_{\boldsymbol{\theta}} p(y \mid \boldsymbol{\theta}, \mathbf{X}) p(\boldsymbol{\theta} \mid \mathbf{X}) d\boldsymbol{\theta} \\ &= \int_{\boldsymbol{\theta}} p(y \mid \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \mathbf{X}) d\boldsymbol{\theta}. \end{aligned}$$

To summarise, the predictive distribution can be expressed as

$$\textbf{Predictive distribution} : p(y \mid \mathbf{X}) = \int_{\boldsymbol{\theta}} p(y \mid \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \mathbf{X}) d\boldsymbol{\theta}. \tag{2.4}$$

Neural Networks 3

This chapter is based on the sources [Aggarwal, 2015], [Hastie, 2001], and [Stanford, 2023]

The deep learning method Neural Networks, abbreviated NN, takes inspiration from the human nervous system, composed of cells referred to as neurons and connected at contact points referred to as synapses. Changing the strength of these synaptic connections between neurons is the basis of learning in living organisms. Therefore, neural networks can be considered a simulation of this biological process.

Individual nodes in artificial Neural Networks, similar to those in biological networks, are referred to as neurons. These neurons are computation units that receive input from other neurons, perform computations on these inputs, and feed it back into other neurons. The computation function of a neuron is defined by the weights on the neuron's input connections, simulating the strength of a synaptic connection. The computation function can be learned by appropriately changing these weights, which is analogous to learning synaptic strength in biological neural networks. The training data serves as the "external stimulus" in artificial neural networks for learning these weights. The idea is to incrementally modify the weights whenever the current set of weights makes incorrect predictions. The architecture used to arrange connections between nodes is critical to the neural network's effectiveness.

Several different architectures are accessible, depending on the network type being evaluated. A *recurrent neural network*, abbreviated RNN, is introduced in this chapter with the purpose of extending it to an *echo state network*, abbreviated ESN. An introduction to neural networks can be found in [Hastie, 2001] and [Stanford, 2023].

3.1 Recurrent Neural Networks

This section is based on [Pra, 2020], [Liu, 2020], [IBM, 2023], [Faik, 2021], [Engati, 2023c], [Madhan, 2020], and [McDermott and Winkle, 2017]

Given their fundamental design principles, classical neural networks, especially *feed-forward neural networks*, see [Hastie, 2001], lack inherent means to manage time

dependence. No matter the temporal order in which inputs are delivered, these networks are designed to operate independently on input data. Therefore, traditional neural networks do not explicitly model relationships or temporal dependencies between subsequent inputs but instead treat each input as a distinct unit. This shortcoming, however, becomes apparent when dealing with sequential or time-dependent data, when the timing and order of inputs matter. In certain situations, standard neural networks might perform less well because they cannot accurately capture the temporal dependencies. Recurrent neural networks (RNNs) provide a solution to this problem.

The recurrent neural network distinguishes itself from the above-mentioned traditional neural network by its *memory*. RNN uses information from preceding inputs to influence current input and output, in contrast to classical neural networks, which assume inputs and outputs are independent of one another. Thus, RNNs eliminate the independence between input and output. This particular memory architecture has a stronger analogy to how living beings' brain works, namely that it tries to find correlations, also known as *long-term dependencies*, between past situations to better understand an event happening in the present. As a result, RNNs offer the possibility of processing data that takes the form of a sequence, such as time series.

Note that only *unidirectional* RNNs are considered for this project; therefore, only past events will be considered. In other words, forecasts won't be based on what will happen in the future.

In an RRN, opposite to NNs, the nodes are assigned a fixed time step, and the hidden layers are forwarded in a time-dependent direction.

The architecture of an RNN is illustrated in Figure 3.1 below, with a following definition of a basic RNN. Neither the illustration nor the definition include a bias term, but it is still possible to include it.

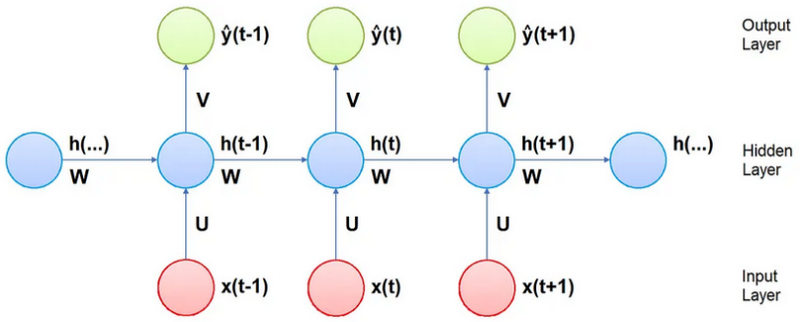

Figure 3.1: Illustration from [*Pra, 2020*]. The architecture of a recurrent neural network. A fixed time stamp is given to each node in the RNN, progressing from left to right. Hence, it can be thought of as numerous feed-forward network replicas, each of which sends a message to a descendant. The input layer is represented by the red circles, the hidden layer by the blue, and the output layer by the green. Indicated by the letters U, W, and V are the weight matrices that connect the input and hidden layers, the hidden layers at various times, and the hidden layer and the output layer, respectively.

Definition 3.1 (Basic Recurrent Neural Network).
For time $t = 1, \ldots, T$:

$$\text{response:} \quad \mathbf{Y}_t = f_o\left(\mathbf{o}_t\right) \tag{3.1}$$

$$\text{output:} \quad \mathbf{o}_t = V\mathbf{h}_t \tag{3.2}$$

$$\text{hidden state:} \quad \mathbf{h}_t = (1-\kappa)\mathbf{h}_{t-1} + \kappa\tilde{\mathbf{h}}_t, \tag{3.3}$$

$$\tilde{\mathbf{h}}_t = f_W\left(W\mathbf{h}_{t-1} + U\mathbf{x}_t\right), \tag{3.4}$$

Here $\boldsymbol{Y}_t$ is an n_y-vector of responses at time t, $\boldsymbol{x}_t$ is an n_x-dimensional input vector (typically, assumed to include a one in the first position for an "intercept"), $\boldsymbol{o}_t$ is an n_y-vector of outputs that are associated with a linear transformation of the n_h-dimensional hidden unit vector $\mathbf{h}_t$, $\tilde{\mathbf{h}}_t$ being its update. The hidden layer weight matrices W, U are $n_h \times n_h$ and $n_h \times n_x$, respectively, and V is the $n_y \times n_h$ output weight matrix. Lastly, $f_o(\cdot)$ and $f_W(\cdot)$ are specified activation functions, and the κ parameter, known as the *leaking rate* takes a value $(0, 1]$.

An activation function determines and controls the transformation of the output from the weighted sum of inputs in a certain layer of the network. In other words, it determines the node's output based on the supplied input. Three types of activation functions exist; a binary step function, a linear activation function, and a non-linear

activation function, hereunder

$$\text{Tanh Function (Hyperbolic Tangent)} : \quad f(x) = \frac{\left(e^{x} - e^{-x}\right)}{\left(e^{x} + e^{-x}\right)}, \tag{3.5}$$

$$\text{ReLU Function (Rectified Linear Unit)} : \quad f(x) = \max(0, x), \tag{3.6}$$

$$\text{Sigmoid Function (Logistic} : \quad f(x) = \frac{1}{1 + e^{x}}, \tag{3.7}$$

are commonly used activation functions.

As illustrated in 3.1, the input-to-hidden connections are parameterised by U, and hidden-to-hidden recurrent connections are parameterised by W. Hidden-to-output connections are parameterised by V, and all these weights (U, V, W) are shared across time. The before-mentioned memory of the RNNs is kept in the hidden state, as seen in Equation 3.4, where its calculation is based on the current input and information from the previous hidden state. The weight matrices are fixed and are subject to change through the process of *backpropagation*, BP, and *gradient descent*, i.e., an algorithm that updates the weights. For a small introduction to the backpropagation algorithm see [Ognjanovski, 2019]. The BP used in RNNs distinguishes itself by considering the time as a factor and is referred to as *backpropagation through time* (BPTT). It can be interpreted as a generalisation of the backpropagation algorithm but specifically applied to RNNs. This is not introduced further in this project, due to its drawbacks.

As with the BP, a loss function $\mathcal{L}$ must be minimised to train the network. However, not only V but also W and U need to be updated. Since W considers all the previous hidden states, all the gradients must be accumulated to update W. This results in long calculi that have to be repeated for each time-step to obtain all the partial losses $\mathcal{L}_t$, and since input sequences often consist of thousands of time steps, it would require a huge number of derivatives to perform a single weight update. Consequently, BPTT results in a quite inefficient algorithm in terms of computation time. However, this is not BPTT's main drawback. BPTT can suffer from two severe problems, which could significantly affect the training process. Consequently, BPTT, hereunder RNNs, are rarely used in practice.

Firstly, the *vanishing gradient problem* involves training and modifying the network's initial layers. Gradients tend to get smaller until they approach zero as the backward propagation algorithm moves from the output to the input layer. This tendency results from the gradients of the first layers being derived by multiplying the gradients of the later layers. As a result, their multiplication diminishes at a particularly high rate if the gradients of the later layers are smaller than one, which entails that the initial or first layers' weights will essentially stay the same. As a result, the gradient

descent never converges to the ideal state. Furthermore, since the first layers are crucial for identifying the main components of the input data, the network as a whole could become inaccurate if the weights and biases are not appropriately updated. A network may exhibit symptoms of a vanishing gradient problem when parameters in the later layers change quickly, but parameters in the first layers change very little or not at all, the weights of the model may zero out, or the model learns slowly and stagnate in its early stages.

Secondly, *the exploding gradient problem*, which may be seen as the opposite of the vanishing gradient problem, is another restriction associated with the gradient. The gradients, in this case, are always greater than one. As a result, this issue arises when significant error gradients accumulate, which leads to substantial modifications to the network's weights during training. These values can get extremely large to the point that they overflow and produce NaN values, which could lead to an unstable model that cannot learn from the training data.

RNNs have many extensions that were developed to address the drawbacks mentioned above. The *Echo State Network*, often known as ESN, is one such approach. The ESN is simple to implement and does not suffer from vanishing or exploding gradients.

3.2 Echo State Networks

This section is based on [Engati, 2023a], [Engati, 2023b], [Ciortan, 2019], [REU, 2021], [McDermott and Winkle, 2017], and [Klein et al., 2023]

Echo state networks and *liquid state machines*, LSMs are two approaches often labelled more generally as *reservoir computing* methods, where reservoir refers to a dynamic system which is identified by a mathematical function that explains how a point in space behaves over time. Reservoir computing considers sparsely connected hidden layers, typically less than 10% connectivity, that allow for sequential interactions and can be viewed as a "black box". In addition, a crucial component of such reservoir models is that the connectivity and the weights for the hidden units are fixed yet randomly assigned. That is, the input data goes into a hidden fixed reservoir that contains sequential linkages.

The reservoir is typical of a higher dimension than the input, so there is a dynamical expansion of the input, thus adding model flexibility. The reservoir states are then mapped to the desired output, and importantly, training has been limited to this output step since only the weights at this mapping phase are estimated. In a classical setting, the reservoir must possess two qualities: it must be made up of distinct, non-linear units and be able to retain data. Reservoir computing is a strategy designed

to enable machine learning algorithms to analyse data more quickly and at reduced learning costs.

The architecture of an ESN is illustrated in Figure 3.2, consisting of an input layer, a hidden layer - now the reservoir with the appropriate properties discussed above - and an output layer.

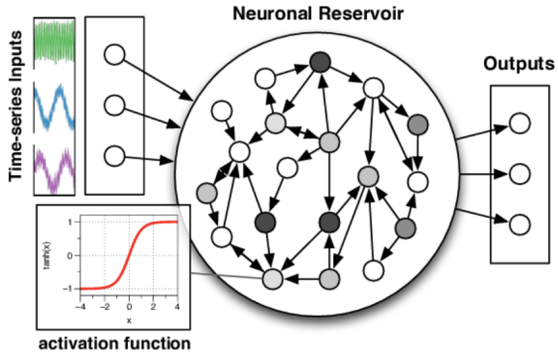

Figure 3.2: Illustration from [*Demiris, 2023*]. Architecture of an ESN. The input layer feeds the network with the input time series. Calculations are then performed in the reservoir, which represents the input stream in a higher dimension. The reservoir and output layer are connected by a weight matrix V.

Considering the basic RNN from Definition 3.1, the ESN version of this simple RNN considers the hidden layer matrices U and W (the reservoir weights) to be fixed. They are drawn once from a distribution centred around zero, with added sparsity. Only the output matrix V is estimated. Herein lies the reduced learning cost since there are only relatively few output weight parameters, which can be estimated through standard regularisation-based statistical estimation approaches. E.g. if $f_o(\cdot)$ is the identity function, then a simple ridge regression estimation of V is typically used.

The hidden units in the reservoir act as a nonlinear expansion of the input vector, $\mathbf{x}_t$, and as a way to establish memory or account for the sequential nature of the dependence in the input vectors and, ultimately, the response. The idea of a nonlinear expansion in a high dimension helps to magnify potentially important dynamic features of the input, and the output weights provide a way to select those expanded states that are important for the response.

The name 'echo state' refers to the *echo state property* i.e., the spectral radius (largest eigenvalue) of W must be less than one. This property allows the hidden states to lose dependence on the initial input conditions with large enough time increments. However, suppose the spectral radius is not less than one. In that case, the hidden state can experience complex nonlinear dynamics, e.g., multiple fixed points, periodicities,

and chaotic behaviour, which destroys the echo state property. A rule of thumb is that a smaller spectral radius should be used if the responses are more dependent on the input at recent times, and a larger value (but still less than one) should be used if the responses depend more on the past.

Copulas 4

This chapter is based on [Ruppert, 2011], [Smith, 2023], [Hofert et al., 2018], and [McNeil et al., 2005]

Copulas play a fundamental role in statistical modelling, mainly when dealing with multivariate data. They provide a powerful tool for understanding and analyzing the dependence structure between random variables independent of their marginal distributions. Copulas have gained significant attention in various fields, including finance, insurance, environmental sciences, and engineering.

When comparing two bivariate data sets, (X_1, X_2) and (Y_1, Y_2), in terms of their underlying variables, the linear correlation coefficients can be estimated between the parameters in each data set. If the data sets have distinct marginal distributions, this will undoubtedly affect how the potential differences in dependence are perceived. Therefore, a comparison would be more accurate if the two data sets were transformed to be comparable regarding the underlying marginal distributions. This comparison is accomplished by transforming the marginal distributions of the two bivariate data sets in this scenario to match a common distribution, such as a uniform distribution. Copulas is one such method that can be used to isolate and describe marginal behaviour and the dependence structure.

Definition 4.1 (Copula).
A distribution function in the range $[0,1]^d$ with uniform standard marginal distributions is called a d-dimensional copula.

A function $C : [0,1]^d \mapsto [0,1]$, where

$$C(\mathbf{u}) = C(u_1, \ldots, u_d),$$

with $\mathbf{u} = (u_1, \ldots, u_d)^\top$ denoting the marginal distribution functions, is a copula if the following three properties are fulfilled:

1. $C(u_1, \ldots, u_d)$ is increasing in each element u_i.

2. $C(1, \cdots, 1, u_i, 1, \cdots, 1) = u_i$ for all $i \in \{1, \cdots, d\}$ and $u_i \in [0, 1]$.

3. The rectangle inequality: For all $(a_1, \ldots, a_d)$, $(b_1, \cdots, b_d) \in [0, 1]^d$, where $a_i \leq b_i$, it follows that

$$\sum_{i_1=1}^{2} \ldots \sum_{i_d=1}^{2} (-1)^{i_1+\ldots+i_d} C\left(u_{1i_1}, \ldots, u_{di_d}\right) \geq 0, \tag{4.1}$$

with $u_{j1} = a_j$ and $u_{j2} = b_j$ for $j \in \{1, \cdots, d\}$. In other words if $(U_1, \ldots, U_d)^\top$ is a random vector with distribution function C, then $\mathbb{P}(a_1 \leq U_1 \leq b_1, \ldots, a_d \leq U_d \leq b_d) > 0$.

The first property is required of any multivariate distribution function, and the second is required of uniform marginal distributions. The third property is less obvious, but the so-called rectangle inequality in (4.1) ensures that if the random vector $(U_1, \ldots, U_d)'$ has df C, then $P\left(a_1 \leqslant U_1 \leqslant b_1, \ldots, a_d \leqslant U_d \leqslant b_d\right)$ is non-negative. Note also that, for $2 \leqslant k < d$, the k-dimensional margins of a d-dimensional copula are themselves copulas.

The requirement of standard uniform margins in Definition 4.1 can be regarded as arbitrary. The important message is that the way a multivariate distribution is "standardised" from the point of view of its margins does not alter the philosophy behind the concept of a copula. However, the choice of $U(0, 1)$ margins is sensible due to the following proposition. Proposition 4.1 introduces some important operations, *probability* and *quantile transformation*, used when considering copulas.

Proposition 4.1.
Let F denote a distributional function, and $F^{\leftarrow}$ its generalised inverse. That is, $F^{\leftarrow}(y) = \inf\{x \in \mathbb{R} : F(x) \geq y\}$.

1. Probability transformation: Let Y have continuous univariate distribution function F, then $F(Y) \sim Unif(0, 1)$.

2. Quantile transformation: Let $U \sim Unif(0, 1)$, then $\mathbb{P}\left(F^{\leftarrow}(U) \leq y\right) = F(y)$.

Proof. Let $u \in (0, 1)$ and $y \in \mathbb{R}$.

