Author:
Akyol, Huseyin Burak

Title:
The persistence forecast effect in time-series predictions

The Persistence Forecast Effect in Time-Series Predictions

BY

HUSEYIN BURAK AKYOL

Department of Computer Science
UNIVERSITY OF BRISTOL

A dissertation submitted to the University of Bristol in accordance with the requirements of the degree of DOCTOR OF PHILOSOPHY in the Faculty of Engineering.

JANUARY 2023

Word count: 45013

Abstract

Over the past few years, time-series forecasting has been put into practice for a variety of purposes across various fields. It is regarded as crucial for many types of organisations and applications, as effective decision-making processes and intelligent autonomous systems depend heavily on predictions of the future. Consequently, time-series forecasting has drawn the attention of academic researchers and industry professionals towards its robust implementation and then reliable evaluation. However, due to the many practical challenges and issues involved, these tasks are not always straightforward and achievable. This thesis provides a formal definition and in-depth investigation of an important form of bias that has oftentimes been overlooked in the literature and thus works towards more robust time-series prediction models and their reliable assessments.

The bias, undermining the quality of models and skewing the predictions in a systematic way, arises when the underlying time-series data lacks regularity and certainty. When it occurs, it is observed that forecasting outputs systematically approximate one of the most recently observed values used in the input feature set, resulting in a series of predicted values that is almost identical to the series of observed values but is continuously delayed by a few steps in time. However, this behaviour often cannot be detected by many of the current accuracy assessment methods, which ultimately leads to overconfidence in forecasting models and prediction outputs.

Therefore, with the objective of guarding time-series models and predictions against the bias to achieve robust models and reliable predictions, this thesis provides a formal definition of the bias, explores its characteristics in greater detail, establishes the factors causing the bias, evaluates its potential negative implications, proposes a novel method for quantitative detection of the bias, and finally, investigates the prevalence of the bias in the literature and discusses how the bias may invalidate the outcomes of previously published works. Moreover, it presents experimental studies using various time-series datasets, including domestic electricity consumption data, to demonstrate the bias and the implementation of the proposed method within a realistic setting.

Dedication and Acknowledgements

This thesis is dedicated to my beloved parents, Saadet and Ekrem, and sister, Merve, who have always believed in me and unconditionally supported me in all of my endeavours. I cannot thank you enough for your unwavering support and for always being there for me.

I would like to express my sincere gratitude to my supervisors, Dr. Daniel Schien and Prof. Chris Preist, for their guidance and support throughout my PhD studies and the process of writing this thesis. I am also deeply grateful to my viva panel, consisting of Dr. Matthew Edwards and Prof. Adrian Friday, for their insightful comments and constructive feedback that have contributed to the refinement of this thesis.

Also, I would like to extend my heartfelt gratitude to all of my dear friends for their invaluable support in various aspects. In particular, I would like to thank Sakine Yalman, Mucahit Aydemir, and Yilmaz Cankurtaran for their solidarity and support throughout my journey in the United Kingdom. Their kindness and encouragement have been a constant source of strength and comfort.

Finally, I would like to thank the Ministry of National Education of Turkiye for granting me a scholarship to pursue my PhD studies in the United Kingdom.

Author's Declaration

I declare that the work in this dissertation was carried out in accordance with the requirements of the University's Regulations and Code of Practice for Research Degree Programmes and that it has not been submitted for any other academic award. Except where indicated by specific reference in the text, the work is the candidate's own work. Work done in collaboration with, or with the assistance of, others, is indicated as such. Any views expressed in the dissertation are those of the author.

SIGNED: ..... DATE: .....

TABLE OF CONTENTS

List of Tables

List of Figures

Figure **Page**

ABBREVIATIONS

AMPds2 The Almanac of Minutely Power dataset Version 2.

ANN Artificial Neural Networks.

AR Auto-Regressive.

ARIMA Auto-Regressive Integrated Moving Average.

ARIMAX Autoregressive Integrated Moving-Average with Exogenous Regressors.

ARMA Auto-Regressive Moving Average.

BPNN Back-Propagation Neural Network.

CNN Convolutional Neural Networks.

Corr Correlation Coefficient.

CV Coefficient of Variation.

DPE Double Penalty Effect.

ELM Extreme Learning Machines.

K-NN K-Nearest Neighbour.

LR Linear Regression.

LSTM Long Short-Term Memory.

MA Moving Average.

MAE Mean Absolute Error.

MAPE Mean Absolute Percentage Error.

MASE Mean Absolute Scaled Error.

MIMO Multi-Input Multi-Output.

MLR Multi-Linear Regression.

MPE Mean Percentage Error.

MSE Mean Squared Error.

MSPE Mean Squared Percentage Error.

N-RMSE Normalized Root Mean Squared Errors.

n-SS n-Step-Shifting.

non-LR non-Linear Regression.

PFE Persistence Forecast Effect.

R-MAPE Resistant Mean Absolute Percentage Error.

RAE Relative Absolute Error.

RAE-PM Relative Absolute Error with Persistence Model.

RMSE Root Mean Squared Error.

RMSLE Root Mean Squared Log Errors.

RNN Recurrent Neural Networks.

RRSE Root Relative Absolute Error.

S-MAPE Symmetric Mean Absolute Percentage Error.

SARIMA Seasonal Auto-Regressive Integrated Moving Average.

SGSC Smart Grid Smart City.

SVR Support Vector Regressions.

VARMA Vector Autoregression Moving-Average.

W-APE Weighted Absolute Percentage Error.

W-MAPE Weighted Mean Absolute Percentage Error.

CHAPTER 1

INTRODUCTION

The opening chapter of this thesis will provide the context for the research and introduces the main problem that drives the investigation throughout the remainder of this thesis. It will also outline the research objectives, the core research contributions of the study, and the organisational structure of the thesis.

1.1 Context and Motivation

The rapidly increasing number of information and communication technology devices, along with continued worldwide advances in data storage, processing, and transfer technologies, means more data is now generated than ever before. For example, it has been reported in [1] that the volume of data generated doubles every 18 months. This occurs in every aspect of our daily lives and in myriad sectors, including healthcare, education, finance, manufacturing, agriculture, transportation, entertainment, and security [2, 3]. However, if this tremendous volume of data is not analysed and turned into knowledge, it is nothing more than a collection of some values, and the effort of collecting and storing the data is wasted. On the other hand, extracting information from this data and transforming it into knowledge can make it meaningful and serve as a valuable resource for gaining a better understanding and control of the surrounding world and for making more informed decisions [4, 5].

It is often helpful to record the time associated with data as it provides context, for example, the age of the data. However, for some kinds of data, time is vital as it allows a change of a process over time to be tracked and understood. Each value in such data is stored with a timestamp illustrating exactly when the value was measured and what the order of data is. This is called *time-series data* and includes, for example, a step count recorded hourly by a

smartwatch or smartphone. Time-series data are used variously in different contexts. However, one of the primary objectives of collecting and analysing time-series data is to gain insight into the future based on past recorded values. This is known as *time-series forecasting*. It should be noted that the terms *prediction* and *forecast* are used interchangeably in this thesis, as they are in many other texts available in the literature. Time-series forecasting is about examining patterns and changes in the past in order to find the most likely state or value of a given time point in the future. This is regarded as critical in many types of organisations and application areas since proper decision-making processes and intelligent autonomous systems rely primarily on predictions of the future. For example, the planning of electricity generation and distribution is a vitally important task for a power-generation company in order to achieve the most efficient and effective resource management and maximise profit. For this task, an accurate prediction of future electricity demand is a must, which is possible only with time-series data representing the past electricity demand. Such predictions must be achieved for specific time periods, such as time of day, day of the week, and week and month of the year, to be able to avoid power shortages and optimise the use of renewable energy sources. Furthermore, in order to be able to benefit the most from renewable energy generation and to achieve balance between energy demand and supply, accurate weather prediction (wind speed, wind direction, cloud cover, etc.) is also required.

In order to obtain more accurate and reliable time-series forecasts, researchers from fields such as mathematics, statistics, and computer science have developed many different approaches, methods, and pieces of software. These efforts have resulted in the emergence of several different time-series prediction methods, each of which adopts different approaches and strategies. Based on the approaches they adopt, these methods have been split into two categories: statistical methods and machine-learning methods. These categories will be elaborated on in the next chapter. Briefly, however, they can be described as follows: statistical methods rely on statistical analyses of historically observed data points in a time-series dataset, and they aim to model the relationship between the past and the future through mathematical equations [6, 7], whereas machine-learning methods are computational techniques that mainly create functional relationships between the past and the future [8]. Time-series prediction previously used to be dominated by statistical methods. However, recent studies have shown that machine-learning methods are generally better than statistical methods at producing accurate and robust time-series predictions. Therefore, machine-learning methods have received more attention in recent years [7, 9].

However, as each time-series prediction problem has its own distinctive features and difficulties, there is no single method that is uniformly superior for every time-series prediction problem. Therefore, when tackling a given time-series prediction problem, it is recommended that multiple forecasting methods be tried, and then their prediction accuracy should be evaluated and compared to find the best-performing prediction method among alternatives [10]. Besides that, evaluating the accuracy of forecasts is also essential to validate both the model developed and the resultant predictions before putting them into practice. Therefore, it can be concluded that the

correct and reliable evaluation of accuracy of forecasts is critical to determining the effectiveness and quality of prediction outputs [11]. There are various accuracy metrics suggested in the literature, which will be listed and described in Chapter 2. What these metrics essentially do is compare the prediction outputs with the actual measurements and then convert this information into a quantitative representation of the overall accuracy of predictions based on mathematical notations describing these metrics.

Nevertheless, due to the many challenges and issues that the time-series prediction and accuracy assessment tasks entail, it is not always feasible to develop a model that produces accurate time-series forecasts or to perform reliable prediction assessments in practice. Depending on the context of the time-series prediction application, the deployment of a prediction model that is mistakenly believed to produce accurate predictions due to an inadequate accuracy assessment can have serious negative impacts on the application, including its functionality, cost, safety, stability, and maintenance. For instance, a smart household energy management system that relies on a trusted but nonetheless unreliable prediction model of future domestic electricity demand would probably fail to function properly. Such a smart energy management system is likely to make incorrect decisions in many aspects, which can possibly lead to inefficient energy consumption and misguided energy consumption planning. This means higher bills to pay for residents. Moreover, the flawed consumption planning could eventually result in potentially exacerbated peaks in demand across the power grid, reducing the overall network security and resilience. Such impacts on the grid would furthermore cause ineffective and unsuccessful network management and planning by network operators.

Therefore, in order to guard against the potential negative impacts, it is crucial to have reliably accurate forecasts. This can be achieved through a robust and secure assessment of accuracy of forecasting models and predictions. This is the main focus of this thesis. The research presented in this thesis provides a formal definition and in-depth examination of a phenomenon. This phenomenon undermines the quality of models and skews the predictions in a systematic way, making it an important form of bias. The bias, which is labelled the *Persistence Forecast Effect* (PFE) in this thesis, occurs when the underlying time-series data lack regularity and certainty. When it occurs, it is observed that prediction outputs systematically approximate one of the most recently observed values, resulting in a series of predicted values that is nearly identical to the series of observed values but is continuously delayed one or a few steps in time, see Figure 1.1. In other words, biased prediction results suggest that what will happen next is almost identical to what has recently happened, which is not the case in most of the real-world scenarios. Despite the fact that this is a common form of bias that weakens the reliability and robustness of time-series forecasts and their evaluations, it has frequently been overlooked so far. This is primarily due to insufficient evaluation of prediction accuracy. Most often, the currently available evaluation metrics are likely unable to identify this effect, potentially resulting in overconfidence in time-series predictions and models. The primary motivations of this thesis are

to shed light on this bias – PFE, to explore the causes and potential negative impacts of the PFE, and to propose a method for detecting when predictions are biased by the PFE.

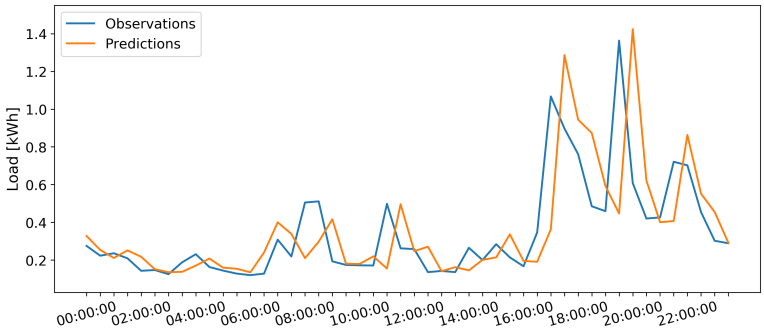


Figure 1.1: Forecasting outputs systematically follow the observed values behind in time due to the PFE.

The bias investigated in this thesis can occur in any time-series forecast in any time-series domain. That is, if the characteristics of the gathered time-series data meet certain conditions, which will be explored in the following chapters, the bias occurs and affects the produced time-series predictions, regardless of the time-series data domain. Nevertheless, this thesis studies the bias more specifically in the context of electricity demand forecasts, which are crucial for tackling a range of global issues such as electricity supply and climate change.

1.2 Research Objective and Contributions

The research presented in this thesis contributes to the study and development of data-driven time-series forecasting and prediction evaluation so as to improve the reliability and robustness of time-series predictions and assessments. In particular, this thesis explores the common but frequently overlooked bias of the PFE, which may affect the reliability of predictions and the robustness of their evaluations. A formal and quantitative method for identifying the existence of the bias in predictions is then provided.

In more detail, as the main contributions of this thesis, it:

1. formally defines the PFE, which refers to a systematic temporal delay observed in time-series forecasts;
2. offers a systematic exploration of high variability and irregularity in data as underlying factors causing the PFE;
3. demonstrates the influence of auto-correlation on the amount of the delay induced by the PFE;
4. thoroughly investigates and evaluates the potential negative impacts of the PFE to motivate the significance of its identification;

5. introduces a novel computational method that facilitates a mechanism for the detection of the presence of the PFE in time-series predictions and illustrates its implementation;
6. presents empirical studies using diverse time-series datasets, such as household-level electricity consumption data, in order to effectively demonstrate the PFE and the application of the proposed detection method within a practical context;
7. demonstrates the pervasiveness of the PFE in the existing literature and discusses its potential to invalidate the outcomes of previously published works.

1.3 Thesis Structure

The remaining chapters of the thesis are organised as follows:

- Chapter 2 gives an overview of the background knowledge associated with the fundamentals of time-series forecasting, including a brief introduction to time-series forecasting methods and forecast accuracy metrics. Furthermore, in order to contextualise the work presented in this thesis, the chapter also provides a review of a selection of previously reported time-series forecasting problems that are most relevant to the PFE, together with the proposed solutions to these issues.

- Chapter 3 provides a formal description of the PFE and a comprehensive analysis of the factors that lead to the PFE phenomenon. This chapter also investigates and evaluates the potential negative implications of the PFE, which motivates its identification prior to taking further steps relying on predictions affected by the PFE.

- Chapter 4 proposes and introduces the *n-Step-Shifting* (n-SS) method as a new generic method for the quantitative identification of the absence or presence of the PFE. Moreover, this chapter also exemplifies how to deploy the n-SS method, guides the reader through determining the value of n and discusses the limitations of the proposed method in order to improve its effective and efficient operation.

- Chapter 5 provides an empirical study conducted with the purpose of illustrating how the PFE manifests itself in single-step time-series forecasts and how the proposed n-SS method can be applied. The experiments presented in this chapter are carried out with a large-scale domestic electricity consumption dataset and advanced machine-learning methods. This chapter also determines and presents the effects of the PFE within a practical context.

- Chapter 6 investigates how irregular time-series data influences multi-step forecasts and how this effect differs from the PFE in single-step forecasts. Besides that, the chapter explores the usefulness of the proposed n-SS method in multi-step cases and also discusses under what conditions it might be used effectively. Finally, the chapter presents an experimental study similar to that of Chapter 5.

- Chapter 7, explores further details on the PFE. To this end, the chapter first poses a set of relevant research questions regarding the characteristics of the PFE and then answers them through carrying out in-depth analysis and experiments using time-series data collected from various domains.
- Chapter 8 provides a survey of a number of published works that implement different prediction methods on different time-series datasets from various domains. The purpose of the provided survey is to inspect their results in order to determine whether or not the time-series predictions presented in these works are affected by the PFE and, thereby, to assess the prevalence of the PFE in the existing time-series prediction literature. The chapter then also discusses how and why the PFE might undermine the conclusions for existing published works.

1.4 Published Work

Some of the text and research presented in different chapters of this thesis were previously published as a journal article in IEEE Transactions on Smart Grid.

> [12] **H. B. Akyol**, C. Preist, and D. Schien, "Avoiding Overconfidence in Predictions of Residential Energy Demand Through Identification of the Persistence Forecast Effect," *IEEE Transactions on Smart Grid*, vol. 14, no. 1, pp. 228–238, 2023.

This publication introduced the preliminary PFE idea, proposed the conceptual PFE detection method, and also provided a case study with household-level electricity consumption data.

CHAPTER 2

BACKGROUND AND LITERATURE REVIEW

This thesis defines and investigates a bias, referring to an effect causing time-series predictions to trail the actual observations one or a few steps behind in time. By doing this, it is aimed to contribute to a better assessment of time-series predictions so that more robust and reliable time-series predictions are achieved. This chapter will, therefore, first present a background review of time-series forecasting, including overviews of time-series prediction methods and prediction accuracy assessment. This is followed by a review of the most frequent and relevant practical issues that have been reported in the time-series forecasting literature.

It is important to note that time-series analysis and time-series forecasting are broad fields with varying research objectives. Attempting to provide an exhaustive description of these concepts within a single thesis chapter is practically impossible. Therefore, this chapter will primarily focus on the fundamental concepts and methods that are directly relevant to the rest of the presented research in order to maintain the scope of this chapter within the context of the research objectives. For readers seeking a more comprehensive and broader understanding of time-series forecasting, it is recommended to consult the following references: [13–19]. These are some well-established and well-regarded sources that offer detailed explanations and comprehensive coverage of time-series forecasting.

2.1 Basics of Time-Series Forecasting

Despite the fact that time-series forecasting is a dynamic research area that has drawn the attention of researchers over the last few decades, the practice of forecasting has always been desirable for people throughout the history of mankind, as mankind has always been curious about their future lives [19, 20]. At times and in certain places, foretelling was considered a crime

and was even strictly forbidden in some cultures and religions. However, this never prevented our collective fascination with predictions. People throughout history have developed new methods to foretell the future and found various notions to rely on, such as the positions of the planets and stars, dreams, hand and face shapes, and crystal balls. However, time-series forecasting is fully distinct from what diviners, fortune tellers, and soothsayers have been doing throughout history. This is because time-series forecasts rely on recorded past observations through time, referred to as *time-series data*, and knowledge extracted from the time-series data rather than mystic notions, spirits, or special powers.

2.1.1 Time-Series Data

In its simplest form, time-series data are no more than a collection of observations, each recorded sequentially at a specific time [21, 22]. Each observation is indexed by the time stamp of when the value was measured, resulting in a notation of $\{x_t, t \in T\}$, where x_t indicates the observed value, t refers to the observation time, and T is the set of discrete integers representing time points at which measurements are made; $T = \{0, 1, 2, ...\}$. Not necessarily always, but mostly, time-series data involves a collection of recorded values with equally spaced time intervals between observations [23, 24], and this thesis restricts its attention to such series. An intrinsic feature of time-series data is the fact that typically, there is a fundamental dependence and correlation amongst the consecutive elements of a sequence of measurements [25–27]. This is due to an assumption that systems and processes are not expected to change quickly between successive measurements, depending on the sampling frequency [17]. Thus, it could be said that the value of x_t is dependent in some way on the previous values of x; $\{x_{t-1}, x_{t-2}, ..., x_{t-n}\}$. As a result, the chronological order of measurements that comprise the time-series data always matters and must always be taken into account properly [28, 29]. Examples of such data are many and varied, for instance, hourly measurements of electrical power consumption, daily measurements of outdoor temperature (see Figure 2.1), sequences of closing stock prices, a weekly series of interest rates, a monthly series of totals of international airline passengers (see Figure 2.2), quarterly observations made on the population of a country, and annual birth rates in a city. The study of time-series is concerned with understanding and modelling the stochastic mechanism of a time-dependent structure of the observed values that is essential for forecasting the future values of the series, which is called *time-series forecasting* [30, 31].

2.1.2 Time-Series Forecasting

Time-series forecasting can be described as the process of predicting the future by comprehending and learning from the past through previously collected time-series data [20, 34]. That is, in order for time-series forecasting, a patterns describing the past is identified in the obtained data, and then the identified repetitive patterns are extended into the future [13]. Even more formally, time-series forecasting problems can be represented as the problem of predicting the value(s) of

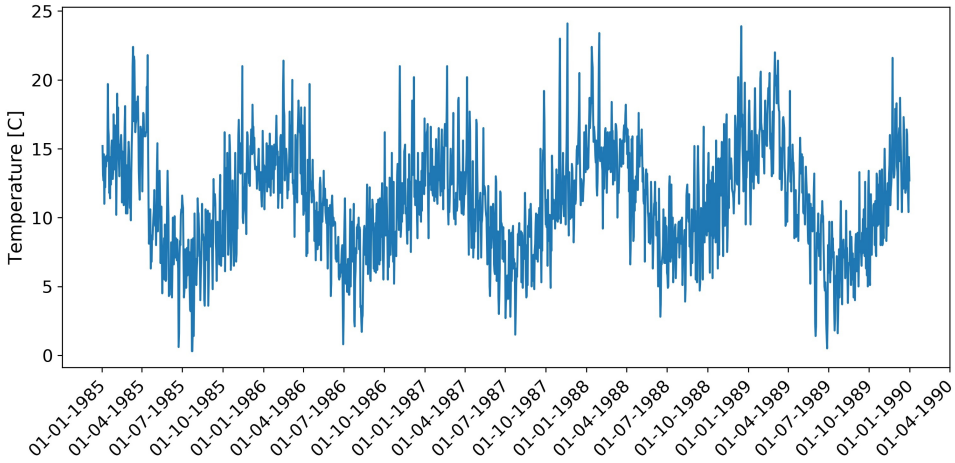


Figure 2.1: Minimum daily outdoor temperatures in Melbourne, Australia from 01 January 1985 to 01 January 1990 [32].

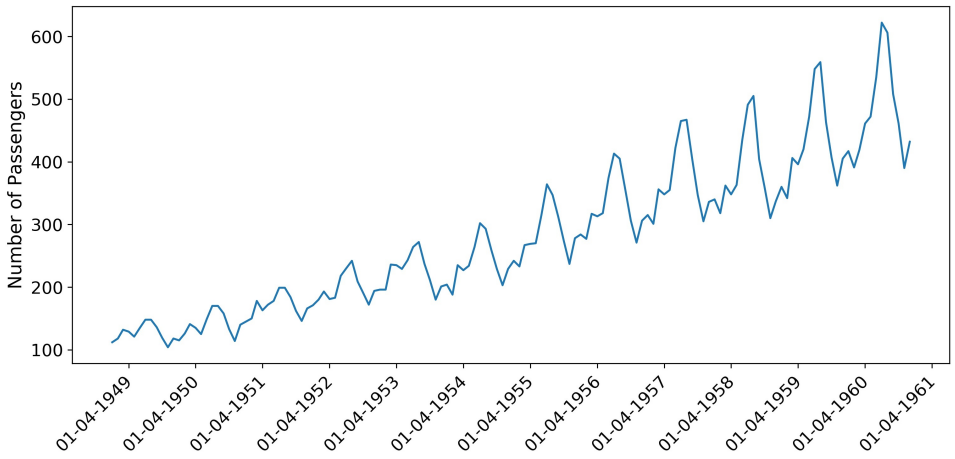


Figure 2.2: Monthly totals of international airline passengers in thousands from January 1949 to December 1960 [33].

x_{t+h}, $h > 0$ based on the previously observed values of x and, possibly, some other related series or factors, given time t is the current time. Time-series forecasting is an indispensably vital tool, holding immense importance in making future strategic decisions, taking precautionary measures, inventory management, and efficient planning and scheduling [13, 18] in a wide range of practical fields, including business, policy, economics, finance, engineering, the environment, telecommunications, meteorology, healthcare, and astronomy [19, 35]. Depending on the discipline and the problem at hand, time-series forecasts might be needed for short-, medium-, and long-term horizons [36]. Automation systems, for instance, often require short-term forecasts, whereas strategic planning and decision-making procedures generally require long-term forecasts [36].

Time-series predictions can also be classified into two types based on the prediction outcomes: point forecasting and probabilistic forecasting. In point forecasting, prediction methods produce a single numerical predicted value for each time point in the forecasting horizon, while probabilistic forecasting aims to estimate a probabilistic distribution of possible future outcomes based on given time-series observations. It should be noted that, in this thesis, time-series forecasting always refers to point forecasting, unless otherwise specified. Point forecasting can further be divided into two types depending on the forecasting horizon: single-step (ahead) forecasts and multi-step (ahead) forecasts. Single-step forecasts (illustrated in Figure 2.3a) can be defined as:

Single-step forecasting is a task of predicting a single value for the immediate next time step.

$$y_{t+1} = f(x_t, x_{t-1}, ..., x_{t-T}, X) \tag{2.1}$$

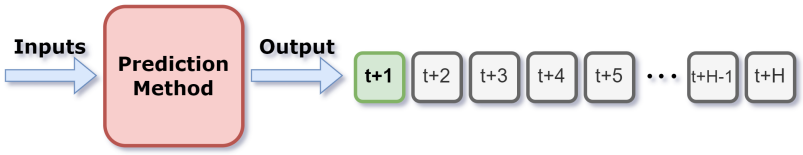
where y_{t+1} is the forecast of x_{t+1}, f is the forecasting model, x_t is the observation at time t, T is the number of past observations used in the input set, and X is the set of other relevant features in the input set. Multi-step forecasts (illustrated in Figure 2.3b) can be defined as:

Multi-step forecasting is a task of predicting a sequence of multiple values for future time steps within the forecasting horizon.

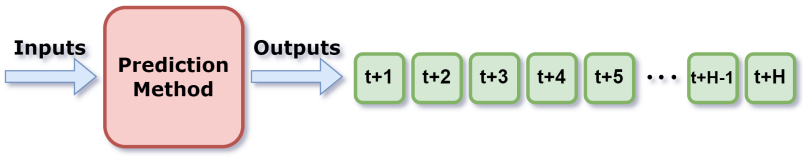
$$(y_{t+1}), (y_{t+2}), ..., (y_{t+H}) = f(x_t, x_{t-1}, ..., x_{t-T}, X) \tag{2.2}$$

where H is the absolute forecasting horizon and y_{t+H} is the forecast of x_{t+H}, f is the forecasting model, x_t is the observation at time t, T is the number of past observations used in the input set, and X is the set of other relevant features in the input set. The bias that is the research focus of this thesis will be investigated in both single- and multi-step forecasts.

The variables (prediction outputs) on the left side of Equations (2.1) and (2.2) are referred to as dependent variables, while the variables (features in the input set) on the right hand side of the equations are referred to as independent variables or predictors. Given this, the task of data-driven forecasting methods can be articulated as calculating the functional relationship between the dependent and independent variables to be able to predict the future values of the dependent variables. In time-series predictions, an independent variable set always includes a certain number of past observations of the time-series of the output: $\{x_t, x_{t-1}, ..., x_{t-T}\}$. However, the set of other relevant features, denoted by $\{X\}$ in the equations, is optional. In the context of building-level electrical load forecasting, for example, $\{X\}$ could include weather-related features, such as outside temperature and humidity, and calendar-related features, such as time of day and week of year. On the basis of the existence of the $\{X\}$, the prediction models can be divided into two kinds: univariate (models that utilise only the past observations of x) and multivariate (models that utilise additional relevant variables along with the past observations of x) [13].



(a)



(b)

Figure 2.3: Diagrams illustrating (a) single-step (ahead) forecasts and (b) multi-step (ahead) forecasts.

The research presented in this thesis does not differentiate between these different types of time-series forecasts. Instead, the general time-series point forecasting problem defined by Equations (2.1) and (2.2) will be considered.

2.1.3 Time-Series Forecasting Methods

As will be seen in the following chapters, the bias that is the main subject of this thesis is triggered solely by the characteristics of the underlying time-series data, making it independent of data-driven forecasting methods. This section, therefore, is not intended to provide a thorough description of particular prediction methods or to be a comprehensive manual of how to apply time-series prediction methods. Rather, the purpose of this section is to give a brief overview of several forecasting methods proposed and used in the time-series point forecasting literature.

Before proceeding further and getting into an overview of time-series forecasting methods, it is important to clarify what is meant by the term *forecasting method* and to distinguish between a forecasting method and a forecasting model, as these terms should be kept clearly distinct [16]. Unfortunately, in the time-series forecasting literature, there is no consensus on the definitions of these terms, and they are sometimes wrongly used interchangeably. In this thesis, the term *forecasting method* is used to describe a mathematical procedure or an algorithmic rule that computes and generates time-series forecasts, while the term *forecasting model* is used to describe a mathematical representation of a time-series data with full specifications and parameters [36, 37]. A forecasting method is, therefore, a technique or an algorithm, whereas a forecasting model is a method fitted to or trained on given time-series data.

The time-series forecasting process can thus be summarised as having two steps. First, the selected time-series prediction method is fitted or trained over the collected time-series data so that a suitable prediction model, providing a plausible description of the relationship between

independent and dependent variables, is developed. The developed prediction model is then used to forecast the future values of the given time-series data.

The selection of a proper prediction method and then adequate model fitting or training are not trivial tasks. They are extremely significant not only in successful and accurate time-series forecasting but also in determining the time, effort, and costs involved with the prediction process [20]. Over many years, researchers and practitioners have put a lot of effort into the development and improvement of successful time-series prediction methods on the basis of the time-series prediction problem they were dealing with. This has resulted in the emergence and evolution of a wide variety of methods for performing time-series predictions in the literature. Some are very simple and basic methods that use transformations and adjustments to predict the future values of the series [19], without the need for model fitting or training. These methods, which include the average method, moving average method, naïve method, seasonal naïve method, and drift method, are mostly used as benchmarks to verify the prediction capability of alternative methods [19, 38]. Methods that require model fitting or training on historically recorded time-series data are referred to as *data-driven methods*. These methods are nowadays, broadly speaking, divided into categories of statistical (traditional) methods and machine-learning (artificial intelligence) methods.

Statistical forecasting methods are simple in terms of development and implementation as they are represented by mathematical equations [6, 7, 39], meaning they are regarded as consistent but relatively inflexible [9, 39]. They are widely known for their distinctive features, which include low computational requirements and short implementation times [6], while also providing a high level of transparency and explainability of how the prediction output is produced. Statistical methods are effective at dealing with time-series data composed of finite, countable, and explainable independent variables by associating one or more independent variables to a dependent variable [6, 40]. They have, therefore, been widely used for linear problems in a variety of contexts. However, statistical methods suffer from generally being unable to identify non-linear patterns in time-series data [9, 41]. This drawback arises from the assumption that the relationship between input variables and the output variable is linear, which is also a major source of inaccurate predictions [7]. The authors of [42], presenting a systematic and critical review of forecasting methods used in 483 studies, report that 28 different statistical methods have been used in the time-series forecasting literature. Some most frequently used examples of statistical time-series forecasting methods are: Regression Methods, which includes Linear Regression (LR), Non-Linear regression (non-LR), Multi-Linear Regression (MLR), Gaussian process Regression, Ordinary Least Squares Regression, Bayesian Regression, and Exponential Smoothings; and Time-Series Methods, including Auto-Regressive (AR), Moving Average (MA), Auto-Regressive Moving Average (ARMA), Vector Autoregression Moving-Average (VARMA), Auto-Regressive Integrated Moving Average (ARIMA), Autoregressive Integrated Moving-Average with Exogenous Regressors (ARIMAX), Seasonal Auto-Regressive Integrated Moving Average (SARIMA), Decision Trees, Wavelet Transforms, Kalman Filters, and Logistic Regression.

Machine-learning methods, by contrast, are computational techniques that do not require any mathematical modelling [7]. Instead, they create functional relationships between dependent and independent variables [8] so that they are able to tackle non-linear and highly dynamic patterns hidden in time-series data [6]. This feature has been responsible for the recent popularity of machine-learning time-series forecasting methods. These methods can handle massive amounts of time-series data involving complex relationships between input and output variables, but are known to be computationally intensive and more demanding in terms of time [43, 44]. Another significant disadvantage of such methods is that their explainability and interpretability may not be fully clear, making it challenging to comprehend the underlying processes and the relationship between input and output variables in many cases. Nevertheless, since observational time-series data generally involve non-linear characteristics and machine-learning methods have been shown to be better than statistical methods at handling the non-linearity in time-series data, most researchers and practitioners nowadays prefer machine-learning methods to statistical methods [9, 45]. In consequence, various types of machine-learning methods have been developed and deployed for a variety of applications and research areas in the time-series forecasting literature [45]. One literature survey [42] found that 22 different machine-learning methods had been deployed in 483 time-series forecasting studies. Examples of the most popular machine-learning methods include Support Vector Regressions (SVR), Artificial Neural Networks (ANN), Fuzzy Logic Models, Evolutionary Algorithms, Extreme Learning Machines (ELM), Gradient Boosting Machines, Neuro-Fuzzy Systems, Regression Trees, XGBoost Regressions, K-Nearest Neighbour (K-NN), Case-Based Reasoning, Recurrent Neural Networks (RNN), Convolutional Neural Networks (CNN), and Random Forest Regressions.

Statistical methods have historically been popular, however, with emergence and advent of machine learning methods and their superior forecasting performance, they have gained the most attention in recent times [7, 9]. Moreover, hybrid methods, which combine two or more complementary methods to benefit from the advantages of each method used, have recently been shown to most often improve forecasting quality and yield higher accuracy compared to stand-alone methods [6, 46].

Overall, each category and each method within these categories has its own set of advantages and disadvantages. Although these two categories of prediction methods have evolved in different scientific fields (machine-learning methods evolved in the computer science community while statistical methods evolved in mathematics and statistics), the line separating the two is becoming increasingly ambiguous due to the collaboration of these scientific communities, and they have a lot in common [36, 47]. For instance, they are both data-driven methods, which means they both use previously recorded time-series values to build models that produce forecasts for future time points, and their overall success is entirely dependent on the characteristics of the underlying dataset providing historical recordings [6]. In case the historically observed values do not involve regular and repetitive patterns, proper model learning/fitting cannot be realised regardless of the

choice of the method used. Such an improperly developed model often results in the PFE, which causes time-series predictions to trail the actual observations one or a few steps behind in time.

2.1.4 Evaluation of Time-Series Forecasts

One of the primary objectives of time-series studies is to be able to predict the future values of the series. What is equally or even more critical is a reliable assessment of the goodness of those forecasts [30], leading to intriguing follow-up questions: “What are good forecasts?”, and "How can the goodness of forecasts be measured?” The main purpose of this subsection is to provide answers to these questions.

In time-series forecasting, the term a *good forecast* means an accurate forecast. That is to say, the goodness of forecasts is determined by the accuracy of overall time-series forecasts. If the accuracy of forecasts is not measured and evaluated properly, time-series forecasting can be perceived as nothing more than generating meaningless arbitrary numerical values in a sense. Therefore, effective accuracy assessment of time-series forecasts is as vital as the forecasting process itself, particularly considering that time-series forecasts are very rarely perfect [39].

The accuracy of point forecasts is measured by forecast error (e_{t+i}), which can simply be defined as the difference between what is forecasted and what is measured [39]. That is, given the current time is t, x_{t+i} is the measured value and y_{t+i} is the predicted value, then the forecast error can be formalised as follows:

$$e_{t+i} = x_{t+i} - y_{t+i} \tag{2.3}$$

where the time t is the current time and $i \in N^+$, and thus the time $t+i$ is a time point in the future.

Based on the definition of forecast error, it could be concluded that the assessment of the accuracy of predictions requires the actual values of the future to have already been measured, which is not possible in a real-world setting. Therefore, it is a general practice to split the data into two portions of a training set and a test set (Figure 2.4). This is a kind of way of treating a part of the past as if it were the future. The logic behind this is that as the test set is not used for model fitting or training, it can be used to achieve a reliable accuracy evaluation [19].

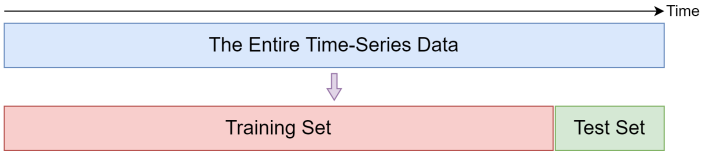


Figure 2.4: Illustration of training-test split of time-series data.

The training set is used for time-series model fitting or training, while the test set is used for accuracy measurement and performance evaluation. The prediction model, fitted or trained on the training set only, produces predictions for each time point in the test set, and then the forecast error is calculated for each prediction over the entire test set. In order to find out the overall accuracy of predictions, individual forecast errors are assembled on the basis of the chosen evaluation metric. This is called an *out-of-sample error*. If the error is calculated on data points in the training set, which is already used for model fitting and training, this is called an *in-sample error*. Among these two, out-of-sample error is generally considered preferable for the overall predictions accuracy assessment since it is calculated on time-series data that has not been used during the model fitting or training [36]. However, the outcome of the in-sample error is also important in some cases, as will be seen and discussed in later sections.

Time-series point forecasts are evaluated by point-wise evaluation metrics, also known as point-wise accuracy metrics. Point-wise evaluation metrics measure the forecast error from a discrepancy between the predicted and observed values at each time step in the form of numeric values, as described in Equation (2.3). A wide range of point-wise evaluation metrics, with their distinct merits and demerits, have been reported in the time-series forecasting literature, and each metric provides different types of information about the prediction outputs. Some of the most commonly used point-wise accuracy metrics are listed below, along with their mathematical formulations:

$$\text{(Mean Squared Error)}\ MSE = 1/m \sum_{t=1}^{m} (x_t - y_t)^2 \tag{2.4}$$

$$\text{(Root Mean Squared Error)}\ RMSE = \sqrt{1/m \sum_{t=1}^{m} (x_t - y_t)^2} \tag{2.5}$$

$$\text{(Mean Absolute Error)}\ MAE = 1/m \sum_{t=1}^{m} (|x_t - y_t|) \tag{2.6}$$

$$\text{(Mean Absolute Percentage Error)}\ MAPE = \left[1/m \sum_{t=1}^{m} (|x_t - y_t|)/x_t \right] \times 100 \tag{2.7}$$

$$\text{(Correlation Coefficient)}\ Corr = \frac{\sum_{t=1}^{m} (x_t - \overline{x})(y_t - \overline{y})}{\sqrt{\left[\sum_{t=1}^{m} (x_t - \overline{x})^2 \right] \left[\sum_{t=1}^{m} (y_t - \overline{y})^2 \right]}} \tag{2.8}$$

$$\text{(Relative Absolute Error)}\ RAE = \frac{\sum_{t=1}^{m}(|x_t - y_t|)}{\sum_{t=1}^{m}(|x_t - \overline{x}|)} \tag{2.9}$$

$$\text{(Root Relative Squared Error)}\ RRSE = \sqrt{\frac{\sum_{t=1}^{m}(x_t - y_t)^2}{\sum_{t=1}^{m}(x_t - \overline{x})^2}} \tag{2.10}$$

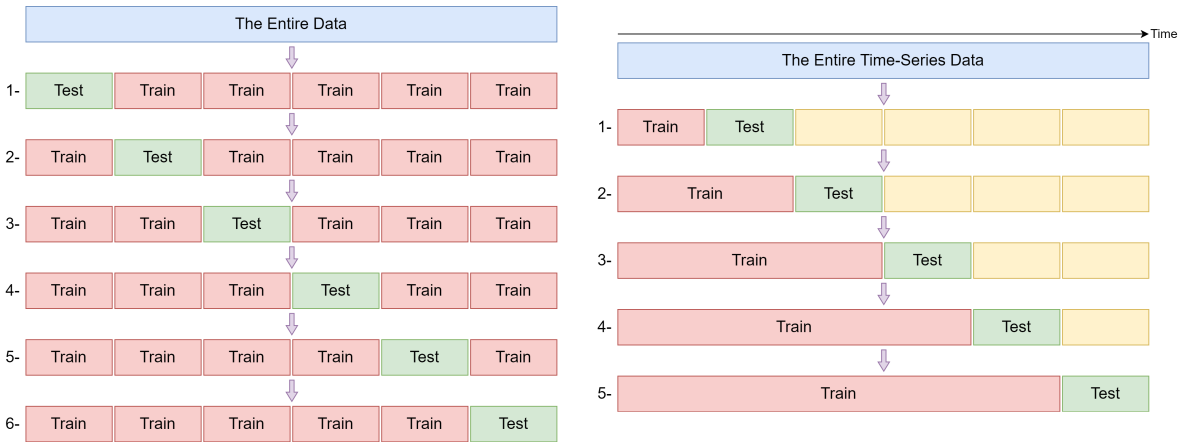
$$\text{(Coefficient of Determination)}\ R^2 = \frac{\sum_{t=1}^{m}(x_t - y_t)^2}{\sum_{t=1}^{m}(x_t - \overline{x})^2} \tag{2.11}$$

$$\text{(Coefficient of Variation)}\ CV = \frac{\sqrt{1/m\sum_{t=1}^{m}(x_t - y_t)^2}}{\overline{x}} \tag{2.12}$$

In all of the equations above, x_t is the actual value observed at time t, and y_t is the predicted value for time t. Furthermore, $\overline{x}$ and $\overline{y}$ represent the mean of the actual values and the mean of the predictions, respectively, and m is the total number of point predictions. RAE and RRSE are metrics that compare the actual forecast error to the forecast error of a simple (naïve) model [48, 49]. To put it differently, these metrics evaluate whether a model outperforms a simple baseline model, which can be an average method, moving average method, seasonal naïve method, or persistence model. In the mathematical expressions defining these metrics, Equations (2.9) and (2.10), the average method is chosen as a baseline model, as is typically done in the literature. It is also important to note that there are many other metrics besides those defined above, including variations of the above-described metrics, such as Mean Squared Percentage Error (MSPE), Mean Absolute Scaled Error (MASE), Normalised Root Mean Squared Errors (N-RMSE), Root Mean Squared Log Error (RMSLE), Weighted Absolute Percentage Error (W-APE), Weighted Mean Absolute Percentage Error (W-MAPE), Symmetric Mean Absolute Percentage Error (S-MAPE), Resistant Mean Absolute Percentage Error (R-MAPE), Adjusted R^2, and others.

Evaluation of the accuracy of time-series predictions based on a collection of unseen observations is critically important. However, considering the dependency among adjacent observations in time-series data, this is only possible with the training-test split of known time-series data, as shown in Figure 2.4. This approach is, however, criticised for not making complete use of available

data and allowing only a single accuracy evaluation on a test set [50–52]. In response to these negative sides of this approach, standard cross-validation, which is based on the assumption that observations in the underlying data are independent of one another, is modified specifically for time-series forecasting. This version of cross-validation is known as time-series cross-validation and also rolling origin. In standard cross-validation, data are split into K folds (parts). Each part is then used in turn as a test set while all the other parts are used for model fitting or training, resulting in K separate error measurements. This is illustrated in Figure 2.5a. However, standard cross-validation is not applicable for time-series forecasts, as the successive measurements are dependent on each other in time-series data. Therefore, it is modified into time-series cross-validation. Similar to the standard cross-validation, the time-series cross-validation also partitions the entire data into K parts. Every part except for the first part is then used as a test set once, with the observations that occurred prior to the test set forming the corresponding training set. As a result, the method eventually achieves $K-1$ independent realisations of the error measure, and almost all parts of the data are used for both fitting/training and testing (the first part is not used for testing, and the last part is not used for fitting or training). Finally, the overall error measure is calculated as the average of the $K-1$ obtained error measures. This is believed to provide more robust assessment results. An example of time-series cross-validation is provided in Figure 2.5b, where $K = 6$. In this example, the entire time-series data is divided into six complementary subsets, which results in five distinct training-test set pairs.



(a) (b)

Figure 2.5: Example of (a) 6-fold standard cross-validation and (b) 6-fold time-series cross-validation.

Even though there are some works adopting cross-validation in the recent time-series forecasting literature, cross-validation has yet to be accepted as a standard procedure in the accuracy assessment of time-series forecasts [52]. Practitioners and researchers still predominantly resort to a simple training-test split approach (represented in Figure 2.4) to accomplish prediction methods' parameter selections, to find the best-performing prediction model that must be deployed in

the final application, and to inspect whether or not the newly proposed method performs better than the alternative prediction methods [51, 53].

On the whole, evaluation metrics and how they are calculated generally have a fundamental significance in analysing the overall performance of a prediction procedure that followed. As will be shown in the following section, they are mostly used as a control mechanism to try to ensure that a model performs optimally and generates time-series predictions without any issues. However, most importantly, the bias resulting from the PFE, which is the main focus of the present research, cannot always be detected effectively by these accuracy metrics. Even though the predictions are affected by the PFE and they systematically fall one or a few steps back of the actual values, the evaluation metrics usually fail to detect the temporal displacement in predictions, and these metric results misleadingly suggest the prediction outcomes are accurate. As a result, the PFE is commonly overlooked, and it is infrequently diagnosed or considered a problem that can have serious repercussions in the end. Examples of the PFE and its implications will be examined and presented in great detail in the upcoming chapters.

2.2 Some Practical Issues in Time-Series Forecasting and Evaluation

Time-series forecasting has been receiving intensive interest from academia and industry. Recently, in particular, time-series forecasting has been put into practice for a variety of purposes in a wide range of domains. While it may seem simple and straightforward from a theoretical perspective, achieving robust and unbiased predictions with a satisfactory level of accuracy and reliable accuracy evaluation can be a complex task as many practical challenges and issues are involved. Some of these issues, such as over-fitting, under-fitting, phase error, and double penalty effect, are akin to and exhibit some degree of relevance to the PFE. Therefore, this section considers the discussion of these practical issues that arise in time-series forecasting and evaluation in order to contextualise the PFE proposed and the work presented in this thesis among related works. This survey provides an understanding of the differences between the proposed bias and the previously reported issues.

2.2.1 Under- and Over-Fitting Phenomena

The two most common practical issues in time-series forecasting are over- and under-fitting. These issues are major factors inhibiting a proper prediction of future values and thus leading to a poor forecasting performance. They essentially occur as a result of inadequate model fitting or learning [54]. Under-fitting is observed when a model is too simple or not trained enough and so not capable of capturing noteworthy regularities in a training set, whether through memorisation or not. Over-fitting, by contrast, which is the opposite of under-fitting, is observed

when a model is too complex or over-trained. In instances where it occurs, a model memorises the data points in a training set, including noise and the peculiarities in the training set, instead of learning the overall structural patterns, and so fails to generalise well outside the training set [36, 55, 56]. Therefore, with under-fitting, as the patterns and regularities in the training set are not sufficiently learnt and the precision of the learning is low, the prediction model cannot perform well on either the training or test set, resulting in large in-sample and out-of-sample errors [57, 58]. With over-fitting, on the other hand, the prediction model performs almost perfectly on the training set, which it has memorised with very weak flexibility and adaptability, and terribly on the test set, resulting in a small in-sample error but a large out-of-sample error [57, 58]. Therefore, the existence of under- and over-fitting phenomena can be easily detected by monitoring the evaluation metrics calculated across training and test sets [59]. As illustrated in Figure 2.6, insufficiently low accuracy for both sets indicates the presence of under-fitting, while a large gap between the accuracy of the two sets (where the in-sample error is significantly less than the out-of-sample error) indicates the existence of over-fitting [57].

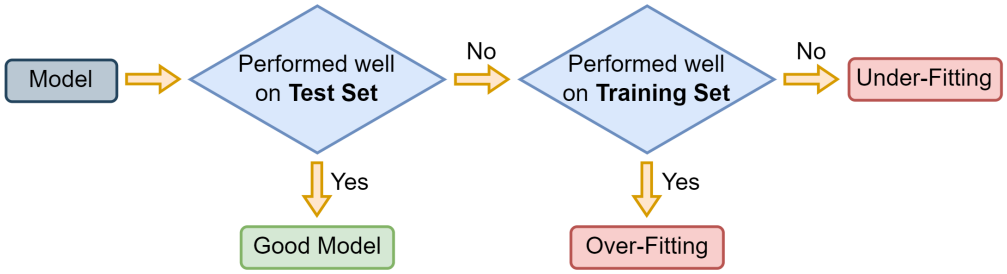


Figure 2.6: Good model vs. over-fitting vs. under-fitting (adapted from [60])

In [19], authors note that neither of these two phenomena is preferable over the other, as a model over-fitting to the training set is just as bad as a model failing to identify and learn systematic patterns in time-series observations. The mainstream view is that under- and over-fitting should both be avoided, although this is, unfortunately, not always completely achievable [54]. In the literature, there are several approaches that have been proposed in order to address the issues of under- and over-fitting. The details of these approaches are beyond the scope of this research; however, the author of [61] provides a succinct summary of the fundamental principles behind those approaches in his lecture notes. These principles include reducing the complexity of the prediction method, training the prediction method with more data, reducing its learning/fitting ability, adding noise to the time-series data, simplifying the time-series data to avoid excessive information, and using cross-validation to obtain better-tuned hyper-parameters to prevent over-fitting. In contrast, prevention of under-fitting requires implementing measures that are the exact opposite of those used to avoid over-fitting. As a consequence, it is suggested to develop a model that is balanced between under- and over-fitting so that it is able to learn the actual relationships between independent and dependent variables but does not memorise the data in the training set [56, 61].