Probability transformation:

$$\mathbb{P}(F(Y) \leqslant u) = \mathbb{P}\left(F^{\leftarrow} \circ F(Y) \leqslant F^{\leftarrow}(u)\right) = \mathbb{P}\left(Y \leqslant F^{\leftarrow}(u)\right) = F \circ F^{\leftarrow}(u) = u.$$

The first equality follows by the fact that $F^{\leftarrow}$ is strictly increasing by applying the property that for an increasing F, it holds that F is continuous $\Leftrightarrow$ $F^{\leftarrow}$ is strictly increasing. The second equality follows by the property that if Y is a random variable with distribution function F, then $\mathbb{P}\left(F^{\leftarrow} \circ F(Y) = Y\right) = 1$. The last equality follows by the property that if F is increasing and $F^{\leftarrow} < \infty$, then F being continuous entails that $F \circ F^{\leftarrow}(y) = y$.

Quantile transformation:
Using

$$F(y) \geq u \Leftrightarrow F^{\leftarrow}(u) \leq y,$$

it follows that

$$\mathbb{P}\left(F^{\leftarrow}(U) \leq y\right) = \mathbb{P}\left(U \leq F(y)\right) = F(y).$$

□

For continuous and strictly increasing distribution functions F, $F^{\leftarrow}$ equals the ordinary inverse F^{-1}. The probability transformation transforms a random variable with continuous distribution function F to a standard uniform random variable. The continuity of F is crucial since, if not, the range of F would not contain $(0, 1)$. The quantile transformation transforms standard uniform random variables into variates from a distribution with distribution function F. In this case, it should be noted that F does not need to be continuous.

Copulas are essential in the investigation of multivariate distribution functions. *Sklar's Theorem* is regarded as the central theorem of copula theory, explaining how copulas play a crucial role in determining the dependence structure among the components of a random vector. Moreover, it explains why copulas determine the dependence between the components of a random vector. Given a univariate df F, $\operatorname{ran} F = \{F(x) : x \in \mathbb{R}\}$ denotes the range of F and $F^{\leftarrow}$ denotes the quantile function associated with F. Recall that the latter is merely the ordinary inverse F^{-1} if F is continuous and strictly increasing.

Theorem 4.2 (Sklar's Theorem).

1. For any d-dimensional df F having univariate margins $F_1, \ldots, F_d$, a d-dimensional copula C exists such that

$$F(\boldsymbol{x}) = C\left(F_1(x_1), \ldots, F_d(x_d)\right), \quad \boldsymbol{x} \in \mathbb{R}^d \tag{4.2}$$

The copula C is uniquely defined on $\prod_{j=1}^{d} \operatorname{ran} F_j$ and is given by

$$C(\boldsymbol{u}) = F\left(F_1^{\leftarrow}(u_1), \ldots, F_d^{\leftarrow}(u_d)\right), \quad \boldsymbol{u} \in \prod_{j=1}^{d} \operatorname{ran} F_j. \tag{4.3}$$

2. Conversely, given a d-dimensional copula C and univariate dfs $F_1, \ldots, F_d$, F defined by (4.2) is a d-dimensional df with margins $F_1, \ldots, F_d$.

Proof. The following proof of existence and uniqueness of a copula is given for the case with $F_1, \cdots, F_d$ being continuous. The converse is given in its general form.

For $x_1, \cdots, x_d \in \bar{\mathbb{R}} = [-\infty, \infty]$, it is inferred that if X has distribution function F, then

$$F(x_1, \cdots, x_d) = \mathbb{P}\left(F_1(X_1), \leq F_1(x_1), \cdots, F_d(X_d) \leq F_d(x_d)\right).$$

Because $F_1, \cdots, F_d$ are continuous, it follows from Proposition 4.1 and Definition 4.1 that the distribution function of $(F_1, \cdots, F_d)$ is a copula, denoted C. Hence the identity (4.2) is obtained.

Evaluating (4.2) at the argument $x_i = F_i^{\leftarrow}(u_i)$, for $0 \leq u_i \leq 1$, with $i = 1, \cdots, d$, and using that if F is continuous then $F \circ F^{\leftarrow}(y) = y$, the following is obtained:

$$C(u_1, \cdots, u_d) = F\left(F_1^{\leftarrow}(u_1), \cdots, F_d^{\leftarrow}(u_d)\right). \tag{4.4}$$

This establishes uniqueness by explicitly expressing C in terms of F and its margins.

Assume that C is a copula in the opposite assertion and that $F_1, \cdots, F_d$ are univariate distribution functions. A random vector is created using (4.2) by setting $X := \left(F_1^{\leftarrow}(U_1), \cdots, F_d^{\leftarrow}(U_d)\right)$, where U is a random vector with distribution function C. It is hereafter verified, using that since F is right-continuous then $F(x) \geq y \Leftrightarrow F^{\leftarrow}(y) \leq x$, that

$$\begin{aligned}
\mathbb{P}\left(X_1 \leq x_1, \cdots, X_d \leq x_d\right) &= \mathbb{P}\left(F_1^{\leftarrow}(U_1) \leq x_1, \cdots, F_d^{\leftarrow}(U_d) \leq x_d\right) \\
&= \mathbb{P}\left(U_1 \leq F_1(x_1), \cdots, U_d \leq F_d(x_d\right) \\
&= C\left(F_1(x_1), \cdots, F_d(x_d)\right).
\end{aligned}$$

□

According to Sklar's Theorem, copulas refer to functions that combine the univariate marginal distribution functions $F_1, \cdots, F_d$ to form a d-dimensional distribution

function F. Essentially, copulas establish a connection or a "coupling" between multivariate distribution functions and their marginal distributions. Consequently, copulas are of significant interest when examining the dependence among the components of a random vector. The copula C is unique if the margins are uniform; otherwise, C is uniquely determined on the ranges of F_i for $i = 1, \ldots, d$. Conversely, F is a joint distribution function with margins $F_1, \cdots, F_d$ if C is a copula and the margins are univariate distributed functions.

The extension of a copula to multivariate distribution functions with continuous margins is shown or demonstrated in equation (4.4). In addition, (4.4) demonstrates how copulas represent reliance on a quantile scale because the value $C(u_1, \cdots, u_d)$ is the joint probability that X_1 sits below its u_1-quantile, X_2 sits below its u_2-quantile, and so forth. Also, according to Theorem 4.2, defining the concept of a distribution's copula in the context of continuous margins makes sense.

A copula can also be expressed in terms of its density, obtained for the continuous case by differentiating Equation (4.2) in the following way:

$$f(x_1, \cdots, x_d) = \frac{\partial^d}{\partial x_1 \cdots \partial x_d} F(x_1, \cdots, x_d) = c\left(F_1(x_1), \cdots, F_d(x_d)\right) \prod_{j=1}^{d} f_j(x_j),$$

with $f_j = \frac{\partial}{\partial x_j} F_j$, and $c(\mathbf{u}) = \frac{\partial^d}{\partial u_1 \cdots \partial u_m} C(\mathbf{u})$ for $\mathbf{u} = (u_1, \cdots, u_d)^\top$ being the *copula density.*

Definition 4.2 (Copula of F).
Let $\mathbf{X}$ be a random vector having a joint distribution function F, with continuous marginal distributions $F_1, \ldots, F_d$. Then the distribution, C, of $\left(F_1(X_1), \cdots F_d(X_d)\right)$ is the copula of F (or X).

Under strictly increasing marginal transformations, a valuable property of the copulas of a distribution is that it remains unchanged.

Proposition 4.3 (Invariance Principle).
Assume $(X_1, \cdots, X_d)$ is a random vector with continuous margins and a copula C. Assume further, that $(T_1, \cdots, T_d)$ is a set of strictly increasing functions. Then $\left(T_1(X_1), \cdots, T_d(X_d)\right)$ has copula C as well.

Proof. The transformed variable $T_i(X_i)$ is first demonstrated to have continuous distribution function $\tilde{F}_i(y) := F_i \circ T_i^{\leftarrow}(y)$. Using that for T increasing and $T^{\leftarrow}(y) < y$

it follows that if T is strictly increasing then $T \circ T^{\leftarrow}(y) = y$, it is first seen that

$$\tilde{F}_i(y) = \mathbb{P}\left(X_i \leq T_i^{\leftarrow}(y)\right) = \mathbb{P}\left(T_i^{\leftarrow} \circ T_i(X_i) \leq T_i^{\leftarrow}(y)\right).$$

Applying the rule that, since $T^{\leftarrow}$ is an increasing transformation, $\mathbb{P}(F(X) \leqslant F(x)) = \mathbb{P}(X \leqslant x)$, where F denotes the distribution function of the random variable X, results in

$$\tilde{F}_i(y) = \mathbb{P}\left(T_i\left(X_i\right) \leqslant y\right) + \mathbb{P}\left(X_i = T_i^{\leftarrow}(y), T\left(X_i\right) > y\right).$$

The second probability on the right hand side equals 0, because F is continuous. C being X's copula makes it possible to determine that

$$\begin{aligned} C\left(u_1, \ldots, u_n\right) &= \mathbb{P}\left(F_1\left(X_1\right) \leqslant u_1, \ldots, F_d\left(X_d\right) \leqslant u_d\right) \\ &= \mathbb{P}\left(\tilde{F}_1\left(T_1\left(X_1\right)\right) \leqslant u_1, \ldots, \tilde{F}_d\left(T_d\left(X_d\right)\right) \leqslant u_d\right), \end{aligned}$$

where $\tilde{F}_i \circ T_i(x) = F_i \circ T_i^{\leftarrow} \circ T_i(x) = F_i(x)$. It then follows from Definition 4.2 that C is a copula of $\left(T_1(X_1), \cdots, T_d(X_d)\right)$. □

The following theorem involving the *Fréchet-Hoeffding* bounds is a cornerstone of copula theory. The upper and lower Fréchet-Hoeffding bounds M and W are referred to as any copula C's pointwise bounds. In other words, for every given set of marginal distributions, M and W are the greatest and smallest possible value, respectively, that the copula can have.

Theorem 4.4 (Fréchet-Hoeffding Bounds).
The limits for each copula $C\left(u_1, \cdots, u_d\right)$ are given as

$$w(\mathbf{u}) = \max\left\{\sum_{i=1}^{d} u_i + 1 - d, 0\right\} \leq C(\mathbf{u}) \leq \min\{u_1, \cdots, u_d\} = M(\mathbf{u}).$$

Proof. The second inequality results because for all i,

$$\bigcap_{i \leq j \leq d} \left\{U_j \leq u_j\right\} \subset \left\{U_i \leq u_i\right\}.$$

The first inequality follows because

$$C(\boldsymbol{u}) = P\left(\bigcap_{1 \leqslant i \leqslant d} \left\{U_i \leqslant u_i\right\}\right) = 1 - P\left(\bigcup_{1 \leqslant i \leqslant d} \left\{U_i > u_i\right\}\right)$$

$$\geqslant 1 - \sum_{i=1}^{d} P\left(U_i > u_i\right)$$
$$= 1 - d + \sum_{i=1}^{d} u_i.$$

□

Section 5.2 constructs a model employing a *Gaussian* copula, which will be elaborated on shortly. It is an *implicit copula* i.e., a copula defined by (4.3). The term implicit copula given in [Smith, 2023] will now be formally introduced since this concept is utilised throughout the application. The term is used for the copula, which is implicit in the multivariate distribution of a continuous random vector $\mathbf{Z} = (Z_1, \ldots, Z_m)^\top$, and are constructed from multivariate distributions that already exist. This copula family is broad and flexible, and they all share an auxiliary representation that makes estimation manageable in high dimensions.

If $\mathbf{Z}$ has distribution function F_Z with marginals $F_Z, \ldots, F_{Z_m}$, then its implicit copula function is given by

$$C_Z(\mathbf{u}) = F_Z\left(F_{Z_1}^{-1}(u_1), \cdots, F_{Z_m}^{-1}(u_m)\right). \tag{4.5}$$

The implicit copula density is obtained by differentiating (4.5) with respect to $\mathbf{u}$;

$$c_Z(\boldsymbol{u}) = \frac{\partial^m}{\partial u_1 \cdots \partial u_m} C(\boldsymbol{u}) = \frac{f_Z(\boldsymbol{z})}{\prod_{j=1}^{m} f_{Z_j}\left(z_j\right)}$$

Here $\mathbf{z} = (z_1, \cdots, z_m)^\top$ is a function of $\boldsymbol{u}$, with $z_j = F_{Z_j}^{-1}(u_j)$, $j = 1, \cdots, m$. The implicit copula model uses Sklar's theorem twice, once to form the joint distribution F_Y with arbitrary marginals and a second time to construct the implicit copula from the joint distribution F_Z.

Copula models can be viewed as transformations from $\mathbf{Y}$ to $\mathbf{U} = (U_1, \cdots, U_m)^\top \in [0,1]^m$. Instead of looking directly at the domain of $\mathbf{Y}$, it is typically simpler to capture multivariate dependence using C on the vector space $[0,1]^m$. As mentioned implicit copulas have a second transformation from $\mathbf{U}$ to $\mathbf{Z} = \left(F_{Z_1}^{-1}(U_1), \ldots, F_{Z_d}^{-1}(U_d)\right)^\top$, capturing the dependence structure using the distribution of $\mathbf{Z}$, F_Z. The vector $\mathbf{U}$ are referred to as the copula vector and $\mathbf{Z}$ the auxiliary vector, where a pseudo code for simulation from an implicit copula model is provided below, in the case where F_Z is tractable. The algorithm provides a draw $\mathbf{y} \sim F_Y$.

Algorithm 2 Random iterate generation from an implicit copula model

1: Generate $\boldsymbol{z} = (z_1, \ldots, z_m)^\top \sim F_Z$
2: For $j = 1, \ldots, m$, set $u_j = F_{Z_j}(z_j)$, and $\boldsymbol{u} = (u_1, \ldots, u_m)^\top$
3: For $j = 1, \ldots, m$, set $y_j = F_{Y_j}^{-1}(u_j)$, and $\boldsymbol{y} = (y_1, \ldots, y_m)^\top$

Notice that the transformation $U_j = F_{Z_j}(Z_j) \sim \text{Uniform}[0, 1]$ removes all features of the marginal distribution of Z_j. This becomes an important observation for establishing parameter identification when constructing implicit copulas.

The before-mentioned Gauss copula belongs to the family of *elliptical copulas*, which is extensively utilised copulas in practical applications. They describe the dependence structure of the multivariate normal distribution using the Gauss copula and the dependence structure of the multivariate t distribution using the (Student) t copula.

As elliptical copulas are implicitly constructed through Sklar's Theorem, thus an implicit copula, their properties are typically derived from the properties of the corresponding elliptical distributions. Consequently, comprehending the construction and properties of elliptical distributions becomes crucial.

Definition 4.3 (Elliptical Distributions).
A d-dimensional random vector X has an elliptical distribution with location vector $\boldsymbol{\mu} \in \mathbb{R}^d$, scale (or dispersion) matrix $\Sigma = AA'$ where $\text{rank}(\Sigma) = k \leq d$ for a matrix $A \in \mathbb{R}^{d \times k}$ and radial part $R \geq 0$ if

$$\boldsymbol{X} \overset{d}{=} \mu + A\boldsymbol{Y}, \qquad \text{for } \boldsymbol{Y} \overset{d}{=} RS,$$

where R and S are independent and where $S \sim \text{Unif}\left(\left\{\boldsymbol{x} \in \mathbb{R}^k : \|\boldsymbol{x}\| = 1\right\}\right)$, $\|\cdot\|$ denotes the Euclidean norm, that is $\boldsymbol{S}$ is uniformly distributed on the unit sphere in $\mathbb{R}^k$. The distribution of $\boldsymbol{Y}$ is known as spherical distribution.

By the implicit construction, elliptical copulas are of the form (4.3), with F denoting a multivariate elliptical df and $F_1, \ldots, F_d$ the corresponding univariate margins. In accordance with Proposition 4.3, marginal location-scale modifications made before using (4.3) will not affect the copula. Hence, F can be assumed such that $\boldsymbol{\mu} = \mathbf{0} = (0, \ldots, 0)$ and Σ is a correlation matrix P. These assumptions imply that $F_1 = \cdots = F_d = F$, that is, the univariate margins of F are identical. When using an arbitrary df (on $[0, \infty)$) for the radial component R, the common univariate marginal df F of X frequently loses tractability.

Due to (4.3), it may be challenging to evaluate an elliptical copula C when the

evaluation of F or its marginal quantile function $F^{\leftarrow}$ is numerically challenging. For example, randomised quasi-Monte Carlo methods are used to evaluate multivariate normal and t dfs in three or more dimensions.

If $\boldsymbol{Y} \sim N_d(\boldsymbol{\mu}, \Sigma)$ is a Gaussian random vector, then its copula is referred to as a so-called Gauss copula. Because the process of standardising the margins entails performing a sequence of strictly increasing transformations, the copula of $\boldsymbol{Y}$ and the copula of $\boldsymbol{X} \sim N_d(\mathbf{0}, P)$, where $P = \rho(\Sigma)$ is the correlation matrix of $\boldsymbol{Y}$, are identical by Proposition 4.3. By Definition 4.2 the Gauss copula is given by

Definition 4.4 (Gauss Copula).

$$C_P^{\mathrm{Ga}}(\boldsymbol{u}) = P\left(\Phi\left(X_1\right) \leqslant u_1, \ldots, \Phi\left(X_d\right) \leqslant u_d\right) \tag{4.6}$$

$$= \boldsymbol{\Phi}_P\left(\Phi^{-1}\left(u_1\right), \ldots, \Phi^{-1}\left(u_d\right)\right), \tag{4.7}$$

with Φ denoting the standard univariate normal df and $\boldsymbol{\Phi}_P$ denoting the joint df of $\boldsymbol{X}$.

The notation C_P^{Ga} emphasises that the copula is parameterised by the $\frac{1}{2}d(d-1)$ parameters of the correlation matrix; in two dimensions we write C_ρ^{Ga}, where $\rho = \rho\left(X_1, X_2\right)$

The Gauss copula does not have a simple closed form but can instead be expressed as an integral over the density of $\boldsymbol{X}$; for$|\rho|< 1$ in two dimensions it follows, using (4.6), that

$$C_\rho^{\mathrm{Ga}}\left(u_1, u_2\right) = \int_{-\infty}^{\Phi^{-1}(u_1)} \int_{-\infty}^{\Phi^{-1}(u_2)} \frac{1}{2\pi\left(1-\rho^2\right)^{1/2}} \exp\left\{\frac{-\left(s_1^2 - 2\rho s_1 s_2 + s_2^2\right)}{2\left(1-\rho^2\right)}\right\} \mathrm{d}s_1 \, \mathrm{d}s_2$$

For $d = 2$ and P having off-diagonal entries $\rho = -1$, such that $C_P^{\mathrm{n}} = C_\rho^{\mathrm{n}}$ equals the lower Fréchet-Hoeffding bound W, and for $d \geq 2$ and P having off-diagonal entries equal to $\rho = 1$, the homogeneous Gauss copula $C_P^{\mathrm{n}} = C_\rho^{\mathrm{n}}$ equals the upper Fréchet-Hoeffding bound M.

Models 5

In this chapter, two of the models from *Klein et al., 2023*, namely the *Gaussian Probabilistic ESN*, and the *Copula model* are set up with the help of the preceding theory.

5.1 Gaussian Probabilistic ESN

This section is based on [Klein et al., 2023] and [McDermott and Winkle, 2017]

Classical ESNs rarely consider uncertainty quantification, and in this section, a probabilistic ESN that does so is outlined.

When predicting nonlinear spatio-temporal processes, it can be useful to include quadratic interactions between hidden processes and the response and embeddings, i.e., lagged values of the input. [McDermott and Winkle, 2017] have shown this will increase predictive accuracy for series with highly nonlinear dependence. These quadratic interactions are represented with simple modifications of the basic ESN, referred to as a basic quadratic ESN, QESN. A type of QESN is stated below in Definition 5.1 for continuous output, i.e., where $f_o(\cdot)$ is the identity function. Since the disturbances in Definition 5.1 are Gaussian, Equation (5.1) are referred to as the Gaussian probabilistic ESN.

Definition 5.1 (Gaussian probabilistic ESN).
Let $\{Y_t\}$ be a stochastic process, then the ESN with Gaussian disturbances $\varepsilon_t \sim \mathcal{N}\left(0, \sigma^2\right)$ takes the following form for $t \geq 1$:

$$\text{Response Equation :} \quad Y_t = \beta_0 + \boldsymbol{h}_t'\boldsymbol{\beta}_1 + \left(\boldsymbol{h}_t'\right)^{\circ 2}\boldsymbol{\beta}_2 + \varepsilon_t, \tag{5.1}$$

$$\text{Hidden State Equation:} \quad \boldsymbol{h}_t = (1-\kappa)\boldsymbol{h}_{t-1} + \kappa\tilde{\boldsymbol{h}}_t \tag{5.2}$$

$$\tilde{\boldsymbol{h}}_t = f_W\left(\frac{\delta}{\lambda_W}W\boldsymbol{h}_{t-1} + U\boldsymbol{x}_t\right) \tag{5.3}$$

In the definition above, '$\circ 2$' denotes the element-wise square of a matrix, and the parameter $0 < \kappa \leq 1$ is the leaking rate. The constant λ_W is the largest eigenvalue of W, and $0 < \delta < 1$ is a scaling parameter so that $\left(\delta/\lambda_W\right) W$ has a spectral radius less than one.

The elements of the matrices $W = \{w_{il}\}, U = \left\{u_{ij}\right\}$ are assumed to be random and distributed independently from mixtures of a uniform distribution and a point mass at zero. If $\mathcal{U}(a, b)$ denotes a uniform distribution over domain $(a, b), \mathcal{B}(\pi)$ denotes a Bernoulli distribution with mean π, and δ_0 is the Dirac function at zero, then the weights are given as,

$$w_{il} = \gamma_{il}^w \mathcal{U}\left(-a_w, a_w\right) + \left(1 - \gamma_{il}^w\right) \delta_0, \quad \gamma_{il}^w \sim \mathcal{B}\left(\pi_v\right), \tag{5.4}$$

$$u_{ij} = \gamma_{ij}^u \mathcal{U}\left(-a_u, a_u\right) + \left(1 - \gamma_{ij}^u\right) \delta_0, \quad \gamma_{ij}^u \sim \mathcal{B}\left(\pi_u\right). \tag{5.5}$$

Following [McDermott and Winkle, 2017] $a_v = a_u = \pi_v = \pi_u = 0.1$ providing a sparse structure.

Equations (5.1)-(5.3) can be written as a linear model,

$$\boldsymbol{Y} = B_\xi(X)\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_T)' \sim \mathcal{N}\left(\boldsymbol{0}, \sigma^2 I\right). \tag{5.6}$$

Here $\boldsymbol{Y} = (Y_1, \ldots, Y_T)'$ for T time series observations of the stochastic process with corresponding $T \times n_x$ matrix of feature values $X = \left[\boldsymbol{x}_1 | \cdots | \boldsymbol{x}_T\right]'$. $\boldsymbol{\xi} = \{W, U, \kappa, \delta\}$, $H_\xi(X) = \left[\boldsymbol{h}_1 | \cdots | \boldsymbol{h}_T\right]'$ is the $T \times n_h$ matrix of hidden state values, $B_\xi(X) = \left[\boldsymbol{\iota}, H_\xi(X) \mid H_\xi(X)^{\circ 2}\right]$, with $\boldsymbol{\iota}$ as a vector of ones, and $\boldsymbol{\beta} = \left(\beta_0, \boldsymbol{\beta}_1', \boldsymbol{\beta}_2'\right)'$.

The hidden state matrix $H_\xi(X)$ is known without errors provided $\boldsymbol{\xi}, X$ and $\boldsymbol{h}_0 = \boldsymbol{0}$ are given. This is because the hidden state vectors may be calculated recursively. Only the two model parameters $\boldsymbol{\beta}$ and σ^2 require estimation. Their Bayesian posterior distribution is used for this purpose. A regularisation of $\boldsymbol{\beta}$ is done by adopting the shrinkage prior.

$$\boldsymbol{\beta} \mid \tau^2 \sim \mathcal{N}\left(\boldsymbol{0}, P\left(\tau^2\right)^{-1}\right), \quad \sigma^2 \sim \mathcal{IG}(a, b) \tag{5.7}$$

with $\mathcal{IG}$ denoting an Inverse Gamma distribution. A ridge prior with $P\left(\tau^2\right) = \tau^2 I$, and hyper-prior $\tau^2 \sim \mathcal{IG}(\tilde{a}, \tilde{b})$ will be used in this case. The regularised linear model's parameters, $\boldsymbol{\vartheta} = \left(\boldsymbol{\beta}, \sigma^2, \tau^2\right)$, can be calculated using the standard MCMC sampler.

5.1.1 Probabilistic Forecasting of Gaussian Probabilistic ESN

This section is based on [Gneiting and Katzfuss, 2014], and [Klein et al., 2023]

This project uses *probabilistic forecasting*, i.e., forecasts taking the form of probability distributions over future quantities or events. Such forecasts are a crucial component of the most effective decision-making because they help to measure the uncertainty in a prediction.

The absence of a mechanism to quantify the uncertainty of model predictions in most traditional Echo State Network (ESN) applications is surprising, considering that the reservoir weight parameters are selected randomly rather than estimated. The expectation is that the model's behaviour would exhibit variation with different sets of W and U weights, particularly when the number of hidden units is relatively small. Although conventional ESN models typically incorporate a large number of hidden units, the inclusion of multiple ensemble members with a reduced number of units is advantageous. This approach offers flexibility by preventing the risk of overfitting, allowing the ensemble members to function as a collective of relatively weak learners, and providing a more realistic estimation of prediction uncertainty for out-of-sample forecasts. Thus, an ensemble of forecasts can be generated.

Thus, instead of only a single set of weights drawn from (5.4) and (5.5) in ESN implementations, this project follows the approach of [McDermott and Winkle, 2017] and [McDermott and Wikle, 2019]. Hence, $K = 100$ matrices $\{W^k, U^k; k = 1, \ldots, K\}$ are simulated from (5.4) and (5.5). The probabilistic forecasts are subsequently constructed by integrating over the weight matrices U and W using an ensemble.

Let $\xi^k = \left\{W^k, U^k, \kappa, \delta\right\}$. The following ensemble is then the density forecast of Y_{T+h} at time T where $h = 1, \ldots, h_1$ in the forecast window, is then the ensemble

$$f_{T+h|T}\left(y_{T+h}\right) \equiv \frac{1}{K}\sum_{k=1}^{K} p^k\left(y_{T+h} \mid X, \boldsymbol{y}\right). \tag{5.8}$$

In this case, the subscript indicates that $f_{T+h|T}$ is conditional on the filtration at time T.

In (5.8), p^k denotes the Bayesian posterior predictive density. The following equation is used to calculate this posterior predictive density for the configuration ξ^k. Given that $X_{(t)} \equiv \left[\boldsymbol{x}_1|\ldots|\boldsymbol{x}_t\right]'$, this posterior predictive density is

$$\begin{aligned} p^k\left(y_{T+h} \mid X, \boldsymbol{y}\right) = \iint p\left(y_{T+h} \mid X_{(T+h)}, \boldsymbol{\vartheta}\right) p\left(\boldsymbol{x}_{T+2}, \ldots, \boldsymbol{x}_{T+h} \mid \boldsymbol{\vartheta}, X, \boldsymbol{y}\right) \\ p(\boldsymbol{\vartheta} \mid X, \boldsymbol{y}) \mathrm{d}\boldsymbol{x}_{T+2} \ldots \mathrm{d}\boldsymbol{x}_{T+h}\, \mathrm{d}\boldsymbol{\vartheta}. \end{aligned} \tag{5.9}$$

Keep in mind that this constitutes an integral over any unobserved feature value and the posterior of the model parameter. The first term in the integrand is by Definition 5.1 the density of a $\mathcal{N}\left(\boldsymbol{h}'_{T+h}\boldsymbol{\beta}_1 + \left(\boldsymbol{h}'_{T+h}\right)^{\circ 2}\boldsymbol{\beta}_2, \sigma^2\right)$ distribution. Here $\boldsymbol{h}_{T+h}$ is determined by the recursion of the hidden state equation. As a result, $\boldsymbol{h}_{T+h}$ is a deep function of $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{T+h}$. By averaging draws from the posterior $p(\boldsymbol{\vartheta} \mid X, \boldsymbol{y})$ acquired by running the MCMC sampler, the outer integral pertaining $\boldsymbol{\vartheta}$ may be computed.

Throughout the application, $\boldsymbol{x}_t$ will provide historical data on the electricity prices in the focal area. This differs slightly from the approach presented in Klein et al. 2023 since their $\mathbf{x}_t$ also contains past values from the other regions. Hence, at or before time T, $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{T+1}$ are observed. Some elements, $\boldsymbol{x}_{T+2}, \ldots, \boldsymbol{x}_{T+h}$, are unobserved. The integral over these feature vectors in (5.9) is, therefore, only with respect to their unobserved components. The integrals are calculated using a Monte Carlo approach by simulating the values of each of the considered bidding areas in turn from their respective predictive distributions.

5.2 Copula Model

This section is based on [Gneiting et al., 2007], [Klein et al., 2023], and [Smith, 2023]

The Gaussian ESN from the previous section has two limitations. First, the feature vector only influences the mean of the response equation. Second, the density forecasts lack calibration. To address these just mentioned drawbacks, this section presents a copula model. This deep distributional time series model incorporates the feature vector to impact the entire predictive distribution. Moreover, it ensures that the probabilistic forecasts are marginally calibrated, which will be elaborated in Section 5.2.3.

A copula model for the joint distribution of $\mathbf{Y}_{(t)} = (Y_1, \cdots, Y_t)^\top$ conditional on $X_{(t)} = \left[\mathbf{x}_1 \mid \cdots \mid \mathbf{x}_t\right]^\top$ and weight configuration $\boldsymbol{\xi}$ is employed. The density decomposition of $\mathbf{Y}_{(t)}$ is given for $t \geq 2$ as

$$p\left(\mathbf{y}_{(t)} \mid X_{(t)}, \boldsymbol{\xi}\right) = c_{ESN}\left(F_Y(y_1), \cdots, F_Y(y_T) \mid X_{(t)}, \boldsymbol{\xi}\right) \prod_{s=1}^{t} p_Y(y_s), \quad \text{for } t \geq 2 \tag{5.10}$$

with $\mathbf{y}_{(t)} = (y_1, \cdots, y_t)^\top$, and $\boldsymbol{u}_{(t)} = (u_1, \ldots, u_t)'$. The deep copula process has t-dimensional density $c_{\text{ESN}}\left(\boldsymbol{u}_{(t)} \mid X_{(t)}, \boldsymbol{\xi}\right)$ specified below.

The density p_Y and distribution function F_Y remain constant over time. These two functions are estimated from the training data without using fixed parameters, called

non-parametric estimation.

In (5.10), it is assumed that the distribution $Y_t \mid \mathbf{x}_t$ is marginally invariant with regard to $\mathbf{x}_t$. In other words, it is required that, regardless of the value of $\mathbf{x}_t$, the distribution of the variable Y_t at a specific time t and a particular feature vector value $\mathbf{x}_t$ remains constant. The joint distribution, however, maintains the relationship between the distribution $Y_t \mid X_t$ and the matrix of feature vector values $X_{(t)}$. As a result, the joint distribution still has an indirect impact on the Y distribution for a particular time step.

Equation (5.10) utilises an implicit copula process with a density c_{ESN}, which is derived from the joint distribution of a second stochastic process $\tilde{Z}_s$. The latter process follows the Gaussian probabilistic ESN introduced in Definition 5.1 with the parameter β integrated out under the prior in (5.6). Aside from being used to specify the implicit copula, $\tilde{Z}_s$ is not directly observed.

Recall from (5.7) given as $\boldsymbol{\beta} \mid \tau^2 \sim \mathcal{N}\left(\mathbf{0}, P\left(\tau^2\right)^{-1}\right)$, which is assumed to be a proper prior. The distribution with β integrated out is guaranteed to be proper by this assumption.

Keeping this in mind, the t observations $\tilde{\boldsymbol{Z}}_{(t)} = \left(\tilde{Z}_1, \cdots, \tilde{Z}_t\right)^\top$ is conditionally distributed as follows

$$\tilde{\boldsymbol{Z}}_{(t)} \mid X_{(t)}, \sigma^2, \tau^2, \boldsymbol{\xi} \sim N\left(\mathbf{0}, \sigma^2\left(I - B_{\boldsymbol{\xi}}\left(X_{(t)}\right) \Sigma B_{\boldsymbol{\xi}}\left(X_{(t)}\right)^\top\right)\right), \tag{5.11}$$

where $\Sigma = \left(B_{\boldsymbol{\xi}}(X_{(t)})^\top B_{\boldsymbol{\xi}}(X_{(t)}) + P(\tau^2)\right)$, and with $B_{\boldsymbol{\xi}}(X_{(t)}) = \left[H_{\boldsymbol{\xi}}(X) \mid H_{\boldsymbol{\xi}}(X)^{\circ 2}\right]$, stated in (5.6). Because *level* in a copula is unidentified, the first column, the intercept, has been omitted. Here 'level' refers to the intercept level in the model and hence represent the level of the dependent variable when all predictor variables are zero.

The variance, or more precisely $\left(I - B_{\boldsymbol{\xi}}\left(X_{(t)}\right) \Sigma B_{\boldsymbol{\xi}}\left(X_{(t)}\right)^\top\right)$, in (5.11) can be simplified using the Woodbury formula, stated in Appendix A. Recalling from (5.7) that $P(\tau^2) = \tau^2 I$, (5.11) can be simplified to

$$\tilde{\boldsymbol{Z}}_{(t)} \mid X_{(t)}, \sigma^2, \tau^2, \boldsymbol{\xi} \sim N\left(\mathbf{0}, \sigma^2\left(I + \frac{1}{\tau^2} B_{\boldsymbol{\xi}}(X_{(t)}) B_{\boldsymbol{\xi}}(X_{(t)})^\top\right)\right). \tag{5.12}$$

By standardising $\tilde{Z}_i$ to have unit variance, the correlation matrix of (5.12), denoted R, is obtained as shown below. Let $\mathbf{Z}_{(t)} = (Z_1, \cdots, Z_t)^\top = \sigma^{-1} S \tilde{\mathbf{Z}}_{(t)}$, with $S =$

$\text{diag}(\psi_1, \cdots, \psi_t)$ being a diagonal scaling matrix with elements $\psi_s = \left(1 + \frac{\mathbf{b}_s^\top \mathbf{b}_s}{\tau^2}\right)^{-1/2}$, where $\mathbf{b}$ denotes the s'th row of $B_{\boldsymbol{\xi}}(X_{(t)})$.