2.2.2 Phase Error in Time-Series Forecasts

Time-series forecasts are defined by both the timing and amplitudes of a series of events. Therefore, it is critically important to predict the timing of an event correctly, along with its amplitude [62]. Within the literature, different terminologies are used to describe an error that arises when an event is predicted to occur either too early or too late. These terms include *phase error*, *phase shift*, and *timing error*; particularly for events predicted to happen too late, the terms *time lag*, *phase lag*, and *delay effect* are also used [62–65].

In [66, 67], the authors effectively illustrate that the phase error, together with bias and amplitude error, is one of the three main components forming accuracy metrics such as RMSE and MSE. As a consequence, in order to achieve relatively better time-series predictions and thus better metric values, the phase error must also be dealt with properly. One approach to address intermittent and discontinuous phase errors, which are out of the scope of this thesis, is to forecast ramp events, which refer to abrupt and significant fluctuations observed in time-series data within a short period of time [64]. This aims to correctly predict the timing of ramp events as well as their amplitude. In the context of electrical power, for example, there has been a growing interest in the predictions of ramp events in solar (see [68–70]) and wind (see [71–73]) power production to guard against the negative impacts of ramp events for greater safety, stability, and economics of power systems and energy storage devices. Further, a novel evaluation metric, the *ramp score*, has been proposed by the authors of [74] for assessing the quality of timing prediction of significant ramp events through time-series data.

The systematic and continuous delay in time-series predictions, which is the primary focus of this thesis, has received limited attention in the existing literature. Despite its widespread occurrence, only a few previously published works explicitly discuss the delay effect or attempted to resolve it [75].

The seminal work [62] can be credited as one of the earliest acknowledgements of delayed time-series predictions. The authors of this work argue that the delay in predictions can be attributed to the use of RMSE as the loss function for their neural network, which fails to incorporate a penalty for temporally delayed predictions. Hence, they assert that any predictive scheme exclusively reliant on RMSE will inevitably suffer from phase lag. As a result, in order to prevent the occurrence of delayed predictions, they propose adapting the neural network training scheme by integrating both the RMSE and a penalty for delayed predictions within the loss function. To further enhance the flexibility of the approach and its robustness against the temporal delay, they augment this approach with the utilisation of a genetic algorithm.

The authors of [65], [75] and [76] attribute the time delay in their time-series predictions to the high correlation between sequential observations and the use of the most recent observations in the input feature set. Therefore, the proposed solutions to the time lag issue in these papers

focus on reducing the auto-correlation and excluding the most recent observations from the input feature set through data transformations. In [65], for example, the authors decompose the time-series data with a multilevel discrete wavelet transform to effectively prevent any correlation between consecutive recordings. In [75], the most recent observed values used in the input set are replaced with a moving average of previous values to avoid the delay effect. Finally, the authors of [76] propose forecasting the relative differences between consecutive data points rather than the absolute values in order to reduce the correlation causing the phase lag in predictions.

Nevertheless, although these remedies seem to manage to reduce the time lag effect to some extent based on the results of these studies, the appropriateness of their implementation remains a subject of debate. This is because each of these solutions has its own drawbacks, i.e., the potential loss of crucial information inherent in observed values and a substantial decrease in overall prediction accuracy. Additionally, it is also noteworthy that these previously published works concur on a key contention: the phase lag error is a constraint or perhaps the major drawback of machine-learning methods, particularly those involving neural networks.

2.2.3 Double Penalty Effect

Considering the phase error in time-series forecasts, the authors of [74, 77–79] argue that point-wise metrics are inappropriate for time-series prediction evaluation. This is due to the fact that point-wise metrics simply compare the observed and predicted values at each time step, and hence they lead to a *double penalty effect* (DPE). The DPE refers to a situation in which point-wise metrics penalise temporally displaced predictions twice: first, where the event actually should be, and second, where the event is predicted to occur – even if the length and the amplitude of an event are correctly predicted in principle. In order to avoid this effect, the general recommendation of research studies in the literature is to tolerate and not penalise small and discontinuous displacements of predictions in time in order to avoid the DPE. In [77], for example, the authors introduce an *adjusted p-norm error measure* that allows for small and discontinuous displacements of predictions in time. This method partially drops the time dimension during evaluation and provides some temporal flexibility to predictions. Based on the results of the tests presented in [77], the authors conclude that the new metric they proposed is more suitable and useful for volatile and irregular time-series data, whereas standard point-wise metrics are better suited for time-series data involving mostly smooth and regular patterns. Another suggestion for avoiding the DPE is to replace point-wise metrics with alignment-based metrics, such as dynamic time warping, longest common sequence, parameterised forecast error metric, and move split merge in the evaluation of time-series predictions [77, 78]. Alignment-based metrics first calculate the optimal alignment between the predictions curve and observations curve and then assess the accuracy of time-series predictions based solely on the optimal alignment between them. Therefore, they do successfully prevent DPE. However, aligning the two curves in order

to find the optimal match between them implies that the importance of the time dimension is neglected. That is to say, the time order of data points is not preserved, and each is treated independently. Such metrics are, therefore, deemed unsuitable for the assessment of time-series predictions, where the time dimension and time order of the data points are inflexible. [77].

Ref.	**Evaluation Metric(s)**	**Ref.**	**Evaluation Metric(s)**
[44, 80]	MAE	[98]	RMSE, MAE, R^2
[46]	Corr, RMSE	[99]	MAPE, CV
[81]	Corr, MAPE, RMSE	[100]	MAPE, MSE, R-MAPE
[82]	MAE, MAPE, Corr	[101]	MAE, RMSE
[83]	RMSE	[102]	MAPE, MAE, R^2
[84–87]	MAPE, RMSE	[103]	MAE, MAPE, MPE, RMSE
[88]	RMSE, MAE, MAPE, R^2	[104]	CV
[89–91]	MAPE	[105]	R^2, RMSE, MAPE, MSE
[92]	S-MAPE	[106]	MAE, MAPE, MSE
[93–95]	MAE, MAPE, RMSE	[107]	Corr, RMSE, R^2
[96]	MAE, RMSE, N-RMSE	[108]	MSE, RMSE, MAE, MAPE
[97]	Corr, MAE, RMSE, RAE, RRSE	[109]	MAE, Corr, RMSE, MAPE, , and SI

Table 2.1: Some of recent studies of electrical load forecasts and applied evaluation metrics.

Given the issues stemming from point forecasts and point-wise metrics, the literature strongly considers switching from point forecasts to probabilistic forecasts in order to obtain robust and reliable predictions mostly free of phase error and/or the DPE [79]. There has recently been an immense increase in the number of works utilising probabilistic forecasting in various domains, including electric load forecasting. A comprehensive review of probabilistic electric load forecasting for various types of buildings and aggregation levels is provided in [47, 110]. Nevertheless, point forecasting methods and point-wise metrics continue to be popular in the time-series forecasting literature. A number of these works, particularly recently published papers studying electrical load forecasting problems for different types of buildings and aggregation levels, are listed in Table 2.1, together with their applied evaluation metrics. Each of the works listed in Table 2.1 has a different strategy and approach to different variants of the electricity demand forecasting problem. From this table, it is evident that it is a common practice to use multiple evaluation metrics and that the most popular metrics in electrical load forecasting are MAPE and RMSE, which were used in 20 and 19 studies, respectively.

2.3 Summary

This review has given an overview of the background knowledge associated with the concepts and features that are relevant to the rest of this research work. In this chapter, the fundamental

concepts of time-series forecasting have been introduced first, and a brief introduction to time-series forecasting methods and forecast accuracy metrics has been provided. Some of the critical problems that were reported in the relevant literature have then been discussed. The discussion of these issues is considered significant, as they seem to be the most relevant and similar to the bias caused by the PFE that will be introduced in the next chapter and examined in the rest of this thesis.

CHAPTER 3

Persistence Forecast Effect: Systematic Delay in Time-Series Predictions

With the integration of advanced metering and monitoring infrastructures and increased computing capability in recent years, time-series forecasts have begun to play essential roles in a wide range of industrial and scientific areas, including finance, medicine, engineering, science, and a variety of others. Time-series forecasts are now commonly used as a primary criterion for critical decision-making processes or as a core component that can bring intelligence into smart systems or applications. This makes prediction accuracy and robustness imperative. In the electrical power industry, for example, electrical load forecasting is a fundamental and vital task for facilitating the successful implementation of a variety of innovative systems or strategies that next-generation modern electrical power systems, known as smart grids, incorporate. These include smart energy management systems, real-time demand response, demand-side management, and dynamic pricing [111]. Consequently, there is an immense literature on anticipating electrical power consumption accurately for all sorts of buildings and aggregation levels, whose execution can be precisely challenging or, even, unachievable due to various reasons. One of these reasons is an important form of bias that causes a systematic delay, particularly in single-step time-series forecasts in point forecasts, that associate each future time step with a single prediction outcome. When this bias occurs, it can jeopardise prediction validity and robustness. Even though there are numerous published electrical energy demand forecasting studies whose results display some level of systematic and continuous delay in predictions (some of them will be presented in Chapter 8), this phenomenon has yet to be explored, investigated, or commented upon in this domain.

This chapter will first describe the bias, which is labelled the *Persistence Forecast Effect* in this thesis, that causes a systematic and continuous delay in time-series predictions. After that, the

underlying reasons and conditions for the PFE to occur will be investigated. This will be followed by an exploration of its implications, motivating its identification and evaluation, primarily in the context of electricity demand forecasting.

3.1 The Persistence Forecast Effect

Time-series forecasts are defined by both timings and amplitudes of a series of events. Therefore, in time-series predictions, it is critically significant to predict the timing of an event precisely along with its amplitude. However, under some certain circumstances, which will be elaborated on in Section 3.2, it is observed that particularly in single-step forecasts, predictions regress towards one of the most recently observed values and thus predictions are displaced in time. In this case, forecasting methods consistently predict the next value to be similar to one of the current observations used in the input feature set, resulting in a series of predicted values that is nearly identical to one of the most recently observed values. In more formal terms, given Equation (2.1) defining single-step forecasts, prediction output (y_{t+1}) rigorously approximates one of the most recent observations (x_i, $i \in \{t, t-1, ..., t-T\}$) used as an input variable in the input feature set. Ultimately, the main characteristic of affected predictions is the shape of the predictions is remarkably identical to the observed values curve, with the exception that the predicted values are displaced one or a few steps in the future direction of time. In other words, the model returns predictions approximating one of the most recently observed values, and as a consequence, predictions trail actual data systematically and continuously one or a few steps behind in time. The amount of delay is determined according to the past observation data, which the prediction results approximate.

An example of time-series predictions affected by the PFE is given in Figure 3.1, showing observed values of 24-hour electricity consumption of a household along with a series of forecasts that follows the actual consumption values one step behind in time systematically and continuously due to the PFE ($y_{t+1} \approx x_t$). Also shown in Figure 3.1 is the curve of prediction outputs of a persistence model. The persistence model is a well-known naïve method that is mostly used as a baseline method for testing the prediction ability of machine-learning algorithms [74, 92]. It simply returns the value of the most recent observation (x_t) as a forecast outcome for the next time step (y_{t+1}). In other terms, it relates the present and future values via a linear equation as follows;

$$y_{t+1} = x_t \tag{3.1}$$

In the illustrative example depicted in Figure 3.1, the very simple and naïve persistence model has a lower absolute error (6.526 kWh) than that of the advanced machine-learning method (6.723 kWh). This comparative result is particularly important as it shows how the PFE can

impact the robustness and validity of time-series predictions, even when they are produced by one of the most successful time-series prediction methods.

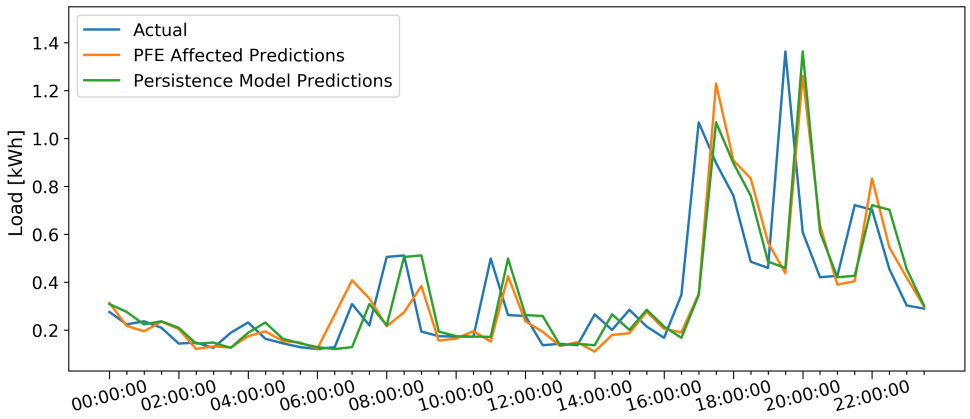


Figure 3.1: A comparison of PFE-affected predictions produced by a machine-learning method and predictions produced by a naïve persistence model throughout a day.

To describe such phenomena in predictions, terms such as *time lag*, *phase lag*, and *delay effect* are used in the literature. However, these terms are inadequate to properly express the continuity of the delay observed in time-series predictions. Hence, based on the striking resemblance between the curve of the predictions of the persistence model and the curve of the predictions suffering from the above-described bias, which can be seen in Figure 3.1, the term *Persistence Forecast Effect* has been created to precisely describe this effect.

3.2 Reasons Behind the Persistence Forecast Effect

At present, in order to solve time-series forecasting problems, including electrical load forecasts, various computationally intensive prediction methods utilising a variety of different input variables on which the output values are assumed to depend are applied. These methods generally exploit underlying correlations and regular patterns in the observed data [111]. Therefore, high volatility in historical data and a lack of regular patterns, often referred to as *uncertainty* and *inconsistency* in time-series data within this thesis, considerably reduce not only the accuracy of time-series forecasts but also their robustness and reliability. For instance, it has been reported many times in the literature that achieving accurate and robust electricity consumption forecasts at the household level is considerably difficult just because of the high volatility and the lack of regularity in load demand [89, 90] that arises from a number of complex factors, such as dweller's habits and lifestyle, family income, cultural background, occupancy, location, weather conditions, and more [81, 91, 93].

Given the importance of the observed data for time-series forecasts, the input feature set usually includes a certain number of most recently observed values from the output domain, see for example [90, 93, 112]. More formally, the most recent observations can be described as

follows: given a time t, x_{t+1} refers to the value to be predicted in single-step forecasts, and x_t denotes the most recently observed value. Hence, the most recent observations can be expressed by x_{t-i}, $(i = 0, 1, \ldots, K)$, where K is the number of data points. These past observations from the output domain are used as input variables in the input feature set with the purpose of helping prediction methods as they infer future values from historical data [112, 113]. However, even though the most recent observations from the output domain can be very powerful predictors of future values, the data irregularity and volatility can make them significant contributing factors to the occurrence of the PFE. The explanation for this is that due to high volatility and pattern irregularity in historical data, prediction methods cannot learn enough about the past and instead extrapolate from one of the most recently observed values used in the input feature set.

When the PFE occurs, the particular past observation that the forecasting method will extrapolate the predictions from is determined by the correlation between the successive observations from the output domain. This is because, owing to the irregular patterns in the historical data, the only thing that prediction methods can do is to learn the superior correlation (similarity) between observations from the output domain. Therefore, the number of time steps of the delay caused by the PFE is determined by the past observations used in the input feature set and the superior correlation between these points. If this is represented with mathematical notations, given that the overall superior correlation is calculated to be between x_{t-q} and x_{t+1} throughout the entire data, and x_{t-q} is one of the input variables used for predicting y_{t+1}, then, when data is irregular and volatile, each predicted y_{t+i}, $i \in N^+$, value is extrapolated from x_{t+i-q}. Hence, each of the prediction outputs approximate x_{t+i-q}, resulting in a time-series of predictions that is temporally displaced (delayed) by $q+1$ steps in time. Predictions thus systematically follow the observed values $q+1$ time steps behind. For instance, by the nature of electricity consumption, the supreme correlation is predominantly between two consecutive data points (x_t and x_{t+1}) throughout the data. The prediction (y_{t+1}) is, therefore, almost always extrapolated from the most recently observed value (x_t) when dealing with volatile and inconsistent electricity consumption data. Consequently, each electrical load prediction output approximates the most recent observation (x_t), and that is why the predictions trail the actual data one step behind in time ($y_{t+1} \approx x_t$), as shown in Figure 3.1.

To sum up, the occurrence of the PFE is determined by the characteristics of the underlying data. It is mainly triggered by the uncertainty and irregularity in data, and the number of time steps of delay is determined by the past observations used in the input feature set and the superior correlation between these points. Conversely, if the data has strong enough regularity in patterns or if it has a set of feature inputs explaining the irregularity in data, the PFE is quite unlikely to take place.

It is of note once again that in this thesis, the PFE is mostly investigated in the context of electricity demand forecasts. Given that the superior correlation in electricity consumption

time-series data is mostly seen between consecutively recorded values, this research will usually examine and address predictions delayed only one time step when they are affected by the PFE. However, instances of larger PFE delays from different contexts will also be analysed in Chapter 7 (Section 7.6).

3.3 Implications of the Persistence Forecast Effect

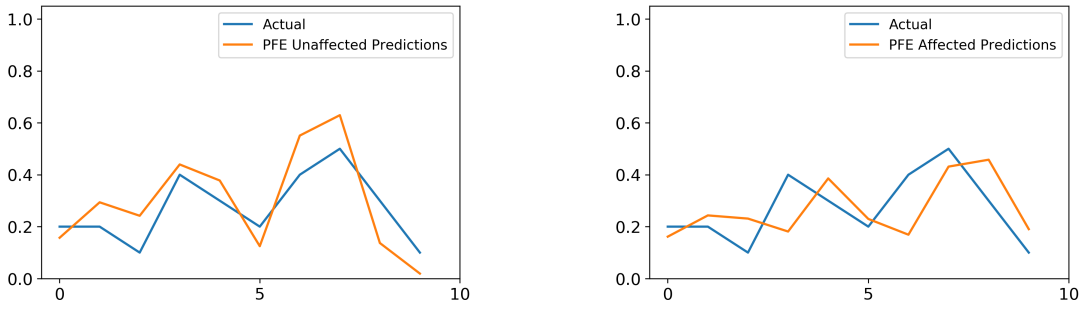
One of the primary objectives of time-series forecasts is to be able to predict the future values for a given time-series. However, the PFE assumes that the future will be *similar to one of the currents* in the general sense, which ultimately ends up with a systematic temporal displacement of predictions. Given the increased computing capability along with the amount of effort, money, and time invested in producing accurate and robust time-series predictions, time-series prediction methods, these days, are expected to be able to say more than *the future value is going to be strictly similar to one of the current observations*. Even more importantly, such predictions affected by the PFE and so delayed in time may also have major detrimental effects that risk the reliability of final applications and studies.

The impact of the temporal displacement of predictions caused by the PFE on applications varies for different applications depending on their tolerance to accommodate temporal displacement of predictions. Applications that do not strictly require temporal accuracy can easily tolerate a few steps of delay in predictions, but the functionality of smart applications that require precise timing and temporal accuracy can be jeopardised by the PFE. Peak shaving, for instance, can be given as an example from the electrical power context. In this context, peak shaving refers to reducing power consumption for a short period of time to eliminate demand spikes in electricity consumption that threaten network stability and the cost of energy [114, 115]. Peak shaving is, therefore, critically important for network security and resilience, as well as the economy. Proposals for reducing the amplitude of peaks in consumption to have a smooth and stable demand curve include use of energy storage devices (batteries) and load shifting [116]. The ideas behind these strategies are quite simple: use from batteries at peak times and recharge them at off-peak times, or shift some of the tasks forming the peak demand to off-peak times. Hence, these strategies require precise forecasts while allowing no temporal flexibility. However, in the presence of the PFE, forecasts naïvely reproduce the current load, ending up with the predictions of peaks and troughs are displaced in time. Such temporally displaced predictions may cause batteries to be charged/discharged at a suboptimal time or tasks to be incorrectly scheduled, potentially resulting in exacerbated peaks rather than shaving them. Eventually, similar to the dramatic effect known as the *avalanche effect* in demand-response terminology [117, 118], exacerbated peaks are highly likely to have noticeable negative consequences, such as causing network infrastructure breakdowns, grid congestion, or necessitating more power production to accommodate the extra demand at peak times. The authors of [77] point out that the delay

in predictions of peaks may also result in uncharged batteries prior to the actual time of peaks, ultimately ending up with unsuccessful peak shaving.

The PFE is a risk to all prediction contexts that exhibit volatile and uncertain load patterns, including small grids (also called *weak grids*) [115]. Such grids supply electricity to a small group of customers, and they generally serve areas with a limited number of consumers, such as islands or wild and remote places. This type of grid is consequently known to have distorted and unstable loads. Therefore, the PFE is quite likely to be observed in the predictions for such grids. This could, in turn, negatively influence energy suppliers' decision-making for strategies with limited temporal flexibility, such as dynamic pricing, tariff adjustment, supply-demand balancing, and network-level peak reduction.

Furthermore, unfortunately, many commonly-used point-wise evaluation metrics, such as MSE, RMSE, MAE, and MAPE, described in Chapter 2 (Subsection 2.1.4), are often fail to identify the PFE in predictions. The point-wise evaluation metrics quantify the prediction error from the discrepancy between the predicted and observed values at each time step in the form of numeric values (*metrics*). Therefore, even though the predictions are biased by the PFE and follow the observed data one or a few steps behind in time, most of these popular point-wise evaluation metrics cannot capture the delay in predictions and may mistakenly suggest that the accuracy of predictions is considerably good. This might result in deceptive evaluation metric results and overconfidence in predictions and prediction models. For instance, Figure 3.2 illustrates two different hypothetical forecasts for synthetically created data. One of these two sets of predictions is not impacted by the PFE (Figure 3.2a), while the other is and follows the actual data one step behind in time due to the PFE (Figure 3.2b). Nevertheless, the evaluation metric results of these predictions, listed in Table 3.1, are virtually identical, which suggests the accuracies of these two predictions are almost equivalent. Such deceptive metric results might cause misplaced confidence not only in predictions affected by the PFE but also in the forecast model producing such predictions. In the context of electrical power, for example, smart



(a) (b)

Figure 3.2: Two hypothetical forecasts produced for a synthetically created data: (a) predictions do not suffer from the PFE ($y_{t+1} \approx x_{t+1}$), and (b) p redictions exhibit the PFE ($y_{t+1} \approx x_t$).

Scenario	RMSE	MAPE
Predictions without the PFE (Figure 3.2a)	0.117	48.232
Predictions with the PFE (Figure 3.2b)	0.119	48.264

Table 3.1: MAPE and RMSE results of two hypothetical forecasts produced for a synthetically created data.

systems or applications (such as dynamic pricing, peak shaving, and consumption scheduling) built on top of such temporally delayed forecasts simply because of the deceptive evaluation metric results suggesting the model can produce accurate time-series predictions could lead to severe troubles for smart grid concepts and result in substantial losses for system operators and energy consumers. According to [119], a 1% rise in forecast error translated to a roughly £10 million increase in annual operating costs for the United Kingdom in 1984. Given the increased level of automation in all parts of the economy, building automated and smart systems on top of predictions missing the temporal perspective owing to an unnoticed PFE will most certainly cost much higher today.

The accuracy measured by evaluation metrics is not always the overriding factor in comparing alternative time-series prediction models, but it is usually the most compelling criterion that can be considered [16]. However, in addition to being a sign of poor time-series prediction generally, the PFE is also a direct threat to the validity of comparative studies that rank methods in-between. This is due to the fact that the PFE often creates inconsistent rankings or even rank reversal of prediction methods. This, in turn, undermines the transferability of findings from studies aiming to optimise forecasting methods or to compare the performance of forecasting methods for a certain type of problem. Figure 3.3, for instance, visualises synthetically constructed datasets (Dataset A and B) and predictions generated by two hypothetical prediction methods (Method A and B), whose evaluation metrics are compared in Table 3.2. This example also has two assumptions: i) Method A has already been scientifically concluded to be better than Method B at learning time-series data and producing accurate time-series predictions; and ii) Dataset A provides regular time-series data, whereas the data irregularity and volatility are significantly higher in Dataset B. In consideration of these assumptions, therefore, Method A is expected to outperform Method B in terms of the accuracy of predictions. For Dataset A, consisting of regular patterns, Method A follows expectations and produces much better predictions compared to those produced by Method B. The difference between the temporal accuracy of predictions produced by Methods A and B is recognisable in Figures 3.3a and 3.3b and can also be seen in the evaluation metric results shown in Table 3.2. The irregular and volatile patterns within Dataset B, however, give rise to the PFE, and prediction outputs trail the actual data behind in time similarly for both datasets (Figures 3.3c and 3.3d). As a result, the evaluation metrics of the Dataset B predictions are calculated to be almost equal to one another with an indiscernibly small difference as a

matter of course (Table 3.2). The results also show that the PFE removes the superior learning capability of Method A. Consequently, even though it is already known that Method A is better than Method B at learning time-series data and yielding accurate forecasts, this can no longer be identified from the metric outcomes when the PFE occurs.

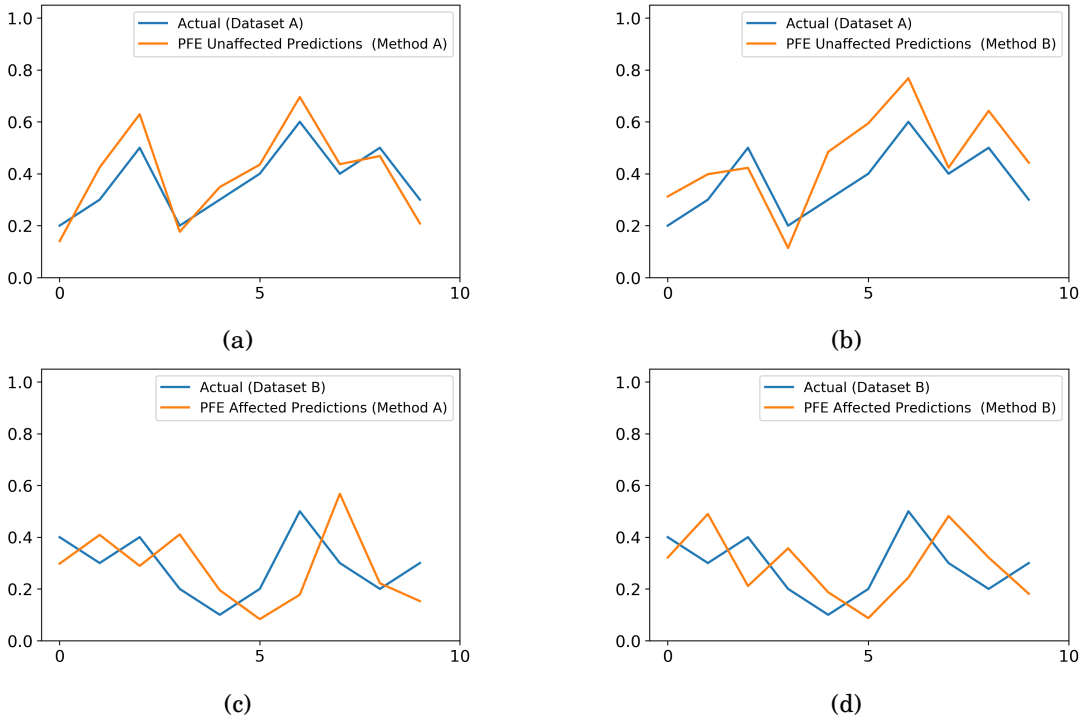


Figure 3.3: Forecasts of two hypothetical prediction methods (Method A and B) for two synthetically created datasets (Dataset A and B). Predictions produced for Dataset A do not exhibit the PFE ($y_{t+1} \approx x_{t+1}$) while the predictions produced for Dataset B are biased by the PFE ($y_{t+1} \approx x_t$).

Dataset	**Pred. Method**	**RMSE**	**MAPE**
Dataset A	Method A (Figure 3.3a)	0.077	19.567
	Method B (Figure 3.3b)	0.132	36.726
Dataset B	Method A (Figure 3.3c)	0.167	56.223
	Method B (Figure 3.3d)	0.164	56.294

Table 3.2: MAPE and RMSE results for hypothetical forecasts produced by hypothetical prediction methods (Method A and B) for synthetically created datasets (Dataset A and B).

Considering the aforementioned potential problems that the PFE may cause, which will also be studied and discussed comprehensively in Chapters 5 and 6 through observational electrical energy demand data and contemporary prediction methods, it is vital that those using time-series forecasts should be aware of whether their predictions are influenced by the PFE before proceeding with their final applications or decision-making.

3.4 Summary

This chapter has provided a formal description of an important form of bias that causes a systematic and continuous delay in single-step time-series forecasts in particular. This bias is called the *Persistence Forecast Effect*, which is inspired by its resemblance to a naïve persistence model. In addition, this chapter has presented the underlying reasons for the phenomenon of the PFE. It has been shown that the presence or absence of the PFE is determined by the characteristics of the underlying data and that volatility and irregularity in data are the major drivers of the PFE. The PFE manifests itself when prediction methods cannot find regular patterns in historical observations to learn due to strong volatility in data, so they learn the superior correlation between past observations from the output domain instead. As a result, predictions approximate one of the most recently observed values used in the input feature set, resulting in time-series predictions delayed one or a few time steps.

It is also worth highlighting that discontinuous temporal displacements of predictions occurring at any point in time at random for all sorts of reasons with varying time steps backwards or forwards should not be construed as the PFE. As a matter of fact, the PFE refers to predictions that systematically trail the actual data one or a few steps behind in time since they are extrapolated from one of the most recent observations used in the input feature set owing to the volatile and inconsistent patterns in historical data.

Finally, this chapter has also evaluated the potential problems of the systematic temporal displacement of predictions introduced by the PFE. An unnoticed PFE in predictions can negatively affect the decision-making of experts or smart systems, as well as intelligent applications that do not have the flexibility to manage temporally displaced predictions. Aside from that, most of the popular evaluation metrics are not always able to capture the PFE, it might, therefore, cause misleading metric results, paving the way for misplaced confidence in predictions. Consequently, it has a direct impact on comparative studies aiming at ranking the methods based on their performance for a certain type of problem and threatens the transferability of studies. Thus, it is recommended that stakeholders take this bias into consideration in order to prevent undermining the robustness and validity of methodological decisions made and to avoid its negative effects on the operation of final applications and systems.

The next chapter will continue with the introduction of a novel shifting method that is proposed for detecting the presence of the PFE in time-series prediction outputs.

CHAPTER 4

THE N-STEP-SHIFTING METHOD TO DETECT THE PERSISTENCE FORECAST EFFECT

The PFE has been defined in the preceding chapter as the bias resulting in a systematic delay in time-series forecasts, and its causes and implications have been explored. However, its detection has yet to be discussed. This chapter, therefore, will now discuss how the PFE can be detected and also introduce a new generic method that is proposed for the detection of the PFE in time-series forecasts.

4.1 Introduction

Despite the challenges and problems discussed in Chapter 2, many time-series forecasting research still mainly focus on point forecasts and utilise point-wise metrics [47, 110]. Given the issues that the PFE may cause, discussed in Chapter 3 (Section 3.3), it is of great importance to determine whether or not the predictions are influenced by the PFE before taking any action, such as making critical decisions or developing smart applications that rely on such time-series predictions with delays.

A time plot, which plots prediction outputs and actual measurements against time, is a very efficient tool for analysing results and spotting anomalies, as discussed in [16]. Similarly, the simplest way of detecting the PFE can be a visual inspection of a time plot, which is probably the first approach that comes to mind. If time-series predictions are affected by the PFE, one should see in time plot that the predictions curve is nearly identical to the actual measurements curve but is systematically and continuously delayed by one or a few steps in time, as exemplified in Figure 3.1. However, although the temporal delay caused by the PFE is explicit in the example presented in Figure 3.1, the delays caused by the PFE may not always be so easily identifiable

in plots, depending on various factors such as the data type, length, granularity, domain, plot size, aspect ratio, resolution, graph type, and so on. For instance, the time-series predictions demonstrated in Figure 4.1 are actually biased by the PFE and trail the actual data by one time step delay. The delay in predictions, however, is not clearly visible upon visual inspection, and instead, they deceptively appear to be perfectly aligned with the actual values. The reason for this is not related to the nature or characteristics of the PFE, but rather the scale of the plot being inadequate to clearly display a one-step delay in the predictions. Thereby, as it has been seen in this example, visual inspection alone is not always sufficient for practical PFE detection, in spite of the fact that it is still an essential supplementary tool. Furthermore, as discussed in the previous chapter (Section 3.3), most of the well-known and widely used standard evaluation metrics are unable to identify the presence of the PFE. Therefore, a computational approach facilitating a mechanism for the detection of the PFE in time-series predictions is required. Given that, to achieve quantitative detection of the bias, a novel *n-Step-Shifting* (n-SS) method, drawing inspiration from the simple technique employed in [62], is presented below in this chapter.

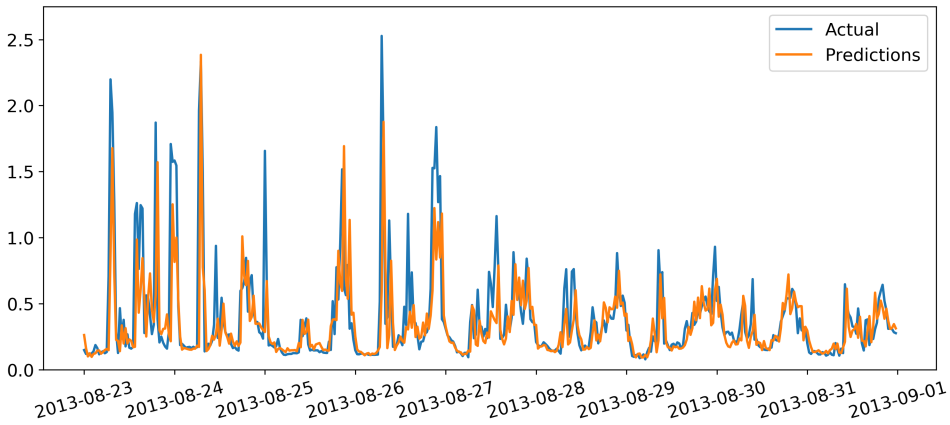


Figure 4.1: An example of the PFE being visually unrecognisable.

This chapter will start with a formal definition of the proposed n-SS method. It will then reformulate the n-SS method specifically for one step delays in predictions, which is the most usual PFE in the electrical energy demand forecasting domain (instances of PFEs causing delays more than one time step will be offered in Section 7.6). Finally, the weaknesses of the proposed method will be discussed.

4.2 The n-Step-Shifting Method

As it has been explained in Chapter 3, the PFE occurs mainly as a result of a lack of regular patterns in the underlying data, and a strong correlation between the sequential observations determines the amount of the delay. When the effect arises, prediction outputs approximate one of the values recorded most recently, which are also used as input variables in the input feature set, and as a consequence, they trail the actual values one or a few steps behind in time.

Accordingly, to detect the delay in predictions caused by the PFE, this thesis proposes the n-SS method. This method suggests the recalculation of standard evaluation metrics after shifting the predictions a few steps back in time (shift predictions to the past) or after shifting the time-series of actual observations a few steps forward in time (shift observations right to the future). It should be noted that in this thesis, all the visualisations and formulations utilise the former strategy, which shifts the predictions a few steps back (*n steps back*) in time to the past. The main idea behind the n-SS method is to show that shifting the predictions back in time yields considerably better evaluation metric results, which approves the systematic and continuous temporal delay in predictions by reason of the PFE.

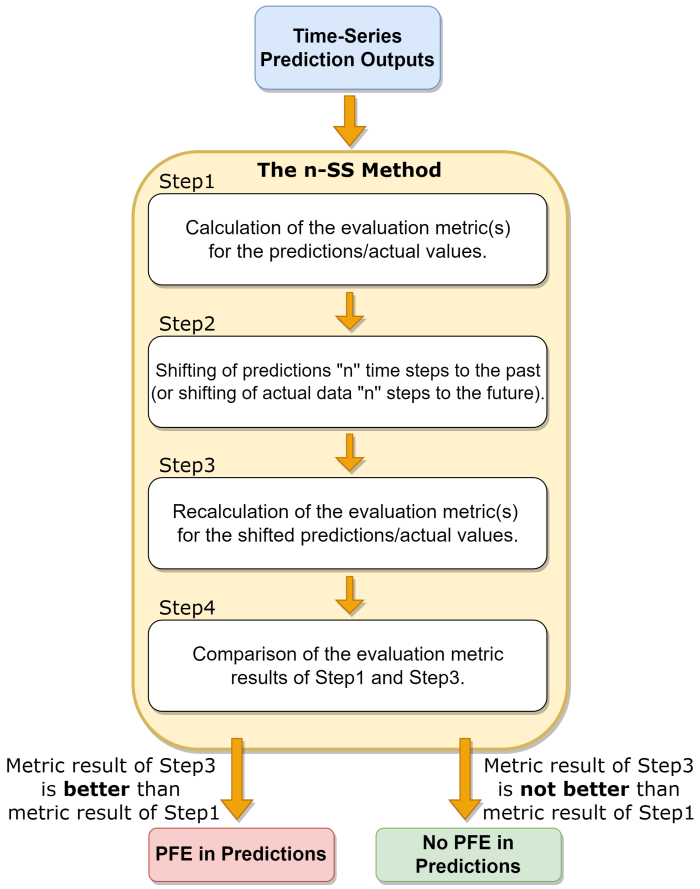


Figure 4.2: Overall flow chart of the proposed n-SS method.

The proposed n-SS method, whose flow chart is also illustrated in Figure 4.2, contains four steps:

Step 1: Calculate evaluation metric(s) for predictions/actual values as usual.

Step 2: Apply shift of predictions 'n' time steps to the past (or shift actual data 'n' steps to the future).

Step 3: Recalculate evaluation metric(s) for shifted predictions/actual values.

Step 4: Compare the evaluation metric results of Step 1 and Step 3. If the n-SS method results in considerable improvements in accuracy (metric result), then it can be claimed that the predictions exhibit the PFE. This can be summarised as in Equation (4.1).

(4.1)

$$\text{n-SS} = \begin{cases} \text{PFE-affected}, & \text{if metric result of Step 3 is } \textbf{better} \text{ than metric result of Step 1.} \\ \text{PFE-free}, & \text{if metric result of Step 3 is } \textbf{not better} \text{ than metric result of Step 1.} \end{cases}$$

This four-step n-SS method is independent of a specific evaluation metric and, in principle, is compatible with various widely recognised and extensively utilised point-wise evaluation metrics, some of which have been outlined in Chapter 2 (Subsection 2.1.4). What is actually meant by *shifting the predictions a few steps back in time* in a formal sense is that it assesses the prediction error from a discrepancy between x_t and y_{t+n} rather than a discrepancy between x_t and y_t. The n-SS method, therefore, modifies the well-known point-wise evaluation metrics, previously defined by Equations (2.4) to (2.12), as follows:

(4.2) $$MSE^* = 1/m \sum_{t=1}^{m} (x_t - y_{t+n})^2$$

(4.3) $$RMSE^* = \sqrt{1/m \sum_{t=1}^{m} (x_t - y_{t+n})^2}$$

(4.4) $$MAE^* = 1/m \sum_{t=1}^{m} (|x_t - y_{t+n}|)$$

(4.5) $$MAPE^* = \left[1/m \sum_{t=1}^{m} (|x_t - y_{t+n}|)/x_t \right] \times 100$$

(4.6) $$Corr^* = \frac{\sum_{t=1}^{m} (x_t - \overline{x})(y_{t+n} - \overline{y})}{\sqrt{\left[\sum_{t=1}^{m} (x_t - \overline{x})^2 \right] \left[\sum_{t=1}^{m} (y_{t+n} - \overline{y})^2 \right]}}$$

$$RAE^* = \frac{\sum_{t=1}^{m}(|x_t - y_{t+n}|)}{\sum_{t=1}^{m}(|x_t - \overline{x}|)} \tag{4.7}$$

$$RRSE^* = \sqrt{\frac{\sum_{t=1}^{m}(x_t - y_{t+n})^2}{\sum_{t=1}^{m}(x_t - \overline{x})^2}} \tag{4.8}$$

$$R^{2^*} = \frac{\sum_{t=1}^{m}(x_t - y_{t+n})^2}{\sum_{t=1}^{m}(x_t - \overline{x})^2} \tag{4.9}$$

$$CV^* = \frac{\sqrt{1/m \sum_{t=1}^{m}(x_t - y_{t+n})^2}}{\overline{x}} \tag{4.10}$$

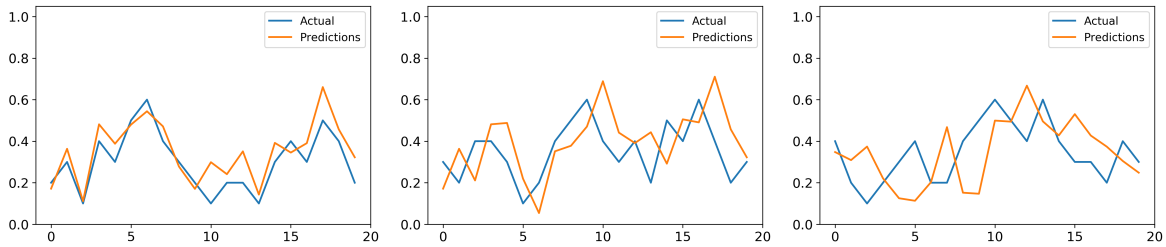
where n is the number of time steps by which the predictions are shifted to the past. The metrics modified by the n-SS method are denoted by (*). For example, MAE* is MAE modified by the n-SS method.

The n-SS method does not actually shift the predictions to the past but simply adjusts evaluation metric formulas as shown. However, the term *shift* and discussing visualising predictions *shifted* in plots are helpful when describing the working principle of the n-SS method and the notions behind it.

When the application of the n-SS method does not change the metric result at all, which is very unlikely, or results in only a negligibly minimal increase or decrease in the metric result, the decision on the existence of the PFE may become unclear or subject to doubt. In these cases, implementing the n-SS method with one or more additional evaluation metrics that assess the predictions from a different perspective can help to strengthen and confirm the presence or absence of the PFE. This will be exemplified in the next chapter.

4.2.1 A Worked Example of the n-Step-Shifting Method Implementation

In order to illustrate how the proposed four-step n-SS method is supposed to be implemented rigorously, a worked example demonstrating all the steps of the n-SS method through hypothetical predictions is presented below for three different synthetic scenarios.



(a) (b) (c)

Figure 4.3: Three hypothetical forecasts produced for three different synthetically created scenarios: (a) Scenario 1, (b) Scenario 2, (c) Scenario 3.

Each of these three synthetically built scenarios with hypothetical predictions, visualised in Figure 4.3, has a different PFE status; predictions for Scenario 1 (Figure 4.3a) are not affected by the PFE, predictions for Scenario 2 (Figure 4.3b) follow the actual data one step behind in time due to the PFE, and predictions for Scenario 3 (Figure 4.3b) are also affected by the PFE, but they are delayed by two time steps. The purpose of this worked example is to show the practical application of the n-SS method for computationally identifying the declared PFE statuses.

Step 1: *Calculate evaluation metric(s) for predictions / actual values as usual.*

It is of note once again that the n-SS method can be applied with any preferred evaluation metric. However, as it is impractical to go through every evaluation metric available in the literature, three of them are arbitrarily chosen for use with the worked example here. These metrics are MSE, MAE, and RRSE, which are defined by Equations (2.4), (2.6), and (2.10), respectively.

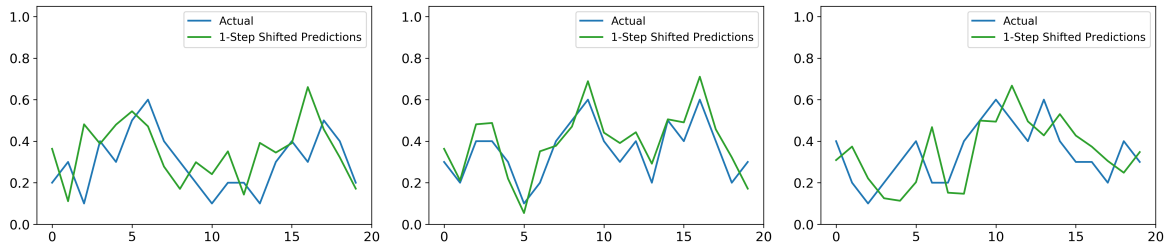
The results of evaluation metrics of the three independent scenarios are listed in Table 4.1.

Scenario	**MSE**	**MAE**	**RRSE**
Scenario 1 (Figure 4.3a)	0.007	0.074	0.645
Scenario 2 (Figure 4.3b)	0.029	0.150	1.290
Scenario 3 (Figure 4.3c)	0.033	0.148	1.341

Table 4.1: MSE, MAE, and RRSE results of three hypothetical forecasts produced for three synthetic scenarios built for the worked example.

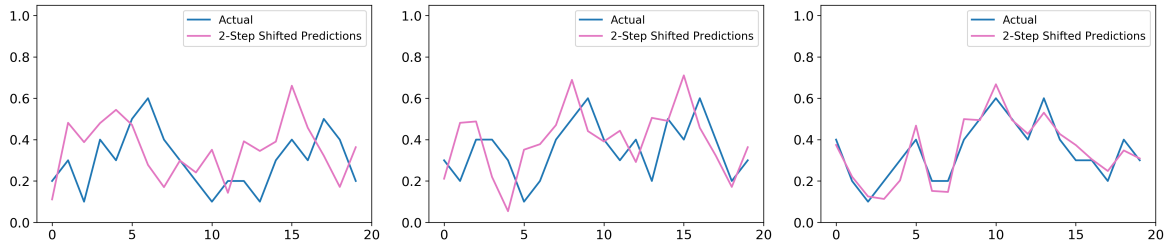
Step 2: *Apply 'n' time steps shift of predictions to the past.*

In the second step of this worked example, two different values of n ($n = 1$ and $n = 2$) will be considered. That is, the hypothetical predictions will be shifted one step back in time ($n = 1$) and then shifted one more step to the past ($n = 2$). The objective of applying the n-SS method with these two different n values is to demonstrate how the n-SS method changes the predictions



(a) (b) (c)

Figure 4.4: One step shifted ($n = 1$) hypothetical forecasts produced for the three independent, synthetically created scenarios illustrated in Figure 4.3: (a) Scenario 1, (b) Scenario 2, and (c) Scenario 3.



(a) (b) (c)

Figure 4.5: Two steps shifted ($n = 2$) hypothetical forecasts produced for the three independent, synthetically created scenarios illustrated in Figure 4.3: (a) Scenario 1, (b) Scenario 2, and (c) Scenario 3.

and evaluation metrics when the PFE status varies with different n values. The one step shifted and two steps shifted versions of the hypothetical forecasts are visualised in Figures 4.4 and 4.5, respectively.

Step 3: *Recalculate the evaluation metric(s) for the shifted predictions.*

After shifting the predictions by n time steps (here: $n = 1$ and $n = 2$) to the past, the three arbitrarily chosen evaluation metrics are recalculated, but for the shifted predictions this time. The results of these evaluation metrics for the predictions that shifted by one step and two steps are displayed in Tables 4.2 and 4.3, respectively.

Scenario	**MSE***	**MAE***	**RRSE***
Scenario 1 (Figure 4.4a)	0.028	0.132	1.226
Scenario 2 (Figure 4.4b)	0.006	0.072	0.620
Scenario 3 (Figure 4.4c)	0.021	0.129	1.070

Table 4.2: MSE*, MAE*, and RRSE* results of one step shifted ($n = 1$) hypothetical forecasts produced for the three synthetic scenarios created for the worked example.