This result in $\mathbf{Z}_{(t)} \mid X_{(t)}, \sigma^2, \tau^2, \boldsymbol{\xi} \sim N(0, R)$, with

$$R = S\left(I + \frac{1}{\tau^2} B_{\boldsymbol{\xi}}(X_{(t)}) B_{\boldsymbol{\xi}}(X_{(t)})^\top\right) S$$

As a result, the copula has density $c_{ESN}\left(\mathbf{u}_{(t)} \mid X_{(t)}, \boldsymbol{\xi}, \tau^2\right) = \frac{\phi(\mathbf{0}, R)}{\prod_{s=1}^{t} \phi_1(s)}$, with $\phi(\mathbf{0}, R)$ denoting the density of a $N(\mathbf{0}, R)$ distribution, ϕ_1 denotes standard normal density, $z_t = \Phi_1^{-1}(u_t)$, and $\mathbf{z}_t = (z_1, \cdots, z_t)^\top$. Because σ^2 is not featured in R and remains unidentified in the copula, it is safe to assume that $\sigma^2 = 1$ throughout.

The dependence structure in $\{Z_t\}$ is captured by the copula; hence $\mathbf{x}_t$ contains past values of this process. This can be computed as $Z_t = \Phi^{-1}\left(F_Y(Y_t)\right)$. Note that since c_{ESN} is conditional on $X_{(t)}$, this Gaussian copula is said to be a process on the feature space.

5.2.1 Estimation

The sole unknown copula parameter for a configuration ξ is τ^2, for which the Weibull prior in [Klein and Kneib, 2016] with scale parameter $b_{\tau^2} = 2.5$ is employed.

Because the evaluation of $c_{ESN}\left(\mathbf{u} \mid X, \boldsymbol{\xi}, \tau^2\right)$ necessitates inversion of the correlation matrix R, which is computationally demanding for all but small sample sizes, direct estimation using the likelihood in (5.10) is challenging, even though the likelihood is given on closed form.

By stating the likelihood conditional on $\boldsymbol{\beta}$ as follows, this can be circumvented effectively. Consider, therefore, a sample of size T. Indicate the feature matrix as $X \equiv X_{(T)}$ and the observations as $\mathbf{y} \equiv \mathbf{y}_{(T)}$. Changing the variables from $\mathbf{y}$ to $\mathbf{z} = (z_1, \cdots, z_t)^\top$, where $z_t = \Phi_1^{-1}\left(F_Y(y_t)\right)$ yields the following conditional likelihood

$$p\left(\mathbf{y} \mid X, \boldsymbol{\beta}, \boldsymbol{\xi}, \tau^2\right) = p\left(\boldsymbol{z} \mid X, \boldsymbol{\beta}, \boldsymbol{\xi}, \tau^2\right) \prod_{t=1}^{T} \frac{p_Y(y_t)}{\phi_1(z_t)} = \phi\left(\mathbf{0}, S B_{\boldsymbol{\xi}(X), S^2}\right) \prod_{t=1}^{T} \frac{p_Y(y_t)}{\phi_1(z_t)}. \tag{5.13}$$

Since S is a diagonal matrix, (5.13) can be evaluated in $O(T)$ operations. This computationally effective method enables estimation to be completed in an acceptable amount of time, even for large datasets.

This is estimated by an MCMC sampler which generated draws from the augmented

posterior of $(\boldsymbol{\beta}, \tau^2)$, such that $\boldsymbol{\beta}$ is integrated out and avoiding direct computation of R.

5.2.2 Probabilistic Forecast of the Copula Model

This section is based on [Klein et al., 2023] and [Smith, 2023]

Following a similar approach as in Section 5.1.1, the forecast is once again provided by (5.8), which is repeated below by using an ensemble to integrate over the distribution of U and V.

$$f_{T+h|T}\left(y_{T+h}\right) \equiv \frac{1}{K} \sum_{k=1}^{K} p^k\left(y_{T+h} \mid X, \boldsymbol{y}\right). \tag{5.8}$$

To recall, equation (5.9) is shown below.

$$\begin{aligned} p^k\left(y_{T+h} \mid X, \boldsymbol{y}\right) = \iint p\left(y_{T+h} \mid X_{(T+h)}, \boldsymbol{\vartheta}\right) p\left(\boldsymbol{x}_{T+2}, \ldots, \boldsymbol{x}_{T+h} \mid \boldsymbol{\vartheta}, X, \boldsymbol{y}\right) \\ p(\boldsymbol{\vartheta} \mid X, \boldsymbol{y}) \mathrm{d}\boldsymbol{x}_{T+2} \ldots \mathrm{d}\boldsymbol{x}_{T+h} \, \mathrm{d}\boldsymbol{\vartheta}. \end{aligned} \tag{5.9}$$

The copula model is used to obtain the ensemble components p^k or the Bayesian posterior predictive densities in equation (5.9), where $\boldsymbol{\vartheta} = \left\{\boldsymbol{\beta}, \tau^2\right\}$.

By switching the variable y_{T+h} to $z_{T+h} = \Phi_1^{-1}\left(F_Y(y_{T+h})\right)$, the first term in the integrand of (5.9) is derived, such that

$$\begin{aligned} p\left(y_{T+h} \mid X_{(T+h)}, \boldsymbol{\vartheta}\right) &= p\left(z_{T+h} \mid X_{(T+h)}, \boldsymbol{\beta}, \tau^2\right) \frac{p_Y(y_{T+h})}{\phi_1(Z_{T+h})} \\ &= \frac{1}{\psi_{T+h}} \phi_1\left(\frac{\Phi_1^{-1}\left(F_y(y_{T+h})\right) - \mu_{T+h}}{\psi_{T+h}}\right) \frac{p_Y(Y_{T+h})}{\phi_1\left(\Phi_1^{-1}\left(F_Y(y_{T+h})\right)\right)}. \end{aligned} \tag{5.14}$$

In the above $\psi_{T+h} = \left(1 + \frac{1}{\tau^2} \mathbf{b}_{T+h} \mathbf{b}_{T+h}^{\top}\right)^{-1/2}$, $\mu_{T+h} = \frac{1}{\psi_{T+h}} \mathbf{b}_{T+h} \boldsymbol{\beta}$, where $\mathbf{b}_{T+h} = \left(\mathbf{h}_{T+h}^{\top}, \left(\mathbf{h}_{T+h}^2\right)^{\top}\right)$ denotes a row vector.

In this case, predictive distribution in (5.14) of the variable Y_{t+h} is a non-linear function of the feature vector $\mathbf{x}_{T+h}$, which result in Y_{T+h} not being marginally invariant of $\mathbf{x}_{T+h}$.

The posterior mean $\mathbb{E}(\boldsymbol{\vartheta} \mid \mathbf{y})$ is obtained using Monte Carlo samples and is subsequently inserted as $\boldsymbol{\beta}$ and τ^2 in (5.14), which can then be used to estimate the response variable at a future time point.

5.2.3 Marginal Calibration

The Gaussian Probabilistic ESN model described in Section 5.1 offers probabilistic forecasts, but it also has significant limitations, including the fact that the feature vector only impacts the mean of the response equation and that no calibration has been applied to the density forecasts.

Different types of calibration exist, including probabilistic, exceedance, and predictive marginal calibration. Marginal calibration will be used throughout the project when considering the copula model; hence, only marginal calibration will be described in this part. An introduction of the various calibration techniques can be found in [Gneiting et al., 2007].

Definition 5.2 (Marginal Calibration).
Let $t = T+1, \cdots, \infty$. Assume that a stochastic process Y_t has a future observation with true distribution $H_{t|T}(y)$ and forecasting distribution $F_{t|T}(y)$. Here the subscript shows that both distributions are conditional on the filtration at time T. Define the limits as

$$\bar{H}(y) \equiv \lim_{h_1 \to \infty} \left\{ \frac{1}{h_1} \sum_{h=1}^{h_1} H_{T+h|T}(y) \right\}$$
$$\bar{F}(y) \equiv \lim_{h_1 \to \infty} \left\{ \frac{1}{h_1} F_{T+h|T}(y) \right\}.$$

The forecast distribution is said to be marginally calibrated if and only if

$$\bar{H}(y) = \bar{F}(y).$$

Note that $H_{t|T}(y)$ is unknown for $t > T$ and hence a direct comparison of $\bar{H}(y)$ and $\bar{F}(y)$ cannot be made. Instead, an approach would be to compare the empirical distribution function of the actual observations over the forecast period to the empirical distribution function of the corresponding forecast distribution.

Application 6

The models described in Chapter 5 are currently being applied to the data provided by Nord Pool. Initially, a concise overview of the data is presented before utilising the models introduced in Chapter 5 with the given data.

6.1 Data Introduction

This section presents the data used throughout the application and a preliminary analysis.

The information used has been provided by Nord Pool and spans the period between 01/01/2019 and 05/05/2023. The data includes hourly intraday details regarding trades and, consequently, prices measured in EUR/MWh made in the Nordic regions. Coordinated universal time, or UTC, is used for all timestamps provided in the data.

This project examines the electricity price information from 12 distinct Nordic regions located in the countries - Denmark, Norway, Sweden, and Finland. The regions are designated as DK1, DK2, NO2, NO3, NO4, NO5, SE1, SE2, SE3, SE4, and FI. They each represent a separate geographical location. A map showing the regions is provided in Figure 6.1 below to give a visual idea of their geographic location.

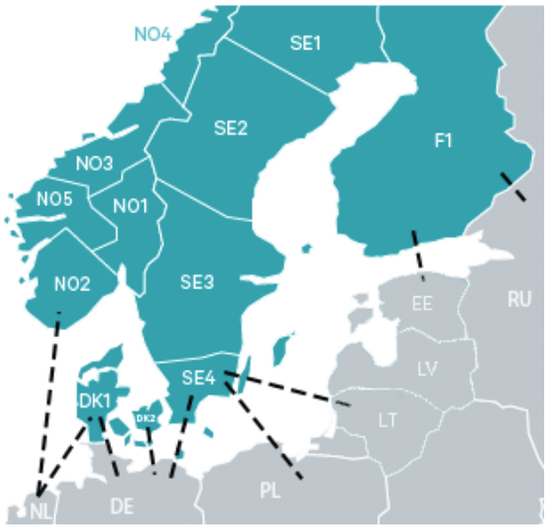

Figure 6.1: A map displaying the Nord Pool market coupling. That is, this maps shows the geographical location of the regions in the Nordic countries, Denmark, Norway, Sweden, and Finland. Illustration from [*Li and Becker, 2021*]

DK1 represents Denmark's western region, whereas DK2 represents Denmark's eastern region which includes the nation's capital city of Copenhagen. The five regions in Norway cover the nation's extensive coastline and mountainous landscape. Oslo, the nation's capital, is included in NO1's coverage of the southeast, while NO2 includes the southwest. The three regions, NO3, NO4, and NO5, represent Norway's northern and central regions. Sweden comprises four regions, with SE1 covering the northernmost portion and the centre covered by SE2. The metropolitan regions of Stockholm and Gothenburg are included in SE3. The southwestern region of Sweden, including the surrounding area of Malmö, is a part of SE4. Last is the region FI, which covers the entire nation of Finland.

The time series of DK1 is shown in Figure 6.2 as an example of price evolution and movements of one of the investigated areas. In Appendix C, plots of all the considered areas for the same period are shown.

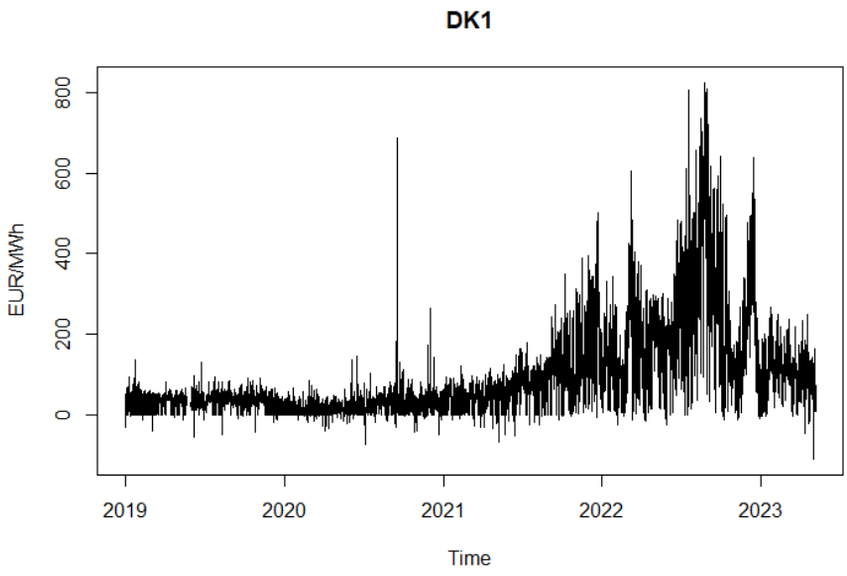

Figure 6.2: Illustration of the DK1 time series, that is price movements for the period between 01/01/2019 and 05/05/2023. The time is depicted along the y axis, while the price of electricity is displayed along the y axis in EUR/MWh.

All regions share a common characteristic: an increase in volatility in their respective time series during the latter half of 2021 and throughout 2022. The economic recovery following Covid-19 and the relaxations of travel restrictions may be the cause of the increased volatility in 2021. Another aspect of the observed increase in volatility could potentially be climate changes, an example could be the increased demand for energy used for cooling during summer heat waves experienced across Europe. The Russian invasion of Ukraine on February 24, 2022, is another event that could have contributed to increased volatility, particularly in 2022. As a result of the European Union's sanctions on Russia, Russia opted to fully stop supplying gas to a number of European nations, creating supply instability. Given that electricity and heat generation account for 31.4% of the European gas supply, this may have had a direct impact on electricity prices [Council, 2023c]. [Council, 2023a], and [Council, 2023b]

NO3 and NO4, see Appendix C are a couple of the time series that stand out. These series show that a significant portion of the EUR/MWh values are exactly 0. The cost of power has historically been relatively low in Norway. The fact that traditionally electricity in Norway has been generated almost entirely by hydropower explains these remarkably low electricity prices. Wind power and thermal plants are a couple of additional sources, but their shares are fairly insignificant. Less rain than usual fell in the southern part of Norway in 2022, whilst more rain fell in the northern portion of Norway, particularly around Trondheim. Less rain implies that the reservoirs of

hydropower are not as full, which results in less electricity being produced, pushing the price of electricity up. On the other side, due to the additional rain, some areas of central and northern Norway produce more electricity than usual, creating a power surplus that is causing some electricity stations to operate at a loss. Furthermore, the electricity generated in northern Norway cannot be transmitted to southern Norway, where less electricity is being produced, because the infrastructure does not have enough capacity for north-south transmission. [*in Norway Editorial Team, 2022*]

A preliminary analysis of the 12 different time series is conducted to determine whether they are stationary. The Dickey-Fuller test is used to determine whether this is the case by comparing the null hypothesis - that a unit root exists - with the alternative hypothesis - that the time series is stationary or trend-stationary. These tests showed that neither time series contained any unit roots, and as a result, all of the time series for the areas under consideration are stationary. Even though each region is examined and applied independently, their correlation is also examined, and is displayed in Appendix D.

To provide further insight into the data, a table of summary statistics is presented, specifically for DK1 in Table 6.1 below. A table containing the summary statistics for all 12 considered areas can be found in Appendix E.

	Min	Max	Mean	SD	Skew	Q1	Q2	Q3
DK1	-111.54	825.51	89.04	106.97	2.41	28.24	48.07	110.6

Table 6.1: Summary statistics for DK1's electricity prices, expressed in EUR/MWh, between January 1, 2019, and May 5, 2023. In other words, the three quantiles (Q1, Q2, and Q3) are shown along with the minimum, maximum, mean, standard deviation, and Pearson skew.

The price difference between the minimum and maximum is 937.05 EUR/MWh, considering the whole period, as can be seen from the summary statistics of DK1 in Table 6.1 above indicates the wild fluctuations seen in Figure 6.2. The standard deviation, which assesses the spread or dispersion of the data, also displays the vast range of price fluctuations. The distribution of prices is skewed to the right, according to a positive skewness value of 2.41, which is also displayed in the histogram in Figure 6.3 below. In other words, the distribution may have a longer right tail, suggesting substantially higher prices or price spikes.

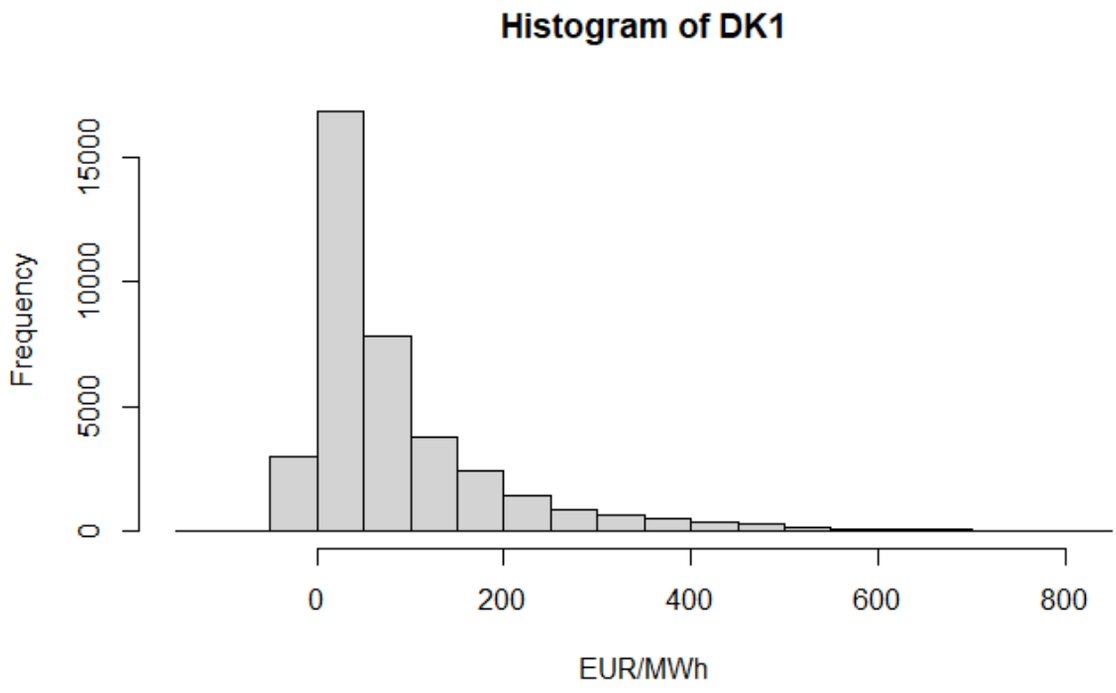

Figure 6.3: Histogram of the prices for DK1 during the entire considered period.

The data is hereafter divided into a training and a test period to assess how well the constructed forecasts are performing since this project aims to make probabilistic forecasts of intraday electricity prices. More specifically, the training period will run from 01/01/2019 to 31/03/2023, and the test period will run from 01/04/2023 to 05/05/2023. That is, the test period spans 35 days.

Hence, a probabilistic forecast is made for each of the 12 considered areas. A feature matrix, where each row $\mathbf{x}_t$ at time t is constructed as follows

$$\mathbf{x}_t = (1, \underbrace{\boldsymbol{Y}_{\text{All},t-1}, \boldsymbol{Y}_{\text{All},t-2}, \ldots, \boldsymbol{Y}_{\text{All},t-24}}_{\text{Prices in the previous 24 hours}}, \underbrace{\boldsymbol{Y}_{\text{All},t-48}, \boldsymbol{Y}_{\text{All},t-120}, \ldots, \boldsymbol{Y}_{\text{All},t-168}}_{\text{Prices at the same hour 2 to 7 days prior}}),$$

with

$$\boldsymbol{Y}_{All,t} = \left(Y_{DK1,t}, Y_{DK2,t}, Y_{NO1,t}, Y_{NO2,t}, Y_{NO3,t}, Y_{NO4,t}, Y_{NO5,t}, Y_{SE1,t}, Y_{SE2,t}, Y_{SE3,t}, Y_{SE4,t}, Y_{FI,t}\right).$$

To put it another way, $\mathbf{x}_t$ has intercept 1, time series of values all lagged over the preceding 24 hours, and lagged corresponding to the same time t 2 to 7 days prior. And each deep time series models are trained independently for each price region.

Due to the computational burden of the feature matrix, principal component analysis

(PCA) is performed on the data before applying the models. The following provides a brief overview of PCA; for more information, see [Nielsen et al., 2022] and [Hastie et al., 2009].

Principle component analysis (PCA) is a feature extraction technique where the original features are transformed into new, more pertinent features called principal components (PCs). PCA transforms the data into orthogonal components, with the first PC containing the most information. Hence, PCA is a dimension reduction method where only a few PCs are kept while preserving most information. Overall, PCA seeks to find linear combinations of these constructed PCs that explain most of the variance in the data. [Nakagome, 2019] [Dwivedi, 2021]

It should be noted that the intercept 1, which is a constant, is not included in the PCA because there is no variability to capture because the intercept has no variation. Figure 6.4 displays the results of the PCA, specifically, how much of the total variation in the data the various PCs account for, which is used to determine how many PCs to use in further analysis.

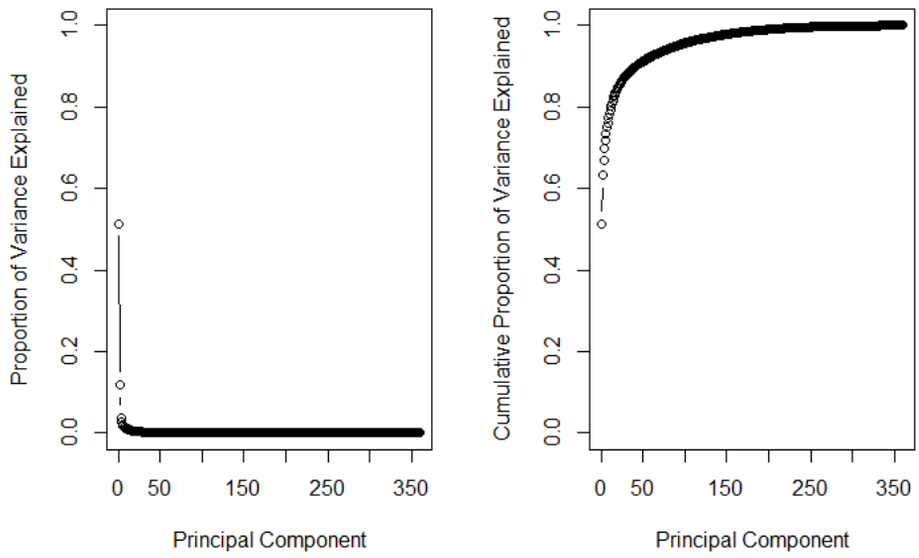

Figure 6.4: After performing PCA on the data, the figure on the left displays the percentage of variance explained by each principal component. The cumulative percentage of variance is depicted in the graphic on the right.

The precise number of PCs to include in the subsequent investigation can be determined using various techniques, some of which are introduced in [Pramoditha, 2022]. In this project, the number of PCs to be used as features in the resulting reduced

feature matrix X is determined by PCs that take into account 80% of the variation in the initial data. In this instance, the first 12 PCs make up 80% of the variation, or in more precise terms, 80.13% of the variance.

The models presented in Chapter 5 will be applied to the reduced feature matrix X.

6.2 Application of the Gaussian Probabilistic ESN Model

The data in the reduced feature matrix constructed by PCA will now be applied in the Gaussian Probabilistic ESN to obtain probabilistic forecasts of the 12 considered regions. The following procedure is only presented for one region, DK1, and the models are then run separately for the remaining 11 regions.

The posterior distributions of the models parameters, $\boldsymbol{\vartheta} = \left\{\boldsymbol{\beta}, \sigma^2, \tau^2\right\}$, are first calculated using a standard MCMC sampler. This MCMC sampler is run over 10000 iterations. MCMC samplers are made to converge to the target distribution, in this instance, the posterior distributions for the model parameters. The MCMC samples are therefore ensured to have converges, that is, to be reflective of the real posterior distributions, by employing a large number of iterations.

The produced posterior distributions are hereafter employed to obtain the predictive distributions, using a Monte Carlo approach, specifically for $K = 100$ configurations $\xi^1, \cdots, \xi^K$, the following steps are performed:

1. Use a standard MCMC sampler to obtain the posterior distribution for a considered region. That is, compute $\boldsymbol{\vartheta}$.

2. For iterations $i = 1, \cdots, N$ and times $t = T+1, \cdots, T+h_1$, where h_1 denotes the number of hours in the test period, the following is computed.

   a) $\mathbf{x}_{i,t}$ is called out from the feature matrix X.

   b) The hidden state $\mathbf{h}_{i,t}$ is computed from $\left\{\mathbf{h}_{i,t-1}, \mathbf{x}_{i,t}\right\}$. This computation is made using Definition 5.1 in Chapter 5.

   c) Lastly, $y_{i,t}$ is drawn from a Gaussian distribution $\mathcal{N}\left(\boldsymbol{h}'_{T+h}\boldsymbol{\beta}_1 + \left(\boldsymbol{h}'_{T+h}\right)^{\circ 2}\boldsymbol{\beta}_2, \sigma^2\right)$, where the values of $\boldsymbol{\beta}$ and σ^2 were found in step 1, and the hidden state determined in step b.

The final samples of step c above are saved in an array and are the final draws of the Gaussian probabilistic ESN model.

These samples are now used to plot the out-of-sample forecasts for DK1, as illustrated in Figure 6.5 below, along with a box plot of the out-of-sample observations for the test period. In Appendix F plots for the 12 different regions are displayed.

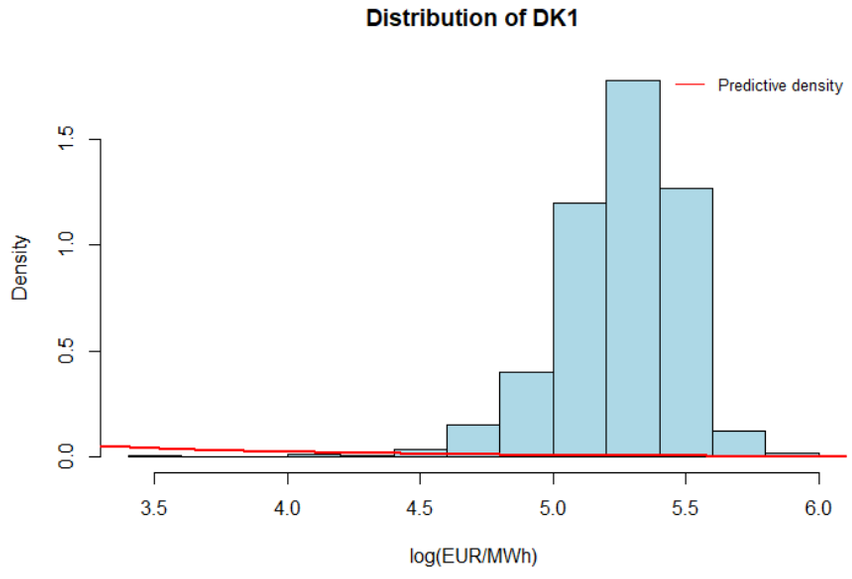

Figure 6.5: The light blue histogram is the logarithm of the out-of-samples observations for Y, here DK1, during the test period from April 1, 2023, to May 5, 2023. The deep time series predictive density obtained by the Gaussian probabilistic ESN model is represented by the red line.

From a visual point of view, the probabilistic forecast in Figure 6.5 shows that the observed values during the test period, represented by the DK1 variable, are located in the right-most tail of the forecast density. This visual analysis indicates that the true values tend to exceed the upper bounds of the forecasted distribution. The discrepancy between the observed data and the forecasted distribution in the tail region suggests that the model may have underestimated the occurrence of high values or failed to capture the extreme events of the DK1 variable. As a result, the forecasted density might not fully encompass the variability and extreme behaviour exhibited by the true values. This observation highlights the potential limitations of the probabilistic forecast in accurately capturing and representing the tail behaviour of the DK1 variable.

Hereafter the forecast accuracy is measured using the mean absolute error (MAE) and the root mean squared error (RMSE). The quantile scores corresponding to the lower and upper quantile, $\alpha = 0.05$ and $\alpha = 0.95$, are considered measures of tail accuracy. These measures of forecast accuracy are presented in Table 6.2 below. In

Appendix G a table containing these measures of forecast accuracy for all 12 distinct regions is presented.

	MAE	RMSE	$\alpha = 0.05$	$\alpha = 0.95$
DK1	94.52	100.88	15.82	146.60

Table 6.2: This table shows the mean absolute error (MAE, root mean squared error (RMSE), along with the lower and upper quantile, $\alpha = 0.05$ and $\alpha = 0.95$, respectively, for the area DK1 as a measure of forecast accuracy for the Gauss probabilistic ESN model.

Table 6.2 provides a comprehensive overview of the forecast accuracy for the Gauss probabilistic ESN model applied to the DK1 area. The Mean Absolute Error (MAE) of 94.52 represents the average absolute difference between predicted and true values. The Root Mean Squared Error (RMSE) value of 100.88 is the square root of the average squared difference between predicted and true values. Similar to MAE, a lower RMSE suggests higher accuracy in the forecasted values. In addition to MAE and RMSE, the table presents the lower and upper quantiles, $\alpha = 0.05$ and $\alpha = 0.95$, respectively. For example, the $\alpha = 0.05$ quantile of 15.82 suggests a 5% probability that the true values will fall below this lower bound. Similarly, the $\alpha = 0.95$ quantile of 146.60 indicates a 95% probability that the true values will be below this upper bound. These quantiles help assess the uncertainty and provide a range of values within which the true observations are expected to lie.

In conclusion, Table 6.2 and the visual analysis of Figure 6.5 show the Gauss probabilistic ESN model's shortcomings in catching extreme values and making reliable projections for the DK1 region. These conclusions highlight the necessity of future model improvement to enhance the model's ability to capture tail behaviour and conduct probabilistic forecasting.

6.3 Application of the Copula Model

Returning to the feature matrix constructed by PCA, the second model will now be applied, that is, the Copula model. Again the purpose is to obtain probabilistic forecasts for the 12 considered regions. This chapter presents the procedure for the region DK1, and this model is also run separately for the 11 remaining regions.

The conditional likelihood in (5.13) is estimated using an MCMC method. Specifically, β and $\log(\tau^2)$ are computed using a Metropolis-Hasting algorithm. As with the previous model, this MCMC sample is also run 10000 times to achieve convergence.

The produced values of β and $\log(\tau^2)$ are hereafter used in a Monte Carlo approach

to obtain the predictive distribution. As with the previous model, this approach is run over $K = 100$ configurations $\xi^2, \cdots, \xi^K$. Hence for each configuration ξ^k the following steps is performed:

1. β and $\log(\tau^2)$ are computed as explained above.
2. For iterations $t = 1, \cdots, T$, $z_{i,t} = \Phi_1^{-1}\left(F_Y\left(y_t\right)\right)$ is computed.
3. For iterations $i = 1, \cdots, N$ and $t = T+1, \cdots T+h_1$, with h_1 denoting the number of hours in the test period, the following steps are performed:
   a) $\mathbf{x}_{i,t}$ is called out from the feature matix X.
   b) Compute the hidden state $\boldsymbol{h}_{i,t}$ from $\left\{\boldsymbol{h_{i,t-1}}, \boldsymbol{x}_{i,t}\right\}$, again using Definition 5.1 presented Chapter 5.
   c) Let $\boldsymbol{b}_{i,t}^r = \left[\left(\boldsymbol{h}_{i,t}^r\right)^\top, \left(\left(\boldsymbol{h}_{i,t}^r\right)^{\circ 2}\right)^\top\right]$. Then compute $\psi_{i,t} = \left(1 + \frac{1}{\tau^2}\boldsymbol{b}_{i,t}\left(\boldsymbol{b}_{i,t}\right)^\top\right)$.
   d) $z_{i,t}$ is drawn from a Gaussian distribution, $\mathcal{N}\left(\psi_{i,t}\boldsymbol{b}_{i,t}\hat{\boldsymbol{\beta}}, \left(\psi_{i,r}\right)^2\right)$.
   e) Lastly, set $y_{i,t} = F_Y^{-1}\left(\Theta_1\left(z_{i,t}\right)\right)$, and store the samples in an array.

The samples obtained in step e above are the final draws of the Copula model. Throughout this application, $N = 100$ and $T = 840$ have been used.

As with the Gauss probabilistic Model, the samples obtained for this Copula model are used to plot out-of-sample forecasts for DK1, which is displayed in Figure 6.6 below. This probabilistic forecast is plotted on top of a box plot representing DK1 for the test period. In Appendix H these probabilistic forecasts are displayed for the 12 regions considered.

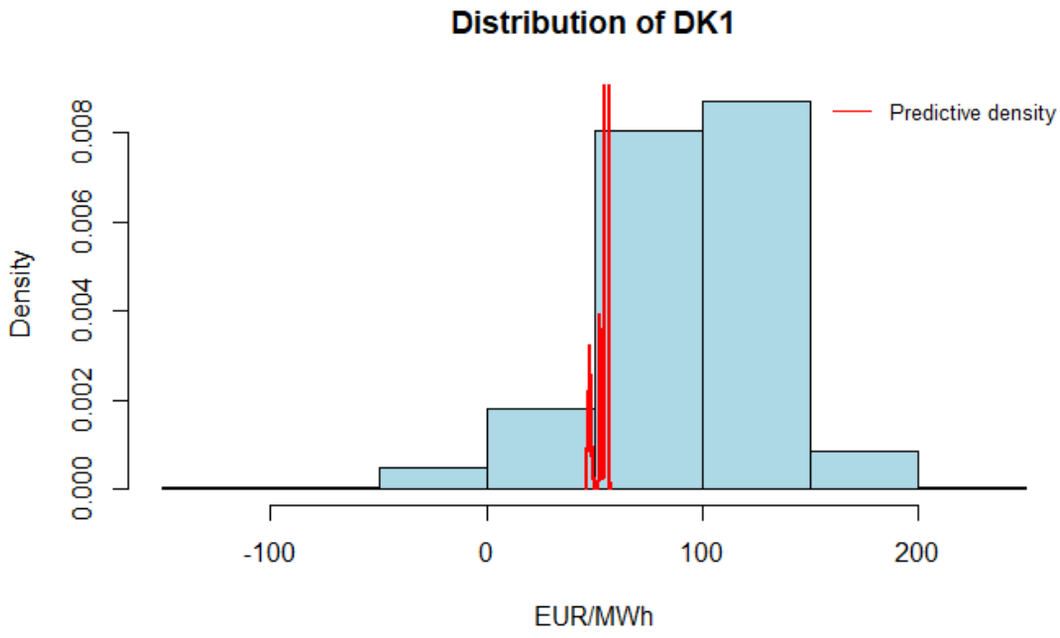

Figure 6.6: The out-of-samples observations for Y, here DK1, during the test period between April 1, 2023, and May 5, 2023, are represented by the light blue boxplot. The deep time series predictive density obtained using the Copula model is displayed as the red line.

Upon visually examining Figure 6.6, it becomes evident that the probabilistic forecast generated by the copula model falls short of capturing the true density during the test period. Notably, the forecasted distribution appears to be concentrated in the range of 40-60, failing to capture the tail behaviour of the true distribution observed in DK1. This observation highlights a limitation in the copula model's ability to accurately capture the extreme values and tail events in DK1's distribution.