Scenario	**MSE***	**MAE***	**RRSE***
Scenario 1 (Figure 4.5a)	0.036	0.166	1.379
Scenario 2 (Figure 4.5b)	0.030	0.146	1.311
Scenario 3 (Figure 4.5c)	0.002	0.045	0.402

Table 4.3: MSE*, MAE*, and RRSE* results of two steps shifted ($n = 2$) hypothetical forecasts produced for the three synthetic scenarios created for the worked example.

Step 4: *Compare the evaluation metric results of Step 1 and Step 3.*

The final step of the provided worked example is to compare the evaluation metric results obtained in Step 1 and Step 3 to identify whether shifting the predictions brings any improvement in accuracy, which would indicate the PFE. The evaluation metric results are shown and visually compared with each other in Figure 4.6. The results of these comparisons for each scenario can be described as follows:

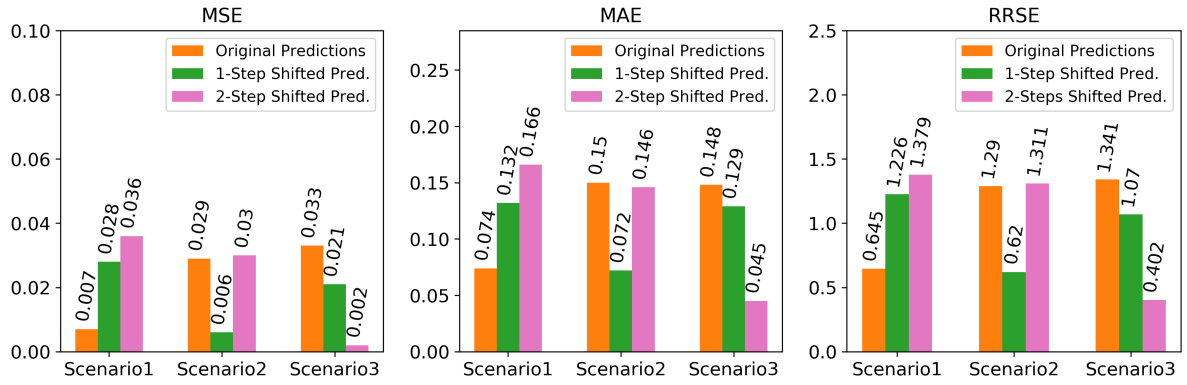
- *Scenario 1:* The n-SS method shows that the hypothetical predictions of this scenario are not affected by the PFE, as values of MSE, MAE, and RRSE increase, meaning greater error, with one-step shifting from 0.007, 0.074, and 0.645 to 0.028, 0.132, and 1.226, respectively, and with two-step shifting to 0.036, 0.166, and 1.379, respectively. Overall, the best evaluation metric values belong to the initial hypothetical forecasts.

- *Scenario 2:* For this scenario, one step shifted predictions yield by far the best accuracy, with metric values of MSE=0.006, MAE=0.072, and RRSE=0.620. Therefore, the n-SS method shows that the predictions of this scenario suffer from the PFE, and the predictions follow the actual load one time step behind due to the PFE.

- *Scenario 3:* The two steps shifted predictions yield the best accuracy for this scenario, which shows the PFE is present and causing predictions to be delayed by two steps. After shifting the predictions by two steps to the past, the metric values of MSE, MAE, and RRSE improve to 0.002, 0.045, and 0.402, respectively, from 0.033, 0.148, and 1.341.

To summarise, it can be deduced from the n-SS method outcomes that there is no PFE in predictions in Scenario 1 ($y_t \approx x_t$), whereas predictions in both Scenario 2 ($y_{t+1} \approx x_t$) and Scenario 3 ($y_{t+2} \approx x_t$) suffer from the PFE, which confirms the in-built PFE status of each scenario.

It is also interesting to determine what value(s) of n should be considered. The choice of n is largely subjective and data-specific. Considering that the number of time steps of the delay caused by the PFE is determined by the correlation between the successive observations from the output domain used in the input feature set, auto-correlation analysis can serve as a reliable tool, together with visual inspection where possible, to ensure the accurate selection of the value of n.



(a) (b) (c)

Figure 4.6: Metric results of original predictions and n-SS applied predictions with $n = 1$ and $n = 2$ using MSE, MAE, and RRSE metrics.

Auto-correlation calculates the degree of similarity between a time-series variable and a lagged version of itself across successive time intervals [17, 24]. In other words, it measures how the value observed at time t is related to the other values recorded at earlier points in time from the same time-series variable. If the strongest auto-correlation of the time-series data is found to be between x_t and x_{t-e} (represented by auto-correlation value at lag e), then the value of e must be assigned to n ($n = e$). Another way of determining the value of n could be to use a brute-force approach. This approach refers to the repetition of the n-SS method implementation for each time point of each historically observed value used in the input set. For instance, if the currently observed values ranging from x_t to x_{t-f} are used as input variables in the input feature set to predict y_{t+1}, then the n-SS method should be applied for each value of n from $n = 1$ to $n = f + 1$ independently to explore whether any of these values brings a considerable improvement in metric results, indicating the existence of the PFE.

4.2.2 The 1-Step-Shifting Method

As previously stated, the main focus of the present research is the PFE in the domain of electricity demand forecasting, in which superior auto-correlation tends to be between consecutive data points, as is often observed in time-series data. This means that any temporal delay in the domain of electricity demand forecasting arising from the PFE overwhelmingly displays itself as a systematic one time step delay. As a result, the forecasting results approximate the value of the previous data point and hence follow the actual load values one step behind in time. Most of this thesis, therefore, will deploy *1-Step-Shifting* (1-SS), where $n = 1$. However, various values of n (where $n > 1$) as part of the n-SS method will be considered in later chapters (Section 7.6) with some data gathered from diverse domains.

4.3 Limitations of the n-SS Method

The proposed four-step n-SS method facilitates a conceptually straightforward mechanism for detecting the PFE in time-series forecasts. It is a generic method that is compatible with many of the evaluation metrics commonly used in the time-series literature, expediting and easing its deployment. Nevertheless, it does have a few constraints that restrict its applicability. These limitations will be explored and discussed in this section. Each of these limitations will then be exemplified with the experimental results in the upcoming chapters.

First, the proposed n-SS method necessitates that a time-series forecast has already been performed in advance, implying that it does not prevent investing time, resources, and effort in making a prediction. As the n-SS method relies on how shifting the predictions by a few steps to the past affects the accuracy metric results, it is unable to detect the potential occurrence of the PFE before the time-series forecast is carried out. Therefore, it can only be used as a control mechanism once the time-series predictions are produced in order to make sure the model and the prediction results are robust and reliable before proceeding with any further actions that rely on time-series predictions.

Additionally, it is a common practice (see Table 2.1) and is recommended by many to deploy multiple accuracy metrics [94] as part of a prediction evaluation, as each evaluation metric has individual advantages and disadvantages and evaluates the accuracy of predictions from different perspectives. However, although rare, the use of multiple metrics might result in conflict between metrics regarding the existence of the PFE since each metric has different aspects. In other words, while one metric might indicate there is a PFE in predictions due to improved metric values after the n-SS method is applied, another metric could reflect that predictions do not exhibit the PFE owing to increased metric result with the n-SS method deployment. This results in the n-SS method producing an inconclusive PFE investigation and requires further scrutiny.

The proper functioning of the n-SS method requires a certain level of variance in the time-series curve, which can be measured by the standard deviation. The standard deviation indicates how much the data deviates from the mean of the series. If the standard deviation of either a series of observed values or a series of predicted values is equal to or close to zero, the n-SS method does not work as intended. This is because the n-SS method relies on how the metric results change when predictions are shifted in time, but a flat curve, and near-zero standard deviation, will mean that metrics will remain approximately constant as the predicted values are shifted. This means that the n-SS method will be unable to yield significant changes to metric values and will therefore be unlikely to indicate the existence of the PFE properly. Figure 4.7, for example, represents three synthetically created scenarios with increasing levels of fluctuations in observed values and hypothetical forecasts. Table 4.4 shows the standard deviation values of these series and the accuracy metric results before (MAPE and RMSE) and after (MAPE* and RMSE*) the n-SS method is applied to each of the hypothetical forecast series. The hypothetical predictions produced for each of these scenarios exhibit PFE and follow the actual data one time

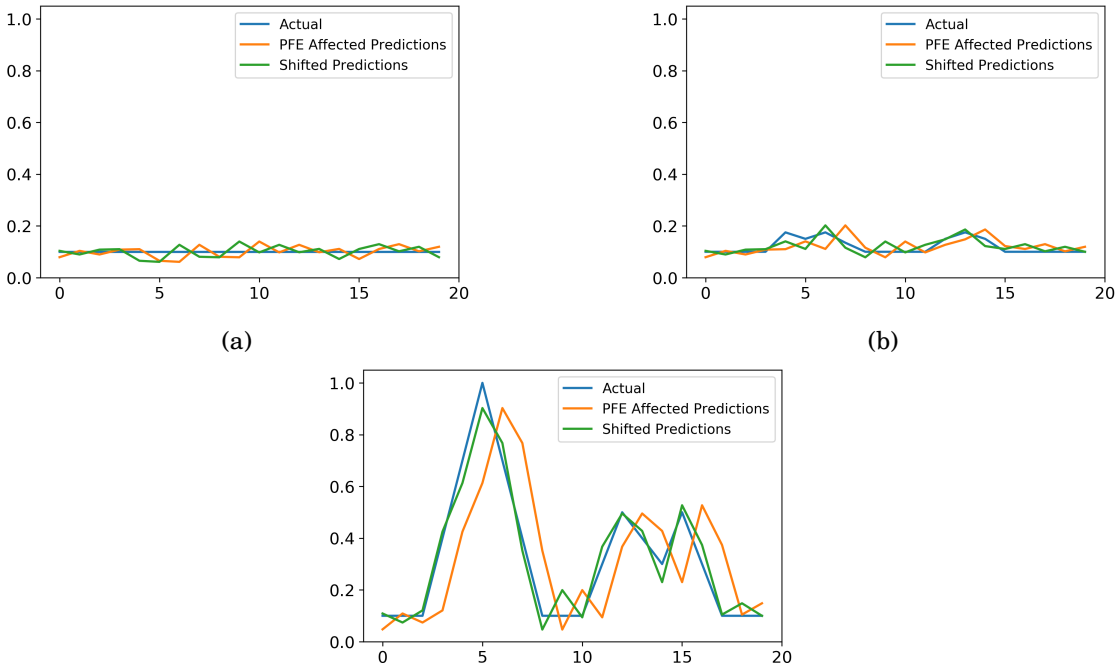


Figure 4.7: Three hypothetical forecasts applied to synthetically created scenarios of increasing data variance: (a) Scenario1, (b) Scenario2, and (c) Scenario3.

Scenario	**Std. Dev.**	**RMSE vs. RMSE***	**MAPE vs. MAPE***
Scenario 1	0.000	0.021 ↔ 0.021	18.289 ↔ 18.289
Scenario 2	0.030	0.031 ↓ 0.021	19.515 ↓ 14.283
Scenario 3	0.258	0.205 ↓ 0.053	68.885 ↓ 20.019

Table 4.4: Comparison of changes in the evaluation metric outcomes with the application of the n-SS method ($n = 1$) for three different synthetic scenarios with different levels of fluctuations (standard deviations) in observed values (Figure 4.7)

step behind. Therefore, the 1-SS method has been applied here, which also depicted in Figure 4.7. The observed values of Scenario 1 are constant, and their standard deviation is 0.000, so the 1-SS method makes no change in metric results and cannot capture the existence of the PFE, even though it is known that it is present in the hypothetical predictions. Changes in metric results produced by the 1-SS method for Scenario 2 and Scenario 3 increase in direct proportion to the standard deviation values of the scenarios, which are 0.030 and 0.258, respectively. To sum up, the provided synthetic example confirms that the n-SS method requires a certain level of variance (peaks and troughs) in both actual data and predictions to perform effectively.

Furthermore, in some particular cases, the n-SS method might suggest that the entire time-series predictions is biased by the PFE despite only a part of the series being affected by the bias, or vice versa. For instance, the PFE is not present in the first half (the first ten time points) of the

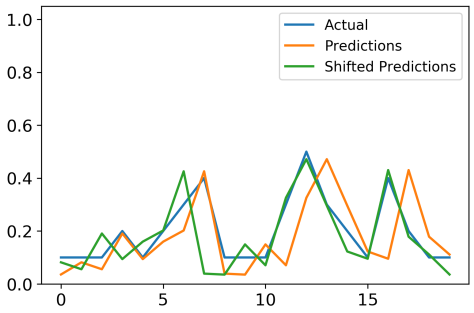


Figure 4.8: Synthetic time-series data along with hypothetical forecasts whose first half is biased by the PFE whilst the second half is not affected by the PFE.

Data	**RMSE vs. RMSE***	**MAPE vs. MAPE***
First Half (without the PFE)	0.051 ↑ 0.133	32.313 ↑ 51.369
Second Half (with the PFE)	0.166 ↓ 0.037	56.937 ↓ 18.232
Whole Data	0.123 ↓ 0.098	44.625 ↓ 34.800

Table 4.5: Comparison of the evaluation metric results of original predictions and shifted predictions illustrated in Figure 4.8.

hypothetical predictions illustrated in Figure 4.8 but is present in the second half (the last ten time points) of the hypothetical predictions. This is seen in the predictions for early time points being aligned with the synthetic time-series data but a delay appearing later. However, when the n-SS method ($n = 1$) is applied to the whole series, the change in the accuracy metric results (shown in Table 4.5) suggest that all the predictions are affected by the PFE. This is principally because, in this case, the improvement in metric results brought about by the delay in the second half is greater than the degradation in metric results seen from predictions of the first half of the data series, as shown in Table 4.5. The improvement in the overall metric results when the n-SS method is applied is directly related to the number of peaks and their amplitudes in the parts with and without the PFE, and a different pattern in the actual data could reverse the conclusion, leading to no PFE being indicated. The reasons for such scenarios along with how to deal with them when they occur will be discussed later in Chapter 7.

Finally, as the n-SS method is proposed to detect a systematic and continuous temporal delay in predictions, if the predictions are uncorrelated to the actual data, the proposed method might not work properly. In such cases where there is no similarity or correlation between the actual data and prediction outputs, it does not appear conceivable to talk about either the temporal accuracy or inaccuracy or the PFE. For instance, Figure 4.9 represents a synthetically created time-series data along with utterly uncorrelated predictions and one step shifted predictions. In this fictitious example, applying the n-SS method would end up with either an increase or decrease in metric results at random. However, this would not indicate the existence or absence of the PFE, as the predictions curve and observations curve are completely dissimilar to each other.

The n-SS method therefore requires a certain degree of similarity between the predicted and observed series to be applicable for detecting the presence or absence of the PFE in predictions.

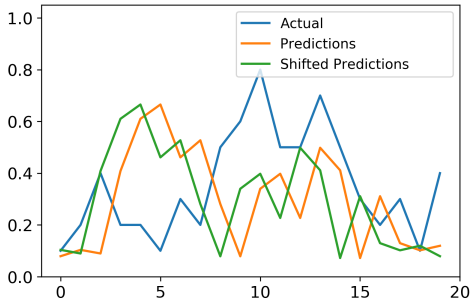


Figure 4.9: An example of a synthetically created time-series and hypothetical forecasts that are very dissimilar, showing an instance of where the n-SS method is inapplicable. The plot also shows the 1-SS applied predictions.

4.4 Summary

The n-SS method that is proposed for detecting the existence of the PFE in time-series forecasts has been introduced in this chapter. The n-SS method is a conceptually straightforward method, whose worked example has also been presented in this chapter. The proposed method relies on the recalculation of evaluation metrics after shifting the predictions 'n' time steps back to the past or shifting the observed data 'n' time steps to the future. This allows the existence or absence of the PFE to be determined quantitatively. The chapter has also presented the 1-SS method as a specific form of the n-SS method, where $n = 1$. The 1-SS method has been particularly described as the PFE causes a one step delay in predictions in the context of electricity demand forecasts, in which the PFE is mainly investigated in this thesis. Finally, this chapter has provided discussions of the limitations of the proposed generic n-SS method and described under what conditions the method may not operate effectively.

The following chapters will now expand upon the descriptions and definitions given in Chapters 3 and 4 to explore how the data irregularity influences single-step forecasts (Chapter 5) and multi-step forecasts (Chapter 6) in more detail.

CHAPTER 5

THE PERSISTENCE FORECAST EFFECT IN SINGLE-STEP AHEAD FORECASTS

In Chapter 3, the PFE has been defined along with how and why it might occur and the potential problems it might cause, highlighting the significance of why it should be taken into consideration by stakeholders. Chapter 4 then has introduced the n-SS method proposed for identifying the existence of the PFE in time-series predictions. These chapters have drawn upon several hypothetical time-series examples to explain and illustrate various aspects of the PFE and the n-SS method. The present chapter will now expand upon all the definitions and descriptions by conducting an experimental study to exemplify the PFE and the n-SS method deployment in a practical setting.

The experimental study, which will be presented in this chapter, is carried out with a large-scale dataset and contemporary machine-learning methods. The purpose of the study is to explore the PFE in single-step point forecasts, defined by Equation (2.1) and also visualised in Figure 2.3a, and the application of the n-SS method. This chapter will also present a study of the repercussions of the PFE, but now with observational time-series data, and finally, investigate further the direct relationship between the occurrence of the PFE and underlying data characteristics.

5.1 Introduction

Previous chapters have given a theoretical treatment of the PFE and a formal introduction of the n-SS method. This chapter now evaluates and demonstrates them within a practical context. This thesis is concerned mainly with the PFE in the context of electricity consumption, so this chapter uses residence-level electricity demand data. This type of data is exceptionally known to

consist of volatile and irregular patterns [120, 121], which are the origins of the PFE occurring, as discussed in previous chapters.

Household-level load forecasting has recently been receiving significant attention from both scientific and industrial researchers seeking the greatest benefits of a wide range of modern power system features, including demand response, demand-side management, and energy management systems [122]. Accurate and precise energy demand forecasting at the household level is desirable and offers numerous social, environmental, and economic benefits for both system operators and dwellers [85]. However, it is often challenging to achieve due to the inherent uncertainty and volatility of influencing factors [89, 90] and the small scale of household loads, which are affected by various consumption activities [94, 123].

In this empirical study, therefore, the volatility and uncertainty within the household-level electrical demand data are exploited in order to provide an in-depth exploration of the systematic and continuous delay caused by the PFE and the suitability of the n-SS method to detect the existence of the PFE.

5.2 Methodology

This section describes the empirical setup, including the large-scale dataset, prediction methods, data analysis methods, and the n-SS method, deployed for the experiment that is conducted to shed light on the identification of the PFE and the application of the proposed n-SS method.

5.2.1 Experimental Setup

The experimental study offered in this section replicates one of the previously published works to avoid the assumption of choosing the set of data or methodologies supporting its arguments. This approach also enables this thesis to establish that the PFE was already present but had gone unnoticed in some published works. Therefore, the study presented in this chapter is based on [89] and replicates relevant results from this paper, which has already been cited by thousands of studies, emphasising its popularity in the electrical energy consumption prediction literature. The authors of this paper publish their single-step forecast results on a publicly available dataset of electricity consumption data for several households and deploy multiple prediction methods from the Keras library [124] with Theano back-end [125]. Their use of multiple prediction methods alongside data from various individual residences is highly relevant to the purpose of the present experimental study and also provides comparisons of different prediction methods and households with varying electricity consumption behaviours. Additionally, this thesis conducts clustering and auto-correlation analyses on the household consumption data from several residences to explain the association between the PFE and data irregularity.

5.2.1.1 Dataset

In [89], the work whose methodology is followed, the authors use a dataset from a customer trial conducted as part of the Smart Grid Smart City project (SGSC), which was a four-year project initiated by the Australian Government together with participating industry partners [126]. The SGSC dataset is a standardised dataset providing electrical energy consumption data at 30-minute intervals for a large number of individual households in the state of New South Wales, Australia. One of the strengths of this database is that it provides smart meter data collected with the same instrumentation and methodology from various independent individual households. Thus, it captures the variability of residential activities and diverse electricity consumption profiles of households, which is of critical importance for the purpose of the empirical study presented in this chapter.

Although the SGSC dataset serves half-hourly smart meter data for thousands of individual residences, the authors of [89] utilise just a reasonably-sized subset of the SGSC dataset as they find it unrealistic and time-consuming to use the full dataset. The subset they considered in their study corresponded to 69 households (identified by 7- or 8-digit numeric IDs), for the proof of concept. Therefore, identical to [89], the current work also uses the same subset of 69 households with 92 consecutive daily load profiles from a 3-month time span (01.06.2013 – 01.09.2013), covering the whole winter season in New South Wales, Australia. This three-month period of time was initially chosen as it includes complete half-hourly electrical load data for all the 69 individual residences in the subset. Besides, the same training-validation-test split is applied as they do in [89]: 67 days of data for training, 16 days for validation, and the remaining nine days for testing.

In the study that is reproduced in this chapter, the authors test their prediction methods (explained in Subsection 5.2.1.2) with different numbers of recently recorded electricity consumption values in the input feature set. They feed their methods with the two, six, and 12 most recent observations separately. However, their results show that increasing the number of most recent observations in the input feature set does not bring a considerable improvement in the average accuracy of the subset. In fact, based on their outcomes, for some time-series prediction methods, an increased number of current recordings in the input feature set results in even poorer accuracy metric results. Therefore, for the experiment conducted in this thesis, it has been decided to use only the two most recent observations in the input feature set, as this appears to be sufficient for the proof of concept.

The other input features used in the experimental study are identical to those in [89]. It is worth noting that the same input feature set is utilised for all 69 households, and it includes:

- Electricity load data recordings of the two most recent time steps (x_t, x_{t-1}).
- Day-of-week indicator (ranges from 0 to 6).
- Time-of-day indicator (ranges from 0 to 47).
- Weekend indicator (ranges from 0 to 1).

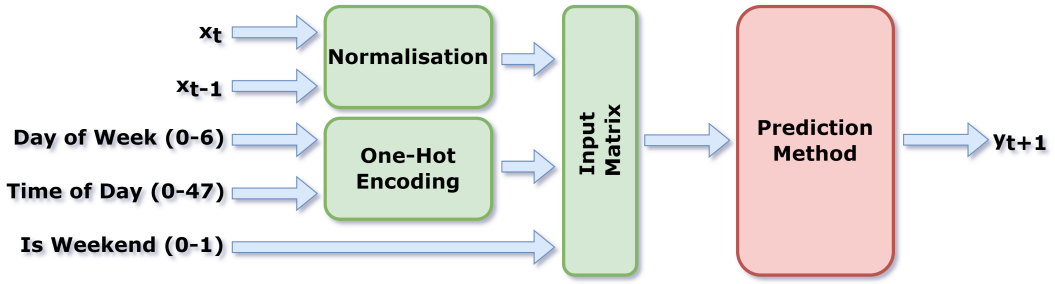


Figure 5.1: A diagram illustrating the input feature set and forecasting framework.

The data preparation is also carried out in an identical manner to [89]: all the input features are transformed to a standard scale (0, 1) independently, This is crucial to reducing the impact of outliers and achieving consistent scaling across the features. To do this, one-hot encoding is applied to time-of-day and day-of-week data, and min-max normalisation is performed for the two most recent electricity consumption values, as illustrated in Figure 5.1.

However, it is of note that for one of the 69 houses (House 8568209) from the subset, electricity consumption is approximately zero for a substantial part of the dataset, potentially because the property was vacant. This includes the majority of the validation set and the entire time

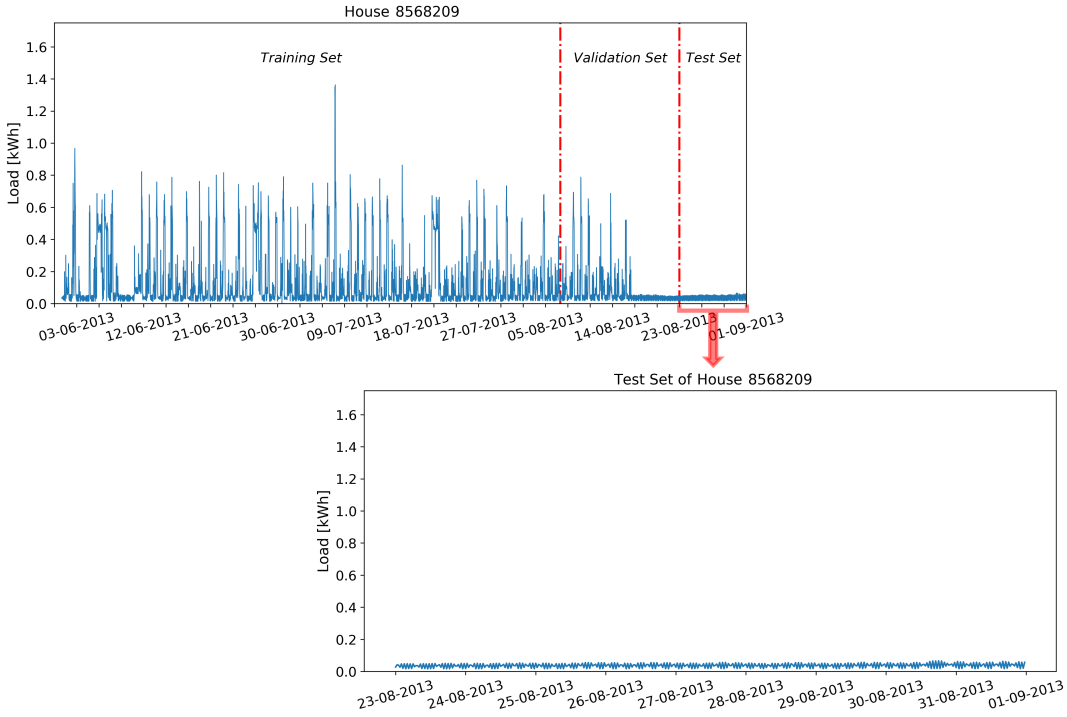


Figure 5.2: Electrical load profile of House 8568209 over the three-month span, and the test set in which all the values are approximately zero. Apparently, the consumption profile in the test set differs significantly from that in the training set.

span that defines the test set (see Figure 5.2). The electricity consumption taking place during the test set, thus, seems entirely independent and irrelevant to the consumption profiles in the training set. In order for the proper evaluation of the PFE for this household, therefore, a different three-month date range would be needed, which would be a deviation from the reproduction of [89]. For this reason, the PFE will not be investigated in this household's predictions. However, the prediction results of this household are discussed in Subsection 5.3.1 to clarify the reason why it is unreasonable to investigate the PFE in such cases.

5.2.1.2 Prediction Methods

Electrical energy consumption forecasting has gained popularity among researchers and practitioners over the last few decades. So far, dozens of methods have been developed, used, and reported to address the numerous variations of load forecasting problems with varying time horizons, aggregation levels, and building types. Among these methods, data-driven methods, i.e., statistical and machine-learning methods, most of which are listed in Chapter 2 (Subsection 2.1.3), have recently gained immense popularity for electrical energy consumption forecasting research because of their ease of use, practicability, adaptability, and high forecasting accuracy [127, 128]. In [9, 113, 129], the authors offer comprehensive reviews of most of these methods used for forecasting building-level electrical energy consumption. In recent years, the interest in deep learning methods, including RNN and CNN, has been constantly increasing. As a consequence, in the literature, there have been a great number of research works employing deep learning methods [130].

The study [89], from which relevant results are reproduced, is one of those studies developing a deep learning method. The authors of this paper propose Long Short-Term Memory (LSTM) RNN. Then, in order to validate the method they propose, they compare the accuracy of their method with some benchmark methods, including Back-Propagation Neural Network (BPNN), K-NN, EML, a sophisticated input selection scheme combined with a hybrid forecasting framework introduced in [131], and two empirical methods (i.e., empirical mean and empirical MAPE minimisation). Of these methods, only two contemporary machine-learning techniques, LSTM RNN and BPNN, are considered in the current study, as these methods achieve overwhelmingly better accuracy compared to the other methods, according to the results reported in [89].

The technical details of these two machine-learning methods, such as their working principles and mathematical representations, are out of the scope of this study provided in this chapter. Therefore, these details are not explained here, but their brief descriptions are provided instead. Nonetheless, interested readers can find full details of BPNNs in [132–134] and of LSTM RNNs in [83, 96, 135, 136].

BPNN is one of the multi-layer feed-forward neural networks [132]. It contains forward propagation of the input signal and reverse propagation of an error signal [133]. The idea behind reverse propagation is to minimise the difference between the output and expected output by

fine-tuning the weights and thresholds of the nodes in the hidden layers based on the error calculated on the output layer [132, 134]. Therefore, the learning process of BPNN divides into two key phases of forward propagation, during which the output is calculated, and then back propagation, during which the weights and thresholds of the nodes are adjusted, based on the calculated difference between the output and the desired output [137].

RNN is a powerful method for successive data where the order of data matters as it has cyclic backward connections feeding the network with the previous outputs to model the sequence [138]. As in BPNN, the learning process of RNN also has phases of forward and back propagation [139]. The fundamental difference is that RNN uses the output of the previous epoch as an input for the current epoch with the purpose of modelling the effect of the value of the previous time step to the value of the current time step [136]. As to LSTM, it is a specific RNN architecture that was designed to model sequences and their long-range dependencies accurately [140] in order to deal with the vanishing gradient problem that can occur in simple RNNs [46]. LSTM RNNs consider the relations between successive data and have been successfully deployed in various tasks dealing with successive data, such as language modelling, speech recognition, learning word embeddings, audio modelling, handwriting recognition, image generation, and time-series analysis [140–142].

In the present study, both of these methods are built with the same architecture and hyper-parameters, mimicking [89]. The common architecture and hyper-parameter settings of the methods are summarised in Table 5.1. The specific values for batch size, drop-out rate, and loss function were not mentioned in [89]. Therefore, a default batch size of 32 and no drop-out are applied. However, the Keras framework does not provide a default choice of loss function, and hence a grid search has been carried out and MAE has been selected as it is the function that resulted in the best reproduction of the results from [89].

Parameter	**Setting**
Hidden Layers	2
Nodes on Each Layer	20
Epoch	150
Optimizer	Adam
Learning Rate	0.001
Decay	0.0
Batch Size	32
Dropout Rate	No Dropout
Loss Function	MAE

Table 5.1: Hyper-parameter settings of the LSTM RNN and BPNN methods.

5.2.1.3 Evaluation Metrics

As previously stated in Chapter 4 (Section 4.2), the n-SS method, proposed for detecting the PFE in time-series forecasts, is functionally compatible with various widely utilised point-wise evaluation metrics. Therefore, in order to confirm that the n-SS method facilitates a generic mechanism that is compatible with different evaluation metrics, in this empirical study, multiple evaluation metrics are applied, as has been recommended elsewhere [94] and is commonly practised in the literature (see Table 2.1).

Among many options, three of the most popular evaluation metrics (see Table 2.1) in the electricity consumption forecasting literature are chosen for the present study: RMSE, MAPE, and Correlation, which are defined by Equations (2.5), (2.7), and (2.8), respectively. However, being three of the most frequently used metrics is not the sole reason. MAPE and RMSE are chosen as they complement each other in many aspects (see Table 5.2), and Correlation is included as it measures the linear correlation (similarity) between two sets of variables. Besides, MAPE is the only accuracy metric used in [89] whose results are partially reproduced in this chapter. As a result, these three metrics are used to evaluate predictions and test the n-SS method from many angles. The fundamental properties of these three metrics are summarised in Table 5.2.

RMSE
• Unit dependent (unit of data)
• Not always easy to interpret
• Not affected if some of the actual values are equal to zero
• Vulnerable to extremely high errors
• More punishment for larger errors (squares the errors)
• Depends systematically on the magnitude of the error
• Penalises the negative and positive errors equally
• The smaller RMSE, the better prediction
• Can not be used to compare different series because of their scale dependency

(a)

MAPE
• Unit independent (percentage)
• Easy to interpret
• It fails if some of the actual values are equal to zero
• Vulnerable to close-to-zero actual values
• Penalises the errors equally (not squares the errors)
• Depends systematically on the level of the time-series
• Penalises the negative and positive errors equally
• The smaller MAPE, the better prediction
• Appropriate for comparing different series since it is scale independent.

(b)

Correlation
• Unit free (-1 ,1)
• Easy to interpret
• Measures the strength of the relationship between the relative movements of two sets
• Reasonable metric to compare the shape and synchronicity of two set of variables
• Depends on the direction of the correlation (whether positive or negative)
• The higher correlation, the better prediction.

(c)

Table 5.2: Properties of (a) RMSE, (b) MAPE, and (c) Correlation as evaluation metrics.

As three independent metrics evaluating prediction accuracy from various perspectives are used, it can be said that for a given individual household, predictions exhibit the PFE if $MAPE^* < MAPE$, $RMSE^* < RMSE$, and $Corr^* > Corr$. Conversely, predictions do not suffer from the PFE, if $MAPE^* > MAPE$, $RMSE^* > RMSE$, and $Corr^* < Corr$. Any other combinations other than these two is considered inconclusive. This is also outlined by Equation (5.1).

(5.1)

$$\text{n-SS} = \begin{cases} \text{PFE}, & \text{if } (MAPE^* < MAPE) \text{ and } (RMSE^* < RMSE) \text{ and } (Corr^* > Corr) \\ \text{No PFE}, & \text{if } (MAPE^* > MAPE) \text{ and } (RMSE^* > RMSE) \text{ and } (Corr^* < Corr) \\ \text{Inconclusive}, & otherwise \end{cases}$$

5.2.1.4 Statistical Data Regularity Analysis

Clustering analysis and auto-correlation analysis are performed to show that the PFE is directly associated with the irregularity and volatility in the underlying time-series data and how the number of time steps of the delay caused by the PFE depends upon the strong correlation between the past observations used in the input feature set. In particular, the clustering and auto-correlation results of households suffering and not suffering from the PFE are compared to find and quantify the common characteristics of datasets subject to the PFE.

Clustering is an unsupervised learning method that groups objects based on their relative similarity given some distance metric [143]. Clustering algorithms are widely employed in the electricity demand prediction domain with the aim of improving the prediction accuracy, as in [96, 100, 143]. In this study, however, clustering is used as a data analysis tool rather than as a supporting tool that helps to obtain predictions with better accuracy. The clustering is used to group the 92 daily load profiles (between 01.06.2013 and 01.09.2013) based on their resemblance to one another. The total number of groups provides a measure of the regularity level of load patterns from a household on a day-by-day basis based on clustering in line with [89, 90]. The clustering, therefore, provides both visual and numeric representations of the regularity of daily electricity consumption of households, allowing for easy comparison of the regularity levels across multiple buildings.

Among available clustering methods, hierarchical clustering (see [144, 145] for technical details) is chosen as: i) it requires only a few hyper-parameters to be selected; ii) it does not need a predetermined number of clusters; iii) it explicitly identifies outliers; and iv) it is easy to interpret.

Since the objects to be clustered in the current study are not single data points but a successive data array of 48 data points representing a day, the correlation coefficient is chosen as the distance

metric. To calculate the distance between individual objects or clusters, the average method is performed. The threshold for assignment to the same cluster is set to $corr = 0.75$. In other words, two daily electricity demand profiles are assigned to the same cluster if their correlation coefficient is equal to or greater than 0.75. Table 5.3 lists the hyper-parameter selection of the hierarchical clustering used in this work.

Parameter	**Setting**
Similarity Metric	Correlation Coefficient
Method	Average
Threshold Value	0.75

Table 5.3: Hyper-parameter settings of the hierarchical clustering.

In addition to the hierarchical clustering, auto-correlation analysis is also performed to evaluate the similarity between successive observations of the same time-series variable as a function of the delay between the observations. In auto-correlation analysis, the distance between observations is called lag. For example, in the current study with data points separated by half an hour, lag 1 refers to the observation 30 minutes prior to the most recent one, and lag 48 to the observation 24 hours earlier. Auto-correlation, thus, calculates the correlation of each value in the series to lagged values across the entire dataset, offering a way to measure and understand similarity within the successive observations. This also determines the number of time steps of delay created by the PFE.

5.2.1.5 The n-SS method for the PFE Investigation

Once time-series forecasting is completed for each building, the n-SS method proposed in the previous chapter is applied to identify which electricity consumption predictions are influenced by the PFE. This method requires determining the value of n for each building beforehand.

In Subsection 4.2.1, two different approaches have been proposed to choose the value of n: auto-correlation and the brute-force approach. Among them, the former strategy – auto-correlation – is used here. Because the input feature set includes only the two most recently recorded electricity consumption values, it is enough to compare the auto-correlation values at lag 1 and lag 2, with the lag having the highest auto-correlation value becoming the value of n. The auto-correlation values at lag 1 and lag 2 of each building are compared in Figure 5.3 side-by-side. It is evident from this figure that the auto-correlation values at lag 1 are always greater than the values at lag 2 for all households, meaning that 1 must be assigned to n; ($n = 1$). This means that the 1-SS method, introduced in Chapter 4 (Subsection 4.2.2) is deployed for each of the 68 households to determine the existence of the PFE in predictions.

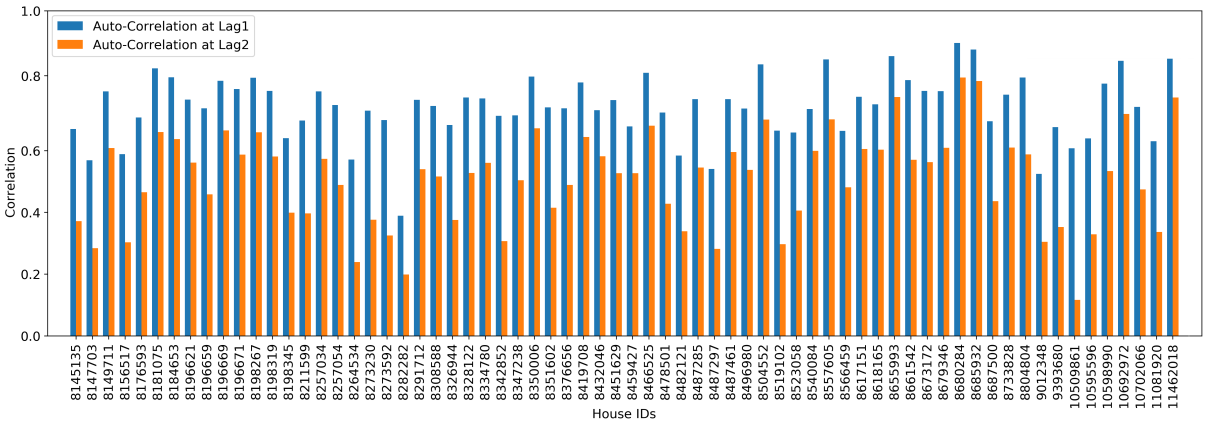


Figure 5.3: Comparison of auto-correlation values at lag 1 and lag 2 for each of the 69 houses studied. For all households, the auto-correlation values at lag1 are always greater than the ones at lag2.

5.3 Results

This section will present the findings of the empirical study that is conducted with an electricity consumption dataset and advanced time-series prediction methods, which are described above in Section 5.2.

5.3.1 Prediction Results and Evaluation of the PFE Existence with n-SS

In order to accomplish the objectives of this study, the LSTM RNN method is deployed on each of the electricity consumption data of 69 residences independently. The described LSTM RNN is deployed for each building with identical architecture, hyper-parameter settings, and input features, including electricity load recordings of the two most recent time steps. Following that, the evaluation metrics are calculated from the original predictions, and then the 1-SS method is applied to the predictions in order to investigate the PFE.

However, prior to moving on to the PFE results on a building-by-building basis, it is important to take an in-depth look at the prediction results of the household with approximately zero electricity consumption for a substantial part of the dataset, including the whole test set (demonstrated in Figure 5.2). For this household, the electricity consumption patterns in the training set are completely dissimilar to the consumption patterns in the test set. This means that the predictions produced are markedly different from the actual consumption values in the test set, as can be seen in Figure 5.4. This disparity between the data curves renders discussion of the temporal accuracy or the existence/absence of the PFE here inappropriate, as already explained in Chapter 4 (Section 4.3). This building is, therefore, omitted from the wider PFE evaluation, and the study continues with the remaining 68 individual households.

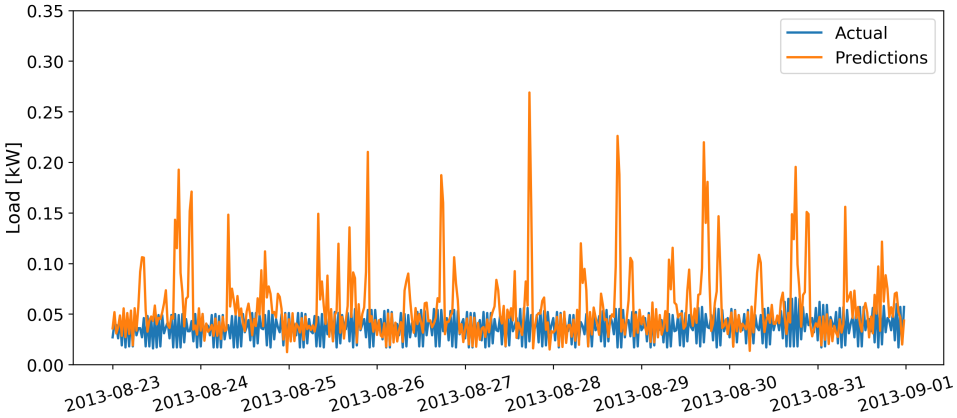


Figure 5.4: Prediction outputs for House 8568209, whose electricity consumption values are approximately zero across the entire test set.

Coming back to the PFE investigation, the default (MAPE, RMSE, and Corr) and the 1-SS method applied (MAPE*, RMSE*, and Corr*) evaluation metric results are presented together in Section A.1 in the appendix in Table A.1. Additionally, Figure 5.5 shows how the accuracy metrics of each building change with the application of the 1-SS method. This change formed the basis for investigating the possible presence of the PFE. The metric results and changes in them are evaluated for each building independently; the comparison of metric results from two or more households is not necessary for the PFE investigation here and thus outside the scope of this work. For clarity alone, it is worth elucidating that for MAPE and RMSE, a lower metric value corresponds to better accuracy, while the opposite is true for Correlation. Therefore, the prediction accuracy improves for any given household when MAPE and RMSE values increase and correlation value decreases after the application of the 1-SS method.

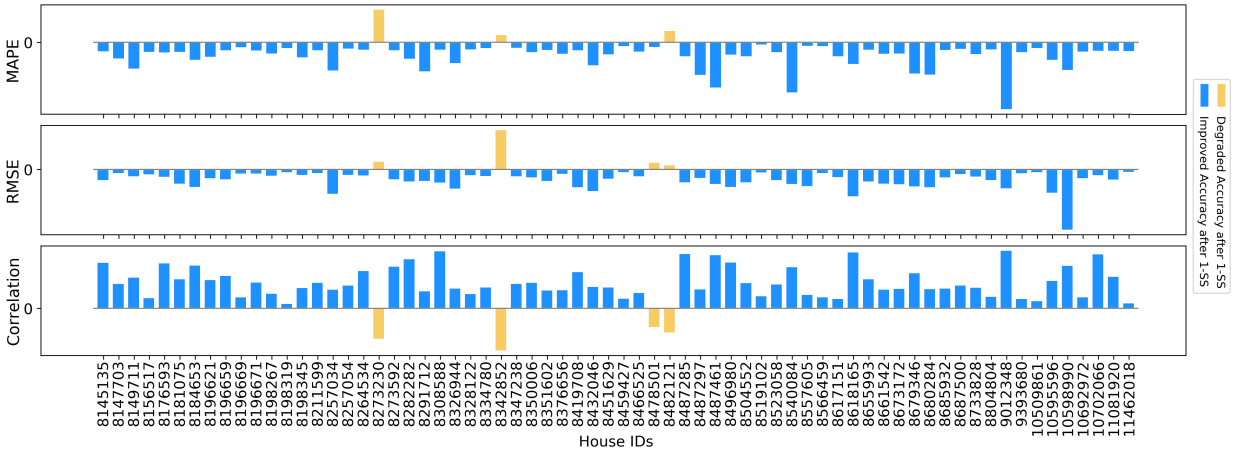


Figure 5.5: Difference between the default and the 1-SS method applied evaluation metrics. The bars are vertically aligned for each House ID. Results indicate that only the predictions of Houses 8273230, 8342852, and 8482121 are PFE-free, and only one household (House 8478501) has PFE-inconclusive predictions.

Based on the positive or negative changes in accuracy metrics produced by the 1-SS method and Equation (5.1), the 68 households can be split into three groups in terms of the existence of the PFE:

- *PFE-free:* All three metrics worsen after the 1-SS method is applied. This occurs for three households: Houses 8273230, 8342852, and 8482121.
- *PFE-inconclusive:* One or two metric(s) improve while another worsens after the 1-SS method is applied. This occurs for only one household: House 8478501.
- *PFE-affected:* All three metrics improve after the 1-SS method is applied. This occurs for all the remaining 64 households, which is the large majority of the subset used in this study.

For instance, the 1-SS method indicates that the predictions of House 8342852 are not affected by the PFE, as MAPE and RMSE values worsen (error increases) from 38.704 and 0.227 to 46.294 and 0.550, respectively. Meanwhile, the Correlation worsens (less alignment) from 0.941 to 0.643. On the other hand, the predictions of House 8184653 are affected by the PFE, as MAPE and RMSE improve (error decreases) to 20.150 and 0.107 from 38.724 and 0.252, respectively, and the Correlation improves (better alignment) from 0.643 to 0.941. Finally, the MAPE value of the only PFE-inconclusive building (House 8478501) drops from 51.373 to 46.553, indicating the PFE, but the RMSE and Correlation values from the same building deteriorate from 0.167 to 0.220 and from 0.834 to 0.702, respectively, indicating the absence of the PFE. An extensive and detailed examination of the PFE in the predictions of this PFE-inconclusive household is provided in Chapter 7.

An example of the 1-SS method deployment is shown in Figure 5.6, illustrating the original and one step shifted predictions over 24 hours along with the actual consumption values of two households: one from the PFE-free group (Figure 5.6a) and the other from the PFE-affected group (Figure 5.6b). In Figure 5.6a, the curve of the original predictions matches with the curve of the actual values significantly better than it does with the curve of the shifted predictions. The opposite is the case in Figure 5.6b, where the temporal alignment between the shifted predictions and the observed load values is substantially better than the temporal alignment between the original predictions and the observed load values.

Taken together, according to the 1-SS results, only three of the households do not suffer from the bias, whereas 64 of them explicitly exhibit the PFE. This reveals how widespread the PFE is in electrical load forecasting at the household level and how difficult it is to make reliable and robust forecasts for such time-series datasets.

The household which has the best (lowest) default MAPE value (17.555) is House 8685932 out of all the households in the dataset. If the performance of the prediction method and the accuracy of its predictions were solely assessed according to the default evaluation metrics and the PFE was completely ignored, as has commonly been done in the time-series literature so

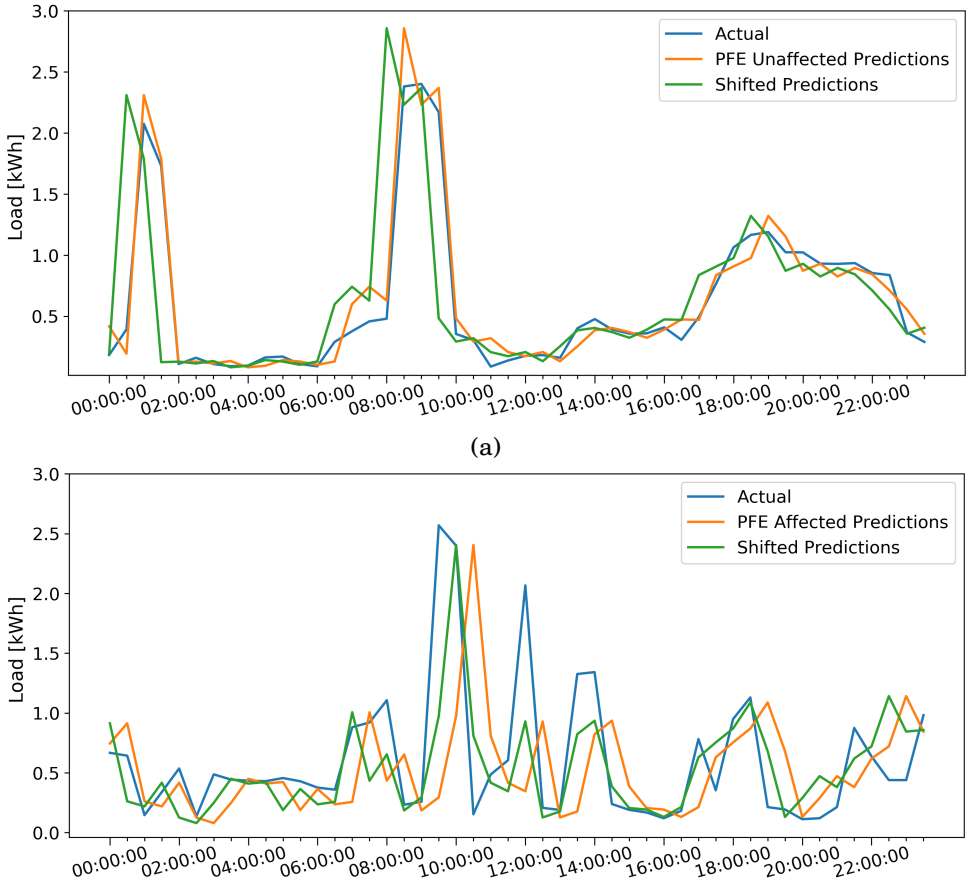


Figure 5.6: Illustration of the 1-SS method application on two predictions for two separate electricity consumption datasets: (a) without the PFE (House 8342852) and (b) with the PFE (House 9012348).

far, it could be inferred from the MAPE values in Table A.1 in Appendix A that the prediction method performs the best on the electricity consumption data of House 8685932 and learns the consumption patterns of this building much better than all the other buildings' patterns. However, in fact, the situation is actually the reverse. The MAPE value decreases to 9.443 with the application of the 1-SS results, which shows that the prediction method does not learn the historical patterns in this case but, instead, returns predictions approximating the most recent consumption value. As a result, the predictions are temporally displaced, and they trail the actual load values one step behind in time due to the PFE.

5.3.2 Evaluation of the PFE Implications

The potential implications of the PFE have already been described and demonstrated in Chapter 3 (Section 3.3). However, it is still valuable to explore them further through the prediction outcomes of the empirical study using the observational time-series data introduced above and LSTM RNN, one of the most popular and successful time-series prediction methods.

Even though the 1-SS method is conceptually very simple, it can detect the presence or absence of the PFE in all but one of the cases above. The default evaluation metrics by themselves, however, are not able to capture such behaviours, causing a delay in predictions, or poor learning of the training data by the prediction method. Therefore, this produces deceptive metric results, which can result in misplaced confidence not only in predictions but also in models. Figure 5.7, for example, juxtaposes MAPE and MAPE* values of 6 houses in three pairs of houses with PFE-free and PFE-affected predictions. These pairs are chosen purely on the basis of their similar default MAPE values, and in each pair, the default MAPE values are very similar to one another (e.g. 25.357 and 24.815 for houses 8273230 and 8487285, respectively). The default MAPE results suggest the prediction method performs relatively similarly for the pairs 8273230 and 8487285, 8342852 and 8184653, and 8482121 and 8661542. However, when the 1-SS method is applied, the evaluation metric values improve for the households with PFE-affected predictions and degrade for those not subject to the PFE – as shown by MAPE* in Figure 5.7. Even though the absolute overall prediction error in each pair of households during the test interval is very similar, for predictions exhibiting the PFE, this error is almost exclusively determined by the cumulative difference of the observed energy consumption between two time steps. Therefore, although the prediction method is unable to perform proper learning from the electrical energy consumption data of PFE-affected houses and predictions approximate the most recent consumption value, default evaluation metrics misleadingly suggest that the predictions produced for those buildings are as accurate as their pairs, leading to misleading determinations on predictions and models.

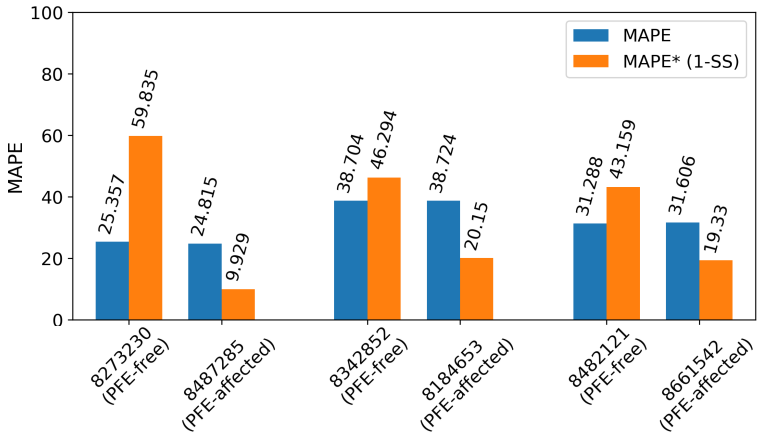


Figure 5.7: Paired MAPE and MAPE* values of three PFE-free predictions (Houses 8273230, 8342852, and 8482121) and three PFE-affected predictions (Houses 8487285, 8184653, and 8661542).

The PFE also raises the risk of misinterpretation of evaluation metrics when comparing the success of prediction methods. This is tested with a comparison of the above-described LSTM RNN and BPNN methods for six residential buildings, three of which have PFE-free predictions, and the other three randomly chosen from those that have PFE-affected predictions. The LSTM

RNN and BPNN are partly stochastic, so every run of them results in slightly varying forecasts. Therefore, each is deployed five times independently for each of these buildings. The default MAPE results are listed in Table 5.4. For buildings with PFE-free predictions, LSTM RNN always performs consistently and significantly better than BPNN. However, for households with PFE-affected predictions, the difference between default MAPE results of LSTM RNN and BPNN is not significant, and the rank order between these two methods is not stable, preventing the identification of the best-performing method. This is because, when the underlying data lacks regularity, both methods are affected by the PFE to a very similar extent. Hence, both methods extrapolate predictions from the most recent consumption value in the PFE-affected houses and deliver very similar accuracy metric results for such PFE-affected households. To sum up, when using the default evaluation metrics to compare alternative methods, the PFE can result in evaluation metrics that separate methods only poorly – or even reversing the rank order between them.

HouseID	PFE	Method	MAPE1	MAPE2	MAPE3	MAPE4	MAPE5
8273230	No	LSTM RNN	25.210	24.617	24.971	25.029	24.688
		BPNN	31.084	29.507	32.054	29.876	30.804
8342852	No	LSTM RNN	37.824	38.576	37.560	38.625	38.785
		BPNN	63.641	68.311	66.885	65.307	63.558
8482121	No	LSTM RNN	31.303	30.839	30.941	30.471	31.434
		BPNN	39.858	40.926	41.360	40.082	39.211
8181075	Yes	LSTM RNN	18.920	18.916	18.610	19.104	18.938
		BPNN	19.206	18.610	19.778	18.888	18.821
8196621	Yes	LSTM RNN	36.807	35.709	36.355	36.765	35.907
		BPNN	35.174	36.523	36.624	34.440	35.672
11081920	Yes	LSTM RNN	31.785	30.708	31.739	31.671	31.474
		BPNN	31.532	30.275	32.569	30.476	31.587

Table 5.4: Evaluation of repeated runs of LSTM RNN and BPNN for a selection of houses from the dataset.

5.3.3 Regularity Analyses Results

Hierarchical clustering and auto-correlation analyses have been deployed to obtain the quantitative expressions of the differences between the regularity level of electrical load data from the PFE-free and PFE-affected households. This allows the direct relationship between the irregularity in the underlying data and the PFE to be studied in detail. Hierarchical clustering is used to group the daily load profiles of a household throughout 92 days to examine the resemblance of daily load profiles to each other. Auto-correlation is used to identify how successive data points are correlated to each other, which also represents the correlation between electricity consumption of successive days.

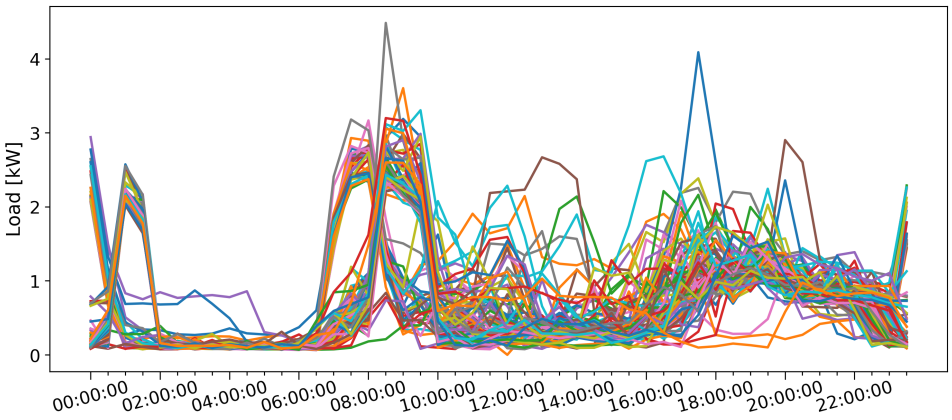
5.3.3.1 Clustering Results

Hierarchical clustering, previously specified in Subsection 5.2.1.4, is performed for each building in the dataset individually, and the results are shown in Table 5.5. In this table, the households are sorted by the number of clusters, with fewer clusters indicating a greater similarity between the daily energy consumption profiles over the 92 days of a specific household and thus better regularity. It is important to note that outlier daily demand profiles that do not have any similar daily profiles to be in the same cluster with according to the correlation distance metric are each put in a separate cluster.

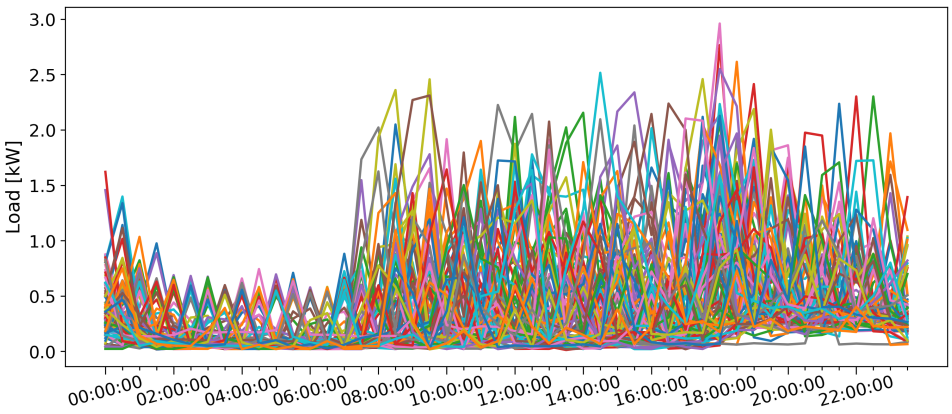
HouseID	PFE	No. of Clusters	HouseID	PFE	No. of Clusters	HouseID	PFE	No. of Clusters	HouseID	PFE	No. of Clusters
8342852	No	10	8198319	Yes	45	8347238	Yes	59	8308588	Yes	74
8273230	No	12	8196659	Yes	45	8504552	Yes	60	8487461	Yes	75
8482121	No	12	8519102	Yes	46	8376656	Yes	60	8679346	Yes	75
8804804	Yes	20	8617151	Yes	46	8655993	Yes	60	8487285	Yes	76
11462018	Yes	20	8350006	Yes	46	8198345	Yes	60	8264534	Yes	77
8478501	Yes	26	8459427	Yes	47	8184653	Yes	60	8432046	Yes	77
8466525	Yes	30	10509861	Yes	50	8145135	Yes	61	8496980	Yes	77
8680284	Yes	34	8257054	Yes	50	8326944	Yes	62	10598990	Yes	77
8334780	Yes	36	8196671	Yes	53	8147703	Yes	63	8257034	Yes	77
8557605	Yes	38	8685932	Yes	54	8351602	Yes	64	9012348	Yes	80
10692972	Yes	38	8273592	Yes	54	8673172	Yes	65	11081920	Yes	81
8291712	Yes	40	8149711	Yes	54	8211599	Yes	66	10702066	Yes	84
8198267	Yes	40	9393680	Yes	56	8566459	Yes	66	8540084	Yes	84
8419708	Yes	40	8661542	Yes	57	10595596	Yes	67	8618165	Yes	86
8181075	Yes	42	8328122	Yes	57	8156517	Yes	67	8487297	Yes	87
8687500	Yes	44	8196621	Yes	57	8451629	Yes	68	8523058	Yes	88
8196669	Yes	45	8733828	Yes	58	8176593	Yes	72	8282282	Yes	91

Table 5.5: Hierarchical clustering results of 68 residences, sorted by the number of clusters.

The clustering results show significant differences between the regularity levels of household demand profiles. The most regular household has only ten clusters for its 92 daily load profiles, whereas the household with the most variable load profiles has 91 clusters. This means that in the most irregular household, only two daily load profiles are similar, and the other 90 daily load profiles are judged to be dissimilar to each other. Figure 5.8 provides a visual representation of 92 daily load profiles for these two households. In Figure 5.8a, for instance, overlapping daily load profiles demonstrate that they follow very similar patterns throughout 24 hours, highlighting the regularity and consistency in the data. Conversely, in Figure 5.8b, it is hardly possible to identify recurring patterns, revealing the lack of regularity in the data. The difference between the regularity of the electricity consumption data of these two buildings is also apparent in Figure 5.9, which shows the electrical load patterns for each building in 3D plots. In Figure 5.9a, the load demands from House 8342852 exhibit noticeable patterns over the three-month period,



(a)

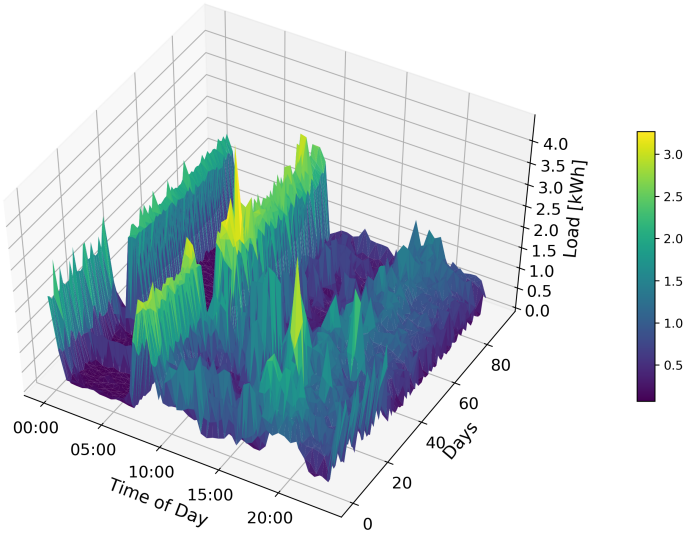


(b)

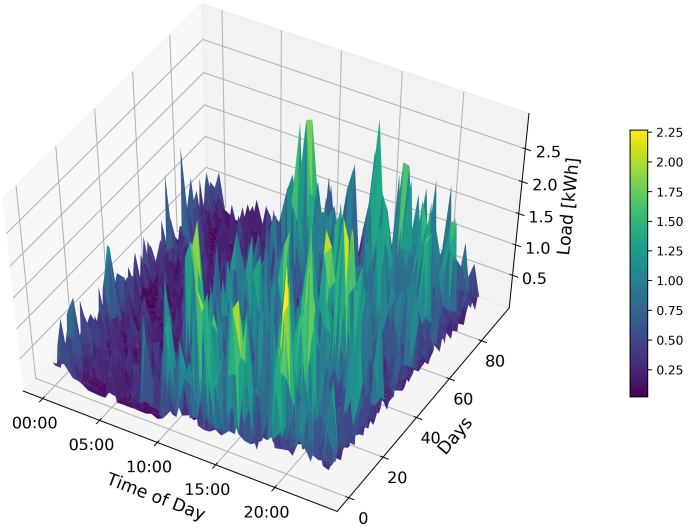
Figure 5.8: 92 daily load profiles of (a) the most regular household (House 8342852) and (b) the most irregular household (House 8282282). The complex data of House 8282282 manifests itself here as chaotic daily consumption patterns, revealing the lack of regularity. The daily loads of House 8342852 are more regularly structured.

explaining the lower number of clusters associated with this building. In contrast, the equivalent data in Figure 5.9b for House 8282282, which had 91 clusters, show no explicit repetitive load demands, implying almost every single day appears to have a unique load pattern.

It is noteworthy that the three houses, namely Houses 8342852, 8273230, and 8482121, which exhibit PFE-free predictions, are positioned at the top of Table 5.5, showing that these houses have the most self-similar and regular daily demand profiles among the 68 houses considered. The considerable difference between the regularity of the daily load profiles of Houses 8273230 and 8487285, whose MAPE and MAPE* values were previously compared as a pair above in Figure 5.7, is clearly evident in Figures 5.10a and 5.10b. The daily load profiles of House 8273230 exhibit a consistent and structured pattern, whereas those of House 8487285 are considerably disorganised. The difference in their regularity is further supported by the dendrograms presented in Figures 5.10c and 5.10d, which visualise the clustering results of these two residences. Based on the clustering results, there are 12 distinct patterns in the 92 daily load



(a)



(b)

Figure 5.9: Load profiles of (a) the most regular household (House 8342852) and (b) the most irregular household (House 8282282) throughout the 92 days.

profiles of House 8273230, whereas there are 76 distinct patterns in those of House 8487285, which quantitatively confirms the difference between the regularity of the daily load profiles of these buildings.

Due to the irregularity and variability in the consumption data of House 8487285, prediction methods are not able to learn the patterns from such data unless there are some extra features in the input feature set that may explain the irregularity, making the data unlearnable. Therefore, it seems once again that the only reason for the accurate default metric results (e.g., MAPE) for House 8487285 is the PFE in predictions.

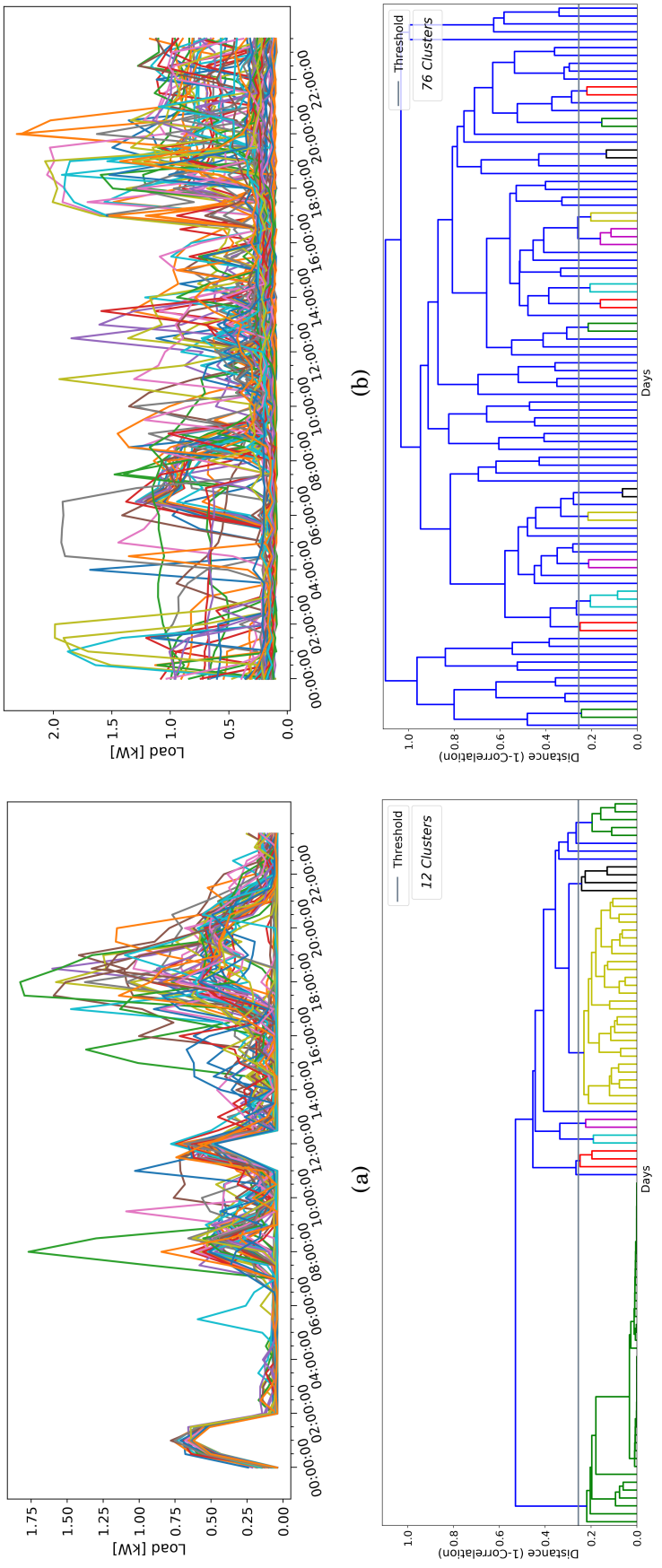
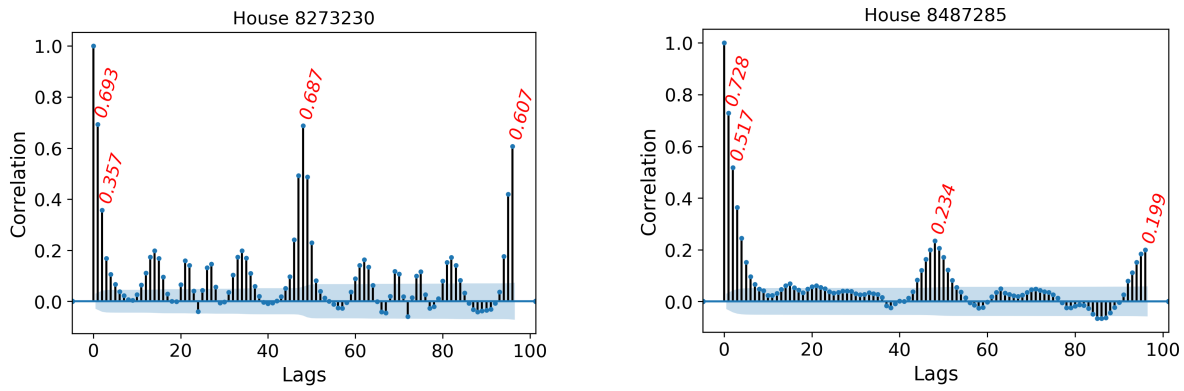


Figure 5.10: 92 daily load profiles of (a) a PFE-free building (House 8273230) and (b) a PFE-affected building (House 8487285). The daily loads of the PFE-free residence follow similar pattern throughout the three months, in contrast to the PFE-affected residence, which exhibit chaotic patterns. This is accompanied by dendrograms showing hierarchical clustering results for (c) House 8273230 and (d) House 8487285. Every colour group, with the exception of blue, corresponds to a distinct cluster, while the blue lines indicate outlier daily profiles.

5.3.3.2 Auto-Correlation Results

Auto-correlation analysis is performed to further study the differences between the PFE-free and PFE-affected household-level electricity consumption data. Figure 5.11a shows significant auto-correlation values at various lags throughout 48 hours for House 8273230, whose predictions are shown to be PFE-free. In particular, the correlation values at lag 1 (0.693), lag 48 (0.687), and lag 96 (0.607) are quite similar, showing the similarity between observations at the same time on successive days. In contrast, Figure 5.11b shows the auto-correlation results for House 8487285, whose predictions are shown to be PFE-affected. Here, the correlation values are notably decreasing towards lag 48 (0.234) and lag 96 (0.199), and are dominated by the correlation value at lag 1 (0.728), ignoring the lag 0, which is self-correlation and always 1. This points out the variability of observations recorded at the same time on consecutive days.



(a) (b)

Figure 5.11: Auto-correlation analysis of (a) a PFE-free building (House 8273230) and (b) a PFE-affected building (House 8487285).

Furthermore, for the PFE-affected household, the auto-correlation value at lag 1 is much higher than values at any other lag, including lag 2, which is also used in the input feature set (see Figure 5.11b). This explains why prediction methods tend to extrapolate predictions from the most recent value. In other words, in datasets with inconsistent and volatile patterns, the current value is always most correlated to the value of the previous time step, and thus prediction methods are most likely to forecast that *the future will be similar to the present*. Similarly, as can be seen in Figure 5.11a, the auto-correlation value at lag 1 is higher than lag 2 for the PFE-free household as well. However, the regular patterns in electricity consumption from this house mean that prediction methods can exploit knowledge of the load at the same time of previous days and so do not necessarily fall back on extrapolating from the most recent observation.

5.4 Discussion

Given the strong within-household variability of the load demand profiles, as shown through clustering and auto-correlation analyses – it is not surprising that prediction methods fail to provide robust and temporally accurate predictions. Such instances should ideally be detected during the prediction evaluation before investing more time and money in systems or applications that rely on such unreliable predictions or models. Nevertheless, many of the accuracy metrics available to the community are generally not able to do this.

The n-SS method provides a conceptually straightforward method to detect the PFE in time-series predictions. It is important to note that although the ranking in Table 5.5 explicitly reveals the relationship between the irregular patterns in the underlying data and the PFE, clustering cannot be used as a tool to identify the PFE before the time-series forecasting is performed. This is because, first, hierarchical clustering requires a choice of hyper-parameters such as the threshold value (here, corr=0.75), and, second, clustering results rely on dataset features, such as dataset length, granularity, and the like. While the ranking of houses is supposed to be independent of these hyper-parameters, the absolute number of clusters is not. Therefore, there cannot be a particular number of clusters that might be used as a threshold value to distinguish datasets that are or are not subject to the PFE.

It is also worth considering the importance of the most recently recorded values used in the input feature set overall. To this end, the resulting accuracy of the LSTM RNN approach is compared when using and not using current observations in the input features for five randomly selected houses from the subset. The accuracy results can be found in Table 5.6. Based on these metric results, the models with the two most recent recordings in the input feature set always yield more accurate predictions compared to the predictions produced by the models with no currently recorded consumption value in the input feature set. This indicates that the most recently observed values are still important predictors, and having them in the input feature set is generally effective, although they are one of the important factors behind the PFE. To sum up, similar to [112], it can be concluded that having a certain number of most recent observations in the input feature set brings substantial improvement in prediction accuracy. In addition, Figure 5.12 illustrates the prediction results of House 8540084, which is one of the five randomly chosen households, with and without current recordings in the input feature set. It is evident from this figure that the prediction method fails to yield accurate and correlated predictions when there are no current recordings in the input feature set. However, when the most recent recordings are included in the input feature set, the prediction method produces much more accurate predictions that are almost identical to the actual load curve but delayed in time if the PFE occurs. It is clear, then, that the most recent observations are of high value to prediction accuracy but also important factors in the PFE. Therefore, the most recent observations should be included in the input feature set to obtain a high degree of prediction accuracy, and then the n-SS method should be used to identify the PFE when the underlying data provides insufficient regularity to prediction methods.

HouseID	Most Recent Obs. in the Input Set	MAPE	RMSE	Corr
8196671	No	66.988	0.233	0.320
	Yes	28.727	0.122	0.635
8350006	No	35.446	0.204	0.571
	Yes	21.296	0.145	0.751
8432046	No	77.413	0.671	0.466
	Yes	63.266	0.477	0.768
8540084	No	135.099	0.318	0.365
	Yes	78.493	0.226	0.662
9393680	No	172.943	0.527	0.403
	Yes	100.083	0.467	0.557

Table 5.6: Accuracy comparison with and without recently recorded values in the input feature set with the evaluation metrics MAPE, RMSE, and Correlation.

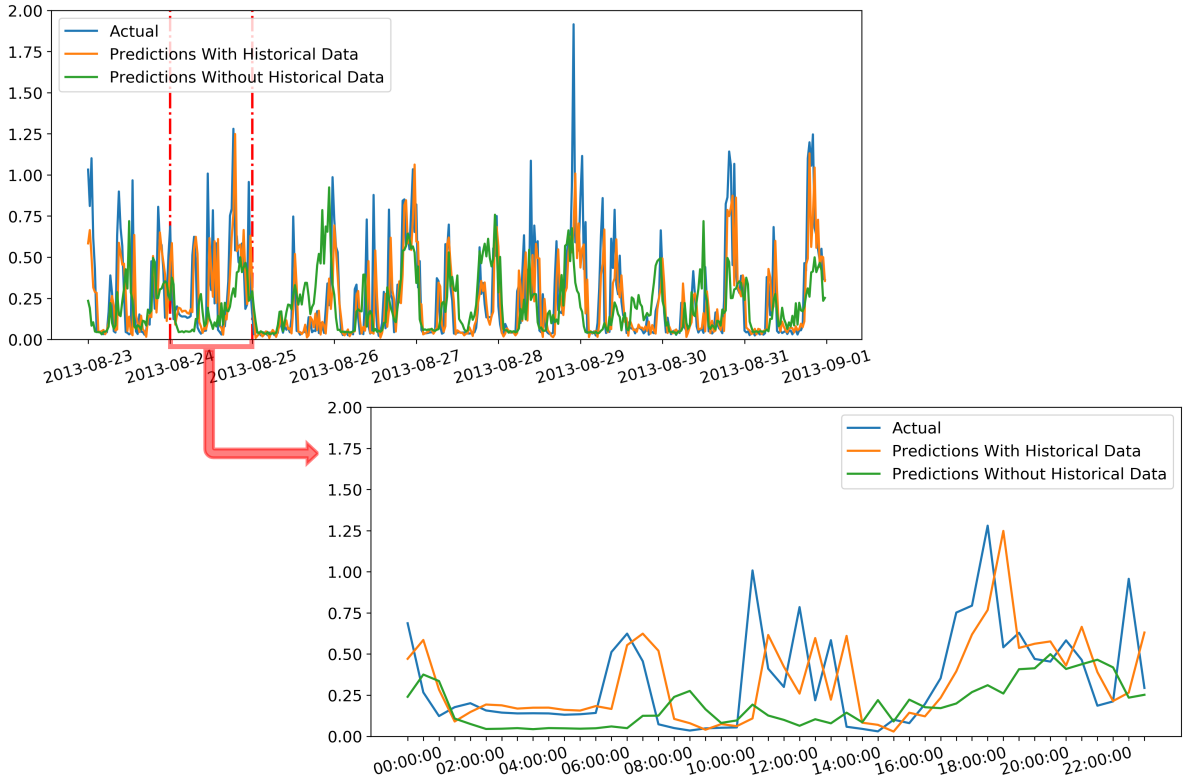


Figure 5.12: The comparison of prediction results of House 8540084 with and without most recent observations in the input feature set. The prediction method fails to yield accurate and correlated predictions with no most recent observations included in the input feature set.

To conclude, it is important to highlight the prominent contributions of the empirical study presented in this chapter. This study replicates the relevant results from [89] with the data and

methods the authors use in this paper. Therefore, the empirical results convincingly demonstrate that the PFE manifestly affects the results of some works previously published in the literature. The study deploys two independent prediction methods for 68 individual households' electricity consumption data. Both methods return predictions either with or without the PFE, and their results are consistent in terms of which houses have predictions that suffer from the PFE and those that do not. This substantiates the fact that the PFE is not related to prediction methods but is directly connected to the characteristics of the underlying data. Additionally, based on the data regularity analyses provided in this chapter, it has been shown that the main reason for the PFE is hidden in the initial data and its volatility and irregularity. Most importantly, only three of the houses have PFE-free predictions, whereas predictions for 64 of them are PFE-affected, revealing how widespread the PFE is at the household-level electrical energy forecasting. Finally, this study demonstrates that the PFE is not always simple and obvious in predictions for observational time-series data, as opposed to the examples presented through synthetic scenarios in Chapters 3 and 4.

5.5 Conclusion

This chapter has provided an experimental single-step point forecast study using a large-scale dataset and recently widely-used machine-learning methods in order to explore further the PFE, its causes, and its effects within a practical setting beyond the idealised cases considered in earlier chapters. The empirical study has shown the PFE in single-step forecasts as well as the application of the n-SS method in identifying the PFE. Furthermore, the implications of the PFE and the relationship between the PFE and data irregularity have been demonstrated on observational electricity consumption data.

The PFE has been investigated using an electrical energy consumption dataset of 68 individual houses by deploying the advanced machine-learning techniques, LSTM RNN and BPNN and following the methodology of a recent, peer-reviewed study. The results have shown that standard evaluation metrics are insufficient to detect the PFE. According to the n-SS results, it has been calculated that only three of the houses (%4.41) have PFE-free predictions, whereas predictions of 64 of them (%94.11) are PFE-affected, and the remaining household (%1.48) is PFE-inconclusive. Finally, the analysis of similarity between-day and within-day through hierarchical clustering and auto-correlation has allowed a more formal description of the PFE to be made.

The next chapter will further look at the PFE in the context of multi-step forecasts. This will be followed by an exploration of the applicability of the proposed n-SS method in a multi-step case. Some of the outcomes of the present chapter will be used as the backbone of the work presented in the following chapter.

CHAPTER 6

THE PERSISTENCE FORECAST EFFECT IN MULTI-STEP AHEAD FORECASTS

In earlier chapters, the PFE has been described as a systematic and continuous delay in time-series predictions caused by volatility and irregularity in the underlying data. Then in the previous chapter, this effect in single-step forecasts and the application of the n-SS method for identifying PFE in single-step forecasts have been examined. Single-step forecasts only predict one time step ahead, and so always have the most recent observations available. However, this is not the case within multi-step forecasts, which is defined by Equation (2.2) and also visually represented in Figure 2.3b. In multi-step forecasts, the forecasting horizon is not limited to only one step in future but rather comprises several time points in multi-step forecasts. Therefore, the most recent observations, on which the prediction outputs merely rely when the PFE occurs in single-step forecasts, are not completely available for all the time points in the forecasting horizon. This suggests that the behaviour of predictions affected by volatility and irregularity in multi-step forecasts may differ from that of single-step forecasts.

Therefore, this chapter will now delve into the examination of how irregularity and volatility in data affect multi-step forecasts, how this effect shows itself, and why. This will be followed by analyses of the applicability of the proposed n-SS method in multi-step forecasts. In addition, an experimental study, similar to that conducted in Chapter 5, will be presented in order to study the bias and the n-SS method with observational data and machine-learning methods.

6.1 Introduction

Time-series forecasting can be performed for both single and multiple steps ahead in time [146]. Although single-step forecasts are quite effective in handling the continuity and periodicity of

time-series data, they fall short of capturing long-term trends [147]. Hence, single-step forecast outputs are not able to offer any information about the time ahead of one time step in the future [148], which may make them insufficient to ensure the reliability and robustness of final applications or systems depending on their ultimate purpose [149]. Robust multi-step forecasts are, therefore, considered desirable and essential in such cases. In the time-series context, multi-step prediction can be described as a task aiming to forecast multiple time steps into the future with no output measurement before the entire forecasting horizon has been predicted [148, 150].

Multi-step forecasts are known to be even more challenging and complex tasks compared to single-step forecasts [135, 146, 151]. This is essentially because the longer the forecasting horizon, the more difficult it is to get accurate predictions [36]. This is mostly owing to some additional issues and complications that multi-step forecasts have to deal with, such as rising uncertainty across the forecasting horizon, accumulating errors in each time step, reducing accuracy along with the greater number of time points to forecast, and reducing flexibility due to a fixed prediction model [146, 147, 152]. If a multi-step time-series forecast is already a challenging task, performing it on volatile and irregular data, which are advocated to be underlying reasons behind the PFE, to achieve steady and accurate results is even more difficult.

In single-step forecasts, irregularity and volatility in the data cause prediction methods to extrapolate the predictions from one of the most recent observations used in the input feature set, resulting in a series of predictions that trails the series of observations one or a few steps behind. However, the most current observations are not yet available for all the time points in the forecasting horizon of multi-step forecasts. More formally, in single-step case, it is always only y_{t+1} to be predicted, and all the historical observations up to time t (i.e. x_t, x_{t-1}, x_{t-2}, *etc.*) are known and recorded, as described by Equation (2.1). On the other hand, in the multi-step case defined by Equation (2.2), since the forecasting horizon is larger than one step, the historical observations to be used in the input feature set are available only for the next few time steps but not known yet for the majority of the time points in the forecasting horizon. For instance, in multi-step case, for the first time step of the prediction horizon (y_{t+1}), the two most recent past observations x_t and x_{t-1} are known and available to be included in the input feature set. However, for the last time step of the prediction horizon (y_{t+H}), where $H \geq 3$, the two most recent observations x_{t+H-1} and x_{t+H-2} are not available yet. That is to say, for y_{t+H}, the most recent past observations to be used in the input feature set are not available. Therefore, the systematic delay observed in single-step forecasts does not necessarily seem possible in multi-step forecasts, even though the predictions are still negatively affected by irregularity and volatility in the data. To further examine this, this section offers an empirical study, similar to the one presented in the previous chapter, to evaluate the effect of data irregularity on multi-step forecasts and to investigate the effectiveness of the n-SS method in a practical setting.

6.1.1 Strategies for Multi-Step Time-Series Forecasting

Before commencing the experimental study, it is helpful to discuss the strategies used for multi-step time-series forecasts. According to [146, 151], there are five different fundamental strategies proposed in the literature in order to deal with multi-step time-series forecasting tasks (see Figure 6.1). Three of them are recursive strategy, direct strategy, and multi-input multi-output (MIMO) strategy. The others are basically hybrids of two of these three strategies: DirRec strategy, which combines the recursive strategy and the direct strategy; and DirMO strategy, which combines the MIMO strategy and the direct strategy. The details of these strategies and how they achieve multi-step forecasts, given by Equation (2.2), are shown below. Please note that alternative terminologies sometimes exist in the literature for these strategies.

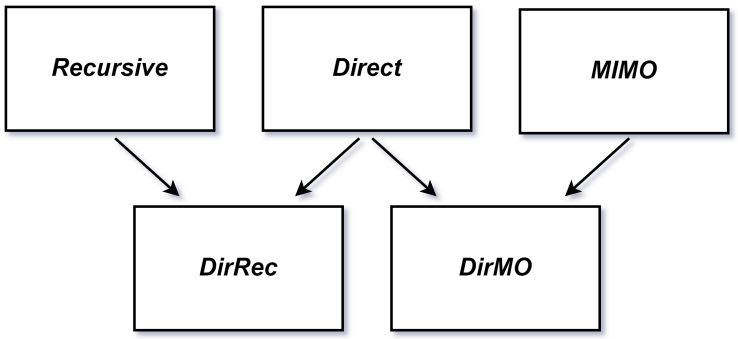


Figure 6.1: Strategies for multi-step forecasts.

6.1.1.1 Recursive Strategy

In the recursive strategy, a single model f is trained to perform a single-step forecast. The same pre-trained model f is then deployed iteratively for each time point in the forecasting horizon of multi-step forecasts (H). That is to say, the pre-trained model f first predicts the next time step (y_{t+1}), then the prediction output is fed into the same single-step model f as an input feature to predict the subsequent time point (y_{t+2}). This procedure is successively repeated until the last value of the desired multi-step forecasting horizon (y_{t+H}) is predicted. The main architecture of the recursive strategy is shown in Figure 6.2.

This strategy takes the conditional dependencies between the successive data points into consideration by using the prediction output of the previous time point as a part of the input variables for producing the next prediction. However, it may potentially yield inaccurate predictions as it is sensitive to the accumulation of forecast error (also known as forecast error propagation) throughout the forecasting horizon [152, 153].

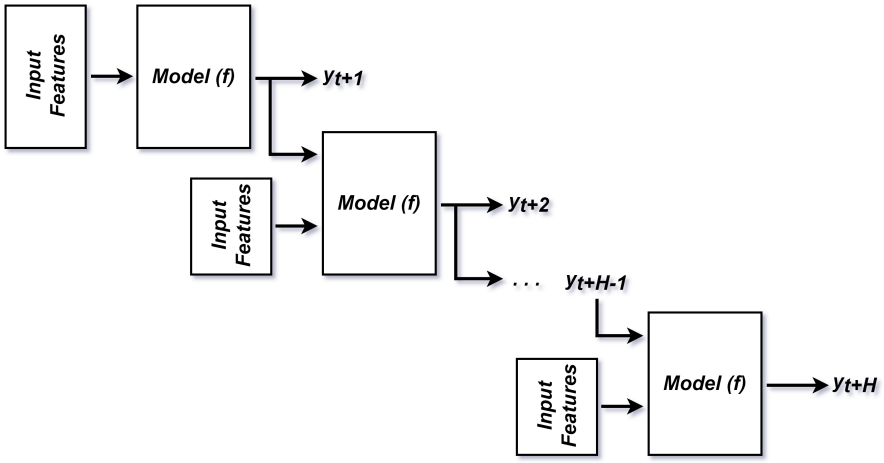


Figure 6.2: The architecture of the recursive strategy.

6.1.1.2 Direct Strategy

The purpose of the direct strategy is to develop independent forecasting models for each time point in the forecasting horizon of multi-step forecasts (H). For example, model f_1 is trained to predict y_{t+1}, model f_2 is trained to predict y_{t+2}, through to model f_H is trained to predict y_{t+H}. This strategy is immune to the forecast error accumulation that the recursive strategy suffers from, as it does not use any predicted value as an input variable for the next time step [135, 153]. However, since it requires multiple models and their independent training, it is computationally expensive [151, 152]. More importantly, every model and prediction output is completely independent of the others in this strategy. This means it ignores the statistical dependencies between data points in time-series data. The architecture of the direct strategy is illustrated in Figure 6.3.

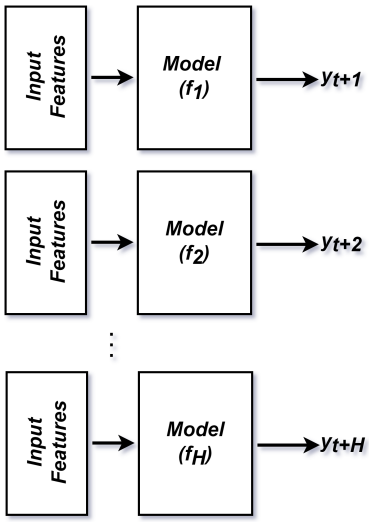


Figure 6.3: The architecture of the direct strategy.

6.1.1.3 Multi-Input Multi-Output Strategy (MIMO)

The MIMO strategy involves developing a single model f that is capable of predicting the entire forecast sequence in a one-shot manner. The previous two strategies can be considered single-output strategies since each model produces only one prediction output. However, in this strategy, the trained model f provides predictions for all the time points in the forecasting horizon (H) at once, as shown in Figure 6.4.

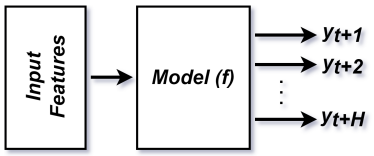


Figure 6.4: The architecture of the MIMO strategy.

This strategy preserves the complex dependencies between the consecutive data points in time-series data while also avoiding the accumulation of forecast error [152]. However, since there is only one fixed model yielding all the predictions at once, this approach might be deemed to be inflexible for the entire multi-step forecasting horizon [151].

6.1.1.4 DirRec Strategy

The DirRec strategy combines the architecture and principles of the direct strategy and the recursive strategy. That is to say, independent forecasting models are developed for each time point in the forecasting horizon (as in the direct strategy), and each model uses the very last predicted value produced for the previous time step (as in the recursive strategy) by another model. The architecture of this strategy is illustrated in Figure 6.5. The major disadvantages of this strategy are that it is costly in time and also computationally expensive [152].

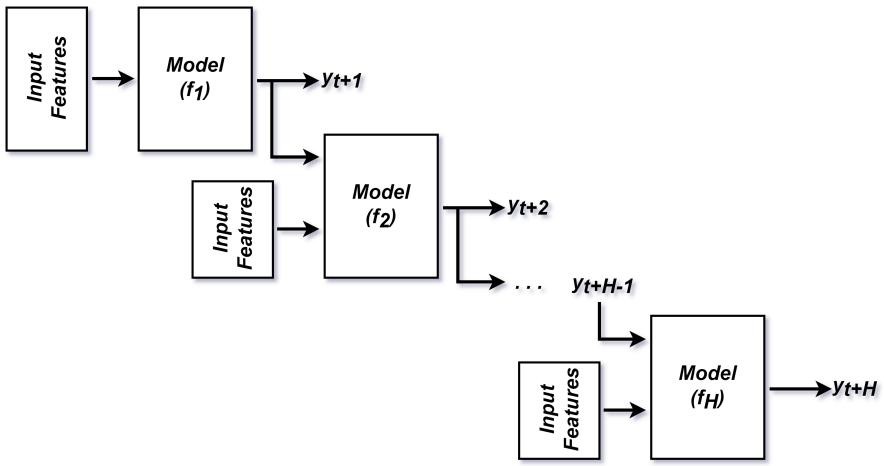


Figure 6.5: The architecture of the DirRec strategy, combining the direct strategy and the recursive strategy.

6.1.1.5 DirMO Strategy

The DirMO strategy is a combination of the architecture and principles of the direct strategy and the MIMO strategy. In the DirMO strategy, separate models are developed (as in the direct strategy) for sub-groups of the entire forecasting horizon (H). In other words, the forecasting horizon is decomposed into v sub-groups including the same number of time points (w) to be predicted, meaning that ($H = v \times w$). For each sub-group, an independent model yielding multiple outputs at a time is developed (as in the MIMO strategy). This architecture is shown in Figure 6.6.

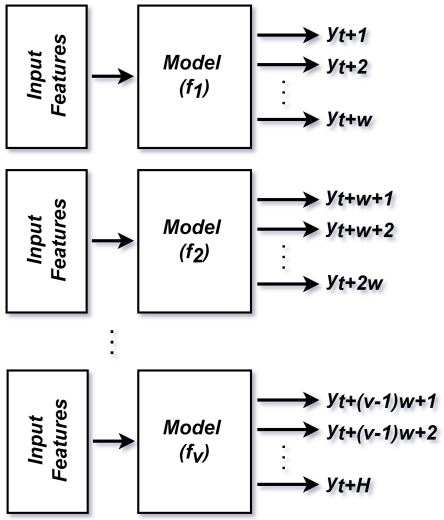


Figure 6.6: The architecture of the DirMO strategy, combining the direct strategy and the MIMO strategy.

When the value of v is 1, the model corresponds to the direct strategy, and when the value of w is 1, it corresponds to the MIMO strategy [151]. Finding the best configuration of v and w values can, therefore, be handled as an optimisation problem based on the requirements of the intended application and the problem domain.

6.2 Methodology

This section describes the experimental settings of the study that is carried out with a large-scale dataset and an contemporary prediction method. This is designed to give insight into how multi-step predictions are affected by volatility and irregularity in data, to illustrate how this effect differs from the PFE and why, and to explore the applicability of the n-SS method in the multi-step case.

6.2.1 Experimental Setup

Considering the purpose of the experimental study presented in this chapter, the study performs a one-day-ahead prediction, yielding 48 predicted values ($H = 48$) for the electricity consumption

of the next 24 hours. This uses household-level data from the same subset of the SGSC dataset that was also used in the study presented in the previous chapter. Using the same household-level electricity consumption data subset for this study as well is useful as the dataset: i) allows the PFE to be investigated in the context of electricity demand forecasts in this thesis; ii) provides data from numerous buildings, some of which are PFE-affected and some of which are PFE-free, and these buildings were already established in the previous chapter; and iii) the statistical regularity analyses of the data from individual buildings were determined in the previous chapter.

To deal with the complex multi-step household-level electrical load forecasting problem, this study considers three alternative strategies: the recursive strategy, the MIMO strategy, and the DirRec strategy. These three strategies are chosen as they are the only ones that consider maintaining the stochastic and intricate dependencies between successive data points, which is vital for obtaining robust and reliable predictions in the electricity consumption domain.

The same single-step prediction method, LSTM RNN, which is already described above in Chapter 5 (Section 5.2.1) is used here as well. For the recursive strategy, the LSTM RNN method is used with the same architecture and hyper-parameters as in the previous chapter. However, the method has to be adapted to work with the MIMO and DirRec strategies, though. With the MIMO strategy, LSTM RNN was trained to learn the daily consumption patterns as a whole starting from midnight (00:00 a.m.), so that it produces 48 predicted values at once, which is the forecast outputs for the next day. Similarly, the DirRec strategy produced independent models for each prediction output, so each was trained with data rows tagged with the corresponding

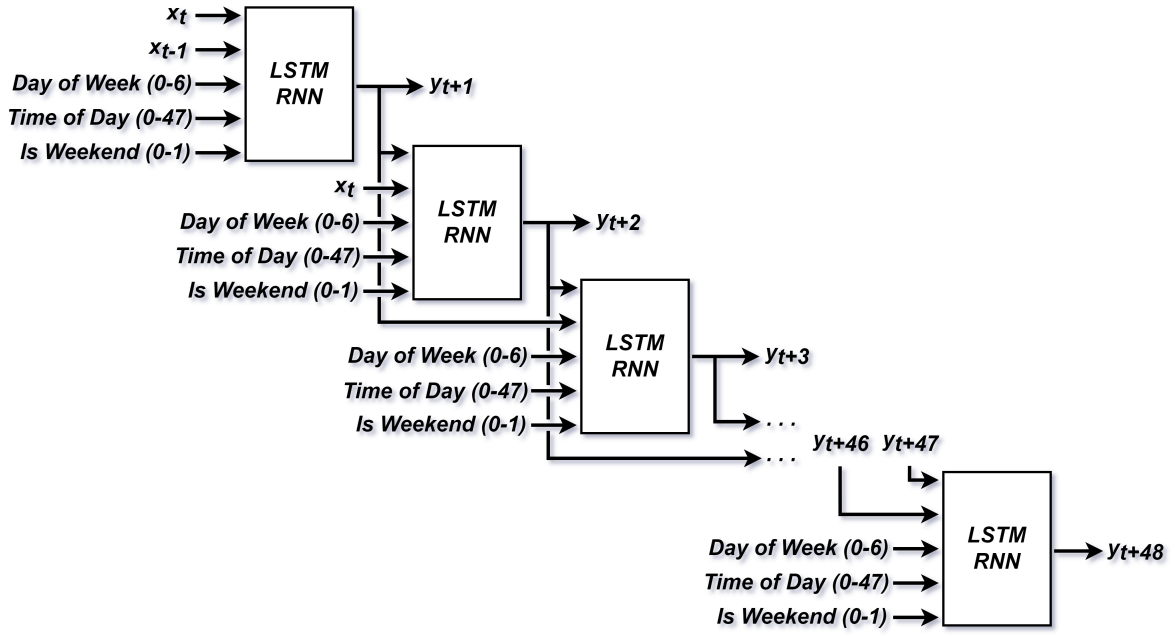


Figure 6.7: A diagram illustrating the architecture of the implemented recursive strategy with the input feature set.