To assess the forecast accuracy, MAE and RMSE are considered. As before, the lower and upper quantiles are furthermore considered to asses the tail accuracy. The result of this analysis is presented for DK1 in the table below. In addition, this analysis of forecast accuracy is presented in Appendix I for all 12 considered regions.

	MAE	RMSE	$\alpha = 0.05$	$\alpha = 0.95$
DK1	47.48	54.76	-40.48	92.83

Table 6.3: This table shows the mean absolute error (MAE, root mean squared error (RMSE), along with the lower and upper quantile, $\alpha = 0.05$ and $\alpha = 0.95$, respectively, for the area DK1 as a measure of forecast accuracy for the Copula model.

The findings from Table I.1 further reinforces the limitations of the Copula model observed in the previous figure (Figure 6.6). The relatively low MAE and RMSE values indicate that, on average, the Copula model performs reasonably well in capturing the central tendency of DK1's distribution. However, when comparing the forecast accuracy with the true density depicted in the previous figure, it becomes evident that the model struggles to capture the tails of DK1's distribution. The lower quantile of -40.48 indicates that the Copula model underestimates the lower tail, and the upper quantile of 92.83 suggests it underestimates the upper tail. These findings align with the visual analysis of the previous figure, which showed that the model's forecast appears concentrated in the middle of the test density, failing to capture the true distribution's behaviour in the tail regions. Therefore, while the Copula model demonstrates satisfactory overall accuracy regarding MAE and RMSE, its limitations in accurately capturing extreme events and estimating the tails become apparent compared to the true density. These results emphasise the need for further refinement of the Copula model to improve its performance in capturing tail behaviour and enhancing its probabilistic forecasting capabilities.

7 Discussion

There are several crucial design decisions and hyper-parameter considerations in the context of echo state networks (ESNs) that might affect the functionality and behaviour of the network. One important consideration is the reservoir's size, which measures the network's hidden states. Large reservoir sizes have historically been preferred in ESN implementations since regularisations were thought to reduce the possibility of overfitting. Smaller reservoir sizes are adequate for spatiotemporal applications when the available time series length may not be very large. An embedding input ($\boldsymbol{x}_t$) compensates for the smaller reservoir size. A small reservoir size also enables the ESN to function as a committee of weak performers, preventing overfitting. Cross-validation is frequently used to determine the suitable reservoir size. The leaking rate (α), which regulates the impact of earlier states on the ESN's current state update, is another crucial element. Strong memory and more information persistence are implied by an α close to 1, as the preceding states have a larger influence on the present state, which may be useful for activities that need the capturing of long-term dependencies or when the dynamics of the underlying system are slow to change. On the other hand, a smaller value of α, closer to 0, lessens the influence of earlier states and increases the network's sensitivity to more recent inputs, which can be helpful for tasks when identifying quick dynamics or short-time patterns are essential. By testing the network's effectiveness using training and validation datasets, experimentation and validation techniques are often used in practice to discover the ideal value of α. Another crucial factor is the scaling of the reservoir weighting matrix (W). For the reservoir to remain stable and to ensure efficient information processing, W must be scaled properly. Signals are increased or suppressed per the scaling factor, which affects how well the network functions overall. To optimise the performance of ESNs for particular tasks and datasets, this discussion emphasises the significance of carefully choosing and modifying hyper-parameters. [McDermott and Winkle, 2017]

Markov Chain Monte Carlo (MCMC) methods have been used in this project. MCMC approaches have several advantages, especially when direct sampling is impractical. In these situations, they offer a useful tool for generating samples. MCMC methods throughout the project are also made possible because their implementation is fairly simple. Furthermore, MCMC methods show dependability when a high number of

iterations is performed, ensuring robustness in the outcomes. It's crucial to keep in mind, nevertheless, that applying MCMC sampling techniques has its drawbacks. One of these drawbacks is the longer computation time relative to situations when direct sampling is possible. The overall computation may take longer using MCMC methods since more samples are needed to obtain reliable estimations. Additionally, analysing the precision and convergence of MCMC algorithms might be difficult, making it hard to assess the quality of the created samples. This assessment is difficult since MCMC approaches, unlike deterministic algorithms, lack well-defined convergence criteria. The assessment becomes more challenging as it becomes increasingly difficult to visualise and analyse the behaviour of the algorithm due to the high-dimensional parameter spaces frequently used in MCMC algorithms. Additionally, the difficulty in precisely evaluating convergence and estimating the accuracy of results is further complicated by elements like inadequate mixing, limited exploration, and the presence of autocorrelation in MCMC chains. These difficulties underline the necessity of rigorous monitoring and diagnostics to guarantee accurate inference and improve trust in the reliability of MCMC-based analyses. [Mehta, 2023]

A copula model is one of the techniques used throughout the project's implementation; thus, here is an overview of the advantages and drawbacks of utilising copula models. One of the benefits of copula models is their ability to distinguish between dependent and marginal structures. As a result, it is possible to modify the modelling approach to account for the unique features of the data, such as the various distributions for each variable. However, the relevance of individual marginal distributions may be reduced due to this adaptability, which would restrict the ability to analyse variables separately. Another advantage is the Copula models' robustness, which allows them to manage complicated or non-linear marginal distributions while accurately reflecting dependencies. Copula models can provide simulated data while maintaining the dependence identified by the copula model. As has been done throughout this study, this copula model feature can be utilised to produce forecasts that offer important insights into how power costs will develop in Nordic nations. Despite these advantages, there are several disadvantages to copula models. The choice of parameter values considerably impacts the model's efficacy and fit, making the selection of complex models problematic. The model selection procedure becomes more challenging as the number of variables or dimensions increases. Copula models also raise issues with over- and underfitting. Poor out-of-sample performance results from overfitting, which occurs when the model grows extremely complex and captures noise. Underfitting, on the other hand, happens when the model is oversimplified and doesn't adequately reflect the real underlying dependent structure. There isn't one 'optimal' copula model that applies to all different kinds of data and dependent structures. As a result, choosing the best copula model frequently necessitates exper-

imentation and meticulous evaluation. The limited interpretation of copula models is another drawback. It might be challenging to offer meaningful explanations or insight based merely on the values of the parameters indicating the dependent structure because they are not necessarily simple or intuitive. Copula models are typically built to capture statistical dependencies rather than causal relationships, further restricting interpretation. Copulas do not reveal the causes or consequences of substantial statistical dependence between variables, despite the possibility of such dependence. Additionally, the sample size plays a crucial role in accurate parameter estimation for Copula models. With limited data points, accurate estimation becomes challenging, affecting the model's performance and dependability. In summary, copula models have advantages in terms of robustness, the ability to simulate, and the separation of marginal and dependence structures. However, they face difficult model selection, constrained interpretation, and sample size sensitivity. Applying copula models requires careful thought, analysis, and interpretation to ensure proper use and interpretation. [Embrechts et al., 2002], and [Cherubini et al., 2004]

Conclusion 8

This project tackled the problem of probabilistic forecasting intraday electricity prices in the Nordic areas using a combination of Bayesian inference, Echo State Networks (ESNs), and Copulas. These theoretical ideas were essential in capturing and simulating the intricate dynamics of data on electricity prices.

By incorporating prior knowledge, observed data, and likelihood functions to estimate the parameters of the models, Bayesian inference established a robust framework for probabilistic modelling. Moreover, this method made it possible to quantify uncertainty and produce probabilistic projections, both crucial for comprehending potential outcomes.

Recurrent neural networks, namely Echo State Networks, provided a flexible and effective method for identifying temporal relationships and irregularities in the electricity pricing data. In addition, ESNs provide a strong modelling approach to capture the different dynamics of the system and produce accurate projections by utilising the reservoir of randomly connected recurrent nodes and just training the output weights. On the other hand, copulas provide a way to simulate the joint distribution of several variables, such as the cost of electricity in various areas. Copulas made it possible to create probabilistic forecasts that considered the relationships between the variables because they could capture the marginal distributions and the dependence structure, which was especially helpful in capturing the complicated relationships between the electricity costs in Nordic countries. The project sought to use a combination of copulas, echo state networks, and Bayesian inference to enhance the precision and dependability of probabilistic forecasts.

The main goal of this study was to look into the approaches described in [Klein et al., 2023], namely the use of neural networks and copulas for probabilistic intraday electricity price forecasting in the Nordic areas. The results revealed the models' advantages and disadvantages in reflecting the intraday dynamics of electricity price changes.

According to the visual analysis of the Gauss probabilistic ESN model, the anticipated density differed from the observed values throughout the test period, especially

for the DK1 variable. The genuine values tended to be higher than the anticipated distribution's upper bounds, suggesting that high values may have been underestimated or extreme events had been missed. As a result, the projected density could not adequately capture the fluctuation and extreme behaviour displayed by the genuine values, revealing that the DK1 variable has some difficulties accurately reflecting its tail behaviour.

In particular, the copula model showed flaws in representing the genuine density in the tail regions. As a result, the extreme values seen in DK1 were not accurately represented by the predicted distribution, which appeared to be centred in a smaller range. This restriction suggests that the copula model faces difficulty identifying extreme occurrences and calculating the tails of DK1's distribution.

Both models did not perform satisfactorily overall when measuring forecast accuracy using measures like Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE), which indicate how well a model captures the central tendency of a distribution like DK1. The models' shortcomings in precisely identifying extreme occurrences and estimating the tails, however, were evident when contrasted with the true density and considered for the lower and upper quantiles. Particularly, the Copula model underestimated the lower tail and gave insufficient predictions for the higher tail.

These findings highlight the need for further model improvement to improve both models' ability to capture tail behaviour and boost probabilistic forecasting accuracy.

In conclusion: While copulas and neural networks are promising methods for probabilistic forecasting of intraday power prices in the Nordic areas, capturing extreme events and estimating the tails of the price distributions was difficult. Additional model refinements that consider regional variances and investigate alternative modelling approaches are required to improve the models' performance and strengthen their probabilistic forecasting skills.

Reflection 9

In the context of this project, it is acknowledged that including a wider variety of exogenous data has the potential to improve the precision and stability of the probabilistic forecasting models. For instance, incorporating weather prediction data may provide insightful information about variables like temperature, wind patterns, and precipitation. Furthermore, it has been demonstrated that these weather-related factors significantly affect the patterns of electricity prices. Therefore, a more thorough grasp of electricity pricing dynamics may have been attained by including such weather forecasts in the models. It's also possible that using probabilistic demand projections would have improved the models' capacity for prediction. These projections offer a probabilistic view of future electricity consumption since they consider a variety of variables, including consumer behaviour and economic data. A more sophisticated and thorough framework for understanding the underlying unpredictability and uncertainty inherent in electricity price dynamics might have been available had these estimates been included as exogenous inputs. Extending the range of exogenous data taken into account by the models can incorporate more data sources and boost the overall effectiveness of the probabilistic forecasting models. The models would have been better equipped to deal with the inherent complexity and uncertainty present in the energy market had they taken into account a larger range of pertinent parameters, such as weather forecasts and probabilistic demand estimates. [Klein et al., 2023]

Exploring different probability distributions might have provided insightful information and enhanced the precision of the probabilistic forecasting models. Even though the described models may have worked well with the chosen distribution, taking into account alternative distributional hypotheses would have given a more complete insight into the underlying data generation process. For instance, non-Gaussian or heavy-tailed distributions may have better-captured electricity market-specific nonlinear dynamics or dramatic price swings. In addition, it might have been possible to acquire a more detailed knowledge of the uncertainties surrounding intraday power price estimates by conducting sensitivity analysis under alternative distributional assumptions. Future research in this area has the potential to identify alternative

modelling frameworks that could result in more solid and trustworthy probabilistic projections in the energy sector.

Energy storage is another intriguing factor that has a lot of potentials to improve the precision and dependability of probabilistic forecasting models. The ability to store excess electricity and release it at times of high demand has recently become achievable due to the rapid advancement of energy storage technologies, such as batteries and pumped hydro storage, in recent years. Data from energy storage could be incorporated into probabilistic forecasting models to provide some intriguing possibilities. It might reduce price volatility, maximise the use of renewable energy sources, increase grid stability, and offer insightful information for business impact and investment choices. Here grid stability refers to the ability of an electrical power system, commonly known as the grid, to maintain a reliable and balanced supply of electricity. The investigation of energy storage's function in this situation has the potential to advance electricity price forecasting. [Masterson, 2021]

Bibliography

Aggarwal, **2015**. Charu C. Aggarwal. *Datamining*. Springer, 2015. ISBN 978-3-319-14141-1.

Bellis, **2018**. Mary Bellis. *The Basics: An Introduction to Electricity and Electronics*, 2018. URL https://www.thoughtco.com/electricity-and-electronics-4072563.

Cherubini et al., **2004**. Umberto Cherubini, Elisa Luciano, Walter Vecchiato and Giovanni Cherubini. *Copula Methods in Finance*. John Wiley Sons, Incorporated, 2004. ISBN 9780470863442.

Ciortan, **2019**. Madalina Ciortan. *Gentle introduction to Echo State Networks*, 2019. URL https://towardsdatascience.com/gentle-introduction-to-echo-state-networks-af99e5373c68.

Council, **2023a**. European Council. *Energy prices and security of supply*, 2023a. URL https://www.consilium.europa.eu/en/policies/energy-prices-and-security-of-supply/.

Council, **2023b**. European Council. *Impact of Russia's invasion of Ukraine on the markets: EU response*, 2023b. URL https://www.consilium.europa.eu/en/policies/eu-response-ukraine-invasion/impact-of-russia-s-invasion-of-ukraine-on-the-markets-eu-response/.

Council, **2023c**. European Council. *Infographic - Where does the EU's gas come from?*, 2023c. URL https://www.consilium.europa.eu/en/infographics/eu-gas-supply/.

Demiris, **2023**. Yiannis Demiris. *Echo-State Network*, 2023. URL https://www.researchgate.net/figure/Echo-State-Network-ESN-In-the-typical-setup-the-inputs-are-fully-connected-to-a_fig1_263124732.

Dwivedi, **2021**. Rohit Dwivedi. *Introduction To Principal Component Analysis In Machine Learning*, 2021. URL https://www.analyticssteps.com/blogs/introduction-principal-component-analysis-machine-learning.

Embrechts et al., **2002**. Paul Embrechts, Alexander J. McNeil and Daniel Strausmann. *Correlation and Dependence in Risk Management: Properties and Pitfalls*. Cambridge University Press, 2002. ISBN 9780521169639.

Engati, **2023a**. Engati. *Reservoir computing*, 2023a. URL https://www.engati.com/glossary/reservoir-computing.

Engati, **2023b**. Engati. *Echo State Networks*, 2023b. URL https://www.engati.com/glossary/echo-state-networks#toc-what-is-the-echo-state-property-.

Engati, **2023c**. Engati. *Vanishing gradient problem*, 2023c. URL https://www.engati.com/glossary/vanishing-gradient-problem.

Faik, **2021**. Lina Faik. *Deep Learning for Time Series Forecasting: Is It Worth It?*, 2021. URL https://blog.dataiku.com/deep-learning-time-series-forecasting.

Gneiting and Katzfuss, **2014**. Tilmann Gneiting and Matthias Katzfuss. *Probabilistic Forecasting*, 2014. URL https://www.annualreviews.org/doi/abs/10.1146/annurev-statistics-062713-085831.

Gneiting et al., **2007**. Tilmann Gneiting, Fadoua Balabdaoui and Adrian E. Raftery. *Probabilistic forecasts, calibration and sharpness.* pages 243–268. Journal of the Royal Statistical Society Series B, 2007.

Hastie, **2001**. Trevor Hastie. *The Elements of Statistical Learning.* Springer, 2001.

Hastie et al., **2009**. Trevor Hastie, Robert Tibshirani and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer, 2009. ISBN 978-0-387-84857-0.

Hofert et al., **2018**. Marius Hofert, Ivan Kojadinovic, Martin Mächler and Jun Yan. *Elements of Copula Modeling with R.* Springer, 2018. ISBN 978-3-319-89634-2. URL https://doi.org/10.1007/978-3-319-89635-9.

Hoff, **2009a**. Peter D. Hoff. *A First Course in Bayesian Statistical Methods.* Springer, 2009a. ISBN 978-0-387-92299-7.

Hoff, **2009b**. Peter D. Hoff. *A First Course in Bayesian Statistical Methods.* Springer, 2009b. ISBN 978-0-387-92299-7.

IBM, **2023**. IBM. *What is recurrent neural networks?*, 2023. URL https://www.ibm.com/topics/recurrent-neural-networks.

in Norway Editorial Team, **2022**. Life in Norway Editorial Team. *Explained: Why Is Electricity So Expensive In Norway Right Now?*, 2022. URL https://www.lifeinnorway.net/why-is-electricity-so-expensive-in-norway-right-now/.

Kenton, **2022**. Will Kenton. *Deregulation: Definition, History, Effects, and Purpose*, 2022. URL https://www.investopedia.com/terms/d/deregulate.asp.

Klein and Kneib, **2016**. Nadja Klein and Thomas Kneib. *Scale-Dependent Priors for Variance Parameters in Structured Additive Distributional Regression*, 2016. URL https://projecteuclid.org/journals/bayesian-analysis/volume-11/issue-4/Scale-Dependent-Priors-for-Variance-Parameters-in-Structured-Additive-Distributional/10.1214/15-BA983.full.

Klein et al., **2023**. Nadja Klein, Michael Stanley Smith and David J. Nott. *Deep Distributional Time Series Models and the Probabilistic Forecasting of Intraday Electricity Prices*, 2023. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/jae.2959.