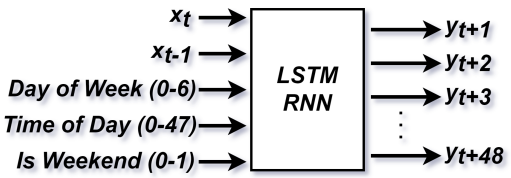


Figure 6.8: A diagram illustrating the architecture of the implemented MIMO strategy with the input feature set.

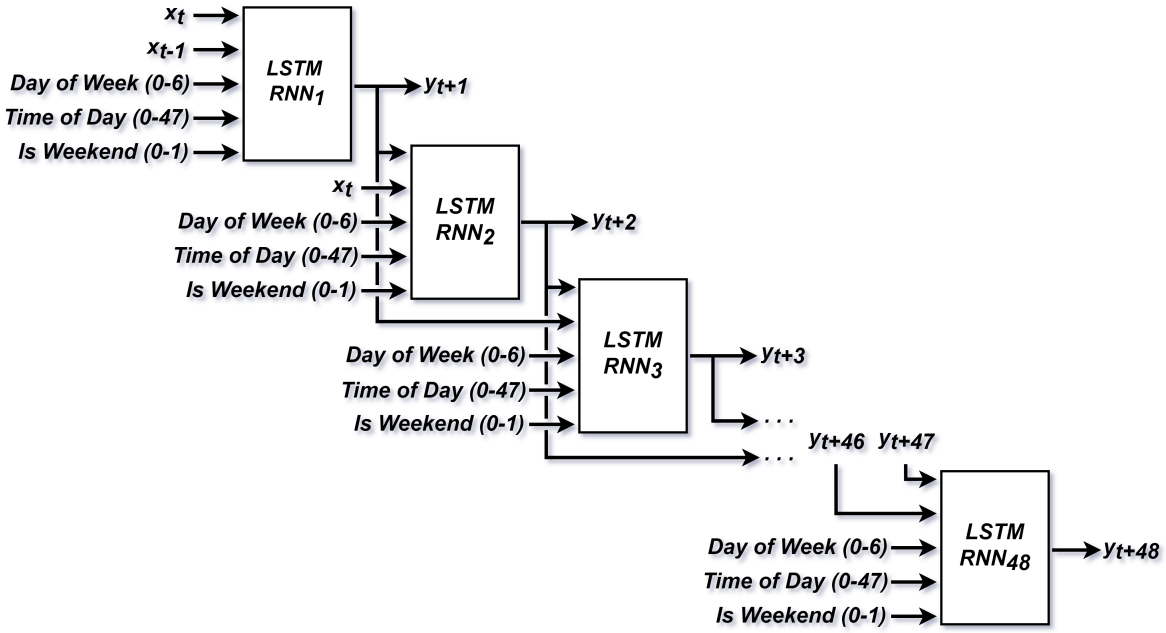


Figure 6.9: A diagram illustrating the architecture of the implemented DirRec strategy with the input feature set.

time stamps. For instance, the model LSTM RNN_{11}, which yields the predicted value of y_{t+11} for the time 05.30, is trained only with the rows having the time stamp "05.30". The architecture of these implementations of the recursive, MIMO, and DirRec strategies with the LSTM RNN method and the input feature set are represented in Figures 6.7, 6.8, and 6.9, respectively.

6.3 Results

This section presents the outcomes of the empirical study that is conducted with an observational electricity consumption data and a prediction method with the objective of exploring the influences of data volatility on multi-step forecasts and the practicability of the n-SS method implementation.

6.3.1 Prediction Results

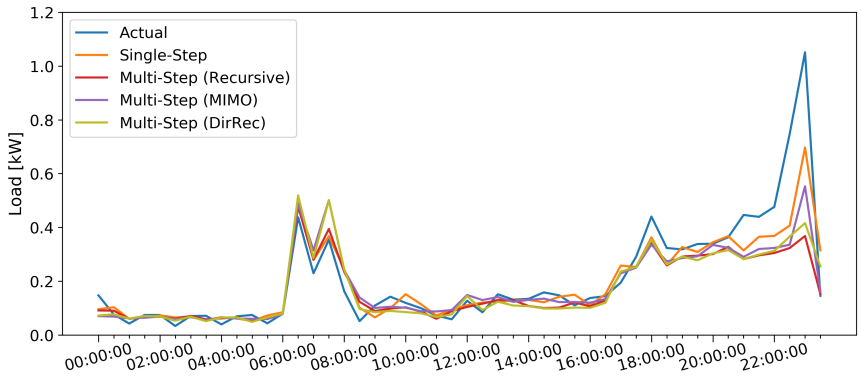
In order to generate day-ahead (midnight to midnight) multi-step forecasts, the method and strategies outlined above are applied to the electrical load data of 68 households, three of which are PFE-free and 64 of which are PFE-affected, as established in Chapter 5.

The multi-step forecast results of the three PFE-free households (Houses 8342852, 8273230, and 8482121) consistently showed no temporal displacement that might imply the presence of the PFE. This is, however, not the case with the data from the 64 PFE-affected households, in which the multi-step predictions curves are completely uncorrelated to the actual data curves. Neither the timings nor the amplitudes of a series of events match with the actual observations properly, resulting in a very poor alignment of predictions and accuracy. However, despite the fact that the predictions are clearly influenced by the irregularity in data, this effect is not in a way that they systematically and continuously follow the actual data one or a few steps behind in time. They are either a flat line with minimal fluctuations or a series of predictions full of arbitrary peaks and troughs.

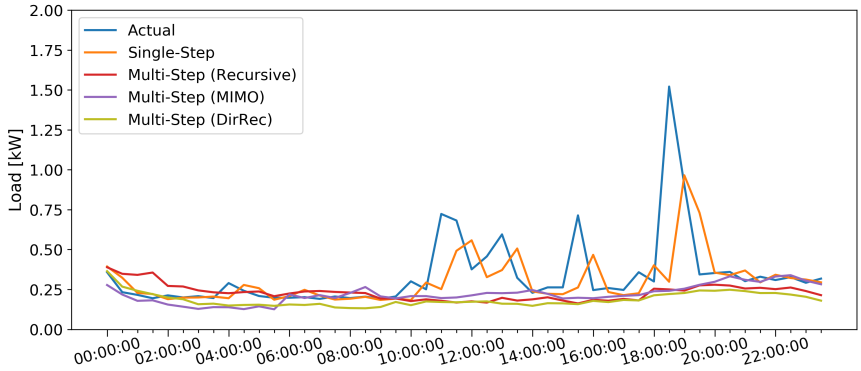
For example, Figure 6.10 shows the day-ahead multi-step forecasts for three households together with their single-step forecasts, which demonstrates their PFE status. One of these three households is randomly chosen from the three PFE-free households, and the remaining two are randomly chosen from the 64 PFE-affected households in such a way that both of the two possible uncorrelated cases are represented. Similarly, the days to be predicted are also chosen randomly and separately for each building from their test set. It is apparent in Figure 6.10 that the day-ahead prediction curves produced for each building by the three different multi-step strategies are quite similar to each other within the building, no matter the building is subject to the PFE or not. For House 8482121, whose predictions have previously been identified to be PFE-free in the single-step case (Figure 6.10a), the multi-step predictions curves show no temporal displacement and are almost identical to both the actual measurements curve as well as the single-step predictions curve. On the other hand, for Houses 8487285 and 8661542, whose single-step predictions have been calculated to be PFE-affected, it is hard to claim that the multi-step ahead predictions closely follow the actual load pattern or are systematically delayed in time. Instead, the multi-step predictions curves are either approximately flat line with minimal fluctuations (Figure 6.10b), or composed of seemingly arbitrary peaks and troughs that are dissimilar to the actual load patterns (Figure 6.10c). Neither of these outcomes is similar to the systematic delay in predictions observed in the single-step case.

As a consequence, although the high level of irregularity in the underlying data causes the PFE, a systematic and continuous delay, in single-step time-series forecasts, the way it affects multi-step forecasts is completely different; the effect now manifests itself as totally uncorrelated and dissimilar predictions rather than an explicit delay, regardless of the multi-step strategy implemented. The difference between multi-step predictions and actual load values is so

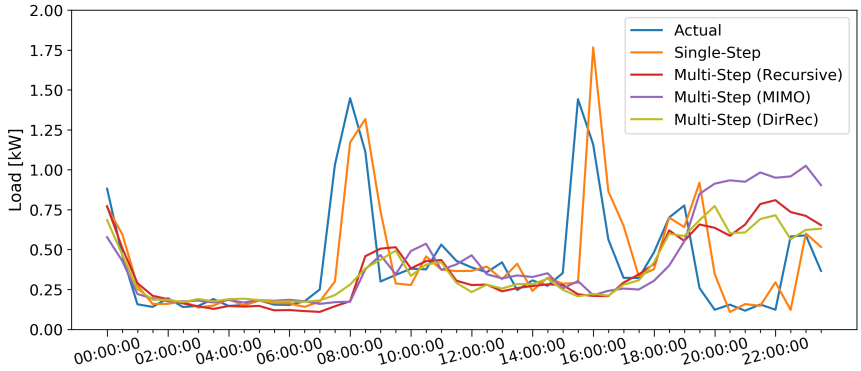
significant that it does not seem appropriate to make a determination regarding the absence or presence of the PFE or temporal displacement in such cases.



(a)



(b)



(c)

Figure 6.10: Multi-step (day-ahead) load predictions using the recursive, MIMO, and DirRec strategies for: (a) a PFE-free building, House 8482121; and PFE-affected buildings, (b) House 8487285 and (c) House 8661542. Each plot also shows the corresponding single-step forecast. The multi-step predictions temporally align with the actual observations in the PFE-free building. However, they exhibit no correlation with the observation curve in the PFE-affected buildings.

6.3.2 Evaluation of the Usefulness of the n-SS Method

The n-SS method proposed earlier in this thesis for the identification of the presence of the PFE basically relies on how the metric results change when the predictions are shifted a few steps back in time. However, as described in Chapter 4 (Section 4.3), in the cases where there is no similarity between predictions curve and observations curve, applying the n-SS method would end up with either an increase or decrease in metric results at random. However, these changes in metric results were meaningless and could not be used to indicate the existence or absence of the PFE. Similarly, the n-SS method can also said to be inappropriate for detecting the effect of data irregularity in uncorrelated multi-step forecasts such as those in Figures 6.10b and 6.10c. Nevertheless, it might be possible for the n-SS method to contribute to confirming the absence of the PFE as PFE-free predictions closely follow the actual measurements curve and are as temporally accurate as single-step forecasts (see Figure 6.10a).

The n-SS results of the multi-step forecasts of House 8482121 data are presented in Table 6.1 with $n = 1$ and $n = 2$. The n-SS method significantly worsens the accuracy of the multi-step predictions of the electrical load of this building by shifting the predictions one or two time steps back, i.e., the RMSE and MAPE values increase. This suggests that the multi-step predictions of this household are not affected by irregular underlying data, and so they are PFE-free. On the other hand, the application of the n-SS method to the PFE-affected data from Houses 8487285 and 8661542 results in only negligible increases or decreases in evaluation metric values. These metric results are listed in Tables 6.2 and 6.3, respectively.

Strategy	**Pred. Type**	**RMSE**	**MAPE**
Recursive Strategy	Orig. Pred.	0.127	26.679
	1-SS Pred.	0.166	44.242
	2-SS Pred.	0.192	59.592
MIMO Strategy	Orig. Pred.	0.108	26.853
	1-SS Pred.	0.160	43.502
	2-SS Pred.	0.190	61.018
DirRec Strategy	Orig. Pred.	0.122	28.002
	1-SS Pred.	0.158	49.285
	2-SS Pred.	0.190	61.446

Table 6.1: Comparison of the RMSE and MAPE evaluation metric results of original and the n-SS method applied (with $n = 1$ and $n = 2$) predictions of electricity load data from House 8482121 (Figure 6.10a), whose predictions do not exhibit the PFE in single-step case.

As a consequence, while the n-SS method is still applicable and useful to confirm the absence of the PFE in multi-step forecasts, it cannot be used to detect the affected multi-step forecasts effectively, as it does not seem appropriate to discuss the temporal accuracy of predictions that are notably uncorrelated to observed values.

Strategy	Pred. Type	RMSE	MAPE
Recursive Strategy	Orig. Pred.	0.266	31.865
	1-SS Pred.	0.263	30.085
	2-SS Pred.	0.260	29.337
MIMO Strategy	Orig. Pred.	0.257	26.749
	1-SS Pred.	0.255	26.389
	2-SS Pred.	0.253	27.349
DirRec Strategy	Orig. Pred.	0.274	35.423
	1-SS Pred.	0.272	35.196
	2-SS Pred.	0.271	35.372

Table 6.2: Comparison of the RMSE and MAPE evaluation metric results of original and the n-SS method (with $n = 1$ and $n = 2$) applied predictions of electricity load data from House 8487285 (Figure 6.10b), whose predictions suffer from the PFE in single-step case.

Strategy	Pred. Type	RMSE	MAPE
Recursive Strategy	Orig. Pred.	0.396	75.328
	1-SS Pred.	0.379	75.116
	2-SS Pred.	0.372	77.946
MIMO Strategy	Orig. Pred.	0.453	99.340
	1-SS Pred.	0.442	97.036
	2-SS Pred.	0.436	97.842
DirRec Strategy	Orig. Pred.	0.384	71.940
	1-SS Pred.	0.371	68.217
	2-SS Pred.	0.365	69.011

Table 6.3: Comparison of the RMSE and MAPE evaluation metric results of original and the n-SS method (with $n = 1$ and $n = 2$) applied predictions of electricity load data from House 8661542 (Figure 6.10c), whose predictions suffer from the PFE in single-step case.

6.4 Discussion

Given that single-step forecasts struggle to yield reliable or robust predictions from electricity load demand datasets with high variability, as discussed in the previous chapter, it is not surprising that multi-step forecasts also fail in these cases. However, unlike the systematic and continuous delays seen in single-step forecasts, the affected multi-step forecasts are discovered to be very different from actual data, appearing in plots either as a flat line with minimal fluctuations (Figure 6.10b) or as a series of arbitrary peaks and troughs (Figure 6.10c). In both cases, they are essentially uncorrelated to the actual observations, and give significantly poor accuracy ultimately. The uncorrelated prediction outputs in multi-step cases seem to arise because of the volatility and pattern irregularity in the underlying data; prediction methods cannot learn enough about the past and also cannot extrapolate the predictions from one of the most recently observed

values as these values are not available for the majority of the time points to be predicted. The behaviour of the affected multi-step predictions causing a flat line can be explained by the model reproducing the same output value throughout the prediction time span. That is, each prediction output approximates the prediction output of the previous time step, resulting in an almost flat line oscillating around the most recently observed value (the value of x_i) throughout the forecasting horizon. However, further analysis is required to understand the reasons for the arbitrary peaks and troughs appearing in the other scenario of affected multi-step predictions.

The proposed n-SS method, which has been shown in the previous chapter to discriminate the presence of the PFE in single-step forecasts extremely well, seems to be insufficient to detect the affected multi-step predictions, which leaves this as an area of further work. The central idea for the n-SS method is the recalculation of standard evaluation metrics after shifting the predictions *a few steps back* in time (shift the predictions to the past) to identify if the predictions are delayed and if so, to determine how many time steps they are delayed. However, since the affected multi-step predictions are completely uncorrelated to the actual observations and there is no similarity between predictions curve and observations curve, the n-SS method does not work properly regarding the determination of whether or not the multi-step predictions are affected. Nonetheless, the n-SS method can effectively be used to show the absence of the PFE in multi-step forecasts anyway and can thus be used to give confidence in evaluation outcomes to stakeholders.

6.5 Conclusion

The main aim of this chapter was to study how the irregular underlying time-series data, which causes the PFE in single-step forecasts, affects multi-step forecasts, how this effect manifests itself, and how useful the n-SS method is in multi-step cases. To this end, an empirical study based on observational household-level electricity consumption data, a state-of-the-art machine-learning method, and three different multi-step strategies has been presented, and its results have been analysed in this chapter.

It has been seen that multi-step predictions that are unaffected by underlying data irregularity are closely aligned with the actual data and show particularly good temporal accuracy. However, this is not the case for the PFE-affected households. In multi-step cases, the use of data with irregular and inconsistent patterns ends up with predictions that are irrelevant and uncorrelated to actual data curves rather than predictions that follow the actual data a few steps behind. The PFE was defined in Chapter 3 as a series of predictions that is almost identical to the series of actual measurements but is slightly delayed in time. However, this does not describe the effect observed on multi-step forecasts when they are affected by irregularity in the data. As such, the term *PFE* is no longer appropriate terminology in this context. Moreover, since there is no similarity between the affected multi-step forecasts curve and the actual observations curve, the n-SS method appears to be unable to identify the affected multi-step predictions. However, it

is still able to demonstrate the multi-step predictions are PFE-free when the predictions are not affected by the data irregularity and volatility, which could be favourable for stakeholders and some sort of smart applications.

The next chapter will further investigate more fundamental questions about the PFE and its characteristics through various experimental studies. It will consider some relevant questions such as whether the PFE is an issue specific to the electricity consumption forecasting domain, whether the delay in time-series predictions caused by the PFE is always one time step, how the training set length and data granularity affect the PFE, and some others.

CHAPTER 7

FURTHER DETAILS ON THE PERSISTENCE FORECAST EFFECT

The preceding chapters have first defined the PFE and then proposed the n-SS method for the detection of the PFE in time-series forecasts. Subsequently, comprehensive empirical studies have been provided in order to investigate the PFE and the implementation of the n-SS method in both single- and multi-step forecasts in the context of electricity consumption forecasting.

This chapter will now carry out comprehensive analyses to gain a deeper understanding of the PFE and provide further details regarding its characteristics. To this end, the work in this chapter will address a set of relevant research questions to interrogate the various aspects of the PFE:

- *Question 1:* How does the number of most recent observations in the input feature set affect the presence of the PFE?
- *Question 2:* Can a time-series prediction have both PFE-affected and PFE-free predictions together?
- *Question 3:* How do the training set length and data granularity affect the PFE?
- *Question 4:* Can the PFE be attributed to prediction methods instead of the characteristics of the underlying data?
- *Question 5:* Is the PFE an issue specific to the electricity consumption forecasting domain?
- *Question 6:* Is the delay in time-series predictions caused by the PFE always one time step?
- *Question 7:* Is there an existing evaluation metric that is potentially resilient to the PFE?

The purpose of this chapter is to address the questions posed above and to explore the reasons behind the answers. Furthermore, it will present some relevant experimental studies,

the results of which will serve as evidence for or against the answers to the questions. However, a time-series dataset whose features are sufficient to properly and comprehensively answer all these questions exploring different characteristics of the PFE is not available. Therefore, the experiments provided in this chapter are conducted using a variety of time-series datasets with different features. It is of note that these datasets have previously undergone preliminary analyses, which confirmed them to be subject to the PFE.

7.1 How Does the Number of Most Recent Observations in the Input Feature Set Affect the presence of the PFE?

The main drivers of the PFE are high volatility and insufficient regularity in the underlying data. When the underlying data involves volatility and irregularity, prediction methods, which mostly rely on regularity in the corresponding data, cannot learn from the data. They rather learn the superior correlation between the successive observations from the output domain. As a result, the produced predictions approximate one of the most recently observed values used in the input feature set. Overall, in single-step forecasts, predictions trail the actual observations one or a few steps behind in time, depending on the most recent observation approximated.

As has been discussed in Chapter 5 (Section 5.4), having a certain number of most recent observations brings substantial improvement in prediction accuracy, and if no recently observed value is used in the input feature set, prediction methods fail to yield accurate and correlated predictions. Therefore, it can be concluded that removing the most recent observations from the input feature set is not a pragmatic approach to avoiding the PFE. However, what about incorporating a greater number of recent observations for feeding prediction methods in order to prevent predictions from the PFE or at least to alleviate the effect of the PFE?

7.1.1 Methodology

The experiment to determine how an increased number of most recent observations in the input feature set affects the existence of the PFE replicates the test conducted with the two most recent observations ($K = 2$) in Chapter 5, but with an input feature set that includes a greater number of most recently observed values this time. To explore the effect of the number of most recent observations on the occurrence of the PFE, four different scenarios are built with the 6, 12, 24, and 48 most recent observations in the input feature set. The input feature sets of these scenarios are as follows:

- Electricity load data recordings of the K most recent time steps; where $K \in \{6, 12, 24, 48\}$.
- Day-of-week indicator (ranges from 0 to 6).
- Time-of-day indicator (ranges from 0 to 47).
- Weekend indicator (ranges from 0 to 1).

In order to mitigate the impacts of varying scales of the input features and anomalies in data, one-hot encoding is applied to time-of-day and day-of-week data, and min-max normalisation is performed for the most recent electricity consumption observations.

These scenarios are tested using the LSTM RNN (defined in Chapter 5) on electricity consumption data from two randomly selected households that have previously been diagnosed to suffer from the PFE in Chapter 5 when using the two most recent observations in the input feature set ($K = 2$). These households are Houses 8196621 and 8733828 from the subset of the SGSC dataset, and predictions yielded for these buildings have been calculated to be delayed one time step by the 1-SS method.

7.1.2 Results

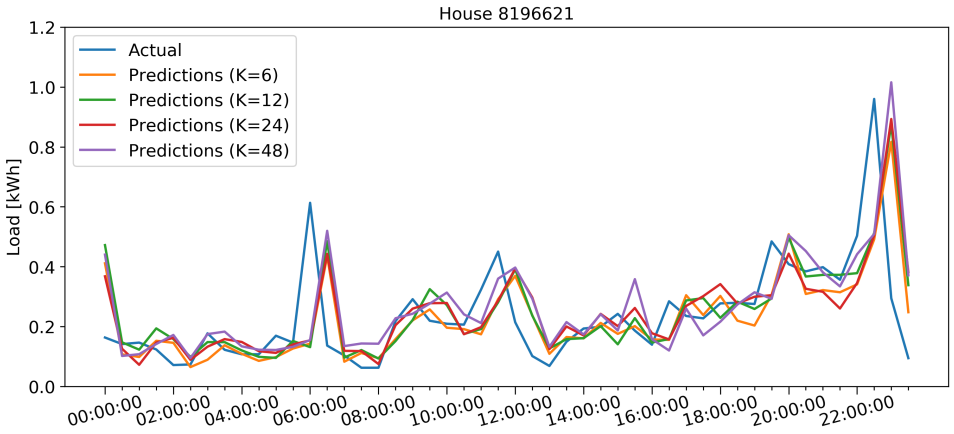
The prediction results of the two chosen households are visualised in Figure 7.1 with different K values, and Table 7.1 gives the metric results of the original and the 1-SS method applied predictions side-by-side. Please note that Figure 7.1 illustrates just 24 hours of a nine-day-long test set for the sake of clarity alone. However, the metrics listed in Table 7.1 have been calculated across the entire time span that constitutes the test set.

House ID	**Metric**	$K = 6$	$K = 12$	$K = 24$	$K = 48$
8196621	MAPE vs. MAPE*	37.185 ↓ 22.904	38.788 ↓ 25.027	36.704 ↓ 26.257	43.517 ↓ 33.623
	RMSE vs. RMSE*	0.162 ↓ 0.094	0.162 ↓ 0.110	0.159 ↓ 0.097	0.170 ↓ 0.113
	Corr vs. Corr*	0.726 ↓ 0.918	0.720 ↓ 0.888	0.735 ↓ 0.917	0.697 ↓ 0.874
8733828	MAPE vs. MAPE*	32.120 ↓ 21.572	33.646 ↓ 23.856	33.747 ↓ 25.840	35.302 ↓ 26.839
	RMSE vs. RMSE*	0.127 ↓ 0.072	0.126 ↓ 0.082	0.127 ↓ 0.083	0.134 ↓ 0.086
	Corr vs. Corr*	0.814 ↓ 0.950	0.817 ↓ 0.938	0.812 ↓ 0.932	0.793 ↓ 0.920

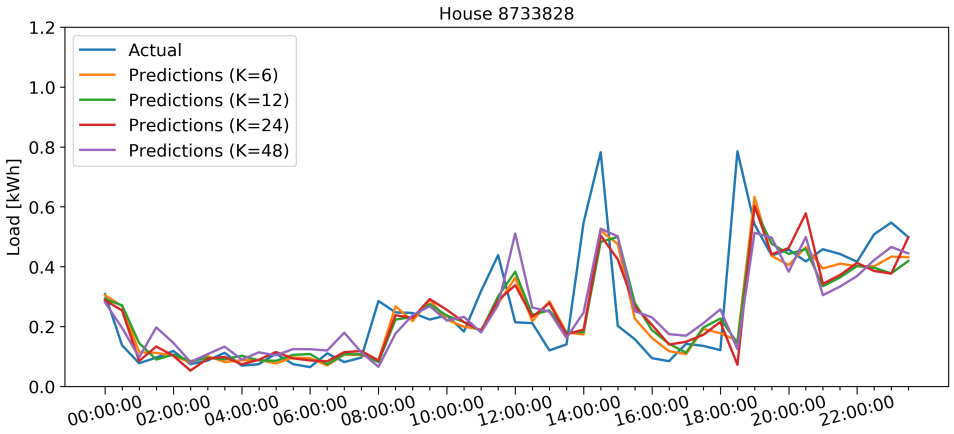
Table 7.1: Default and the 1-SS method applied accuracy metric results of two households with different numbers of most recent observations in the input feature set: $K = \{6, 12, 24, 48\}$

It is evident from Figure 7.1 that all the predictions are temporally aligned with each other for all values of K, representing the number of most recent observations in the input feature set. However, regardless of the value of K, all of the predictions systematically miss the timing of the actual peaks and troughs and fall back of the actual data. Besides that, the evaluation metric results listed in Table 7.1 indicate that applying the 1-SS method substantially improves the accuracy metric results, pointing out that the original predictions are delayed one step in time, regardless of the number of most recent observations in the input feature set.

These results show that, within the parameters used, feeding prediction methods with a higher number of most recent observations does not help to avoid the PFE or alleviate its effect. If the underlying data contains irregular and volatile patterns, the PFE status does not appear to be affected by how many most recent observations are included in the input feature set. Moreover, regardless of the PFE, increasing the number of most recent observations in the input feature set does not appear to improve accuracy.



(a)

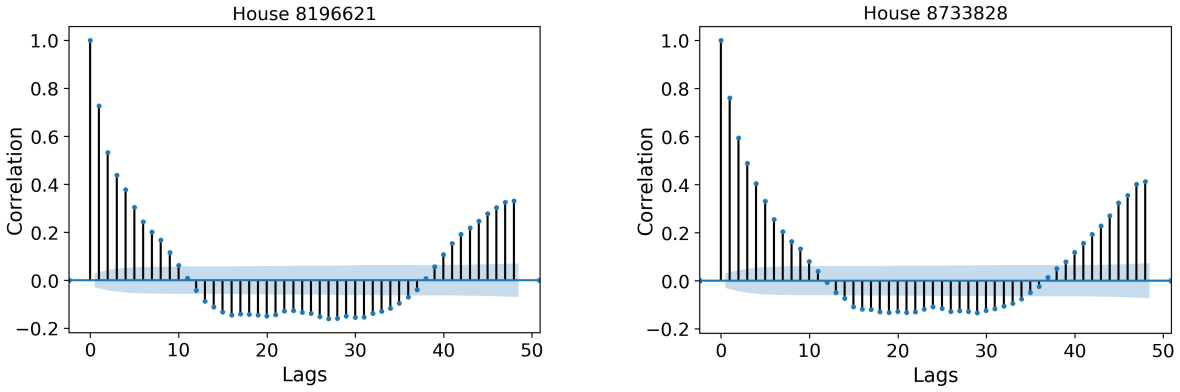


(b)

Figure 7.1: Illustration of the electricity load demand prediction results of (a) House 8196621 and (b) House 8733828 throughout a day with different numbers of most recent observations in the input feature set: $K = \{6, 12, 24, 48\}$

7.1.3 Discussion

Given that the PFE occurs mainly due to the inconsistency and irregularity in the underlying time-series data and that the amount of delay is determined by the superior auto-correlation value, as explained in Chapter 3 (Section 3.2), it is not surprising that the increased number of most recent observations in the input feature set does not help to prevent the predictions from the PFE. This is because the newly added most recent observations do not contribute to improving the data regularity. Besides that, Figure 7.2 shows that the auto-correlation value at lag 1 still remains dominantly higher than the auto-correlation values at all the other lags. The learning of prediction methods is, therefore, not affected or boosted by the newly added most recent observations, and so the predictions keep following the actual load one step behind.



(a) (b)

Figure 7.2: Auto-correlation analysis of two PFE-affected households: (a) House 8196621 and (b) House 8733828.

7.2 Can a Time-Series Prediction Have Both PFE-Affected and PFE-Free Predictions Together?

While analysing the regularity of the data through clustering, Chapter 5 has considered the correlation values between the daily load patterns of households in order to find out how regular daily load patterns each of the 68 households has. Besides that, to explore whether or not predictions of a household are affected by the PFE, the evaluation metrics (default and the n-SS method applied) are calculated over the entire nine-day-long test set. However, this analysis has overlooked what might happen if a household's electrical power consumption was regular between, say, 05:00 and 10:30 but irregular from, say, 13:30 to 18:00. This would require separate clustering analysis and the n-SS method application for each of these periods. Otherwise, in such a case, there is a risk of losing crucial information on the regularity of data and the occurrence of PFE. This possibility motivates the investigation of whether a time-series prediction can have different periods with varying PFE results based on the data regularity of the corresponding period of an underlying time-series.

7.2.1 Methodology

The electricity consumption data gathered from House 8478501 from the subset of the SGSC dataset is used here. This household data was chosen as it was the only household whose predictions have been calculated to be PFE-inconclusive in Chapter 5. Inconclusive PFE investigation refers to a conflict between different evaluation metrics regarding the existence of the PFE – some metrics suggest there is PFE in predictions whilst the other metrics suggest the predictions are PFE-free (also defined by Equation 5.1). Therefore, investigating the PFE across smaller periods of time independently for this building might reveal why there is a conflict between the evaluation metrics.

This PFE in different periods of a time-series study deploys the same LSTM RNN that is used in the previous section. The predictions are produced for the entire test set as usual. However, for evaluation of the existence of the PFE and clustering analysis, a day is divided into three eight-hour-long periods (see Figure 7.3):

1. Morning period (00:00 - 08:00)
2. Daytime period (08:00 - 16:00)
3. Evening period (16:00 - 00:00)

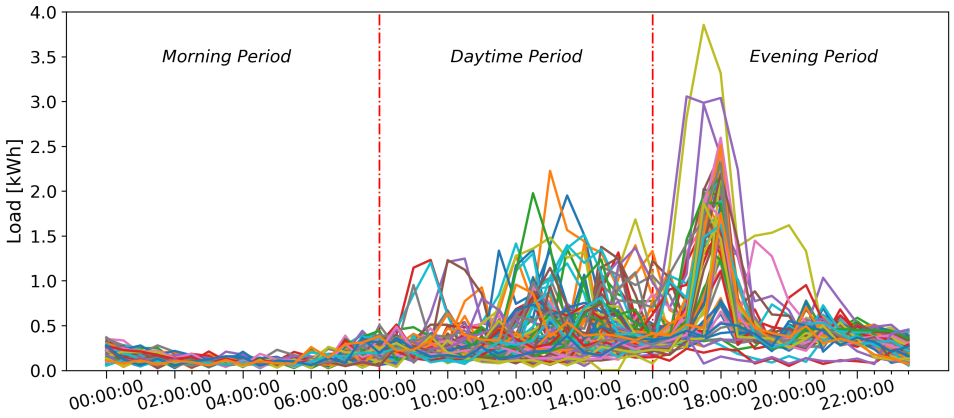


Figure 7.3: 92 daily electrical load profiles of the only PFE-inconclusive household (House 8478501) over the three eight-hour-long periods of a day.

However, the n-SS method cannot be effectively applied when there is not enough signal in the data, as explained in Chapter 4 (Section 4.3). Therefore, considering the limited energy use in the morning period in this household's data, the n-SS method is not applied to the morning period. The focus will be more on the daytime and evening periods in this study. That is, the PFE is investigated during these two eight-hour-long periods separately. Then, their hierarchical clustering results are compared to determine whether there is any difference between the regularity levels of these two periods.

7.2.2 Results

The accuracy metric results of the original and the 1-SS method applied predictions for the daytime and evening periods of House 8478501 data are compared in Table 7.2. These evaluation metrics are calculated over the daytime and evening periods separately over the nine-day-long test set. When the 1-SS method is applied independently for these two periods, the metric results of the daytime predictions significantly improve while the evaluation metric results of the evening predictions get worse. This indicates that there is no PFE in the predictions for the evening period but that the predictions for the daytime period are affected by the PFE. Figure 7.4 illustrates the

Metric	Daytime Period	Evening Period
RMSE vs. RMSE*	0.167 ↓ 0.104	0.227 ↑ 0.367
MAPE vs. MAPE*	47.402 ↓ 35.724	56.110 ↑ 64.929
Corr vs. Corr*	0.493 ↑ 0.811	0.849 ↓ 0.585

Table 7.2: Default and the 1-SS method applied metric results of the daytime and evening periods of the only household (House 8478501) with PFE-inconclusive electrical load predictions.

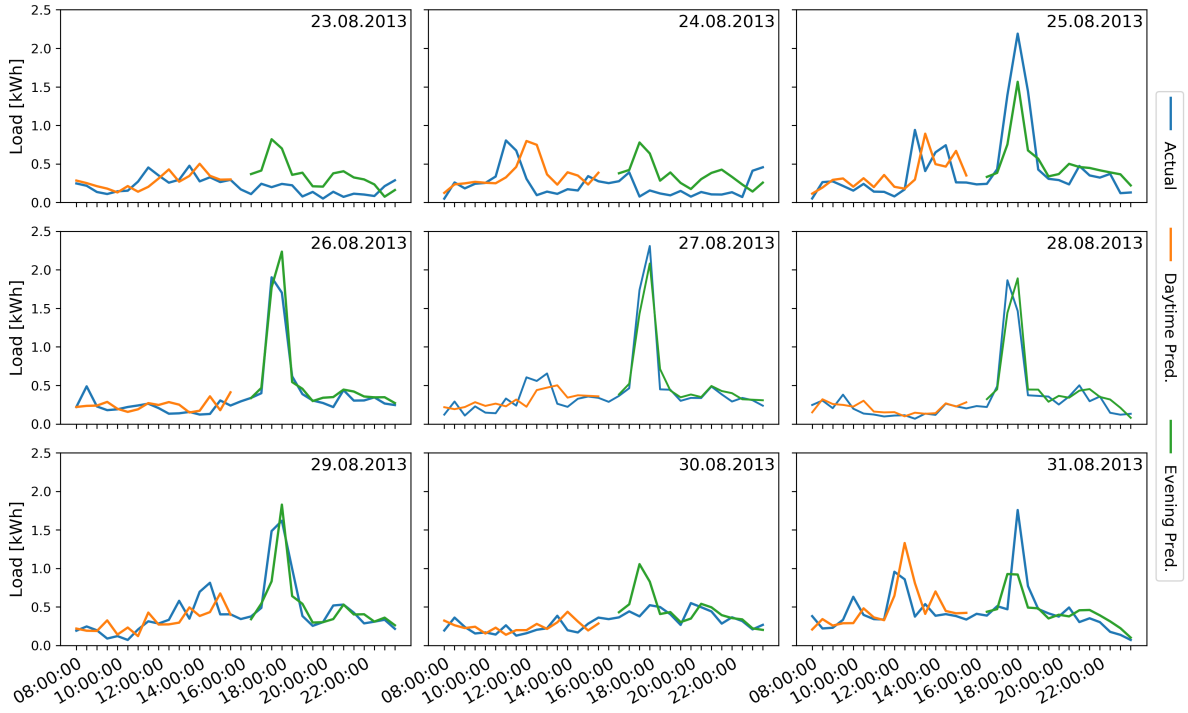


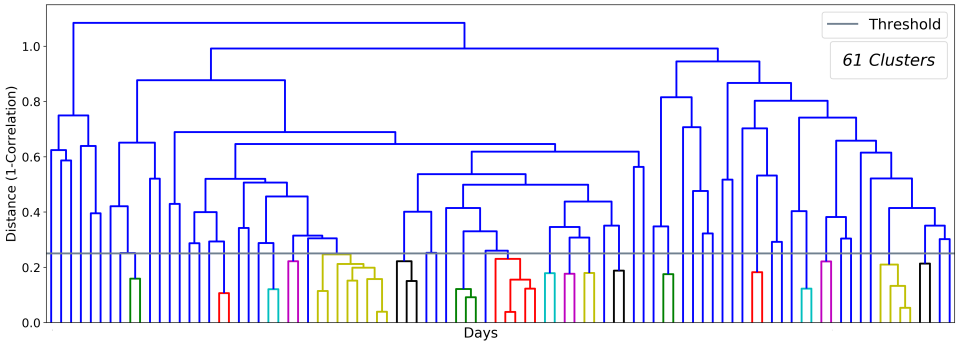
Figure 7.4: Prediction results of the daytime and evening periods of the only household (House 8478501) with PFE-inconclusive electrical load predictions. The predictions for the daytime period exhibit PFE, whilst the predictions for the evening period are PFE-free.

prediction outputs of the daytime and evening periods side-by-side for each day in the test set. A one step delay in the predictions of the daytime period is recognisable through visual inspection. The predictions of the evening period, on the other hand, are much better temporally aligned with the actual observations, consistent with no PFE being present.

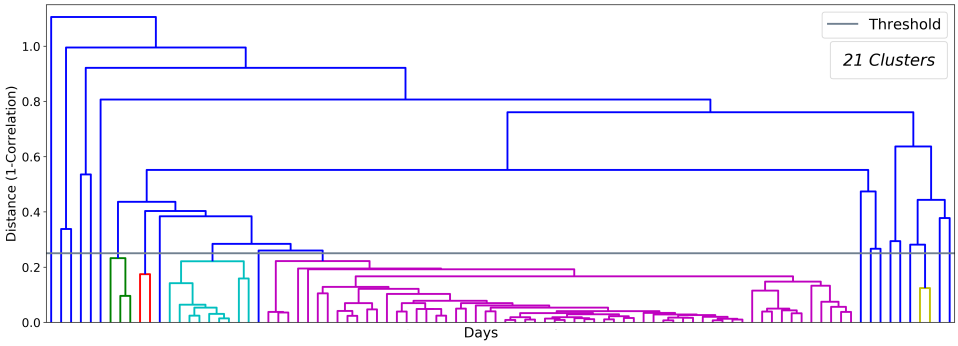
Predictions produced for the entire nine-day-long test set of data from House 8478501 have been calculated to be PFE-inconclusive in Chapter 5. This has been concluded from the 1-SS method application, causing the MAPE value to improve significantly but the RMSE and Correlation values to deteriorate. However, when the prediction outputs for each period are worked separately, and the 1-SS method is applied independently for each period, the results show

that there is no PFE in the predictions yielded for the evening period, whereas the predictions produced for the daytime period are affected by the PFE.

The results of clustering analysis on the partial-day data are shown in Figure 7.5. These shed light on why the predictions of the daytime period suffer from the PFE whereas the predictions of the evening period are PFE-free. Over 92 days, there are 61 distinct clusters for the electricity consumption taking place in the daytime period of the PFE-inconclusive household. However, for the evening period, the number of clusters is only 21. This means that dwellers living in this household have a considerably more consistent and self-similar electrical power consumption routine after 16:00 every day compared to their consumption patterns between 08:00 and 16:00. As a result of the irregularity and volatility in the daytime period, single-step predictions of this period are PFE-affected and trail the actual load data one step behind in time. However, since the evening period has considerably more regular patterns over the 92 days, single-step forecasts of this period are PFE-free and better temporally aligned with the actual load data.



(a)



(b)

Figure 7.5: Dendrograms showing hierarchical clustering results of (a) daytime and (b) evening periods of the PFE-inconclusive household (House 8478501). Every colour group, with the exception of blue, corresponds to a distinct cluster, while the blue lines indicate outlier daily profiles. The clustering results reveal a total of 21 distinct patterns during the evening period, while the daytime period exhibits 61 distinct patterns. This quantitatively confirms the difference in the regularity of these two periods.

7.2.3 Discussion

The presented study, conducted with the electricity consumption data that have previously been identified as PFE-inconclusive, has shown that time-series predictions can have different time periods with varying PFE results. This is due to changes in the regularity level across different periods of the time-series data. However, this does not mean this interesting phenomenon – partial PFE – can only take place within PFE-inconclusive data, as in the case of this household. Even if the prediction series as a whole is identified as PFE-free by the deployed accuracy metrics, there might still be shorter time periods where predictions are PFE-affected, or vice versa. Therefore, based on the final application specifications, it would be valuable to divide the data into shorter periods and investigate the presence of the PFE on these shorter periods independently to obtain more accurate and precise PFE results and to understand more about how the PFE changes and creates an overall response of a whole dataset.

In the presented study, days have been divided into three periods with an equal number of observations, so each eight-hour-long period has 16 time steps. However, in accordance with the time-series domain and the ultimate aim of the final application, the number of periods and the number of time steps within each period could be adjusted further. Aside from that, each period is not required to have an equal number of time steps and can be changed depending on the nature of the investigation or application. Clustering-like methods could be modified and used to determine the split-off point(s) of shorter periods. Moreover, clustering results of these shorter periods could be helpful in judging the probability of the occurrence of the PFE in these shorter periods in advance. However, since this is beyond the scope of this thesis, such works could be regarded as a line of future work.

7.3 How Do the Training Set Length and the Data Granularity Affect the PFE?

It is clear that avoiding or removing the PFE in predictions requires the removal of data irregularity or improving the regularity of data. The data length and data granularity often have an impact on the regularity of time-series data [154]. Therefore, training prediction methods on longer or shorter data and using less granular (coarser) or more granular (finer) data could help to restore the desired data regularity.

The idea of training prediction methods on a longer training set that provides more about the past may initially seem beneficial as it increases the likelihood of having recognisable patterns in data. However, this also means prediction methods have a lot of information to process and learn, which can potentially confuse the methods and make the learning process more challenging. In such cases, a shorter training set might be preferable. Besides that, a shorter training set is to

contain and provide more relevant patterns to the current. On the other hand, the drawback could be a shorter training set being inadequate to provide forecasting methods with sufficient information for proper learning. Therefore, using lengthier data to train forecasting methods might increase the likelihood of having similar patterns in data. This vicious circle suggests that the optimum length of a training set probably depends on the nature of the primary reason behind the irregularity in data. As a result, in theory, tuning training sets to be longer or shorter could be useful to avoid the PFE.

The data granularity refers to the length of the data sampling interval and is an important factor in the learning ability of prediction methods [155]. A shorter interval results in a finer granularity, while a longer interval results in a coarser granularity [43]. Fine-grained data might be able to provide more details, which could help the prediction methods deal with existing irregularity in the data. Coarse-grained data, however, can smooth the patterns in the data, which can end up with improved data regularity. As a consequence, in theory, both finer and coarser granularity could help to avoid the PFE, but in different ways and for different reasons.

Considering the above-explained potential impacts of the training set length and the data granularity level on the learning of prediction methods, this section seeks to understand how the training set length and data granularity affect the PFE. This will be done by analysing the PFE in predictions produced by a prediction method trained on data with different lengths and different sampling intervals.

7.3.1 Methodology

In order to test the impact of the training set length and the data granularity on the PFE in time-series predictions, The Almanac of Minutely Power dataset Version 2 (AMPds2) [156] is used. This dataset provides the electricity consumption data of a residential house located in Vancouver, Canada. The full details of this publicly available dataset have been documented in [157]. For this experiment, the AMPds2 dataset is selected over the SGSC dataset due to its extensive time span of two years (from 01.04.2012 to 01.04.2014) and high-resolution electricity consumption data at one-minute intervals, which enables manipulation of the length of the training set and the data granularity to create different scenarios.

The AMPds2 dataset provides 11 measurement characteristics for various circuits and appliances all across the house. Among these measurement characteristics and circuits, the *apparent energy* data, which represents the total electrical energy consumption of the whole residence, was used here. The apparent energy data provided by the dataset is cumulative, so energy consumption data for each minute was calculated by taking the difference between two successive energy values. Thereafter, any missing values are completed through interpolation to obtain a complete dataset in one-minute intervals of electricity consumption data from a Canadian household spanning over two years.

For this test, scenarios with three different data granularity levels and four different training set lengths are created. For the data granularity part, one-minute interval data are aggregated into 30-, 60-, and 120-minute intervals data. For the training set length part, the 30-minute interval data are sliced down to 1-, 3-, 6-, and 12-month-long data.

Afterwards, these different scenarios are tested by implementing the LSTM RNN method, with architecture and hyper-parameters as already described in the earlier chapters here. The input feature set used for feeding the LSTM RNN, regardless of the data granularity and the data length, can be summarised as follows:

- Electricity load data recordings of the two most recent time steps (x_t, x_{t-1}).
- Day-of-week indicator (ranges from 0 to 6).
- Time-of-day indicator (ranges from 0 to 47).
- Weekend indicator (ranges from 0 to 1).

Similar to the previously presented experiments, the data preparation, including the one-hot encoding application to time-of-day and day-of-week data, and the application of min-max normalisation to the two most recent electricity consumption values, is performed to minimise the influences of outliers in the data and varying scales of the input features.

The above-specified data granularity scenarios are first tested on randomly selected three-month-long data series spanning from 01.07.2013 to 01.10.2013. The train-validation-test split is as follows: the first 67 days of data are the training set; then the next 16 days of data are the validation set; and the remaining nine days of data are the test set. As to the training set length scenarios, the length of the training and validation sets is changed for subsequent testing of training-set-length scenarios, but the test set always remains the last nine days of the data for the sake of comparative study. All these train-validation-test splits can be summarised as in Tables 7.3 and 7.4.

Data Length	**Dates**	**Length**	**Train-Valid.-Test**
3 Months	01.07.2013 – 06.09.2013	67 Days	Train
	06.09.2013 – 22.09.2013	16 Days	Validation
	22.09.2013 – 01.10.2013	9 Days	Test

Table 7.3: Train-validation-test split applied to data for scenarios with three different granularity levels.

Data Length	Dates	Length	Train-Valid.-Test
1 Month	01.09.2013 – 18.09.2013	17 Days	Train
	18.09.2013 – 22.09.2013	4 Days	Validation
	22.09.2013 – 01.10.2013	9 Days	Test
3 Months	01.07.2013 – 06.09.2013	67 Days	Train
	06.09.2013 – 22.09.2013	16 Days	Validation
	22.09.2013 – 01.10.2013	9 Days	Test
6 Months	01.04.2013 – 18.08.2013	139 Days	Train
	18.08.2013 – 22.09.2013	35 Days	Validation
	22.09.2013 – 01.10.2013	9 Days	Test
12 Months	01.10.2012 – 13.07.2013	285 Days	Train
	13.07.2013 – 22.09.2013	71 Days	Validation
	22.09.2013 – 01.10.2013	9 Days	Test

Table 7.4: Train-validation-test split applied to data for scenarios with four different training set lengths.

7.3.2 Results

The results of the tests performed with three different data granularity level settings and with four different training set length settings are presented below.

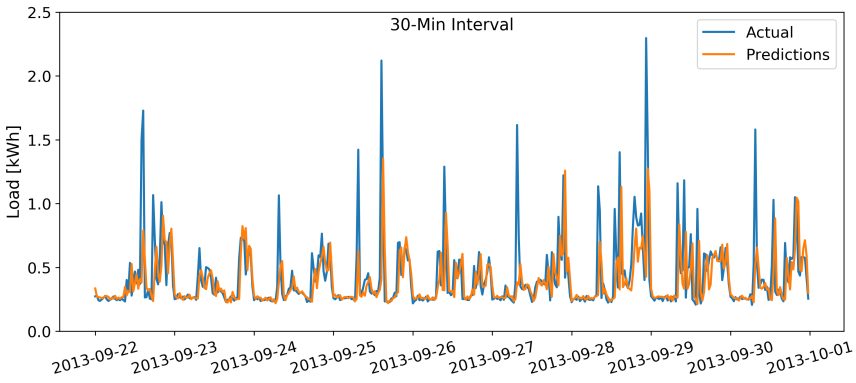
7.3.2.1 Data Granularity Effect on the PFE

The prediction results of the above-specified scenarios with different data granularity levels are shown in Figure 7.6. Additionally, the accuracy metric results of the original and the 1-SS method applied predictions are listed in Table 7.5 along with the clustering results. It is clear from Figure 7.6 and the metric results (Table 7.5) that the predictions follow the actual load values one step behind in time, and all metrics improve when the 1-SS method is applied, regardless of the data granularity.

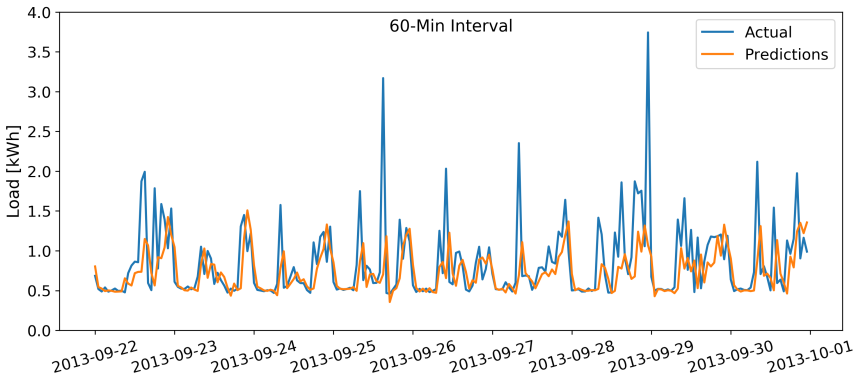
Also clear from Table 7.5 is that coarser granularity improves the data regularity as it smoothes the time-series patterns. The number of clusters reduces from 73 to 52 and then further to 28 with increasing sampling intervals. This, then, provides a means of investigating how improving regularity by manipulating the data granularity impacts the PFE. To this end,

Granularity	Number of Clus.	RMSE vs. RMSE*	MAPE vs. MAPE*
30-Min Interval	73	0.257 ↓ 0.145	23.031 ↓ 10.047
60-Min Interval	52	0.450 ↓ 0.343	24.464 ↓ 13.503
120-Min Interval	28	0.702 ↓ 0.507	25.228 ↓ 16.354

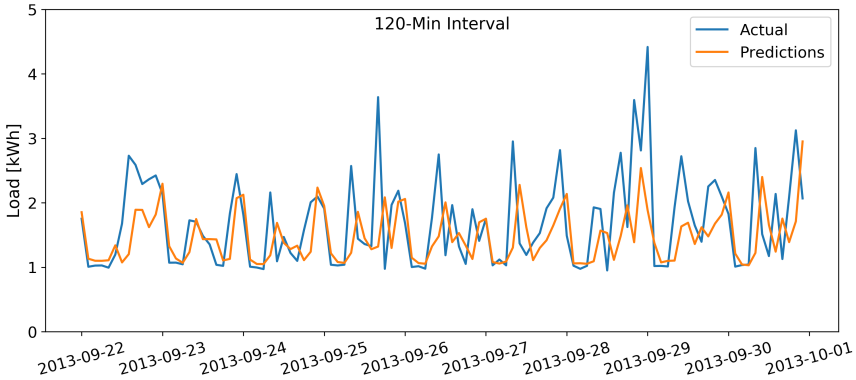
Table 7.5: Clustering and accuracy metric (original and the 1-SS method applied) results of predictions produced for data with different levels of data granularity.



(a)



(b)



(c)

Figure 7.6: Prediction results produced for the same nine-day long test set with different data granularity settings of (a) 30-minute, (b) 60-minute, and (c) 120-minute intervals.

the changes in the evaluation metrics of different granularity level settings before and after the 1-SS method is applied can be compared. However, any comparison of evaluation metrics calculated on different data series on different scales requires the chosen evaluation metric to be scale-independent [158, 159]. Therefore, RMSE is not appropriate for this analysis, but MAPE

can be effectively used (see Table 5.2). The difference between the MAPE and MAPE* values decreases with the reduced number of clusters, from 12.984 to 10.961 when the sampling interval changes from 30-min to 60-min and then to 8.874 when the sampling interval changes to 120-min. This indicates that, even though using coarser data helps to achieve a more regular data series and slightly reduces the change in MAPE (the difference between MAPE and MAPE*), it is not a complete solution to avoid or remove the PFE in time-series predictions.

7.3.2.2 Training Set Length Effect on the PFE

The prediction results of the four scenarios with different training set lengths are shown in Figure 7.7. The evaluation metric results of the original and the 1-SS method applied predictions are presented in Table 7.6 along with the total number of days in the data series and the clustering results. It is noteworthy that the evaluation metrics all change by a similar amount for all data lengths, regardless of the data length. Also, Figure 7.7 shows a strong resemblance between the prediction results of different training set length settings. All the predictions produced here by the identical prediction method that is trained on 17-, 67-, 139-, and 285-day-long training sets to make a prediction for the same nine-day-long test set explicitly suffer from the PFE, and they all trail the actual data one step behind in time. All these indicate that, as long as the irregularity is present in the time-series data, the length of the training set does not really affect the PFE result.

Data Length	**Num. of Days**	**Num. of Clus.**	**RMSE vs. RMSE***	**MAPE vs. MAPE***
1 Month	30	26	0.250 ↓ 0.143	22.460 ↓ 11.346
3 Months	92	73	0.257 ↓ 0.145	23.031 ↓ 10.047
6 Months	183	141	0.256 ↓ 0.140	22.679 ↓ 10.007
12 Months	365	295	0.249 ↓ 0.147	22.061 ↓ 11.046

Table 7.6: Dataset Length, number of clusters, and default and the 1-SS method applied accuracy metric results of predictions produced with different training set length settings.

7.3.3 Discussion

The scenarios tested in this section with different data granularity settings and different training set length settings have revealed that using fine-grained or coarse-grained data and training prediction methods on longer or shorter training sets do not appear to help to avoid the occurrence of the PFE.

The regularity of data improves along with the increasing observation interval, indeed. However, the improvement in regularity achieved by reducing the data granularity does not appear to prevent predictions from suffering from the PFE. Nevertheless, it has also been

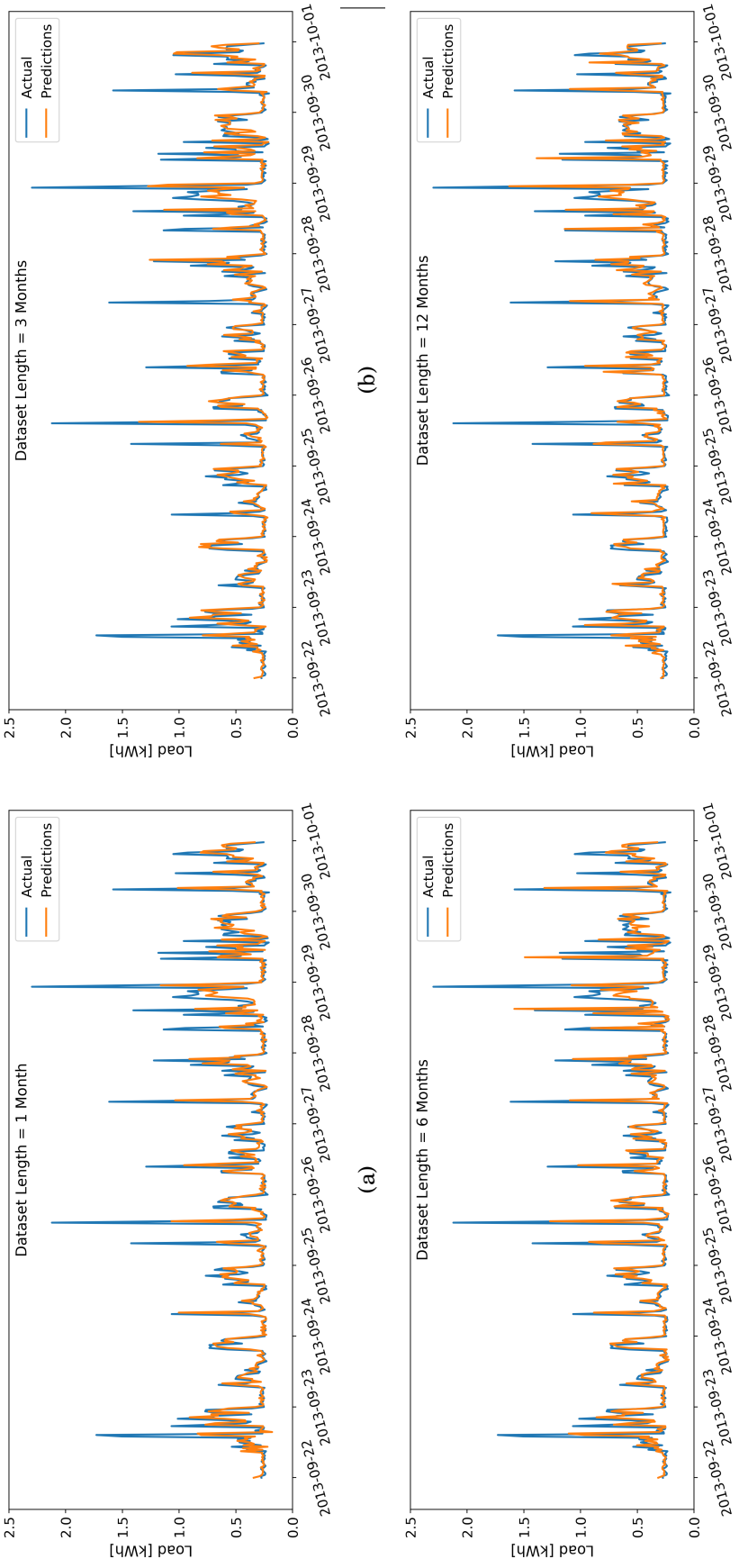


Figure 7.7: Time-series prediction results and actual data of the same nine-day period produced using training sets of different lengths. Total dataset lengths are (a) 1 month, (b) 3 months, (c) 6 months, and (d) 12 months.

observed that the changes in the evaluation metrics resulting from the application of the n-SS method decrease in proportion to the number of clusters, i.e., with increasing data regularity, even if it is a very modest decrease.

The insensitivity of the PFE to the length of the training set indicates that the PFE is still dependent on the volatility and irregularity in data, and, of course, changing the training-set length does not provide regularity to data. As long as the underlying data contains large variations in daily patterns that cause uncertainty, training prediction methods on longer training sets to provide more information to prediction methods or training prediction methods on shorter training sets to focus on the most relevant data does not even mitigate the impact of the PFE.

Overall, regardless of the training set length and the granularity of the data, all predictions are PFE-affected and they continue to trail the actual load values behind. This confirms once again that the PFE directly depends on the irregularity and volatility in the underlying time-series data.

7.4 Can the PFE be Attributed to Prediction Methods Instead of the Characteristics of the Underlying Data?

The investigations of the PFE so far in this thesis have mostly used LSTM RNN as the time-series prediction method to provide a comprehensive definition of the PFE and test different scenarios to explain its features in more detail. The work above presented several quantitative supporting the argument that the PFE is mainly determined by the high irregularity and volatility in the underlying data. However, it is still important to investigate whether the prediction method is one of the factors causing or contributing to the PFE in time-series predictions or whether the PFE is independent of prediction methods.

This section, therefore, deploys some alternative prediction methods to the LSTM RNN method. These have been used in the time-series prediction literature and are deployed here on a dataset with predictions known to be PFE-affected.

7.4.1 Methodology

This investigation uses a neural network method called BPNN, two forms of SVR (nu-SVR and epsilon-SVR), which are popular in the time-series prediction literature, and MLR, which is one of the basic statistical regression models. In this study, these methods are tested on one of the datasets that have also been used in the previous section while analysing the impact of the data granularity and training set length on the PFE. This dataset is a three-month-long electricity consumption dataset (01.07.2013 – 01.10.2013) with half-hourly measurement intervals that has been previously determined to have 73 different daily patterns across a 92-day time-frame. In Section 7.3, the prediction results produced by the LSTM RNN method, which has been

trained-validated-tested using this three-month-long dataset, are found to be affected by the PFE, regardless of the data granularity and training set length.

The architecture and hyper-parameters of the BPNN deployed in this study have already been described in Chapter 5. As for the SVR methods, the technical details of the SVR, including its working principles and mathematical representations, are beyond the scope of this study. Interested readers can find an in-depth description of SVR in [47, 132, 160, 161]. In summary, SVR is a support vector machine that is specifically designed for regression problems and is based on statistical learning theory [102]. The basic idea underlying SVR can be summarised as follows: it introduces a kernel function in order to map the input space into a high-dimensional feature space and then performs regression in this feature space [162]. SVR is widely recognised as an effective technique for regression applications [97], yet its performance strongly depends on the selection of its user-determined hyper-parameters: 'Cost' controls the degree of empirical risk, 'Gamma' controls the width of the Gaussian function in the kernel function, 'Nu' determines the error fraction width, and 'Epsilon' regulates the epsilon-insensitive zone width [160, 163]. To tune the hyper-parameters for the SVR methods deployed in this research work, a grid search has been conducted in order to determine the combination yielding the most accurate predictions. The eventual hyper-parameters are determined as in Table 7.7.

Parameter	**Setting**
Kernel Function	Gaussian RBF
Function Degree	1
Cost (C)	4
Gamma (γ)	Scale
Epsilon (ε) / Nu (ν)	0.01 / 0.75

Table 7.7: Hyper-parameter settings of the epsilon-SVR and nu-SVR methods.

7.4.2 Results

The predictions produced by the above-stated time-series prediction methods are shown in Figure 7.8, including the predictions produced by LSTM RNN for comparison. Evaluation metric results of the original and the 1-SS method applied predictions of these five prediction methods are compared in Table 7.8.

As it can be seen in Figure 7.8, the predictions produced by BPNN, nu-SVR, epsilon-SVR, and MLR all trail the actual load values one time step behind, similar to the predictions produced by LSTM RNN. The metric results in Table 7.8 also make clear that the 1-SS method applied predictions uniformly yield considerably more accurate metric results than the original predictions. Therefore, it can be concluded that the original predictions are affected by the PFE and

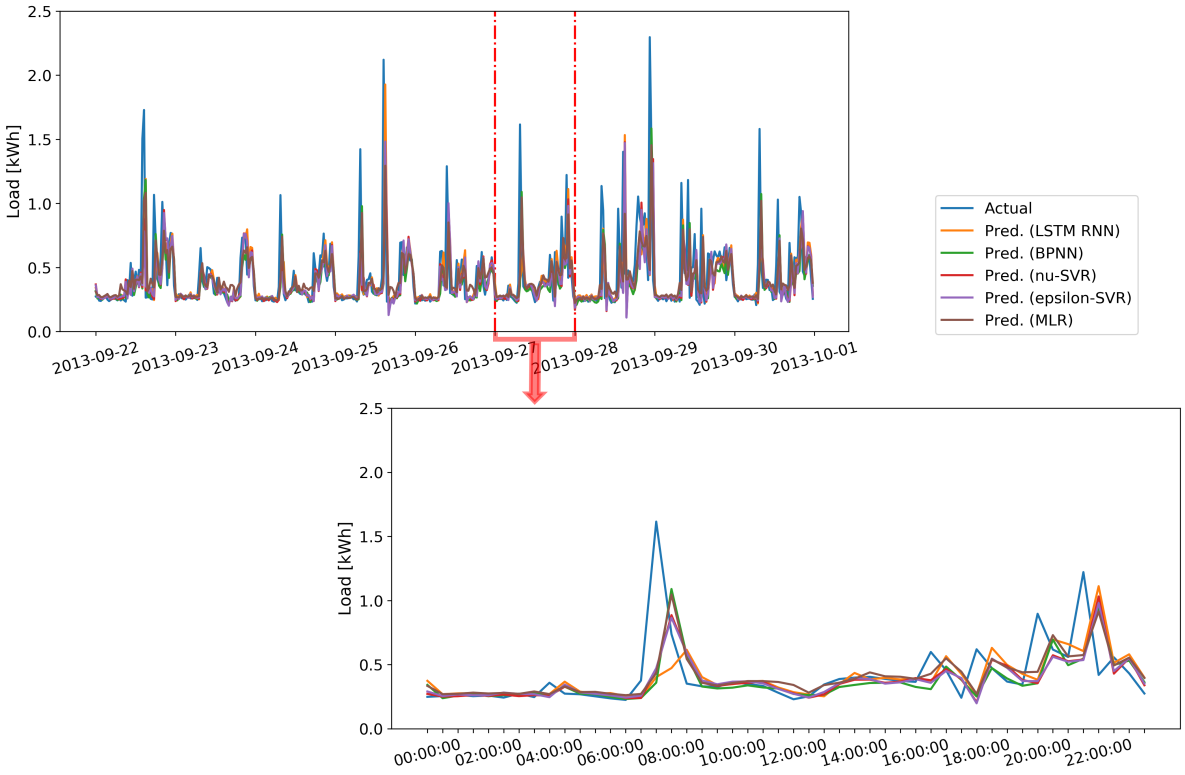


Figure 7.8: Prediction results produced by different time-series prediction methods.

Prediction Method	**RMSE vs. RMSE***	**MAPE vs. MAPE***	**Corr vs. Corr***
LSTM RNN	0.257 ↓ 0.145	23.031 ↓ 10.047	0.438 ↑ 0.854
BPNN	0.257 ↓ 0.117	22.698 ↓ 10.213	0.442 ↑ 0.916
nu-SVR	0.255 ↓ 0.157	22.384 ↓ 10.537	0.446 ↑ 0.853
epsilon-SVR	0.254 ↓ 0.158	22.380 ↓ 11.312	0.447 ↑ 0.853
MLR	0.243 ↓ 0.124	24.520 ↓ 13.778	0.483 ↑ 0.963

Table 7.8: Evaluation metric results of both original and the 1-SS method applied predictions produced by different time-series prediction methods.

thus temporally displaced, regardless of the prediction method. They are one time step delayed. Therefore, the prediction methods do not seem to be one of the factors causing or contributing to the PFE. Furthermore, the default evaluation metric results from the five different forecasting methods are very similar to each other, as all methods have been similarly and almost equally affected by the PFE. That is to say, these evaluation metric results show that the PFE largely neutralises the difference in the learning capabilities of the different prediction methods, similar to what has been found in Chapter 5 (Section 5.4). It is thus not effectively possible to identify the prediction methods producing the most and the least accurate predictions among the five time-series forecasting methods tested.

7.4.3 Discussion

The claim that the PFE is completely independent of prediction methods might ideally require that all the existing methods in the time-series forecasting literature are tested. However, this is impractical and inefficient. Hence, the study above has attempted to compare the PFE in the predictions produced by five different prediction methods, which are very commonly used in the literature and mostly cover different types of approaches.

The results have revealed that the choice of prediction method does not affect the PFE status of predictions. Once the underlying data has a certain level of irregularity and uncertainty, the PFE occurs irrespective of the prediction method choice. Therefore, considering that the main driver of the PFE is the characteristics of the underlying data, the PFE, which causes the predictions to continuously follow the actual values one or a few steps behind in time, cannot be attributed to prediction methods or machine learning methods, in contrast to [62, 65, 75, 76].

Nonetheless, it is also important to note that it seems possible that the magnitude of the change resulting from the n-SS method may vary slightly among alternative forecasting methods even though the data is the same time-series data. However, the quantitative results presented above suggest it is highly unlikely to perform two prediction methods on the same time-series data and observe that the PFE is present in one set of predictions but not in the other.

7.5 Is the PFE an Issue Specific to the Electricity Consumption Forecasting Domain?

It has been established in earlier chapters and sections that the PFE primarily results from the irregularity and volatility in time-series data and manifests itself as one or a few steps of systematic and continuous delay in time-series predictions. Therefore, regardless of the context to which the time-series data belongs, the PFE should be expected whenever the underlying time-series data involves a sufficient level of irregularity and inconsistency. However, given that the main focus of this thesis is to examine the PFE in electrical energy demand forecasting, as stated in earlier chapters, all of the experimental studies until now have investigated the PFE in electrical energy demand forecasting using two separate time-series datasets: SGSC and AMPds2. Therefore, the question of whether the PFE occurs more generally by using data from other domains remains an important area of study.

7.5.1 Methodology

For the purpose of answering the above-asked research question, this section performs time-series prediction on wind-speed data, which is also relevant to the context of electrical power. Wind energy is one of the most important renewable and sustainable sources of energy. The authors of [164] have shown that a 10% reduction in wind-speed prediction error increases the share of wind energy in global electricity generation by 30%. Temporally and spatially accurate wind-speed

forecasting is, therefore, of great significance to the stable operations of power grids, energy generation and consumption scheduling and planning, and the future development of renewable energy [74, 165].

The wind-speed data used in the present study is from the Jena Climate dataset [166]. This dataset comprises information on various meteorological characteristics, including temperature, atmospheric pressure, humidity, wind direction, etc., recorded by the weather station of the Max Planck Institute for Biogeochemistry in Jena, Germany. This publicly available dataset provides 10-minute interval data and spans over 8 years, from 01.01.2009 to 01.01.2016. In this study, five different time-series prediction methods: LSTM RNN, BPNN, nu-SVR, epsilon-SVR, and MLR, are implemented to forecast the future wind speed for four different variations of the wind-speed data:

- Three-month-long 10-minute interval data (3ML–10Min)
- Three-month-long 30-minute interval data (3ML–30Min)
- Six-month-long 10-minute interval data (6ML–10Min)
- Six-month-long 30-minute interval data (6ML–30Min)

The three-month-long variations span from 01.06.2012 to 01.09.2012, and the six-month variations span from 01.03.2012 to 01.09.2012. These time periods do not have any missing or duplicated recordings or any other issues that may require data preprocessing. The applied train-validation-test split is outlined in Table 7.9. Finally, clustering results illustrating the irregularity level of these four different variations of the wind-speed data and the total number of days for each are presented in Table 7.10 along with the number of total days.

Data Length	**Dates**	**Length**	**Train-Valid.-Test**
Three-Month-Long	01.06.2012 – 07.08.2012	67 Days	Train
	07.08.2012 – 23.08.2012	16 Days	Validation
	23.08.2012 – 01.09.2012	9 Days	Test
Six-Month-Long	01.03.2012 – 18.07.2012	139 Days	Train
	18.07.2012 – 23.08.2012	36 Days	Validation
	23.08.2012 – 01.09.2012	9 Days	Test

Table 7.9: Train-Validation-Test split applied to wind-speed data spanning over 3 and 6 months.

Dataset	**Number of Days**	**Number of Clusters**
3ML–10Min	92	71
3ML–30Min	92	69
6ML–10Min	184	128
6ML–30Min	184	126

Table 7.10: Clustering results of four different subsets sliced from the Jena Climate dataset.

7.5.2 Results

The evaluation metric results of the original and the 1-SS method applied predictions produced by the five different time-series prediction methods for four different wind-speed data variations produced from the Jena Climate dataset are presented in Table 7.11. The evaluation metric results suggest that no matter which prediction method is used, all wind-speed predictions are affected by the PFE, irrespective of the data length and granularity, and they systematically follow the actual wind-speed values one time step behind.

Dataset	Pred. Method	RMSE vs. RMSE*	MAPE vs. MAPE*
3ML–10Min	LSTM RNN	0.681 ↓ 0.372	34.956 ↓ 18.938
	BPNN	0.656 ↓ 0.336	32.383 ↓ 14.893
	nu-SVR	0.679 ↓ 0.397	35.406 ↓ 18.880
	epsilon-SVR	0.679 ↓ 0.397	35.447 ↓ 18.899
	MLR	0.624 ↓ 0.223	32.841 ↓ 13.383
3ML–30Min	LSTM RNN	0.873 ↓ 0.497	47.933 ↓ 29.490
	BPNN	0.855 ↓ 0.451	50.825 ↓ 30.880
	nu-SVR	0.921 ↓ 0.565	50.276 ↓ 31.624
	epsilon-SVR	0.925 ↓ 0.572	50.570 ↓ 31.715
	MLR	0.849 ↓ 0.442	46.578 ↓ 26.418
6ML–10Min	LSTM RNN	0.664 ↓ 0.296	34.393 ↓ 16.649
	BPNN	0.625 ↓ 0.218	30.916 ↓ 10.597
	nu-SVR	0.658 ↓ 0.313	33.022 ↓ 14.382
	epsilon-SVR	0.659 ↓ 0.314	33.073 ↓ 14.361
	MLR	0.620 ↓ 0.209	32.503 ↓ 12.955
6ML–30Min	LSTM RNN	0.876 ↓ 0.408	46.196 ↓ 24.293
	BPNN	0.843 ↓ 0.381	44.123 ↓ 20.917
	nu-SVR	0.901 ↓ 0.491	47.455 ↓ 26.108
	epsilon-SVR	0.902 ↓ 0.492	47.259 ↓ 25.729
	MLR	0.837 ↓ 0.397	46.335 ↓ 24.657

Table 7.11: Default and the 1-SS method applied metric results of time-series predictions produced by five different prediction methods for four different wind-speed data variations.

7.5.3 Discussion

The above experiment is the first to consider whether the PFE exists outside of the electrical energy consumption forecasting domain. In fact, this study of wind-speed forecasting showed that the PFE exists and causes a one time step delay with all five prediction methods used and with four variations of wind-speed data. It can, therefore, be deduced that the PFE is not limited to the electrical energy consumption forecasting domain, and it is potentially possible to observe the PFE in any time-series domain if the underlying data is irregular and inconsistent.

Moreover, the outcomes of this study also support the findings of Sections 7.3 and 7.4. Unless the volatility and irregularity in data are eliminated, training prediction methods on a longer or shorter training set, using coarse- or fine-grained data, or deploying an alternative time-series prediction method does not help to avoid the PFE.

7.6 Is the Delay in Time-Series Predictions Caused by the PFE Always One Time Step?

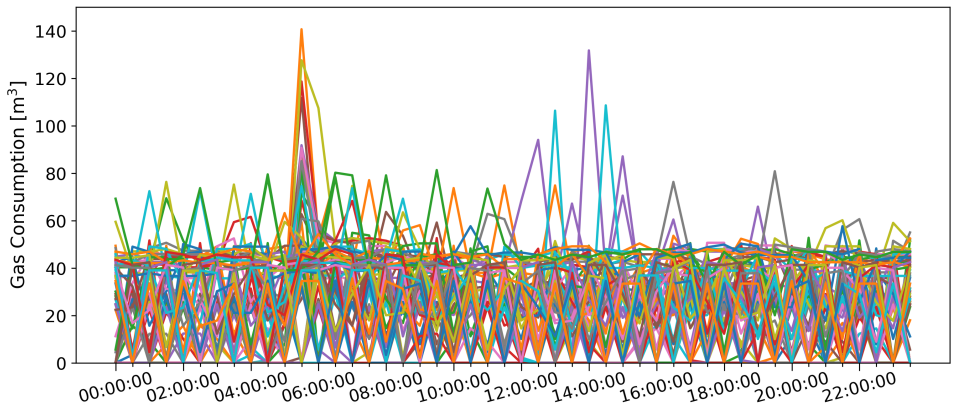
The experiments and discussion within this thesis have explained and highlighted repeatedly that the PFE occurs due to the volatility and irregularity in the underlying data and the number of time steps of the delay caused by the PFE is determined by the past observations used in the input feature set and the auto-correlation values at different lags. Nevertheless, because of the nature of time-series data, particularly that of electrical energy load data, which is mainly focussed on here, the superior auto-correlation is almost always between two successive observations. This in turn means that the auto-correlation at lag 1 is almost always higher than all the auto-correlation values at any other lags except for lag 0, which is self-correlation and always 1. This causes most PFE-affected time-series predictions to be delayed only one step in time, and it is very rare that the PFE leads to time-series predictions being delayed by two or more time steps. Indeed, this explains why all of the empirical studies that have been presented in this thesis show one time step delay behaviour caused by the PFE, and also why the n value of the n-SS method is identified to be 1 in those experimental studies. This section now focusses on whether the PFE can cause two or more time steps delays in time-series predictions or whether the effect is limited to only one step.

7.6.1 Methodology

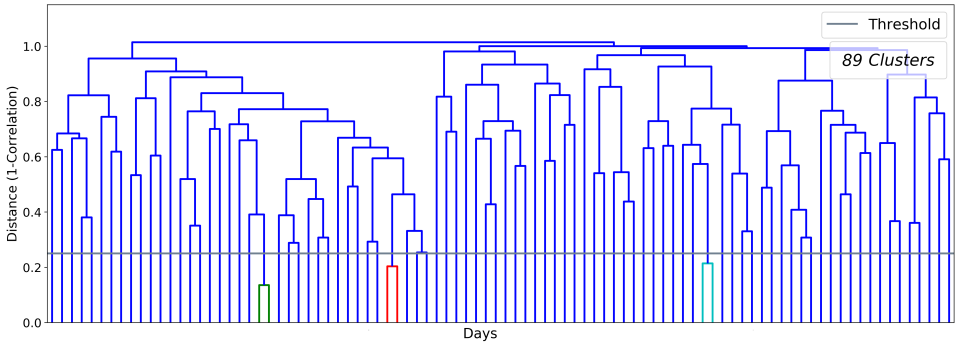
In order to investigate whether the PFE results in two or more time step delays in time-series predictions, this section deploys LSTM RNN on gas consumption data with 30-minute measurement intervals recorded from the Old Park Hill building at the University of Bristol. This building is an administration building with no teaching spaces or laboratories. In this building, gas is only used for space-heating purposes; it is not used for water heating, which is all done electrically. The consumption is measured volumetrically, in units of m^3. A three-month-long period of the data (from 01.10.2021 to 01.01.2022) was used here in this present study. The train-validation-test split is identical to that used with the three-month-long data in the previous sections: the first 67 days of data are the training set, the next 16 days of data are the validation set, and the remaining nine days of data are the test set.

The daily profiles of the gas consumption in the Old Park Hill building over a three-month period are visually represented in Figure 7.9a. The chaotic patterns reveal the extreme irregularity

and remarkable complexity of the daily gas consumption. This conspicuous irregularity in the data is further supported by the dendrogram presented in Figure 7.9b. Notably, there are only three pairs exhibiting notable similarities among themselves.



(a)



(b)

Figure 7.9: (a) Daily gas consumption of the Old Park Hill Building, University of Bristol, which exhibit no clear pattern and (b) dendrogram illustrating clustering results. Every colour group, with the exception of blue, corresponds to a distinct cluster, while the blue lines indicate outlier daily profiles.

Even more importantly, this three-month period of gas consumption data has a very interesting and unusual distribution of auto-correlation values at lags 1, 2, and 3 (Figure 7.10), which is critically significant in accomplishing the aim of the present study. Contrary to the previous scenarios, the auto-correlation value significantly increases across lags 1, 2, and 3. Among these three lags, the highest auto-correlation value is at lag 3, and the auto-correlation value at lag 2 is considerably greater than the auto-correlation value at lag 1.

In order to investigate the relationship between the amount of delay and the auto-correlation values, two different scenarios built with different numbers of the most recent observations in

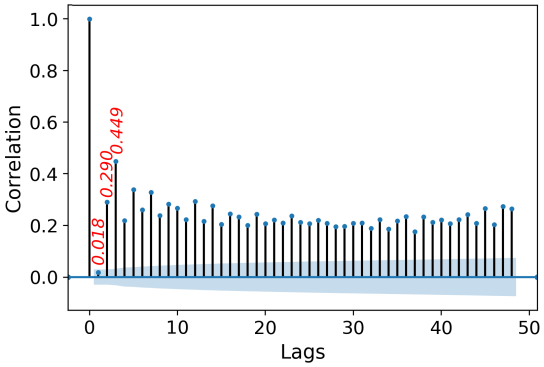


Figure 7.10: Auto-correlation analysis of the gas consumption data across three monts from the Old Park Hill building, University of Bristol.

the input feature set: $K = 2$ and $K = 3$, are tested. The input feature set for these scenarios is, therefore, as follows:

- Gas consumption recordings of the K most recent time steps; where $K \in \{2, 3\}$.
- Day-of-week indicator (ranges from 0 to 6).
- Time-of-day indicator (ranges from 0 to 47).
- Weekend indicator (ranges from 0 to 1).

To standardise the scale of the features to a range of 0 to 1, one-hot encoding is performed to time-of-day and day-of-week data and min-max normalisation is performed for the most recent gas consumption recordings. This ensures consistent scaling across the features.

7.6.2 Results

The first scenario in which the PFE is investigated is the scenario with $K = 2$, i.e., with the two most recently observed gas consumption values in the input feature set. This would imply that the 1-SS and 2-SS methods would be enough to explore the existence of the PFE in predictions, as explained in Chapter 4 (Subsection 4.2.1). However, the 3-SS method is also included to investigate how they change the accuracy metric values independently and gain the fullest possible picture of the conditions under which the PFE causes longer delays in predictions.

The evaluation metric results of the original predictions and the n-SS method applied predictions with different n values are presented in Table 7.12. The evaluation metrics after application of the 1-SS and 3-SS methods worsen significantly, while they undeniably improve following use of the 2-SS method. This explicitly points out that the predictions are affected by the PFE, and they trail the actual gas consumption values two steps behind in time. The explanation of this two time steps delay observed in the predictions produced for the gas consumption dataset, which has irregular and inconsistent patterns, is that only the two most recent observations are used in

the input feature set and the auto-correlation value at lag 2 is significantly greater than that at lag 1.

n-SS	RMSE vs. RMSE*	MAPE vs. MAPE*	Corr vs. Corr*
1-SS	21.262 ↑ 27.485	54.045 ↑ 80.207	0.177 ↓ -0.459
2-SS	21.262 ↓ 15.082	54.045 ↓ 36.346	0.177 ↑ 0.643
3-SS	21.262 ↑ 25.844	54.045 ↑ 79.816	0.177 ↓ -0.313

Table 7.12: Default, and the n-SS method (with $n = 1$, $n = 2$, and $n = 3$) applied metric results of time-series predictions produced with the two most recently observed gas consumption data in the input feature set ($K = 2$).

It is noteworthy that the 3-SS method application results in a deterioration in metric results, despite the auto-correlation value at lag 3 being higher than those at lag 1 and 2. This leads to the interesting question of why the delay is two steps in the above example rather than three if it is the superior auto-correlation determining the number of time steps of the delay caused by the PFE. The answer is very simple – it is because only the two most recent observations are included in the input feature set and not the third one. In order to explore what happens when the three most recent observations are used in the input feature set in such a case where the lag 3 has the superior auto-correlation value, the above-presented experiment is repeated with the second scenario, where $K = 3$.

In this scenario, as the three most recent observations are included in the input feature set, it is necessary to apply the 1-, 2-, and 3-SS methods to ensure the existence or absence of the PFE in predictions. The evaluation metric results of the original predictions and the 1-, 2-, and 3-SS methods applied predictions for the scenario with $K = 3$ are presented in Table 7.13. The 1- and 2-SS methods yield essentially worse evaluation metrics, while only the 3-SS method improves the default evaluation metric results. This is because the auto-correlation value at lag 3 is higher than the auto-correlation values at all the other lags. Therefore, when the PFE occurs, the predictions are delayed by three time steps.

n-SS	RMSE vs. RMSE*	MAPE vs. MAPE*	Corr vs. Corr*
1-SS	16.162 ↑ 27.540	42.821 ↑ 78.280	0.518 ↓ -0.429
2-SS	16.162 ↑ 22.841	42.821 ↑ 63.211	0.518 ↓ 0.020
3-SS	16.162 ↓ 14.831	42.821 ↓ 33.485	0.518 ↑ 0.593

Table 7.13: Default, the n-SS method (with $n = 1$, $n = 2$, and $n = 3$) applied metric results of time-series predictions produced with the three most recently observed gas consumption data in the input feature set ($K = 3$).

The results of these two scenarios tested in this section with two different K values show that the number of time steps of the delay caused by the PFE is determined by the past observations

included in the input feature set and the superior correlation between these points. That is, the PFE-affected predictions approximate one of the most recent observations depending on which lag has the superior auto-correlation value.

7.6.3 Discussion

The study presented in this section has unambiguously illustrated that the delay caused by the PFE is not limited only to one time step and two or more steps delay in time are possible as well. Additionally, it has been shown that the amount of the delay in predictions is determined by the auto-correlation values at different lags depending on the number of most recent observations used in the input feature set. Similar to the previous section, the results of this study have also shown that the PFE is not a sort of issue that is specific to the domain of electrical energy consumption forecasting.

Nonetheless, it is important to highlight once again that the nature of time-series data means the superior auto-correlation is predominantly calculated to be between two consecutive observations, resulting in a superior auto-correlation value at lag 1. Therefore, the PFE can be expected to lead to a one time step delay most of the time.

7.7 Is There an Existing Evaluation Metric That Is Potentially Resilient to the PFE?

Throughout the thesis thus far, a diverse range of widely used evaluation metrics, including MAPE, MAE, RMSE, and MSE, have been deployed for assessing time-series predictions and assessing the existence of the PFE. The outcomes of these metrics have revealed that many of them are not resilient to the presence of the PFE. Consequently, they cause misleading accuracy measurements, thereby leading to overconfidence in both predictions and models. Therefore, the existence of a point-wise evaluation metric that is robust to the PFE and that might still be safe and reliable to use for performance comparisons, even in cases where the presence of the PFE is suspected, is highly significant.

As briefly introduced in Chapter 2, evaluation metrics belonging to the relative error metrics family, such as RAE and RRSE, assess the accuracy of predictions by comparing the actual prediction error with that of a simple model, such as an average method, moving average method, seasonal naïve method, or persistence model. These evaluation metrics offer the potential to exhibit resilience against the phenomenon of the PFE when they are employed with the persistence model as a baseline. Consequently, this section aims to empirically investigate this resilience.

7.7.1 Methodology

In the empirical study conducted within this section, RAE with Persistence Model (RAE-PM) is deployed to evaluate the accuracy of time-series predictions. RAE-PM, which indicates how well a model performs relative to the persistence model, is defined as follows:

$$RAE-PM = \frac{\sum_{t=1}^{m}(|x_t - y_t|)}{\sum_{t=1}^{m}(|x_t - x_{t-1}|)} \tag{7.1}$$

where x_t is the actual value observed at time t, and y_t is the predicted value for time t.

The interpretation of the RAE-PM is straightforward: if the RAE-PM is smaller than one, it indicates that the model outperforms the persistence model. Conversely, if it is greater than one, it means that the model performs worse than the naïve persistence model. Hence, the lower the RAE-PM, the better the model, so in the case of a perfect model, the RAE-PM would be zero.

The RAE-PM metric is employed to evaluate the electricity consumption predictions for House 9012348 within the subset of the SGSC dataset. It is worth noting that this particular household is chosen randomly from the set of households whose single-step forecasts were generated by the LSTM RNN method and previously determined to be affected by the PFE in Chapter 5.

7.7.2 Results

For the original predictions of this household, RAE-PM is calculated to be 0.947 (RAE-PM = 0.947). This indicates that the contemporary deep learning method is only slightly better than the very simple persistence model. When the 1-SS method is applied to these predictions, the RAE-PM value decreases to 0.585 (RAE-PM* = 0.585). The RAE-PM* with 1-SS can be defined as follows:

$$RAE-PM^* = \frac{\sum_{t=1}^{m}(|x_t - y_{t+1}|)}{\sum_{t=1}^{m}(|x_t - x_{t-1}|)} \tag{7.2}$$

Thus, the significant improvement in the accuracy metric with the application of the 1-SS method confirms the presence of the PFE once more, as observed in other metrics employed in Chapter 5.

More importantly, as the RAE-PM metric evaluates prediction accuracy by comparing it to the accuracy of the persistence model, the deceptiveness effect of the PFE-caused systematic delay in predictions is effectively eliminated. The close-to-one RAE-PM value (which is 0.947 in this case) is generally interpreted as indicative of poor and inaccurate time-series predictions for an advanced machine learning methods. Therefore, contrary to many of the other point-wise metrics, the RAE-PM metric does not end up with misleading metric results or misplaced confidence in

predictions or models. In light of these considerations, it can be inferred that the RAE-PM metric exhibits promising resilience to the PFE, specifically compared to other widely employed metrics found in the existing body of literature.

7.7.3 Discussion

It is important to note that a close-to-one RAE-PM value does not necessarily imply that the overall predictions closely resemble those of the persistence model and are systematically displaced in time. For instance, Figure 7.11 presents two distinct hypothetical prediction curves for synthetically generated time-series data: one predictions curve represents PFE-affected predictions in orange, while the other illustrates a heuristic-free straight line in green. Even though these two predictions curves do not resemble each other in any way, their RAE-PM values are nearly equal to each other, specifically 0.938 and 0.937, respectively.

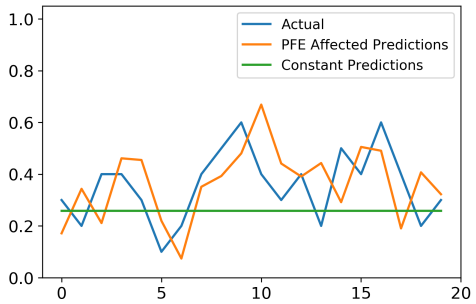


Figure 7.11: Synthetic time-series data along with two hypothetical forecasts: predictions affected by the PFE and predictions based on a constant value.

These close-to-one RAE-PM results indicate that both series of predictions are hardly any better than those of the simplistic persistence model. However, these results do not provide insight into the temporal accuracy of the predictions. Consequently, while the RAE-PM metric appears to be resilient against the deceptiveness effect of the PFE, it cannot always be effectively used for the detection of the PFE alone. Nevertheless, if the n-SS method is applied with the RAE-PM metric, this would provide a robust and reliable mechanism for detecting the PFE and also its quantification. Furthermore, if alternative evaluation metrics are preferred for the n-SS application, once the predictions are identified as being PFE-affected by the n-SS method, RAE-PM can be employed to quantify the impact of the PFE afterwards. These findings can be generalised to other metrics within the family of relative error metrics.

The quantification of the PFE would yield several advantages. It would:

- allow practitioners to determine the statistical significance of the existing PFE, enabling informed decisions regarding the risk associated with continuing with their current model and predictions, even in the presence of the PFE;

- facilitate the comparison of different prediction methods or models in terms of their robustness against the PFE, thereby allowing for the identification of the most suitable method for handling time-series data exhibiting volatile and irregular patterns.
- enable enhanced discrimination between PFE-affected and PFE-free multi-step predictions through the n-SS method application.

7.8 Summary

This chapter has provided many details on the PFE, manifesting itself as a continuous delay in single-step forecasts, by asking and answering some research questions. The overall conclusions of this section can be outlined as follows:

- Increasing or decreasing the number of most recently observed values in the input feature set does not help to avoid the PFE because the PFE is derived from the irregularity in the underlying data. This is because incorporating more or fewer most recent observations into the input feature set does not reduce the data irregularity.
- Different periods of time-series data can have different PFE results, depending on the regularity level. That is, there might be certain time periods in which predictions are PFE-affected, while the predictions for the other periods are completely PFE-free, despite coming from the same time-series data.
- It is not possible to avoid the PFE by training the method on longer or shorter training sets or using finer or coarser data granularity. Making the granularity of time-series data coarser improves the data regularity as a matter of course and helps to reduce the magnitude of the change that results from the n-SS method application. However, it appears to be an insufficient and incomplete solution for removing the PFE from time-series predictions.
- The PFE cannot be attributed to prediction methods and is directly related to the irregularity in the underlying data. As long as the underlying data do not have a certain level of regularity, the PFE occurs and causes systematic temporal displacement of predictions in single-step forecasts, irrespective of the prediction method choice.
- The PFE is seen in domains beyond electricity consumption forecasting. It is domain-independent and appears to be observable in time-series forecasts in every domain as long as the underlying data is irregular and volatile.
- The number of time steps of the delay caused by the PFE depends on the auto-correlation values at different lags and the most recent observations used in the input feature set. Therefore, the PFE can end up with temporally displaced predictions by one or more time steps.

- The evaluation metrics within the relative error metrics family have the potential to exhibit resilience against the deceptiveness effect of the PFE when they are employed with the persistence model as a baseline. Besides, such metrics can also effectively be used for the quantification of the impact of the PFE once the n-SS method confirms that the predictions are PFE-affected and thus temporally delayed.

The next chapter will investigate the presence of the PFE in some recently published works through visual inspections of their plots, illustrating their predictions over actual observations. This will provide a better understanding of the likely prevalence of the PFE in the time-series forecasting literature.

CHAPTER 8

The Prevalence of the Persistence Forecast Effect in Previously Published Works

Time-series analysis and forecasting have grown in popularity in both scientific and industrial fields, and there are now numerous publications on the subject. To the best of the author's knowledge, this thesis is the first attempt to define the PFE, examine its characteristics, identify the causes for its emergence, and evaluate its potentially detrimental impacts. In the literature, there are many research articles from various time-series prediction domains, such as electricity consumption forecasting, with forecasts that appear to be affected by the PFE. However, none of these includes analysis or a broad discussion about the systematic delay in their forecasts, despite its potentially detrimental effects on scientific findings, industrial applications, and smart systems.

This chapter, therefore, will survey a number of published works and attempt to identify whether or not their time-series predictions are delayed in time and exhibit the characteristics of the PFE. Ideally, identifying the presence of the PFE quantitatively in these works would be to reproduce their results with their datasets and prediction methods and then apply the n-SS method, as it has been demonstrated in the earlier chapters. However, this is not always possible because the underlying data is not always available or the required specifics of the deployed prediction methods are not well reported or documented. Hence, this study will instead use visual inspection of their plots to assess whether their time-series predictions are systematically delayed versions of the actual observation values, which indicates the likely presence of the PFE in the forecasts. Finally, the extent to which the continuous delay in forecasts undermines the conclusions of these papers will be discussed, as well as how their conclusions might be invalidated as a result of the PFE.

8.1 The PFE in Published Works from the Electricity Consumption Forecasting Domain

The previous chapter has shown that the PFE can manifest itself in any time-series forecasting domain and purely depends on the characteristics of the underlying data. However, the main domain in which this thesis investigates and explores the PFE is electrical energy consumption forecasting. This section, therefore, reviews several electrical energy forecasting works published in recent years with the aim of determining whether their results seem to be affected by the PFE. That is to say, the published plots that illustrate the prediction outputs over the actual electricity consumption values will be studied to compare the prediction outputs and the actual electricity consumption values gathered from various types of buildings, including domestic, university, and commercial buildings.