Lee, **2004**. Peter M. Lee. *Bayesian Statistics: An Introduction*. Hodder Education, 2004. ISBN 978-0340814055.

Li and Becker, **2021**. Wei Li and Denis Becker. *Day-ahead electricity price prediction applying hybrid models of LSTM-based deep learning methods and feature selection algorithms under consideration of market coupling*, 2021. URL https://www.researchgate.net/figure/A-map-of-the-overview-of-the-Nord-Pool-market-coupling_fig1_348486579/actions#reference.

Liu, **2020**. Davide Liu. *RNN: Recurrent Neural Networks*, 2020. URL https://davideliu.com/2020/02/29/rnn-recurrent-neural-networks/.

Madaleno and Pinho, **2008**. Mara Madaleno and Carlos Pinho. *Some stylized facts in electricity markets: a European comparison*, 2008. URL https://www.researchgate.net/publication/228891266.

Madhan, **2020**. Nithya Prasath Madhan. *Vanishing and Exploding Gradient Problems*, 2020. URL https://medium.com/analytics-vidhya/vanishing-and-exploding-gradient-problems-c94087c2e911.

Marin and Robert, **2014**. Jean-Michel Marin and Christian Robert. *Bayesian Essentials with R*. Springer, 2014. ISBN 978-1-4614-8686-2.

Masterson, **2021**. Victoria Masterson. *How can we store renewable energy? 4 technologies that can help*, 2021. URL https://www.weforum.org/agenda/2021/04/renewable-energy-storage-pumped-batteries-thermal-mechanical/.

McDermott and Wikle, **2019**. Patric L. McDermott and Christopher K. Wikle. *Deep Echo State Networks with Uncertainty Quantification for Spatio-Temporal*

Forecasting, 2019. URL https://ideas.repec.org/a/wly/envmet/v30y2019i3ne2553.html.

McDermott and Winkle, 2017. Patrick L. McDermott and Christopher K. Winkle. *An Ensemble Quadratic Echo State Network for Nonlinear Spatio-Temporal Forecasting*, 2017. URL https://www.researchgate.net/publication/319163998_An_Ensemble_Quadratic_Echo_State_Network_for_Nonlinear_Spatio-Temporal_Forecasting.

McNeil et al., 2005. Alexander J. McNeil, Rüdiger Frey and Paul Embrechts. *Quantitative Risk Management*. Princeton University Press, 2005. ISBN 0-691-12255-5.

Mehta, 2023. Sourabh Mehta. *All you need to know about Markov Chain Monte Carlo*, 2023. URL https://analyticsindiamag.com/all-you-need-to-know-about-markov-chain-monte-carlo/.

Nakagome, 2019. Sho Nakagome. *Principal Component Analysis—Stepping into the details of the math*, 2019. URL https://medium.com/sho-jp/principal-component-analysis-101-part-1-d62aa2b0cc36.

Nielsen et al., 2022. Anders Linnemann Nielsen, Leonora Nielsen and Louise Neema Krogh Pedersen. *The Climate Awareness Effect on Financial Performances*. chapter 3. Aalborg University, 2022.

Norway, 2023. Energy Facts Norway. *The Power Market*, 2023. URL https://energifaktanorge.no/en/norsk-energiforsyning/kraftmarkedet/.

Ognjanovski, 2019. Gavril Ognjanovski. *Everything you need to know about Neural Networks and Backpropagation — Machine Learning Easy and Fun*, 2019. URL https://towardsdatascience.com/everything-you-need-to-know-about-neural-networks-and-backpropagation-machine-learning-made-easy-e5285bc2be3a.

Politiken, 2021. Politiken. *OVERBLIK: Sådan fungerer EU-system for handel med CO2-kvoter*, 2021. URL https://politiken.dk/klima/art8062820/OVERBLIK-S%C3%A5dan-fungerer-EU-system-for-handel-med-CO2-kvoter.

Pra, 2020. Marco Del Pra. *Time Series Forecasting with Deep Learning and Attention Mechanism*, 2020. URL https://towardsdatascience.com/time-series-forecasting-with-deep-learning-and-attention-mechanism-2d001fc871fc.

Pramoditha, **2022**. Rukshan Pramoditha. *How to Select the Best Number of Principal Components for the Dataset*, 2022. URL https://towardsdatascience.com/how-to-select-the-best-number-of-principal-components-for-the-dataset-287e64b14c6d.

REU, **2021**. TREND REU. *Machine Learning to Predict Chaos: Echo State Networks*, 2021. URL https://www.youtube.com/watch?v=RQugL0oNMxU.

Ruppert, **2011**. David Ruppert. *Statistics and Data Analysis for Financial Engineering*. pages 175–195. Springer, 2011. ISBN 978-1-4419-7786-1.

Segal, **2022**. Troy Segal. *Intraday: Definition, Intraday Trading, and Intraday Strategies*, 2022. URL https://www.investopedia.com/terms/i/intraday.asp.

Smith, **2023**. Michael Stanley Smith. *Implicit Copulas: An Overview*, 2023. URL https://arxiv.org/pdf/2109.04718.pdf.

Spot, **2023**. EPEX Spot. *Basics of Power Market*, 2023. URL https://www.epexspot.com/en/basicspowermarket#power-exchanges-organise-trading-and-operate-markets.

Stanford, **2023**. Stanford. *Multi-Layer Neural Network*, 2023. URL http://deeplearning.stanford.edu/tutorial/supervised/MultiLayerNeuralNetworks/.

Stanwell, **2023**. Stanwell. *Negative prices: how they occur, what they mean*, 2023. URL https://www.stanwell.com/our-news/energy-explainer/negative-prices/.

Taboga, **2023a**. Marco Taboga. *Bayesian Inference*, 2023a. URL https://www.statlect.com/fundamentals-of-statistics/Bayesian-inference.

Taboga, **2023b**. Marco Taboga. *Markov Chain Monte Carlo (MCMC) methods*, 2023b. URL https://www.statlect.com/fundamentals-of-statistics/Markov-Chain-Monte-Carlo.

Taboga, **2023c**. Marco Taboga. *Matrix inversion lemmas*, 2023c. URL https://www.statlect.com/matrix-algebra/matrix-inversion-lemmas.

Taillon, **2023**. Jean-Philippe Taillon. *Introduction to the World of Electricity Trading*, 2023. URL https://www.investopedia.com/articles/investing/042115/understanding-world-electricity-trading.asp#toc-trading-electricity.

Ørsted, **2023**. Ørsted. *Energimarkedet*, 2023. URL https://orsted.dk/erhverv/energimarked.

Appendix A

A.1 Woodbury Formula

This section is based on [Taboga, 2023c]

In the set-up of the copula model in Chapter **??**, the Woodbury Formula is used. This formula is presented below.

Theorem A.1 (Woodbury Formula).
Let A denote a $k \times k$ matrix, C an $m \times m$ invertible matrix, and U and V two $k \times m$ matrices. Then if

$$C^{-1} + V^\top A^{-1} U$$

is invertible, then

$$A + UCV^\top$$

is invertible, and its inverse is given as

$$\left(A + UCV^\top\right)^{-1} = A^{-1} - A^{-1}U\left(C^{-1} + V^\top A^{-1}U\right)^{-1} V^\top A^{-1}.$$

The proof of this theorem can be found in Taboga, 2023c.

Appendix B

B.1 Weilbull Prior

This section is based on [Klein and Kneib, 2016]

The Weibull prior is used in the set-up of the copula model in Chapter **??**. This formula is hence presented below.

Theorem B.1 (Weibull Prior).
Let a denote the shape parameter, and b_{τ^2} the scale parameter. Then the Weibull prior is given as

$$p\left(\tau^2\right) = \frac{a}{b_{\tau^2}} \left(\frac{\tau^2}{b_{\tau^2}}\right)^{a-1} \exp\left\{-\left(\frac{\tau^2}{b_{\tau^2}}\right)^{a}\right\}.$$

The proof of this theorem is omitted from the project.

Appendix C

C.1 Time Series of Considered Areas

This section displays a time series plot for each area that was taken into consideration.

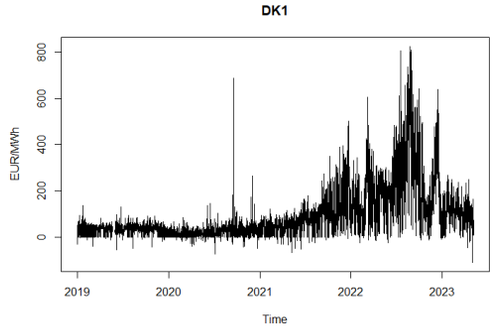

Figure C.1: DK1

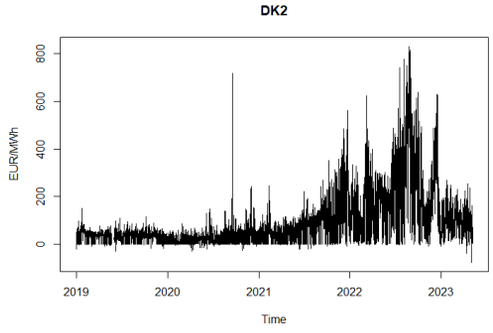

Figure C.2: DK2

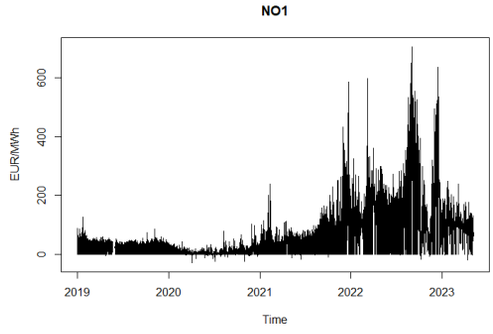

Figure C.3: NO1

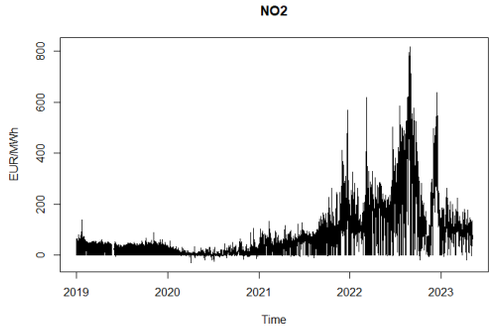

Figure C.4: NO2

Figure C.5: Illustration of the DK1, DK2, NO1, and NO2 time series is price movements between 01/01/2019 and 05/05/2023.

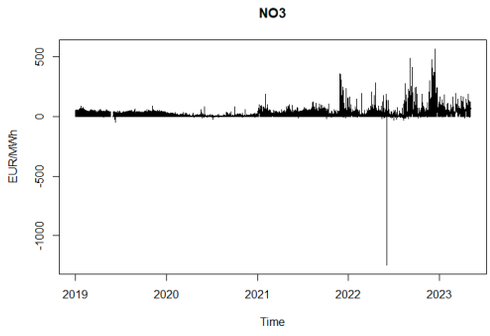

Figure C.6: NO3

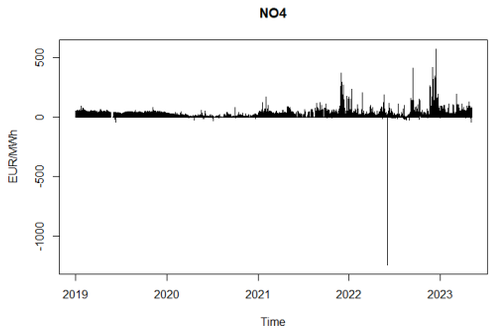

Figure C.7: NO4

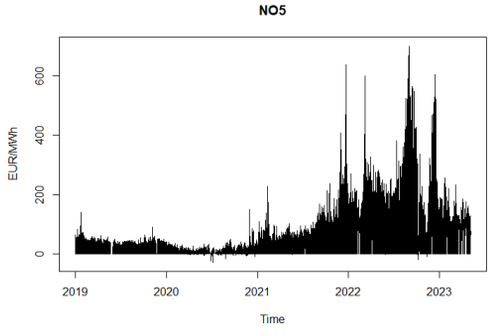

Figure C.8: NO5

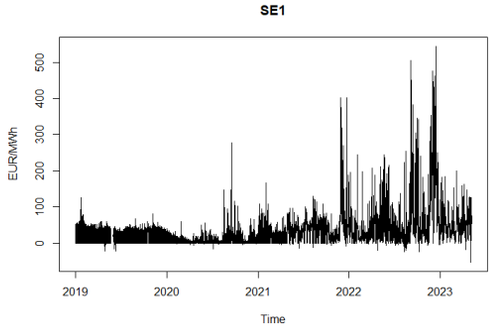

Figure C.9: SE1

Figure C.10: Illustration of the NO3, NO4, NO5, and SE1 time series is price movements between 01/01/2019 and 05/05/2023.

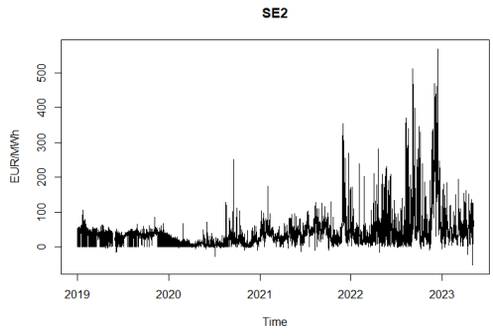

Figure C.11: SE2

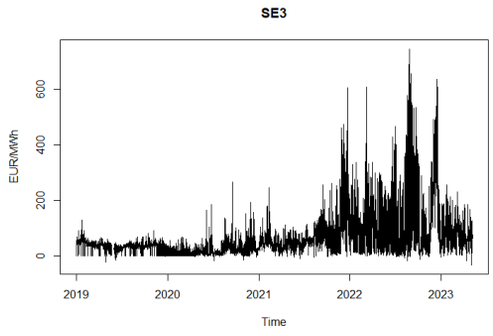

Figure C.12: SE3

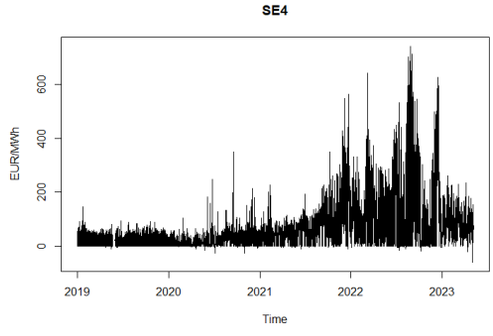

Figure C.13: SE4

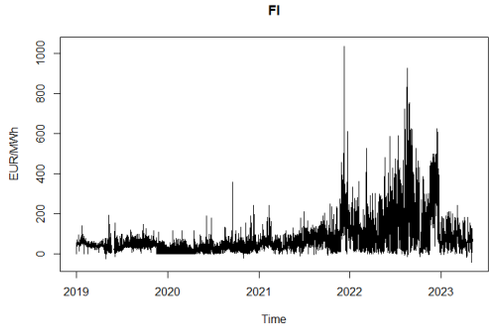

Figure C.14: FI

Figure C.15: Illustration of the SE2, SE3, SE4, and FI time series is price movements between 01/01/2019 and 05/05/2023.

Appendix D

D.1 Correlations of the Considered Regions

Below the correlation matrix for the 12 considered regions are presented.

	DK1	DK2	NO1	NO2	NO3	NO4	NO5	SE1	SE2	SE3	SE4	FI
DK1	1											
DK2	0.974	1										
NO1	0.825	0.807	1									
NO2	0.931	0.904	0.865	1								
NO3	0.348	0.353	0.448	0.373	1							
NO4	0.179	0.186	0.254	0.199	0.563	1						
NO5	0.748	0.731	0.843	0.786	0.416	0.238	1					
SE1	0.440	0.444	0.516	0.450	0.763	0.463	0.464	1				
SE2	0.476	0.481	0.534	0.485	0.759	0.448	0.481	0.950	1			
SE3	0.788	0.802	0.776	0.770	0.518	0.305	0.695	0.630	0.655	1		
SE4	0.814	0.831	0.754	0.774	0.459	0.262	0.675	0.546	0.569	0.913	1	
FI	0.725	0.732	0.684	0.717	0.452	0.261	0.611	0.591	0.631	0.861	0.791	1

Table D.1: Correlations between the 12 considered areas. Only the lower triangular part is shown due to the matrix's symmetry.

Appendix E

E.1 Summary Statistics

Summary statistics for the 12 consideres areas are presented in the table below

	Min	Max	Mean	SD	Skew	Q1	Q2	Q3
DK1	-111.54	825.51	89.04	106.97	2.41	28.24	48.07	110.6
DK2	-77.5	831.89	88.22	106.10	2.46	28.83	47.8	107.37
NO1	-29.67	706.96	68.02	87.39	2.42	3.33	38.85	99.67
NO2	-29.82	817.91	79.86	100.67	2.46	13.68	43.4	108.81
NO3	-1247	569.36	26.86	40.94	3.59	0	16.79	39.26
NO4	-1247	573.87	13.67	29.12	3.39	0	0	20.5
NO5	-28.8	700.17	59.20	86.28	2.47	0	26.71	91.65
SE1	-53.41	544.64	37.20	46.17	4.15	11.68	29.22	44.93
SE2	-53.18	569.26	39.70	45.97	4.13	15.32	31.69	46.09
SE3	-34.81	744.57	62.58	78.47	3.12	20.98	38.88	67.67
SE4	-60.86	743.11	68.55	86.83	2.78	18	41.82	82.02
FI	-44.28	1035.71	72.65	86.65	3.25	26.57	45.65	82.97

Table E.1: Summary statistics for the 12 considered area's electricity prices, expressed in EUR/MWh, between January 1, 2019, and May 5, 2023. In other words, the three quantiles (Q1, Q2, and Q3) are shown along with the minimum, maximum, mean, standard deviation, and Pearson skew.