In [167], the authors aim to address the issue of quantifying uncertainties and fluctuations in electrical energy use. They first conduct statistical modelling analysis, including the Shapiro-Wilk test, the Quantile-Quantile plot normality test, and the Kolmogorov-Smirnov test, to derive a statistical distribution of electrical energy use. They then deploy machine-learning based prediction methods, such as standard Radial Basis Function based SVM, the Least Squares based SVM, and BPN, to achieve accurate energy consumption forecasting. The statistical modelling of energy use through these tests is out of the scope of this thesis, but the prediction outputs and their evaluation are key here. In the study cited above, the authors use hourly household-level electricity consumption data gathered from more than 250 households together with certain pertinent weather data, such as mean, minimum, and maximum temperature and mean and maximum value of wind speed. Instead of including all of these weather data in the input feature set, they perform the *ReliefF Algorithm*, which is introduced in [168], in order to determine the three most significant weather features for the electricity consumption of each building individually, and then they complete the input feature set with the three most recent electrical energy consumption values, resulting in six features in the input feature set in total. To assess the time-series predictions, they consider three point-wise evaluation metrics: MAPE, MSE, and R^2. The forecasts produced for two households from their dataset are shown in Figures 8.1 and 8.2 with their original captions. For a better visual investigation of the temporal delay in predictions, larger versions of these figures are provided in Section B.1 in the appendix in Figure B.1. The systematic and continuous delay in predictions is clearly discernible in these figures, regardless of the prediction method used. The predictions are almost always temporally delayed compared to the actual consumption values, resulting in a series of predicted values that is nearly identical to the series of observed values. However, the paper concludes that the predictions produced, particularly by the SVM based approaches, are exceptionally accurate simply because the blue and red curves are highly consistent with each other and follow the same trend. Based on the assessment in the article, the time-shift between the predictions and actual load curves seems

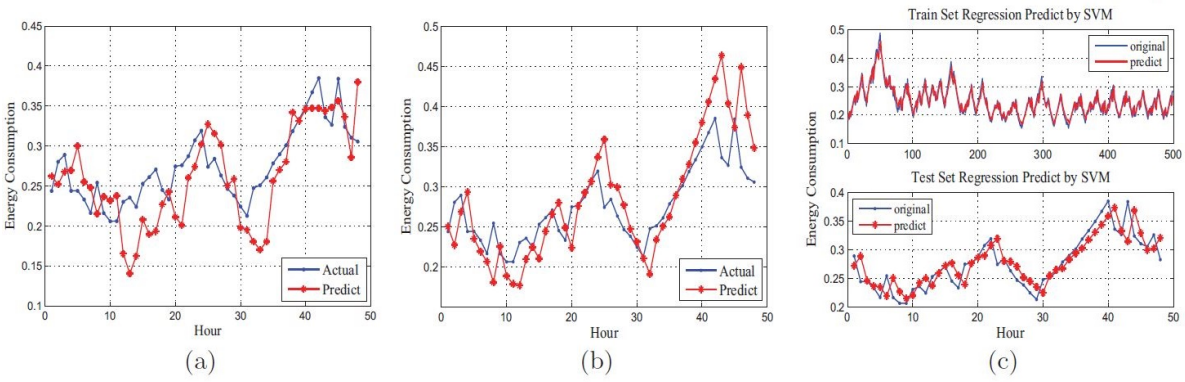


Figure 8.1: (Figure 3 in [167]) Forecasting accuracy of (a) BPNN, (b) LS-SVM and (c) SVM on No. 1002 House.

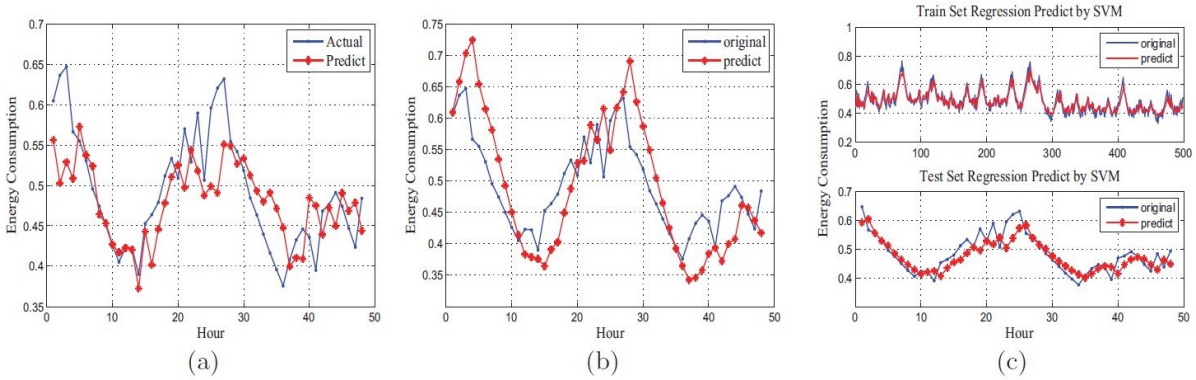


Figure 8.2: (Figure 4 in [167]) Forecast accuracy of (a) BPNN, (b) LS-SVM and (c) SVM on No. 1035 House.

to be overlooked. However, these curves show all of the characteristics of the PFE, with each prediction output approximating the observation recorded at the previous time point. Therefore, the predictions systematically follow the actual load values one step behind in time. Overall, unfortunately, it is very likely that the PFE is causing the apparent similarity between the two curves, and if this is the case, it may undermine the paper's conclusions of good learning by the methods and exceptionally accurate predictions. In other words, if the PFE does exist in these predictions, then the high resemblance between the two curves is not a result of the excellent performance of the methods, but rather because the predictions approximate the most recent values, resulting in a delay in the prediction curve.

In [80], the authors emphasise the difficulty of obtaining accurate forecasts of individual household electricity consumption due to the volatility in consumption patterns and propose a hybrid multi-task multi-information fusion deep learning (MFDL) framework in response. This framework deploys CNN LSTM and combines two concepts: learning from both recent and long-term historical data (multi-information fusion) and learning from and predicting for multiple households simultaneously (multi-task). They first identify three households with

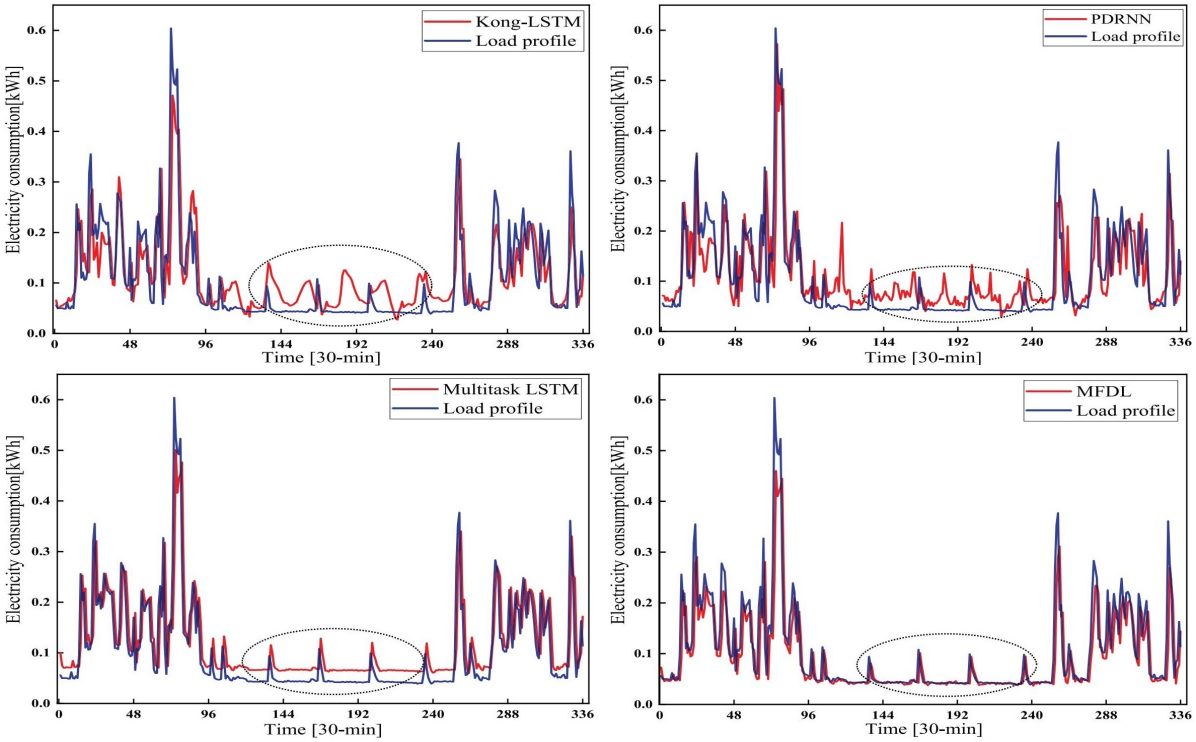


Figure 8.3: (Figure 8 in [80]) One-week prediction results for different methods from household 10006414.

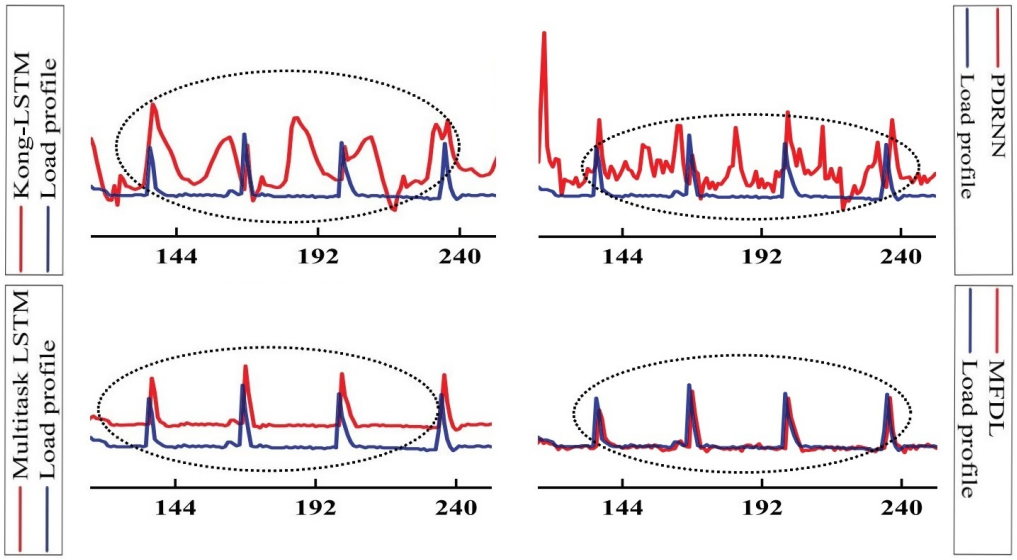


Figure 8.4: A closer comparison of the predictions highlighted by the ellipses in Figure 8.3.

similar load patterns from the same neighbourhood and then use the eight most recent electricity consumption observations together with consumption values from the last seven days from the three identified households. They test the proposed framework with a set of households from the SGSC dataset, which provides 30-minute interval residence-level electricity consumption

data. The full list of IDs of the households included in this study, however, is not mentioned in the research paper. In order to evaluate the performance of the proposed MFDL framework, they compare it with the performance of other methods, including XGBoost, SVR, CNN LSTM, RNN LSTM implemented by the authors of [89] (Kong-LSTM), Pooling Deep RNN proposed by the authors of [96] (PDRNN), and Multitask LSTM (a model proposed in this work as a baseline model that uses multiple households' data but without long-term consumption data) based on the results of three point-wise accuracy metrics: MSE, MAE, and Symmetric MAPE. Nevertheless, a visual comparison of predictions is only provided for Kong-LSTM, PDRNN, multitask LSTM, and the proposed MFDL framework (see Figure 8.3). Larger versions of these figures are provided in Section B.1 in the appendix in Figure B.2 to enable a better visual investigation of the temporal accuracy of the predictions. The article concludes that with more information in the input data from the house itself and its neighbours, the proposed MFDL model achieves significantly better predictions, but they are a little delayed in time. Based on this conclusion, the authors are obviously aware of the delay in their predictions. However, they unfortunately do not look into this further to understand the reasons for this delay or its potential consequences. They also put the emphasis on the time period shown by the dotted ellipses (between min 96 and min 240), when the predictions produced by the proposed method strictly fall behind the actual values and hence have a better-aligned pattern with the actual load values. For the sake of better comparison of the predictions in the ellipses, Figure 8.4 illustrates the zoomed-in version of Figure 8.3. Considering the differences in the predictions in the ellipse regions and unfortunately ignoring the apparent temporal displacement of predictions, the article concludes that the proposed MFDL is far better at capturing the peaks and troughs and so produces better time-series energy consumption predictions compared to the other methods. Nevertheless, taking the systematic and continuous delay in predictions, which closely resembles the phenomenon of the PFE, into account, it might be that the proposed method is simply more strongly affected by the PFE rather than producing more accurate predictions. This would be the explanation of why the predictions implicitly follow the actual observations behind in time and why the proposed method appears to capture peaks and troughs better than others, particularly within the time frame depicted by the ellipses. Outside of the ellipse regions, the prediction results of the four methods are also temporally delayed and appear fairly similar to each other, albeit with some occasional and indiscernible deviations. This could mean that the differences in accuracy metric results of these four methods derive primarily from the part of the predictions within the ellipses, i.e., the PFE could be leading to better apparent accuracy metric results. When these results are re-examined in the light of the systematic delay being shown, it is possible that this has arisen because the proposed method has confused the CNN LSTM with more input data from the household itself and its neighbours, and so its predictions are more strongly affected by the PFE. This would also explain the series of predicted values that is nearly identical to that of the observed values but systematically delayed.

Another instance of predictions that look like they are affected by the PFE and trail the actual observations behind in time can be found in [160]. In this work, the authors aim to develop a prototype of a single hybrid model to produce accurate load forecasts regardless of the building type or the aggregation level. They propose a technique of weighted SVR with differential evolution (DE) optimisation. This proposed method combines two different SVRs: nu-SVR and epsilon-SVR. In the proposed hybrid method, the optimisation technique is used to find out the optimum hyper-parameter selections for the SVRs and to determine the weights corresponding to each SVR. The proposed method is tested on two separate datasets, one with half-hourly measurements and the other with daily measurements, obtained from the same university building composed mainly of laboratories and offices. To evaluate the performance of the proposed model, it is compared with two single SVRs optimised with three different approaches: DE, genetic algorithm (GA), and particle swarm optimisation (PSO). The performance of these methods is compared through point-wise metrics such as ME, RMSE, MAE, MPE, and MAPE. As for the input feature set, the prediction methods are fed only with calendar data and a few most recent electricity consumption values. However, it is not specified how many most recent observations are included in the input feature set. The prediction results produced by the proposed hybrid method and the single SVRs with different optimisation techniques for the half-hourly dataset are shown in Figure 8.5. The authors conclude that, according to some of the metrics considered in the paper, the proposed hybrid method combining two SVRs is slightly better than the single

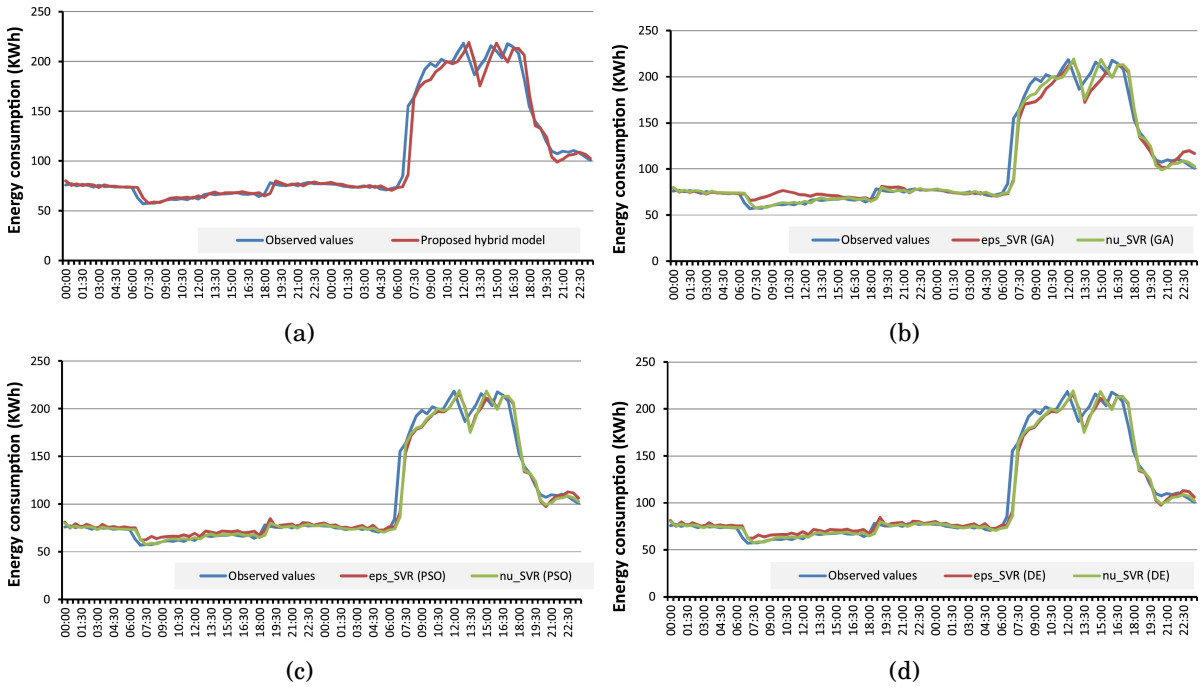


Figure 8.5: (Figures 4 – 7 in [160]) The variation of observed and forecasted values of (a) the proposed model and the two single SVR models with (b) GA, (c) PSO, and (d) DE optimization techniques for testing dataset for half-hourly energy consumption data.

SVR methods, while based on the other metrics, it is not even better or worse than single SVR methods. In other words, the accuracy metrics barely separate the performances of the methods – or even cannot. A succinct explanation for this could be lying in the observation that all the predictions, regardless of the prediction method, trail the actual load patterns behind in time, as can be seen in Figure 8.5. This suggests that if the main reason behind the systematic delay is the PFE, they are all affected almost equally by the PFE, and so, ultimately, the apparent performances of the methods are driven by the PFE and are almost equalised. If these methods were instead tested and compared using a dataset that is not subject to the PFE, one of these four methods might significantly outperform the others. This would provide evidence to judge the superiority or inferiority of the proposed method. The publication actually makes reference to the temporal displacement of the predictions, but only for those produced by the proposed method, which is claimed to fit the overall trend nicely (Figure 8.5a). The article explains the delay as a result of the lack of predictors. The further comment on the delay given in the article is that since there are only a few previous data points in the input feature set, the proposed model cannot do more than producing a prediction output that is very close to the immediate previous value. This explanation aligns with the concept of the PFE, despite the fact that it lacks many details of the PFE. However, unfortunately, this systematic and continuous delay in predictions is not identified as a major issue or investigated further in the publication. A re-evaluation of these results in consideration of the PFE would reveal the following critical points:

- The proposed model fits the overall trend relatively well, as it is claimed in the publication. However, this could be simply because the predictions approximate the most recent value owing to the PFE even though the proposed method has no capability to learn from data and produce accurate predictions.
- It explains the delay in predictions as being simply caused by the lack of predictors, which suggests it is not related to the dataset itself. This means that whenever a few previous data points are included in the input feature set as predictors, prediction methods generally return a prediction output that approximates the previous value alone, regardless of the regularity of the underlying dataset.
- Not only the predictions produced by the proposed method are consistently delayed, but also the predictions generated by the other methods are delayed in a similar manner. As a result, the metric values used to draw conclusions for the performance comparison and the overall conclusion presented in this work might be undermined by the systematic delay in predictions.

In [169], the authors explore how to apply sequence-to-sequence (seq2seq) RNN to electrical load forecasting of commercial office buildings. In their tests and analyses, they use electricity consumption data from four Eastern Washingtonian office buildings, referred to as Buildings A, B, C, and D. They report that these office buildings are mostly occupied from 08:00 to 17:00.

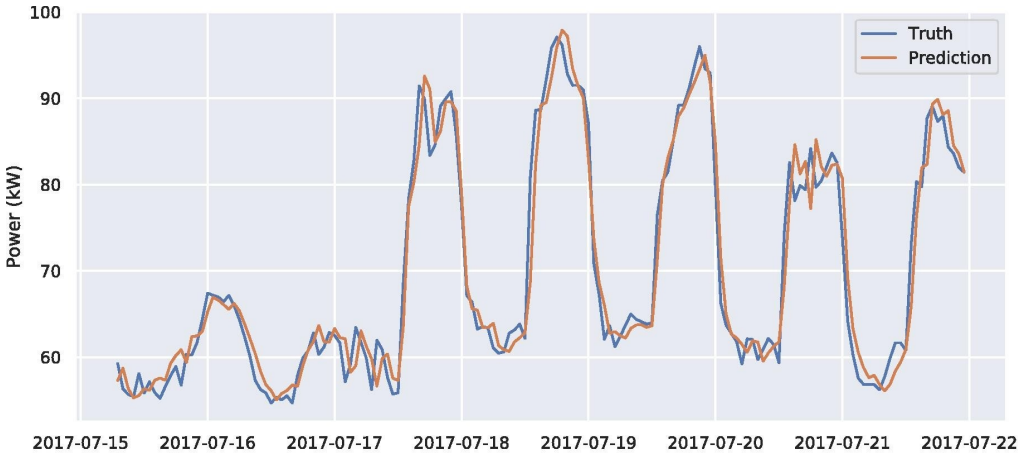


Figure 8.6: (Figure 7(a) in [169]) Plot of predictions vs. ground truth for a model trained and evaluated on Building A. The model was trained on 6 months of training data starting in July, and each hour of predictions was conditioned on the last 6 h of measurements.

The authors test their method on various scenarios built with varying training set lengths, data granularity levels, numbers of most recently observed values in the input set, prediction window sizes, and hyper-parameter settings. Their input feature set includes outdoor temperature, time of day, day of week, and varying numbers of recent consumption measurements. They use RMSE and N-RMSE as accuracy metrics to evaluate the prediction results. However, the article lacks visual representations of prediction outputs generated for different scenarios with different settings. Instead, it solely relies on metric results to evaluate the accuracy of the predictions. From this work, the only plot that allows the investigation of systematic and continuous delay is the one that shows the actual values and the prediction outputs for a scenario with a six-month-long training set and the six most recent readings in the input feature set for an hourly dataset from Building A (see Figure 8.6). The paper interprets the figure as showing that the proposed model captures general patterns effectively, with the exception of a delay in predictions during off-hours and weekends. This delay is mainly attributed to the high level of irreducible noise in electricity consumption during these periods. While certain points here are demonstrably true and valid (e.g., the noise in the data causes the predictions to fall back of the actual observations) from the PFE perspective, others require further examination in light of the understanding of the PFE in time-series predictions:

- Similar to the previous two papers discussed, the model might not be able to capture the overall consumption pattern but may appear to do so due to the predictions approximating the most recent measurement in the input feature set as a result of the PFE.
- The temporal displacement of the predictions is not restricted to off-hour periods and weekends, to use the authors' terminology, but instead appears to apply to the entire test set, spanning over a week. This might be pointing out that irregular electricity consumption takes place not only during the periods of quiescence but also during week-hours.

- Once again, due to the systematically and continuously delayed predictions, which are very likely affected by the PFE, a comparison of metric results calculated on such predictions would not provide a reliable and trustworthy performance ranking for prediction methods.

There are many other examples of peer-reviewed research publications on electricity consumption forecasting in which predictions are delayed based on the visual investigation of plots published in those works. These delayed predictions are consistent with the behaviour expected from the PFE, indicating that they may be impacted by the PFE. For example, Figure 8 in [90], Figure 2 in [91], Figure 13 in [93], Figures 9 and 10 in [101], Figure 10 in [104], Figures 8 and 11 in [106], Figures 9 and 11 in [170], Figures 3 and 4 in [171], Figures 8, 10, and 11 in [172], Figure 5 in [173], Figures 9, 10, and 11 in [174], Figure 5 in [175], and Figures 7 and 10 in [176].

Among these research works, the authors of [175] present spectacular time-series prediction results with regards to the PFE, as one method yields PFE-affected and PFE-free results for two separate datasets. In this work, the authors propose a hybrid method combining LSTM and

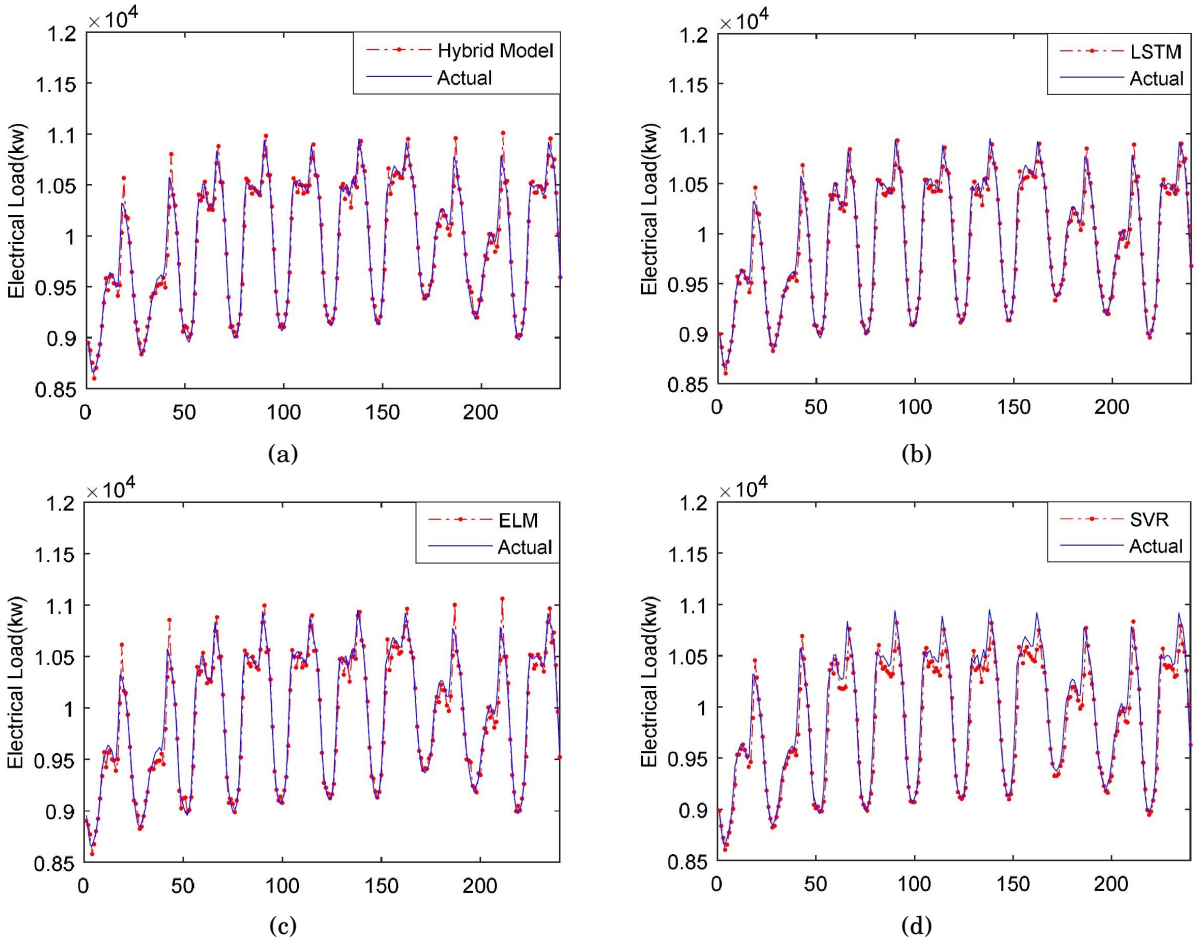


Figure 8.7: (Figure 5 in [175]) Experimental results of the last ten days in 2016 in the first experiment: (a) Hybrid model, (b) LSTM, (c) ELM, (d) SVR.

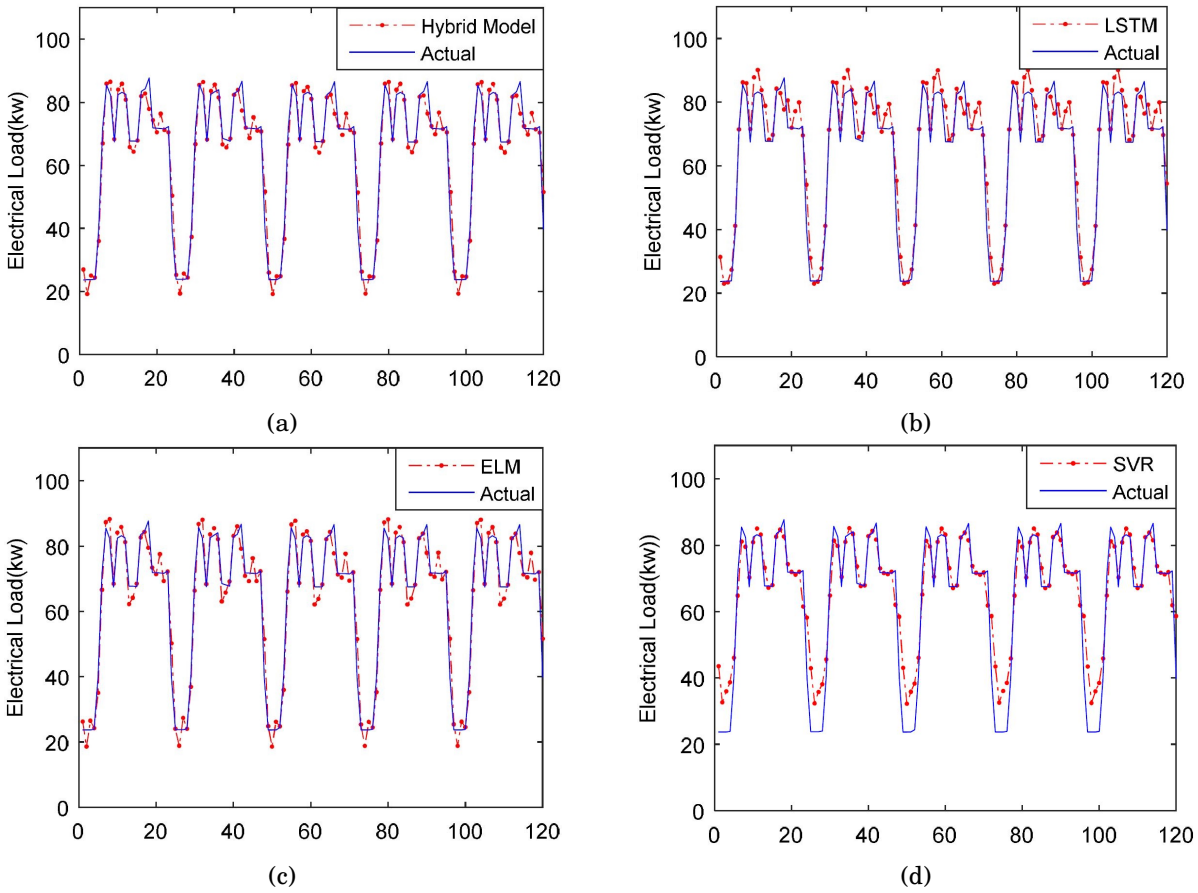


Figure 8.8: (Figure 8 in [175]) Experimental results of the last five days in the second application: (a) Hybrid model, (b) LSTM, (c) ELM, and (d) SVR.

ELM methods together in such a way that each method first predicts the next value individually, then their prediction outputs are ensembled to obtain the final prediction output of the hybrid method. In order to verify the performance of the proposed hybrid method, they test it on two independent hourly electricity consumption datasets and compare its accuracy with those of the LSTM, ELM, and SVR. One of these two electricity consumption datasets is from the Alberta region of Canada, while the other is from a service restaurant. The paper gives the total number of input variables but not the number of the most recently observed values included in the input feature set. The authors calculate three point-wise accuracy metrics (MAE, RMSE, and MAPE) to evaluate and compare the performances of the methods. Based on these three evaluation metrics, the authors conclude that their hybrid method produces sufficiently accurate predictions and outperforms the alternative methods with the most accurate metric outcomes. However, the study does not examine the temporal accuracy of the predictions produced for these two independent datasets. The results of the time-series predictions produced by the four deployed methods for the Alberta area and the restaurant are presented in Figures 8.7 and 8.8, respectively. Even though it is not easy to detect the existence of the temporal delay in predictions visually in the

plots in Figure 8.7, the predictions produced for the Alberta area trail the actual loads behind, regardless of the prediction method used. This is very similar to the behaviour of the PFE, and hence it is likely that the delay seen in these predictions is caused by the PFE. On the other hand, the predictions shown in Figure 8.8 exhibit near-perfect temporal alignment with the actual consumption measurements, which suggests that the PFE is unlikely to be present in these predictions. Assuming it is the PFE causing the delay in predictions of the Alberta area data, the presence or absence of the systematic delays in predictions of these two datasets should be considered in light of certain fundamental aspects of the PFE:

- The PFE can occur and affect the predictions of every level of electricity consumption. The electricity consumption forecasts produced for the Alberta area likely suffer from the PFE, whilst the predictions yielded for the service restaurant are PFE-free.
- The cause of the PFE cannot be attributed to the prediction method but rather to the characteristics of the underlying data. The authors use the same prediction methods for both datasets, yet predictions for one are delayed while those for the other are not. This suggests that the difference lies in the nature of the datasets themselves.
- The PFE largely neutralises the difference in the learning capabilities of the different prediction methods. The authors calculate that the hybrid method they propose outperform the alternatives for both datasets. However, it seems that its supremacy is pretty much wiped away by the delay seen in the predictions of the Alberta area dataset as all the methods are almost equally affected by the PFE. Based on the metric results reported in the work, the hybrid method improves the predictions of the LSTM method – the second most successful method – by 32.48% in MAE, 31.33% in MAPE, and 35.32% in RMSE for the restaurant dataset, whilst these improvement values drop to 4.68%, 4.53%, and 5.42% respectively for the Alberta area dataset.

This section has listed many recently published peer-reviewed works whose results seem to be affected by the PFE, based on a visual investigation of the plots they contain. These works have been explored in detail with the aim of determining their objectives, datasets, prediction methods, strategies, and so on, and then discussed to show why their conclusions may be questionable and how their findings may be invalidated by the temporal delay in their predictions. However, it is also important to note that there are several electrical load forecasting studies that provide predictions that are temporally accurate and hence appear to be completely PFE-free. Here are a few examples among many: Figures 4 and 5 in [87], Figure 10 in [96], Figure 15 in [112], Figures 6, 7, 8, 9, and 10 in [161], Figures 10, 11, and 12 in [177], Figure 10 in [178], Figure 12 in [179], Figure 6 in [180], Figures 6 and 7 in [181], Figures 3, 4, 5, and 6 in [182]. In addition to these studies, the results of [183] and [184] also seem to be PFE-free based on visual inspection. However, these two studies, particularly the regularity of their datasets, should be scrutinised closely from the PFE perspective.

In [183], the authors propose a hybrid model that combines the wavelet transform method and the functionally weighted single-input-rule-modules connected fuzzy inference system (FWSIRM-FIS). In order to test the performance of the proposed method and show its superiority, they compare the three accuracy metric results (MAE, RMSE, and R^2) of the proposed model with those of the four other methods: FWSIRM-FIS, Adaptive neuro-fuzzy inference system (ANFIS), BPNN, and MLR. They test these methods on two sets of 15-minute interval building-level data: one from a laboratory (94 days long) and the other from a retail shop (363 days long). Based on the mentioned evaluation metrics, the authors conclude that the proposed hybrid method outperforms the other methods on both datasets. Plots of the predictions produced by the proposed method are provided in Figures 8.9 and 8.10 for the laboratory building and the retail shop building, respectively. It can be observed that the predictions are not delayed when compared to the actual data. This suggests that there is no PFE present in the predictions, meaning that the metric and comparison results of this research study represent the true behaviour of the prediction models.

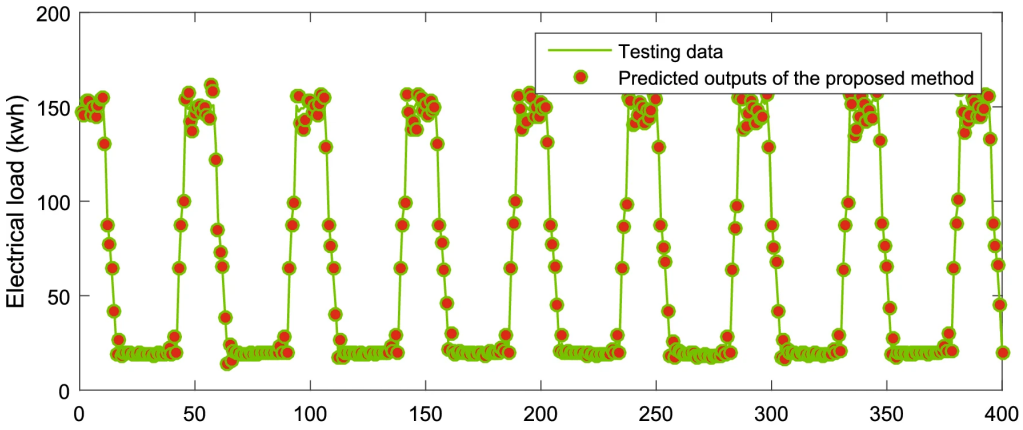


Figure 8.9: (Figure 7 in [183]) Forecasting results of the proposed hybrid model in the first experiment (Laboratory Building).

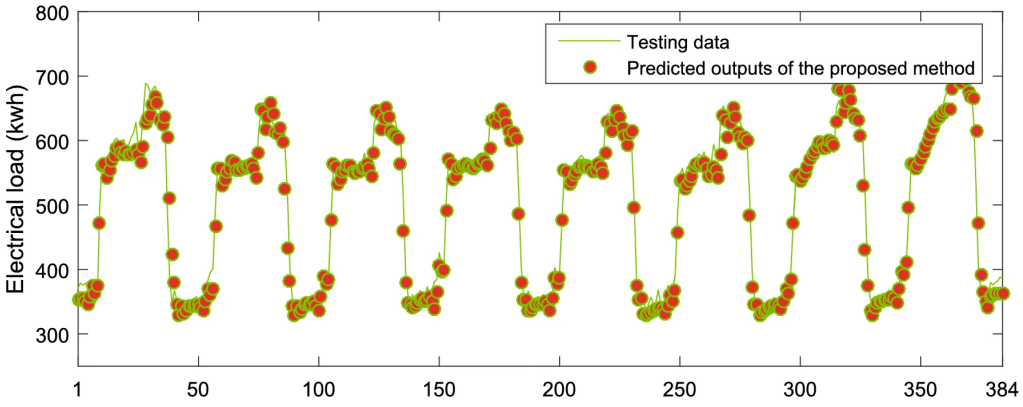
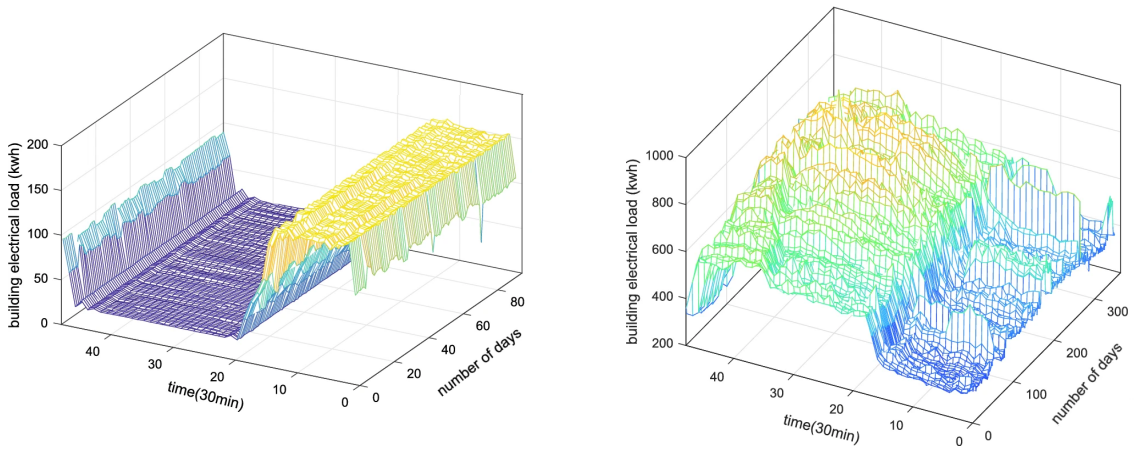


Figure 8.10: (Figure 13 in [183]) Forecasting results of the proposed hybrid model in the second experiment (Retail Shop Building).



(a) (b)

Figure 8.11: (Figures 3 and 9 in [183]) The original building electrical load data in (a) the first experiment (Laboratory Building) and (b) the second experiment (Retail Shop Building).

Most importantly, the authors of this work provide two more plots: Figures 8.11a and 8.11b. These plots illustrate daily electricity consumption patterns throughout the 94- and 363-day periods of the two datasets. Within each building, similar and also regular daily electricity consumption patterns can be observed. The regularity of the daily energy consumption profiles seen in these plots demonstrates the strong connection between data regularity and the PFE, which helps to explain why the predictions made for these two buildings are PFE-free.

Similarly, [184] also presents an example illustrating the significant relationship between data regularity and the PFE. The authors of the paper propose another hybrid method, this time one that combines CNN and RNN with two different architectures: LSTM and GRU. The authors compare the performance of the proposed method against 14 other methods, such as LSTM SVR, Feed Forward NN, LSTM RNN, and GRU RNN, based on three point-wise accuracy metrics: MAPE, MAE, and RMSE. They test all 15 methods individually on two different hourly electricity consumption datasets: the North-American Utility dataset and the New England ISO-NE dataset. In the input feature set used for feeding the prediction methods, they include electrical load and temperature values measured over the previous two days, season of the year, weekday/weekend, and holiday/non-holiday data. As a result of their comparative experiment, they conclude that the proposed method performs better than the other 14 methods on these two datasets. Despite the fact that the visual representations of the predictions by all the methods are not provided in the paper, but fortunately, those for the predictions produced by the proposed method are included (see Figures 8.12 and 8.13). A visual inspection of these plots confirms that the predictions are almost perfectly aligned with actual measurements, which also indicates the absence of the PFE in forecasts. The authors also provide auto-correlation results (see Figure 8.14) for these two datasets. The auto-correlation plots display peaks only at lag 24 and lag 48. This means that the energy loads taking place on successive days are similar and strongly correlated. Additionally, the auto-correlation values at lag 1 do not dominate the auto-correlation values at lag 24 and lag

48. These results indicate that the two datasets contain regular patterns of data, which is the first and most important condition for avoiding the PFE.

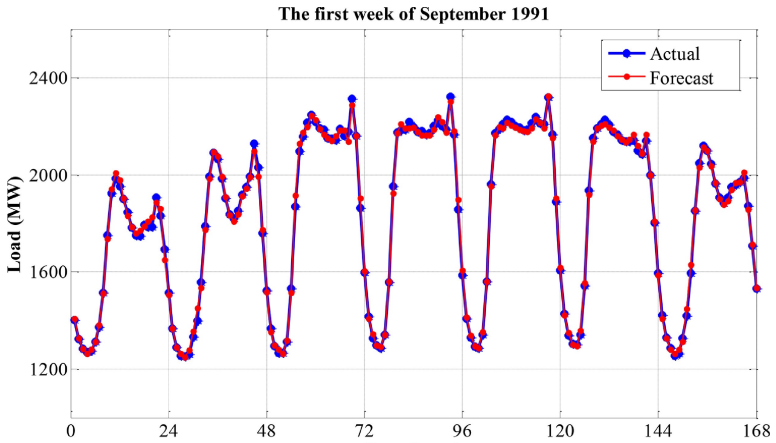


Figure 8.12: (Figure 9 in [184]) The forecast result of the proposed model in the North American Utility dataset.

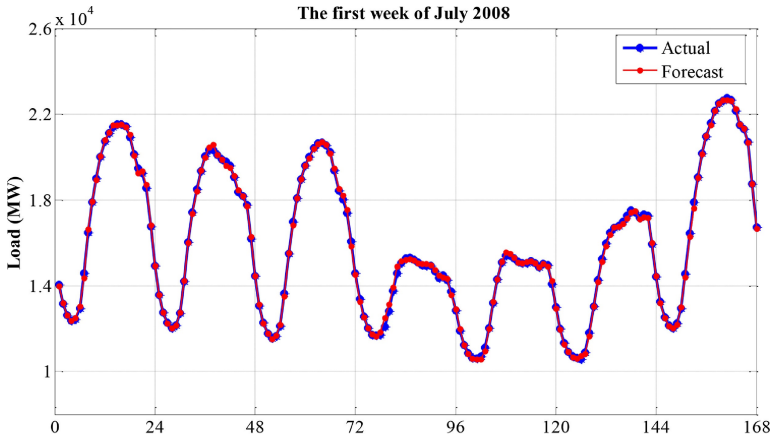
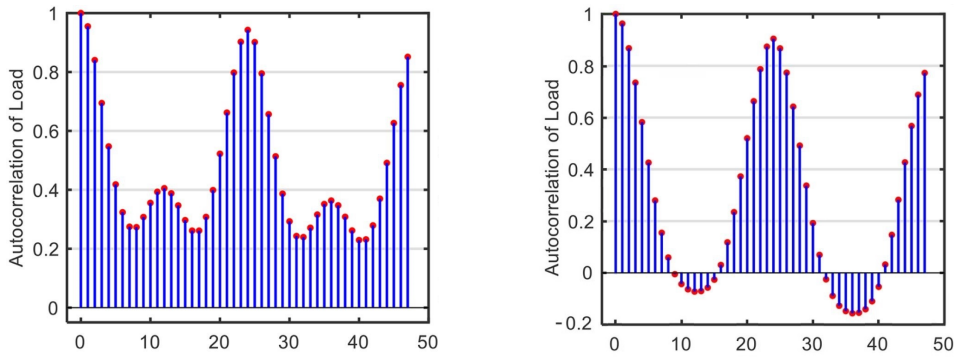


Figure 8.13: (Figure 12 in [184]) The forecast result of the proposed model in the ISO–NE dataset.



(a) (b)

Figure 8.14: (Figures 8 and 11 in [184]) Auto-correlation coefficients of load in (a) North American Utility dataset and (b) New England ISO–NE dataset.

These two research studies with PFE-free forecasts ([183] and [184]) demonstrate again that having a sufficient level of regularity in time-series data effectively prevents predictions from closely approximating one of the most recent measurements used in the input set. In simpler terms, when the underlying data consists of regular patterns, methods are able to recognise and learn these patterns rather than relying on the superior correlation between the successive time points. This prevents individual previously observed values from dominating the individual prediction outputs and avoids predictions following the actual measurements behind in time, thereby avoiding the PFE.

8.2 The PFE in Published Works from the Other Domains

Earlier chapters have demonstrated that the PFE applies to situations beyond electrical load forecasting. The PFE appears to manifest itself within single-step time-series predictions as long as the underlying data consists of patterns with a certain amount of irregularity and volatility. This section broadens the discussion around the PFE to consider several published time-series forecasting studies that utilise time-series data from many different domains outside of electrical load consumption and which have forecasts that appear to be affected by the PFE as observed through visual inspection.

Ref.	**Figure**[a]	**Prediction Method**[b]	**Dataset Domain**[c]
[185]	3, 4, 5	RFF-RMMC	Air Quality ($PM_{2.5}$, PM_{10})
[186]	10, 11, 12	NGCU RNN, LSTM RNN, GRU RNN	Air Quality Index, Hang Seng Index, Gold Price
[187]	6, 7, 8	LSTM-NN, Simple-NN	Groundwater Level
[188]	16, 17, 18, 19	PSO-LSTM, CHSGO-LSTM	Crude Oil Price
[189]	7, 8, 9, 10	ARIMAX, FBProb, XGBoost	Average shared prices invested in BTC
[190]	7	MFRFNN	Wind Speed
[191]	7 (a, b, e, f)	BRKGA-NN	Air Passenger, Death and Injury, Vehicle, Sunspot
[192]	3, 4	RPH with ARIMA and MLP	Shenzhen Integ. Index
[193]	3 (d, e, f), 4 (d, e, f), 5 (d, e, f), 6 (d, e, f), 7 (d, e, f), 8 (d, e, f)	MPNSGA-II KRR EDNN DEC, KRR EDNN DEC, KRR EDNN, MLP, CNN, LSTM	Electricity Price, Wind Speed, Air Quality ($PM_{2.5}$)

Continued on next page

Continued from previous page

Ref.	**Figure**[a]	**Prediction Method**[b]	**Dataset Domain**[c]
[194]	5 (a, c), 6 (a, c), 7 (a, c), 8 (a, c), 9 (a, c)	RNN, LSTM RNN, MLP, SVR, SARIMA	Temperature, Humidity
[195]	6 (a, b), 7 (a, b), 8 (a, b)	PCSA ELM, ELM	Streamflow
[196]	5 (a-l)	ICL with SA	12 financial datasets
[197]	6 (a-d)	ARIMA, LSTM RNN, Simple RNN, SVR	Store System Workload
[198]	13 (a), 14 (a)	Encoder-Decoder LSTM	ACI Finance, Sunspot
[199]	7, 12, 18, 19, 20	MHA, MHA LSTM	Lake, Colorado River, Traffic, Mumps, Chickenpox
[200]	4 (a, b, c)	MTSMFF, RNN, SVR	Air Quality ($PM_{2.5}$)
[201]	6 (a-f), 7 (a-f), 8 (a-f)	SeriesNet, SVR, LSTM, ANN, UFCNN	S&P500 Index, Shangai Composite Index, Hangzho Temperature
[202]	8, 10	ARIMA, SVR, MLP	Colorado River, Exchange Rate
[203]	6 (a-f), 11 (a-f)	EnsemLSTM, ARIMA, ANN, SVR, K-NN, GBRT	Wind Speed
[204]	5 (b), 6(b)	CNN, ELM	Streamflow
[205]	6, 7, 8, 9	DLSTM, LSTM-SAE	Air Quality ($PM_{2.5}$), Rental Bikes

[a] Figure numbers in the corresponding reference.