Appendix F

F.1 Density Plots for the Gauss Probabilistic ESN Model

This section presents the density plots along with the predictive density obtained by the Gauss Probabilistic ESN model for each of the 12 considered areas.

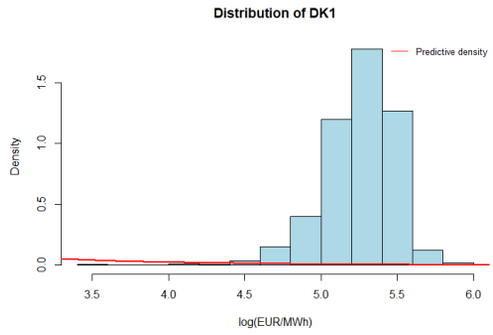

Figure F.1: DK1

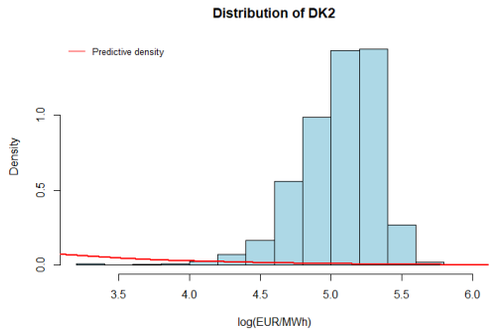

Figure F.2: DK2

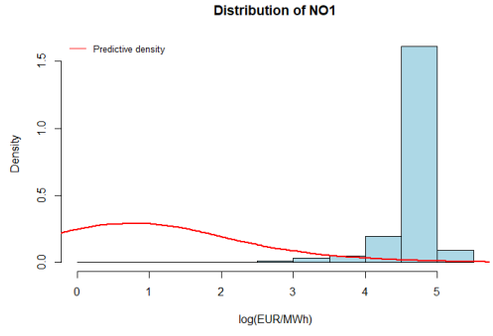

Figure F.3: NO1

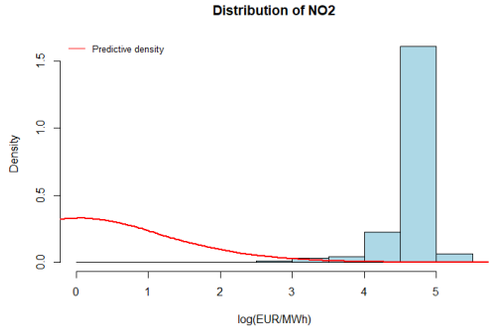

Figure F.4: NO2

Figure F.5: The light blue histograms represent the logarithm of the out-of-sample observations for Y for the time series DK1, DK2, NO1, and NO2, in the four figures. The red line in each plot represents the deep time series predictive density obtained by the Gaussian probabilistic ESN.

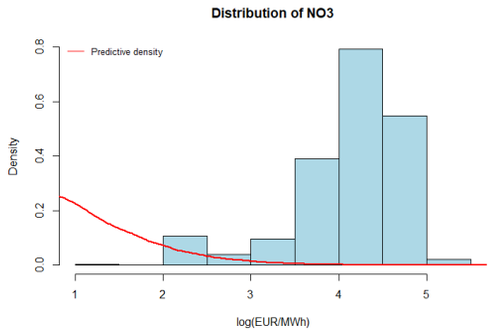

Figure F.6: NO3

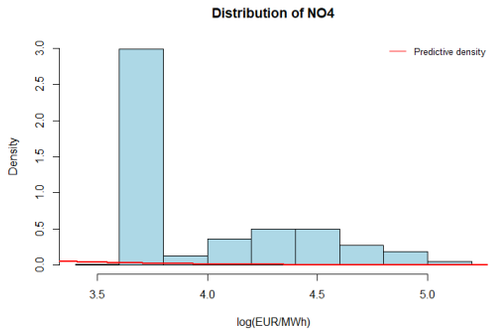

Figure F.7: NO4

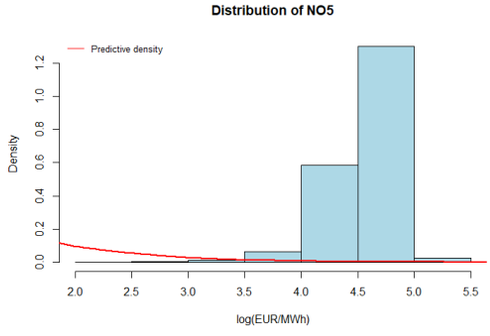

Figure F.8: NO5

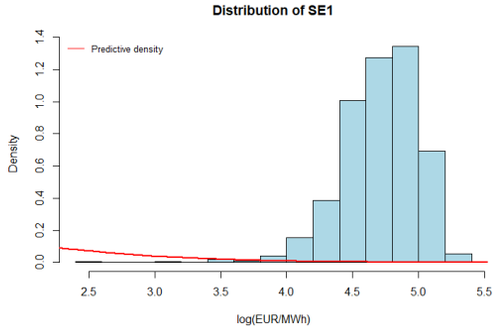

Figure F.9: SE1

Figure F.10: The light blue histograms represent the logarithm of the out-of-sample observations for Y for the time series NO3, NO4, NO5, and SE1, in the four figures. The red line in each plot represents the deep time series predictive density obtained by the Gaussian probabilistic ESN.

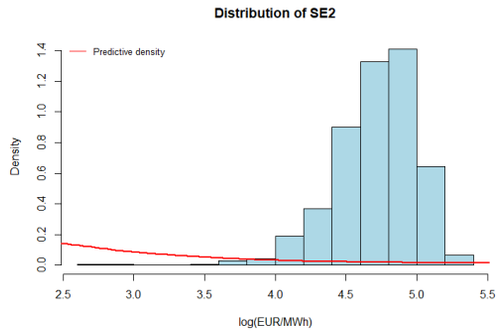

Figure F.11: SE2

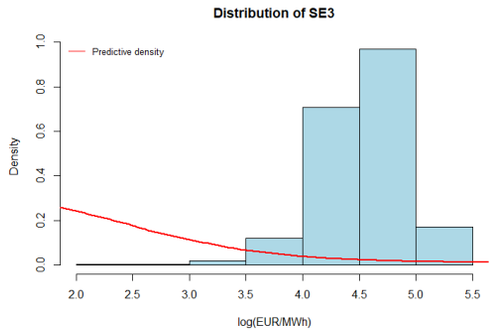

Figure F.12: SE3

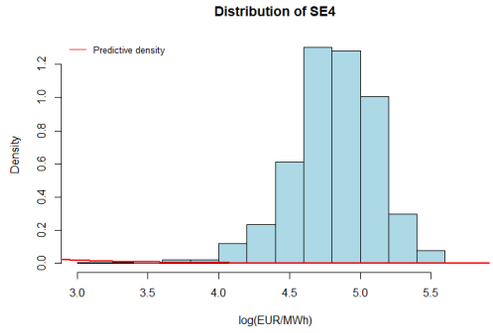

Figure F.13: SE4

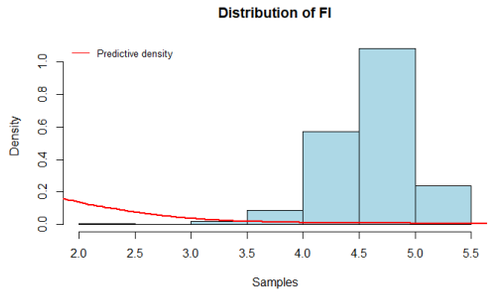

Figure F.14: FI

Figure F.15: The light blue histograms represent the logarithm of the out-of-sample observations for Y for the time series SE2, SE3, SE4, and FI, in the four figures. The red line in each plot represents the deep time series predictive density obtained by the Gaussian probabilistic ESN.

Appendix G

G.1 Forecast Accuracy for the Gauss Probabilistic ESN Model

This section presents the measures of forecasting accuracy, MAE, RMSE, lower-, and upper qunatiles, obtained after having applied the Gauss probabilistic ESN model to the data.

	MAE	RMSE	$\alpha = 0.05$	$\alpha = 0.95$
DK1	94.52	100.88	15.82	146.60
DK2	84.29	93.10	16.14	144.60
NO1	90.52	100.88	15.82	146.60
NO2	91.20	95.02	38.45	123
NO3	60.72	68.78	1.62	111.52
NO4	17.44	31.19	0.51	73.82
NO5	90.47	94.98	14.22	124.78
SE1	61.90	68.56	14.99	111.70
SE2	61.50	68.30	11.87	111.99
SE3	63.51	71.67	12.60	118.81
SE4	67.60	76.81	9.51	129.8
FI	63.06	71.28	9.75	117.64

Table G.1: This table shows the mean absolute error (MAE, root mean squared error (RMSE), along with the lower and upper quantile, $\alpha = 0.05$ and $\alpha = 0.95$, respectively, for the 12 distinct areas as a measure of forecast accuracy for the Gauss probabilistic ESN model.

Appendix H

H.1 Density Plots for the Copula Model

This section presents the density plots along with the predictive density obtained using the Copula model for each of the 12 considered areas.

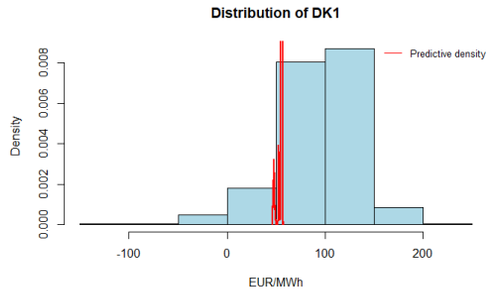

Figure H.1: DK1

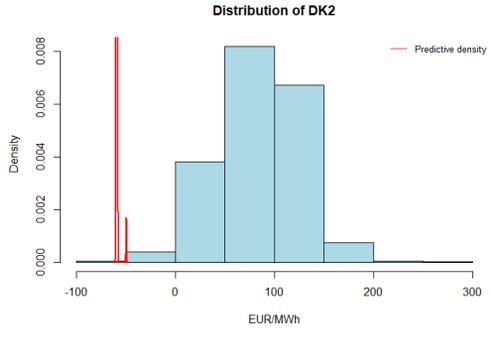

Figure H.2: DK2

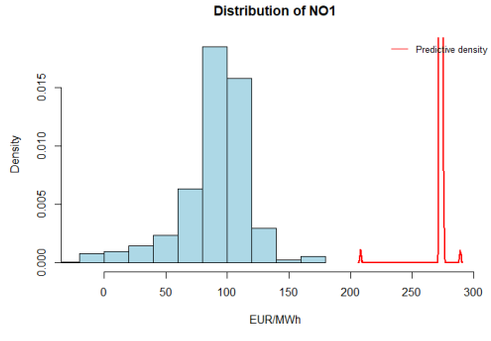

Figure H.3: NO1

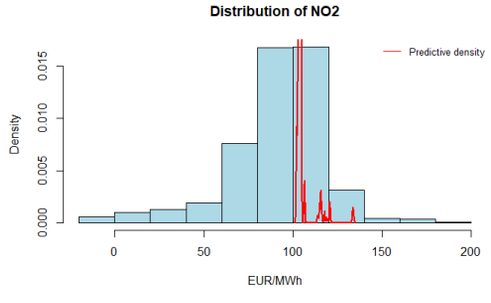

Figure H.4: NO2

Figure H.5: The out-of-sample observations for Y, here DK1, DK2, NO1, and NO2, during the validation period between April 1, 2023 and May 5, 2023, are represented by the light blue boxplot. The deep time series predictive density obtained using the Copula model is displayed as the red line.

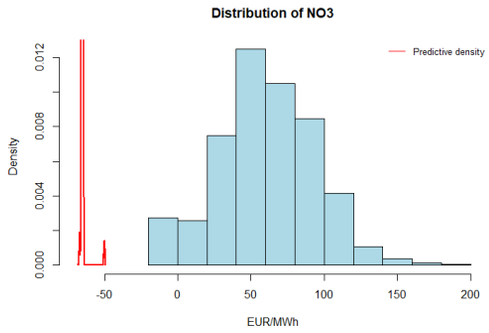

Figure H.6: NO3

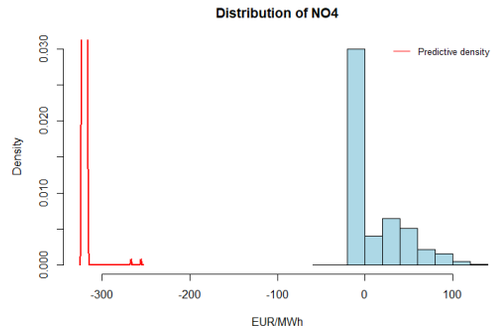

Figure H.7: NO4

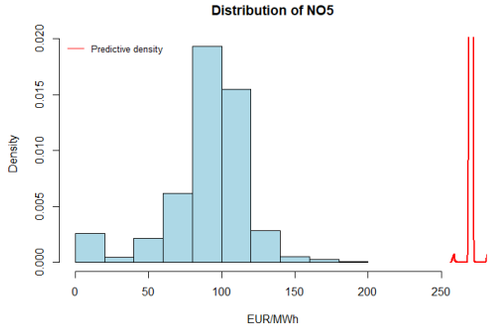

Figure H.8: NO5

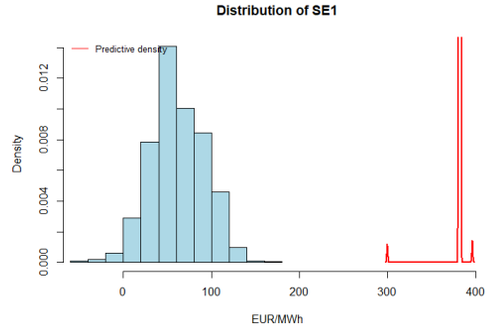

Figure H.9: SE1

Figure H.10: The out-of-sample observations for Y, here NO3, NO4, NO5, and SE1, during the validation period between April 1, 2023 and May 5, 2023, are represented by the light blue boxplot. The deep time series predictive density obtained using the Copula model is displayed as the red line.

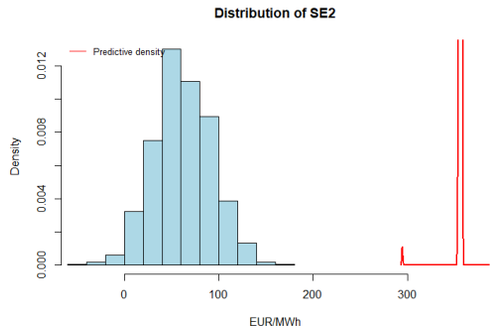

Figure H.11: SE2

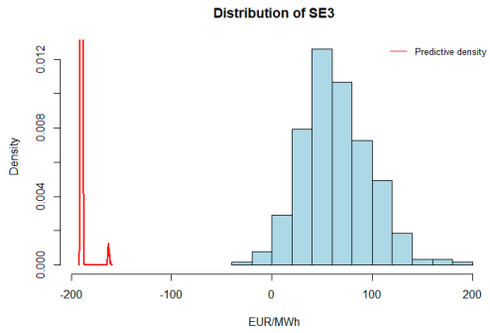

Figure H.12: SE3

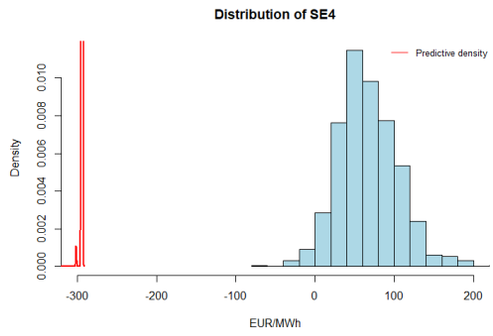

Figure H.13: SE4

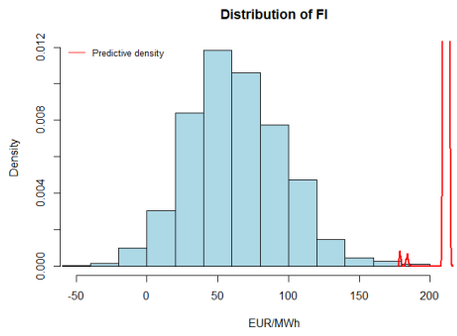

Figure H.14: FI

Figure H.15: The out-of-sample observations for Y, here SE2, SE3, SE4, and FI, during the validation period between April 1, 2023 and May 5, 2023, are represented by the light blue boxplot. The deep time series predictive density obtained using the Copula model is displayed as the red line.

Appendix I

I.1 Forecast Accuracy for the Copula Model

This section presents the measures of forecasting accuracy, MAE, RMSE, lower-, and upper qunatiles, obtained after having applied the Copula model to the data.

	MAE	RMSE	$\alpha = 0.05$	$\alpha = 0.95$
DK1	47.48	54.76	-40.48	92.83
DK2	143.29	149.13	74.83	205.67
NO1	182.72	184.86	-239.16	-147.68
NO2	20.58	30.00	-65.94	20.98
NO3	125.56	129.68	65.37	177.64
NO4	337.65	129.68	65.37	177.64
NO5	179.61	181.95	-256.64	-144.4
SE1	320.35	321.48	-367.48	-269.15
SE2	294.70	296.32	-344.32	-243.04
SE3	253.74	256.02	202.32	310.55
SE4	360.91	362.87	302.64	424.65
FI	148.27	152.22	-201.04	-93.14

Table I.1: This table shows the mean absolute error (MAE, root mean squared error (RMSE), along with the lower and upper quantile, $\alpha = 0.05$ and $\alpha = 0.95$, respectively, for the 12 distinct areas as a measure of forecast accuracy for the Copula model.