[b,c] Methods and datasets visualised by the given figures.

Table 8.1: Recent studies of time-series forecasting with PFE-affected predictions.

Table 8.1 lists a number of peer-reviewed research works that were published particularly in the last three years and that have attempted to solve time-series forecasting problems for various domains, including groundwater level, air quality, crude oil prices, wind speed, and so on. Along with the references to the research works, this table also lists the figure numbers that display the predictions that seem to be PFE-affected based on visual inspection, the prediction methods that yield those predictions, and the domains from which the underlying datasets were gathered. However, it should be noted that some of these publications implement more prediction methods than those listed here on some other unlisted datasets. That is, Table 8.1 lists only the prediction methods and dataset domains that are visually suspicious to be subject to the PFE. The authors of [199], for example, deploy nine different prediction methods on more than 15 time-series datasets. However, the plots they provide in their research only visualise the predictions produced by two

methods: MHA and MHA LSTM, and only five of the datasets appear to be subject to the PFE: Lake, Colorado River, Traffic, Mumps, and Chickenpox. Table 8.1 does not, therefore, include the other prediction methods they deploy or other datasets used in their article.

Figures 8.15 [186], 8.16 [192], 8.17 [197], and 8.18 [205] are a random selection of plots from those listed in Table 8.1, together with their original captions. Each of these plots illustrates time-series forecasts produced for datasets from different domains and produced by different time-series prediction methods. It is visually clear in each of the figures that the time-series predictions are almost identical to but trail the actual observations behind in time, which strongly suggests the existence of the PFE in the predictions. Similar observations can be made from the other figures listed in Table 8.1. This again indicates that the PFE is independent of domain and prediction method and caused simply by the characteristics of the underlying data.

Even more importantly, some of the research works listed in Table 8.1 use the same datasets. For example, [191] and [198] both use the same Sunspot dataset, [196] and [201] both use the same S&P500 Index dataset, [199] and [202] both use the same Colorado River dataset and [200]

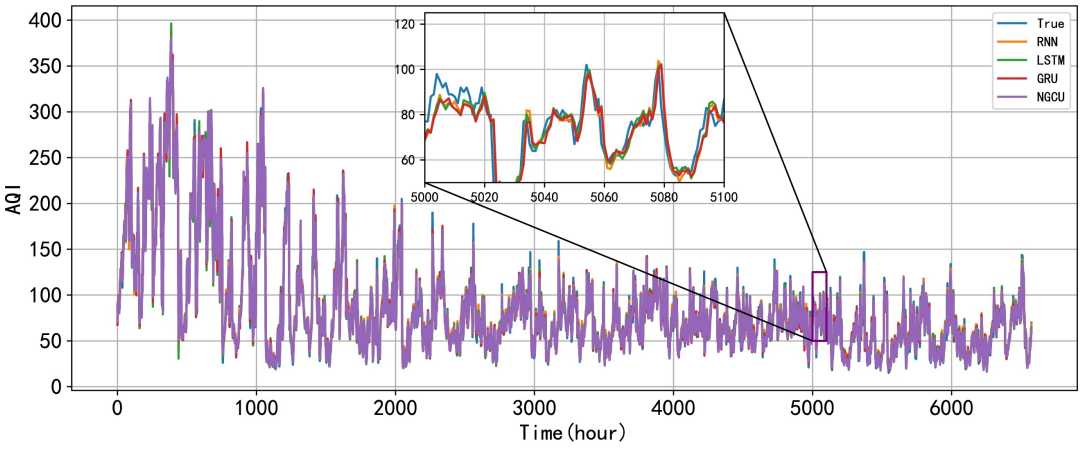


Figure 8.15: (Figure 10 in [186]) Comparison of four models of air quality predicted values and true values.

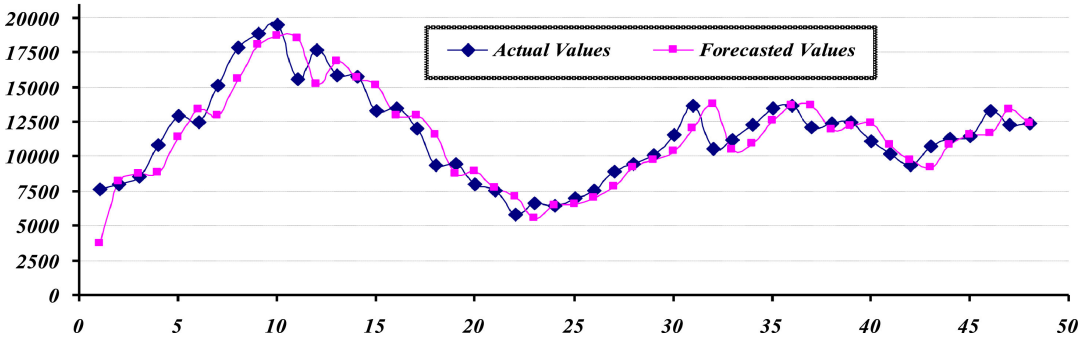


Figure 8.16: (Figure 4 in [192]) Forecasted values of the RPH model for forecasting Shenzhen Integrated index (test data set).

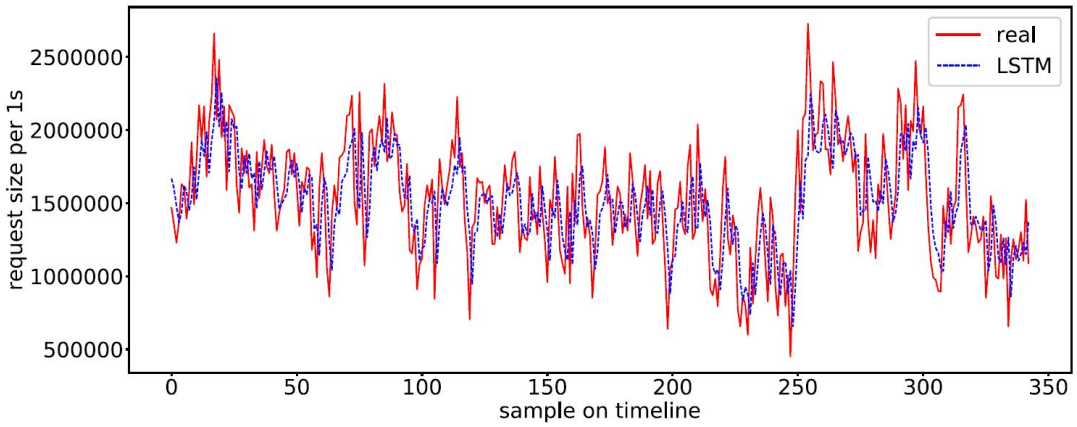


Figure 8.17: (Figure 6(d) in [197]) LSTM Prediction results in test set.

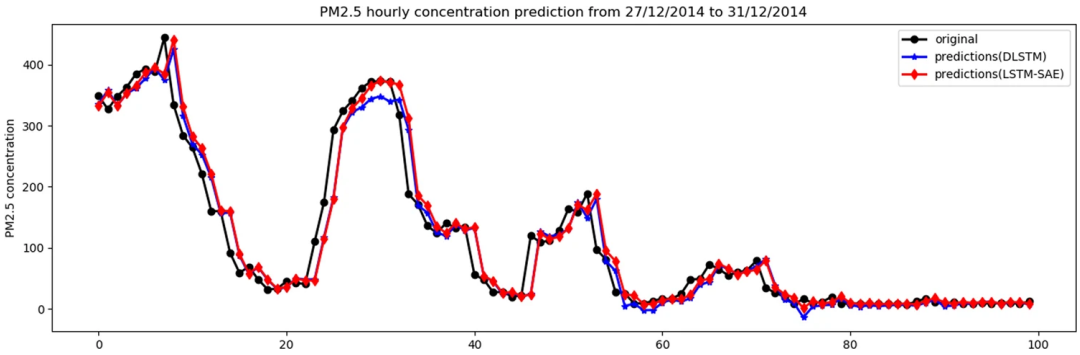


Figure 8.18: (Figure 9 in [205]) Original data vs. prediction for $PM_{2.5}$ with 2 layers.

and [205] both use the same Air Quality ($PM_{2.5}$) dataset. None of these pairs of studies that share the same dataset deploys or tests the same prediction methods. This is important because these shared datasets have prediction outputs produced by many different time-series prediction methods implemented or proposed in two independent works conducted by two distinct groups of researchers, and these prediction outputs appear to be almost identically affected by the PFE. To the best of the author's knowledge, in cases where multiple studies have utilized the same dataset, the predictions have consistently shown either PFE-affected or PFE-free outcomes and there has not been a single instance where one study with one effect and one with the other. This confirms once more that the PFE is caused primarily by the characteristics of the underlying datasets and is not primarily due to the time-series prediction methods.

8.3 Discussion

Plotting the time-series predictions against time together with the actual observations (what is called a time plot) to obtain a simple graphical analysis of the forecast accuracy is highly recommended [17, 29]. A properly structured time plot is also an extremely useful tool for identifying the possible presence or absence of the PFE in time-series predictions. However,

unfortunately, the systematic and continuous displacement of predictions caused by the PFE is not always recognisable through visual inspection. This is not due to the features of the PFE but rather due to various graphical elements, such as data length, granularity, dataset domain, graph size, image quality, resolution, aspect ratio, graph type, and line and point styles.

There are numerous instances in the time-series forecasting literature where plots do not allow judgements to be made in investigating the possible presence of the PFE. For instance, Figures 8.19 [206], 8.20 [207], and 8.21 [208] demonstrates predictions and actual measurements for different dataset domains and are taken from the original publications with no modification affecting their quality or sophistication. Plots in Figure 8.19 include more than 200,000 data points in very small-sized frames, making it impossible to interrogate the possible presence of the delay in predictions caused by the PFE. The scatter plot of almost 1000 data points in Figure 8.20 makes it difficult to judge the temporal alignment of the forecast and measurement datasets, as the forecast and measurement pairings are not clearly visible. Therefore, the temporal delay in predictions cannot be investigated in this plot. Indeed, the inappropriateness of scatter plots for comparing sets of time-ordered variables is explained and exemplified in [16, 17]. Finally, the line style and large datasets of almost 4000 points per series seen in Figure 8.21 means any delay in time-series predictions becomes obscured.

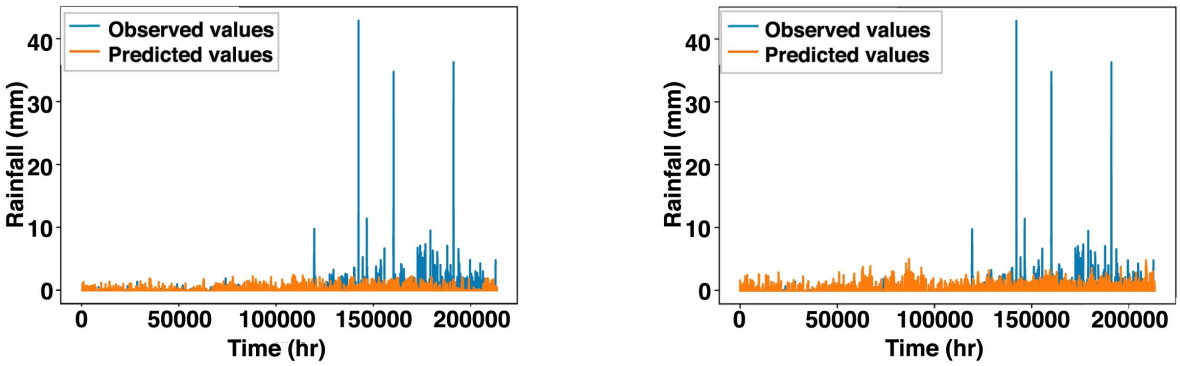


Figure 8.19: (Figure 16(c, d) in [206]) Observed rainfall vs. Predicted rainfall in Bristol (left: Model 4, right: Model 6).

Besides, several publications, including [44, 109, 209–212], do not provide visual representations of their time-series predictions alongside actual measurements. Instead, in such studies, point-wise accuracy metrics alone were used to evaluate the performance of deployed or proposed methods. However, this thesis has shown that such metrics can often be deceptive since they are not fully appropriate or reliable for detecting temporal displacement in time-series predictions. Therefore, in such studies, with inappropriate time plots or without time plots, it is not possible to make a firm statement regarding the existence or absence of the PFE, and the issue remains open.

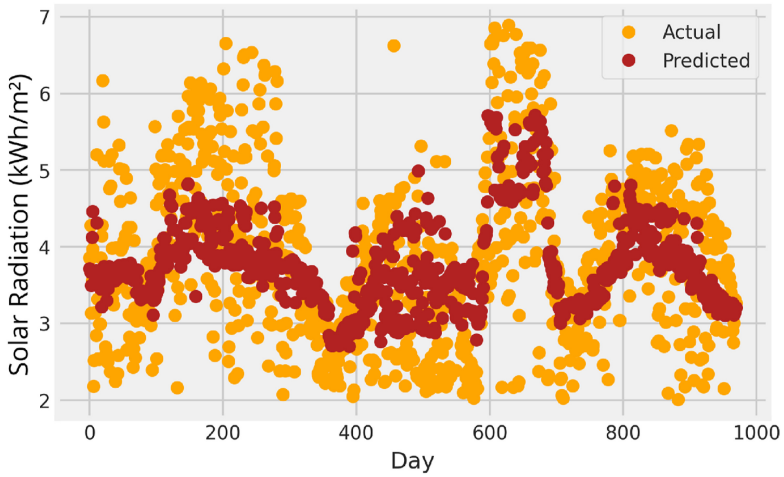


Figure 8.20: (Figure 13 in [207]) Actual vs. Prediction plot of Solar Radiation over time for GRU-based model.

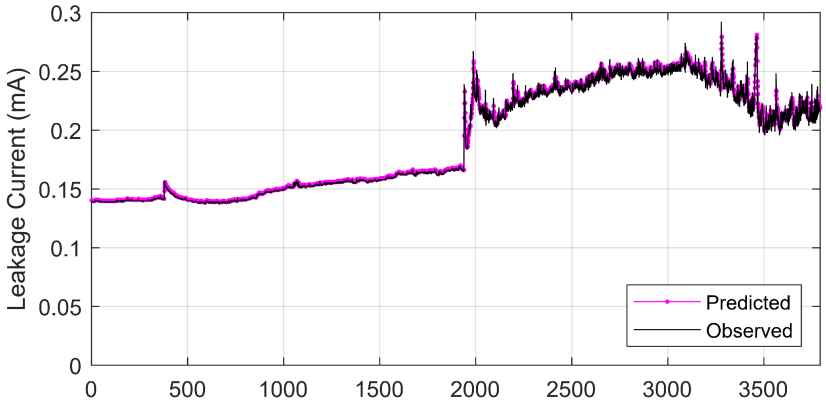


Figure 8.21: (Figure 5 in [208]) Predicted signal compared to observed signal.

8.4 Conclusion

This chapter has surveyed a number of published works that implement different prediction methods on different time-series datasets from various domains, including the electrical energy consumption domain. The prevalence of the systematic delays potentially caused by the PFE in the time-series prediction literature has been examined, as well as how and why the PFE might invalidate the conclusions in the literature or make them questionable in some circumstances. To do this, ideally, relevant time-series prediction results from the literature should be repeated, and then the n-SS method should be applied to achieve quantitative identification of the temporal displacements in forecasts. This has not been possible to do, however, due to some factors, such as inaccessible datasets and a lack of details provided on prediction methods. Therefore, the PFE review here has instead been conducted based on visual inspections of the figures that show forecasting results alongside actual measurements, where possible.

This chapter has examined published time-series forecasts from diverse domains using various time-series prediction methods. It has found many examples of predictions that are likely to be affected by the PFE and some that are not affected by the PFE. Overall, this chapter has demonstrated that i) the PFE is almost certainly prevalent in the time-series prediction literature; ii) the PFE is a bias that is independent of dataset domain and prediction method; and iii) the PFE occurs mainly due to the characteristics of the underlying data. Further, this chapter has identified several published arguments and conclusions that are questionable owing to likely PFE-affected prediction results and also discussed why these arguments and conclusions might be flawed and misled by the delays in predictions. This suggests that, in order to validate published results and arguments, the potentially PFE-affected results should be revisited, with prediction methods tested on different datasets and the n-SS method applied to test for the presence or absence of the systematic delays resulting from the PFE in predictions.

CHAPTER 9

DISCUSSION AND CONCLUSION

9.1 Summary

This thesis has contributed to the study and development of data-driven time-series forecasting and prediction accuracy evaluations in order to achieve reliable, robust, and transferable time-series predictions, which have become a key component applied in many different fields of science and industry.

Chapter 3 has provided a formal description of an important form of bias causing a systematic and continuous delay in single-step time-series forecasts. This important form of bias has been labelled the *Persistence Forecast Effect* (PFE) as predictions affected by it resemble those produced by a naïve persistence model. Furthermore, it has been demonstrated that the presence or absence of the PFE is mainly determined by the characteristics of the underlying data. Inconsistent and irregular patterns in time-series data are the major driver of the PFE. When prediction methods are trained or fitted on time-series data consisting of high irregularity and volatility, prediction models degrade to produce a series of predictions that follow the actual observations one or a few time steps behind. The amount of the delay, on the other hand, is determined by the correlation between the sequential observations, which can be measured via auto-correlation. Finally, in order to highlight the significance of the consideration of the PFE, Chapter 3 has also evaluated the implications of predictions that are systematically displaced in time as a result of the PFE, such as jeopardised smart systems and applications, misled decision-making, deceptively accurate metric results, and misplaced overconfidence in predictions and prediction models.

Determining the existence of the PFE in predictions is of great importance before taking further steps, such as making critical decisions or developing smart applications that depend on the time-series prediction outcomes. A visual inspection of a time plot, which plots prediction

outputs and actual measurements against time, can be seen as the easiest way of doing this; if time-series predictions are affected by the PFE, one should see that the predictions curve trails the actual observations curve one or a few time steps behind. However, visual inspection alone is not always sufficiently objective for practical and repeatable PFE detection. To fill this gap, Chapter 4 has proposed a new generic method, the *n-Step-Shifting* (n-SS) method, that identifies the existence or absence of the PFE quantitatively. The n-SS method facilitates a conceptually straightforward mechanism and relies on the recalculation of evaluation metrics after shifting the predictions n steps back to the past or shifting the observed data n steps into the future. The proposed n-SS method is compatible with a wide range of evaluation metrics commonly used in the time-series literature, relatively expediting and easing its deployment. Nevertheless, it requires pre-determination of the value of n. As it has been observed several times throughout this thesis, the value of n is usually 1 since the superior auto-correlation of underlying time-series data tends to be between adjacent data points by the nature time-series data. However, depending on the features of the time-series data, it is possible, although very rare, that n can be 2 or greater. How the value of n can be determined has been explained in Chapter 4 as well, which has also discussed some constraints that limit the applicability of the n-SS method.

As a means of further elaborating on all the definitive and explanatory statements made in the earlier chapters, Chapters 5 and 6 have provided experimental studies carried out with a large-scale domestic electricity consumption dataset and contemporary time-series forecasting methods. These experimental works have studied how irregularity in the underlying time-series dataset affects both single- and multi-step point forecasts, and how these effects manifest and differ from each other. Besides that, how to apply the n-SS method for PFE identification in time-series forecasts has been exemplified, and its applicability for single- and multi-step forecasts has been discussed in these chapters. The results of these empirical works have demonstrated the direct relationship between the occurrence of the PFE and irregularity in the underlying data, exemplified the implications of the PFE, and pointed the importance of including a certain number of the most recent measurements in the input feature set within a practical context. Furthermore, the n-SS method has been demonstrated to be highly effective in identifying whether single-step forecasts are affected by the irregularity and volatility in data (PFE-affected or PFE-free), but it has appeared to be unable to identify multi-step forecasts that are affected by data irregularity and volatility.

Following the chapters working on the basics and fundamentals of the PFE and the n-SS method, Chapter 7 has provided further analysis and various experimental studies on the PFE phenomenon and its identification. The chapter has established and addressed a set of research questions to interrogate the various facets and aspects of the PFE in single-step forecasts. As a result, this work has revealed that i) changing the number of the most recent observations in the input feature set of predictions based on volatile and irregular underlying data does not prevent the occurrence of the PFE; ii) different periods of the same time-series data can have

different PFE results; iii) it is not possible to avoid the PFE by training the method on longer or shorter training sets, or by using finer or coarser data granularity; iv) the PFE cannot be attributed to prediction methods and is directly related to the irregularity in the underlying data; v) the PFE is domain-independent and can be observed in all sorts of time-series forecasts in every domain; vi) the PFE can end up with temporally displaced predictions by one or more time steps depending on the auto-correlation values at different lags and the set of the most recent observations used in the input feature set; and vii) evaluation metrics within the relative error metrics family have the potential to exhibit resilience against the deceptiveness effect of the PFE and can effectively be used for the quantification of the impact of the PFE when they are employed with the persistence model as a baseline.

Chapter 8 then has examined the prevalence of the PFE in the recent time-series forecasting literature. This chapter has identified several published cases where time-series predictions appeared to be temporally delayed as a result of the PFE and discussed the impacts this could have on the findings and conclusions of these works. A survey of published works that implement different prediction methods on different time-series datasets from various domains again supported the notion that the PFE occurs when there are sufficiently volatile and irregular patterns in the underlying data and is independent of the dataset domain and prediction method used. The survey provided in this chapter has also highlighted the prevalence of the PFE in the time-series prediction literature, although it has mostly been overlooked so far.

9.2 Discussion

The PFE, which has a strong potential to endanger intelligent systems and decision making that rely on time-series predictions, is a consequence of the characteristics of time-series data used for forecasting. Given the strong pattern variability that can be seen in many underlying datasets on which the prediction methods are trained or fitted, it is not surprising that the prediction methods often fail to provide robust time-series predictions that are not affected by the PFE. If this is the case, the PFE causes profound changes to the predictions of methods and should be detected during the prediction evaluation to avoid erroneous predictions being trusted and acted upon. Nevertheless, many of the evaluation methods currently available to the community are not able to do this, as the PFE-affected predictions curve is nearly identical to the observations curve but delayed by only one or a few steps in time. Therefore, it is recommended that developers, practitioners, scientists, and decision-makers apply the n-SS method to examine whether or not their results suffer from the PFE before taking further steps. Furthermore, the n-SS method should be applied with multiple accuracy metrics to give the greatest confidence to the PFE detection results.

Despite the fact that the PFE is independent of prediction methods and derives from the underlying data as described in this thesis, the PFE should not be expected when basic methods, such as the average method, the naïve method, the seasonal naïve method, and the drift method,

are used. This is not because such methods are better, more robust, or more successful than complex statistical or machine-learning methods. It is simply because these basic methods do not require model training or fitting on historically recorded time-series data, in contrast to data-driven methods.

The objective of this study is not to demonstrate that data-driven prediction methods, including statistical and machine-learning time-series forecasting methods, are somehow *broken* in previously unknown ways. Instead, it aims to show the importance of guarding against overconfidence in prediction outcomes and mitigating the risk of developing forecasting models on datasets that do not perform as might be expected. The n-SS method, proposed and described here in this thesis, provides model developers with a tool to detect the PFE in their time-series predictions. If developers find their predictions are affected by the PFE, it is recommended that they review the regularity and consistency of the datasets they use and what additional features they could include in their input feature sets. For example, developers working in the domain of household-level electrical energy demand forecasting, could consider exploring features that determine why, when, and how electrical energy is consumed in a residence, including the lifestyle of occupants and daily routines, power consumption of appliances, and weather data. Such an augmented input feature set could have the potential to bring back the regularity that forecasting methods rely on or might explain the included irregularity so that it contributes to the reduction of the prediction error and the temporal displacements of predictions. Another suggestion in cases where the presence of the PFE is suspected is to use a metric that is potentially resilient against the deceptiveness effect of the PFE in order for more robust and reliable evaluation. This could be any metric from the family of relative error metrics, utilising a persistence model as a baseline. Such metrics may form a possible route for those seeking to differentiate a measure of forecast accuracy independent of interference from PFE.

It is interesting to consider the review of the PFE related to the issues previously reported in the time-series forecasting literature, such as phase error, double penalty effect (DPE), and under- and over-fitting.

The DPE arises from the use of point-wise metrics during the evaluation of predictions that are temporally displaced, commonly referred to as phase error. The phase error that the DPE is concerned with can occur at any point in time and with varying time steps – backwards or forwards – for all sorts of reasons. The working principles of the solutions proposed for the DPE simply rely on dropping the time dimension during evaluation. Hence, they do not aim to effectively resolve the time shift in predictions, but rather to address the issue of penalising forecasts that are temporally displaced twice. Such solutions might be applicable only to application areas that are tolerant of a displacement of events in time. However, not all applications have this temporal flexibility. On the other hand, as a matter of fact, the PFE refers to predictions that are a time-delayed approximation of actual measurements because they are extrapolated from one of the most recently observed values used in the input feature set as a result of data irregularity.

Therefore, the idea behind the solutions proposed for the DPE – dropping the time dimension during prediction accuracy evaluation – cannot be a proper approach to resolve the PFE. Dropping the time dimension would make sense for discontinuous and infrequent temporal displacements but is not appropriate for predictions that consistently follow the actual values behind in a systematic and continuous manner. If dropping the time dimension was suitable for resolving the systematic and continuous prediction displacements caused by the PFE, the simplest prediction method of all, the persistence model, would always become the best-performing method of all time for every time-series prediction problem. There would then be no need to use complex mathematical and statistical formulations and calculations or machine-learning approaches to develop better models. Sadly, this is not the case, and model development is still appropriate. Furthermore, the most common standard accuracy metrics are able to identify under- and over-fitting phenomena through a simple comparison of in-sample and out-of-sample errors. However, this approach is not suitable for detecting the PFE since the PFE affects predictions for training and test sets equally and in a similar way. Also, as the experimental results presented in Chapter 7 have demonstrated, most of the proposed solutions to these phenomena cannot be effectively used to prevent the PFE because the PFE is a result of volatile and irregular patterns in time-series data.

Previously published literature commonly attributes the delay effect in forecasts to two factors: the high auto-correlation in time-series data and the most recently observed values used in the input feature set. Hence, the solutions previously proposed for the continuous time delay in forecasts primarily often focus on eliminating the auto-correlation and the historical data from the input feature set. However, the findings presented in this thesis demonstrate that the main reason behind the systematic and continuous delay in prediction is actually the high volatility and irregular patterns inherent in time-series data. The auto-correlation is just a factor determining the amount of the delay when the delay occurs. When the underlying data exhibits a certain level of regularity that the prediction methods can recognise, the phase lag does not manifest, even if in the presence of superior auto-correlation or historical data in the input set. This likely explains why previously proposed approaches are insufficient in fully remove the phase lag phenomenon. Therefore, in order to effectively and properly overcome the PFE without compromising overall accuracy, it is crucial to enhance the input feature set by incorporating additional features that the prediction outputs may be contingent on. This augmentation of the input feature set could help to improve the robustness of prediction models and time-series forecasts. Another potential solution could be the implementation of techniques such as wavelet transforms with the purpose of mitigating the volatility observed in the time-series data on which the model is constructed.

9.3 Future Work

The research described in this thesis suggests some important future research directions and interesting lines of future work.

First and foremost, the development of a time-series prediction model that is robust against the PFE could be one of the most important contributions. This requires the investigation of whether the n-SS method and a metric belonging to the relative error metrics family using the persistence model as a baseline could be used as part of the prediction model development process. The ultimate goal is to build a model that is effectively capable of avoiding the PFE or at least minimising its effects. That is, instead of the n-SS method being used as a post-hoc control mechanism to check for the presence of the PFE, it could be integrated into the prediction model development process so that the model can progressively learn to produce predictions with no PFE or as little PFE as possible. While this may appear to be improbable for statistical methods, it is conceivable to modify the loss functions of machine learning methods to incorporate the n-SS method, implemented with a metric within the relative error metrics family using the persistence model as a baseline, during the training phase.

Additionally, regarding time-series predictions that could potentially have PFE-affected and PFE-free sub-periods together, which are analysed and discussed in Section 7.2, automating the process of determining the split-off point(s) separating the sub-periods with and without PFE could be another valuable future contribution.

APPENDIX

NUMERIC RESULTS OF CHAPTER 5

A.1 Accuracy Metric Results for 68 Households

Table A.1: Default evaluation metric results (MAPE, RMSE, Corr) and one step shifted evaluation metric results (MAPE*, RMSE*, Corr*) for 68 buildings.

HouseID	MAPE vs. MAPE*	RMSE vs. RMSE*	Corr vs. Corr*	PFE
8145135	47.577 ↓ 37.994	0.206 ↓ 0.117	0.577 ↑ 0.901	Yes
8147703	47.990 ↓ 30.827	0.130 ↓ 0.100	0.625 ↑ 0.798	Yes
8149711	65.248 ↓ 37.330	0.118 ↓ 0.062	0.737 ↑ 0.955	Yes
8156517	127.438 ↓ 117.424	0.450 ↓ 0.408	0.602 ↑ 0.674	Yes
8176593	28.021 ↓ 17.212	0.127 ↓ 0.065	0.601 ↑ 0.920	Yes
8181075	18.507 ↓ 8.408	0.197 ↓ 0.079	0.758 ↑ 0.965	Yes
8184653	38.724 ↓ 20.150	0.252 ↓ 0.107	0.638 ↑ 0.941	Yes
8196621	35.910 ↓ 20.560	0.159 ↓ 0.085	0.740 ↑ 0.941	Yes
8196659	40.227 ↓ 31.864	0.164 ↓ 0.081	0.701 ↑ 0.932	Yes
8196669	24.697 ↓ 19.635	0.192 ↓ 0.158	0.809 ↑ 0.887	Yes
8196671	28.727 ↓ 20.177	0.122 ↓ 0.087	0.635 ↑ 0.819	Yes
8198267	35.769 ↓ 24.038	0.156 ↓ 0.104	0.814 ↑ 0.918	Yes
8198319	23.059 ↓ 17.124	0.141 ↓ 0.118	0.782 ↑ 0.813	Yes
8198345	58.661 ↓ 42.834	0.123 ↓ 0.078	0.681 ↑ 0.825	Yes
8211599	44.857 ↓ 36.606	0.095 ↓ 0.066	0.676 ↑ 0.858	Yes
8257034	95.968 ↓ 66.130	0.413 ↓ 0.212	0.829 ↑ 0.961	Yes
8257054	21.821 ↓ 15.217	0.162 ↓ 0.116	0.764 ↑ 0.926	Yes
8264534	36.324 ↓ 28.641	0.246 ↓ 0.196	0.586 ↑ 0.851	Yes
8273230	25.358 ↑ 59.835	0.080 ↑ 0.143	0.901 ↓ 0.686	No
8273592	37.747 ↓ 29.287	0.228 ↓ 0.146	0.516 ↑ 0.813	Yes
8282282	99.842 ↓ 82.436	0.431 ↓ 0.331	0.377 ↑ 0.725	Yes
8291712	213.888 ↓ 183.257	0.416 ↓ 0.320	0.709 ↑ 0.831	Yes
8308588	24.766 ↓ 16.950	0.222 ↓ 0.113	0.536 ↑ 0.940	Yes
8326944	101.282 ↓ 79.295	0.634 ↓ 0.476	0.678 ↑ 0.819	Yes
8328122	34.642 ↓ 26.993	0.208 ↓ 0.160	0.741 ↑ 0.843	Yes
8334780	26.261 ↓ 20.185	0.158 ↓ 0.103	0.753 ↑ 0.902	Yes
8342852	38.704 ↑ 46.294	0.227 ↑ 0.550	0.941 ↓ 0.643	No
8347238	36.339 ↓ 30.474	0.190 ↓ 0.132	0.698 ↑ 0.872	Yes
8350006	21.296 ↓ 11.010	0.145 ↓ 0.078	0.751 ↑ 0.933	Yes
8351602	23.758 ↓ 15.765	0.235 ↓ 0.139	0.804 ↑ 0.931	Yes
8376656	40.295 ↓ 28.208	0.139 ↓ 0.102	0.765 ↑ 0.895	Yes
8419708	26.055 ↓ 17.694	0.277 ↓ 0.129	0.691 ↑ 0.948	Yes
8432046	63.266 ↓ 38.805	0.477 ↓ 0.299	0.768 ↑ 0.921	Yes
8451629	31.757 ↓ 19.067	0.183 ↓ 0.105	0.783 ↑ 0.932	Yes
8459427	21.150 ↓ 17.217	0.071 ↓ 0.048	0.758 ↑ 0.827	Yes
8466525	29.328 ↓ 19.620	0.179 ↓ 0.121	0.828 ↑ 0.937	Yes
8478501	51.373 ↓ 46.553	0.167 ↑ 0.220	0.834 ↓ 0.702	Inc.
8482121	31.288 ↑ 43.159	0.110 ↑ 0.143	0.785 ↓ 0.614	No
8487285	24.816 ↓ 9.930	0.198 ↓ 0.091	0.565 ↑ 0.950	Yes
8487297	141.440 ↓ 107.062	0.467 ↓ 0.397	0.561 ↑ 0.695	Yes
8487461	105.135 ↓ 57.342	0.225 ↓ 0.104	0.537 ↑ 0.914	Yes
8496980	37.289 ↓ 24.353	0.357 ↓ 0.212	0.532 ↑ 0.857	Yes
8504552	24.409 ↓ 9.628	0.159 ↓ 0.052	0.749 ↑ 0.928	Yes
8519102	78.204 ↓ 76.021	0.153 ↓ 0.128	0.395 ↑ 0.481	Yes
8523058	46.623 ↓ 36.287	0.338 ↓ 0.249	0.635 ↑ 0.805	Yes
8540084	78.493 ↓ 25.351	0.226 ↓ 0.106	0.662 ↑ 0.955	Yes
8557605	50.855 ↓ 47.167	0.386 ↓ 0.248	0.835 ↑ 0.931	Yes
8566459	53.547 ↓ 49.524	0.274 ↓ 0.243	0.715 ↑ 0.794	Yes
8617151	34.169 ↓ 19.291	0.220 ↓ 0.156	0.859 ↑ 0.925	Yes
8618165	39.063 ↓ 16.300	0.315 ↓ 0.094	0.571 ↑ 0.968	Yes
8655993	35.215 ↓ 27.372	0.208 ↓ 0.107	0.701 ↑ 0.908	Yes
8661542	31.606 ↓ 19.331	0.258 ↓ 0.139	0.814 ↑ 0.947	Yes
8673172	47.745 ↓ 35.768	0.337 ↓ 0.214	0.794 ↑ 0.932	Yes
8679346	74.113 ↓ 40.960	0.277 ↓ 0.136	0.668 ↑ 0.918	Yes
8680284	69.775 ↓ 35.542	0.275 ↓ 0.128	0.825 ↑ 0.962	Yes
8685932	17.555 ↓ 9.443	0.111 ↓ 0.044	0.834 ↑ 0.975	Yes
8687500	34.620 ↓ 27.600	0.239 ↓ 0.200	0.667 ↑ 0.830	Yes
8733828	32.093 ↓ 19.590	0.129 ↓ 0.070	0.810 ↑ 0.956	Yes
8804804	36.389 ↓ 28.881	0.396 ↓ 0.307	0.791 ↑ 0.873	Yes
9012348	179.210 ↓ 108.381	0.516 ↓ 0.359	0.282 ↑ 0.690	Yes
9393680	100.083 ↓ 89.805	0.467 ↓ 0.435	0.557 ↑ 0.624	Yes
10509861	28.132 ↓ 22.165	0.242 ↓ 0.219	0.724 ↑ 0.776	Yes
10595596	81.574 ↓ 63.148	0.537 ↓ 0.345	0.676 ↑ 0.871	Yes
10598990	49.664 ↓ 20.330	0.775 ↓ 0.277	0.656 ↑ 0.957	Yes
10692972	48.471 ↓ 38.768	0.251 ↓ 0.177	0.842 ↑ 0.920	Yes
10702066	41.057 ↓ 32.030	0.102 ↓ 0.053	0.479 ↑ 0.862	Yes
11081920	31.636 ↓ 22.638	0.285 ↓ 0.201	0.657 ↑ 0.881	Yes
11462018	27.691 ↓ 18.426	0.099 ↓ 0.078	0.865 ↑ 0.900	Yes

APPENDIX B

BIGGER VERSIONS OF PLOTS IN CHAPTER 8

B.1 Bigger Versions of some Plots Exhibiting the PFE.

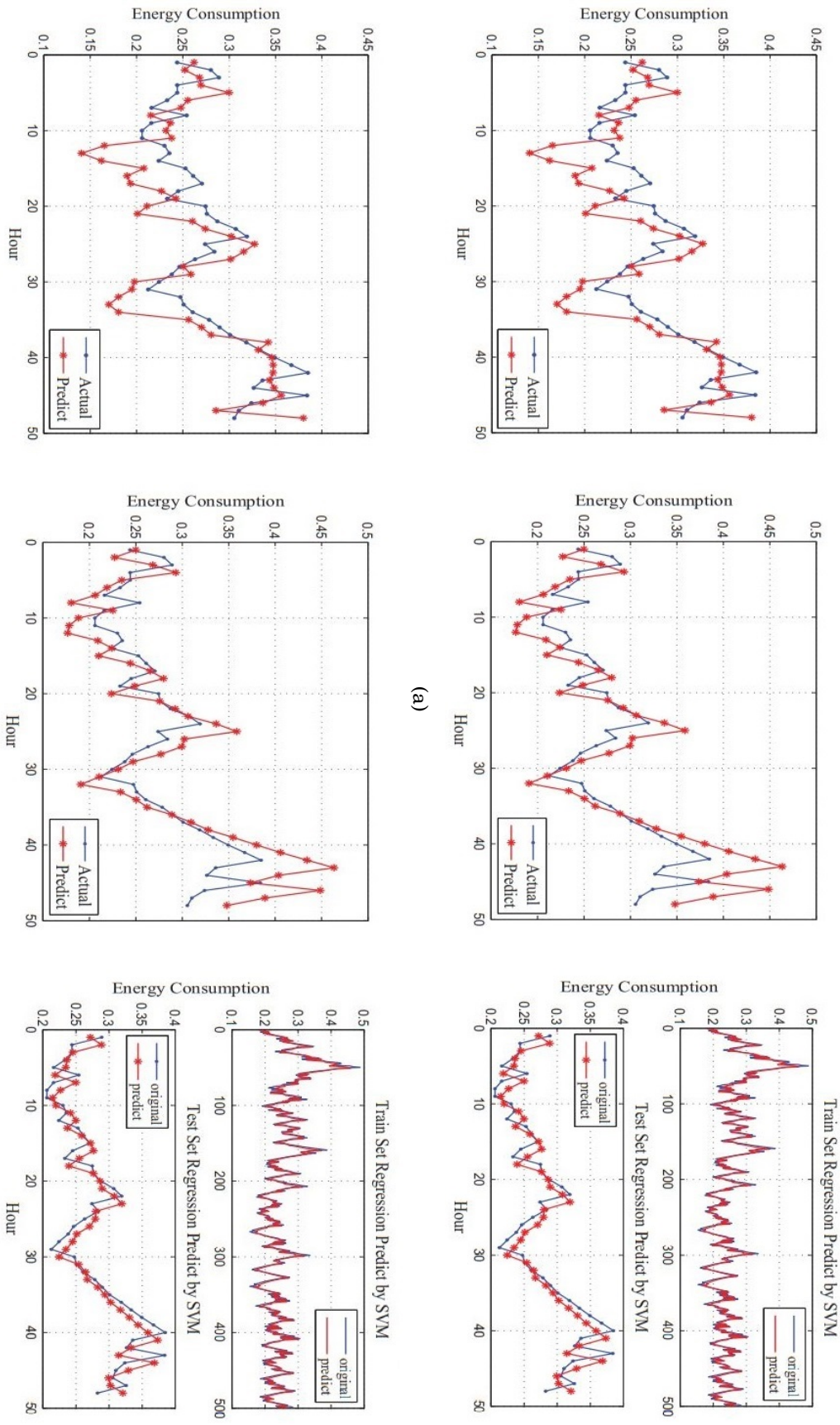
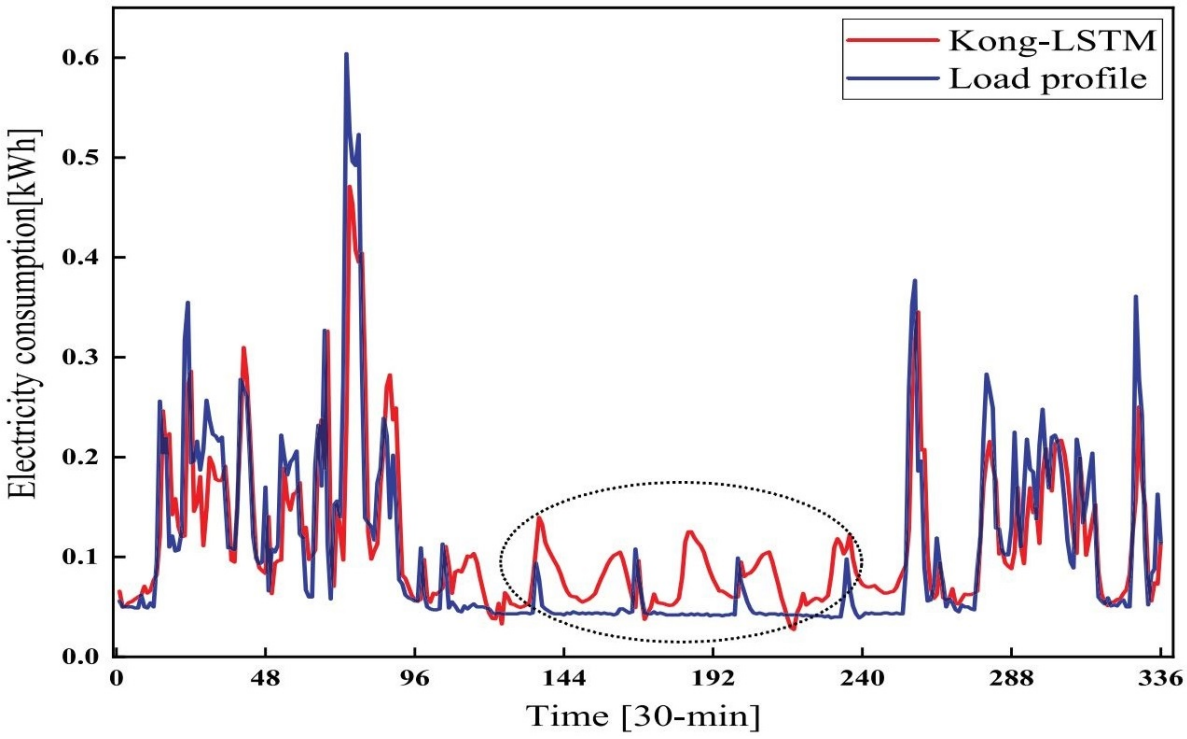
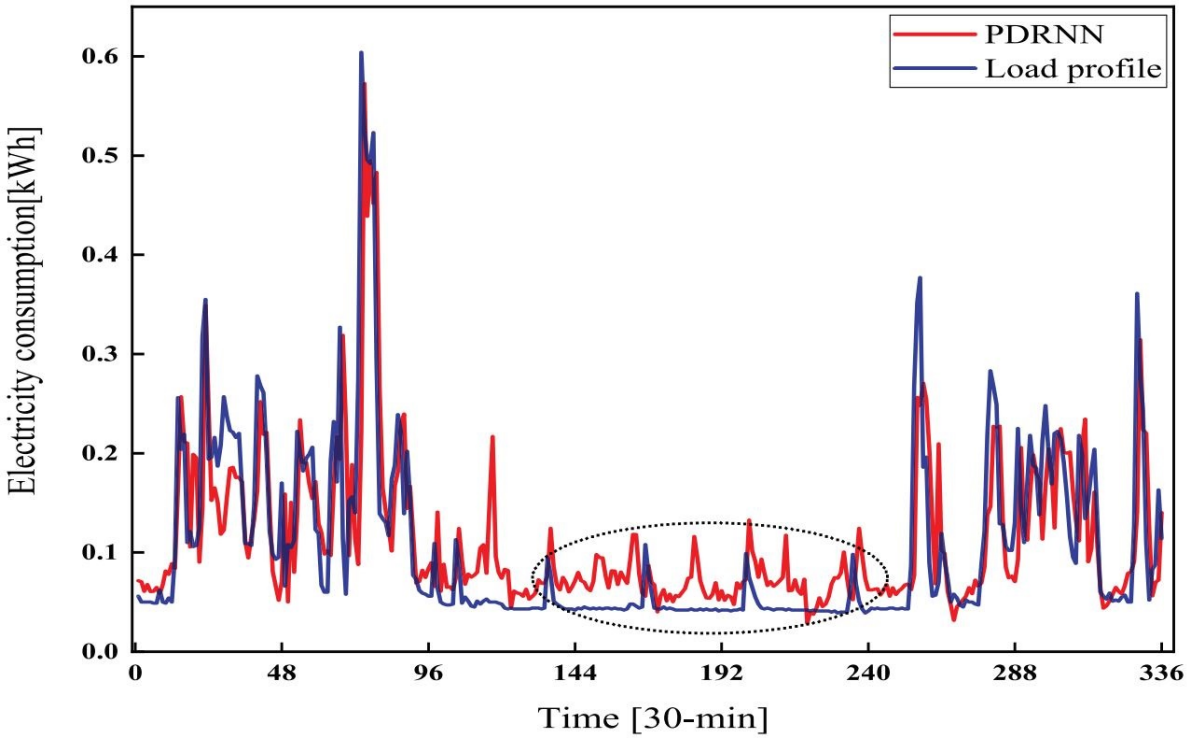


Figure B.1: Bigger version of plots in (a) Figures 8.1 and (b) Figures 8.2, also Figures 3 and 4 in [80].

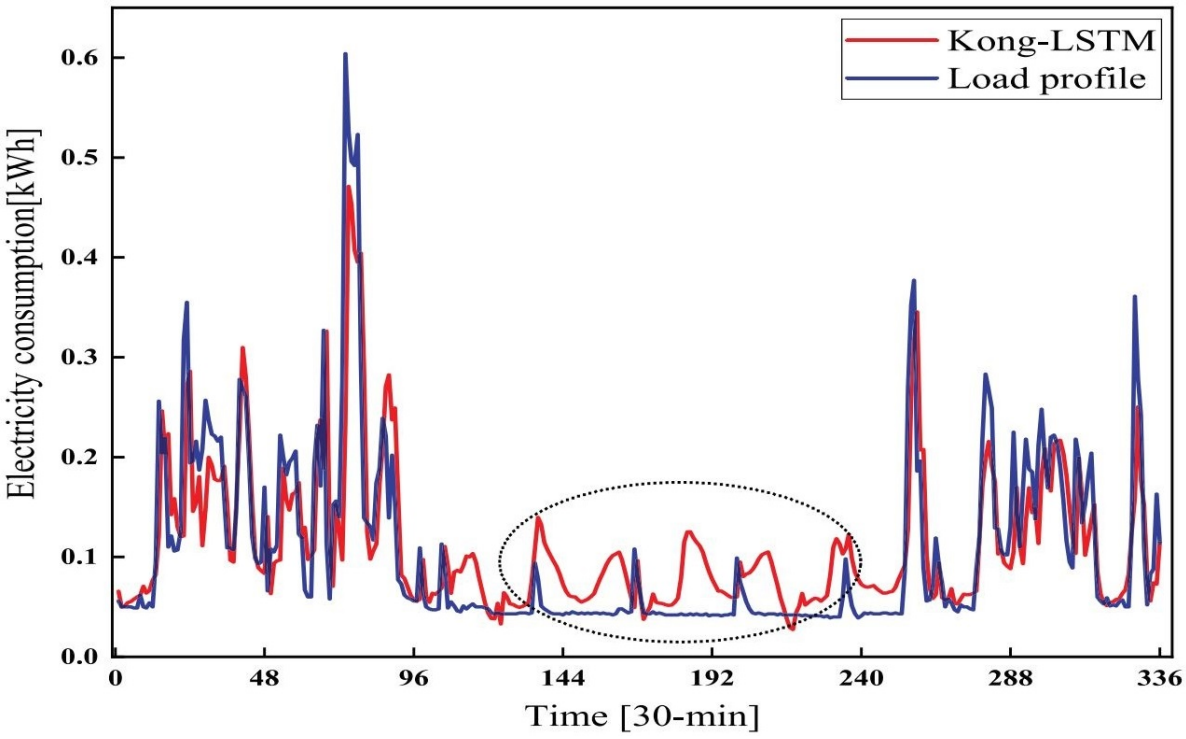


(a)

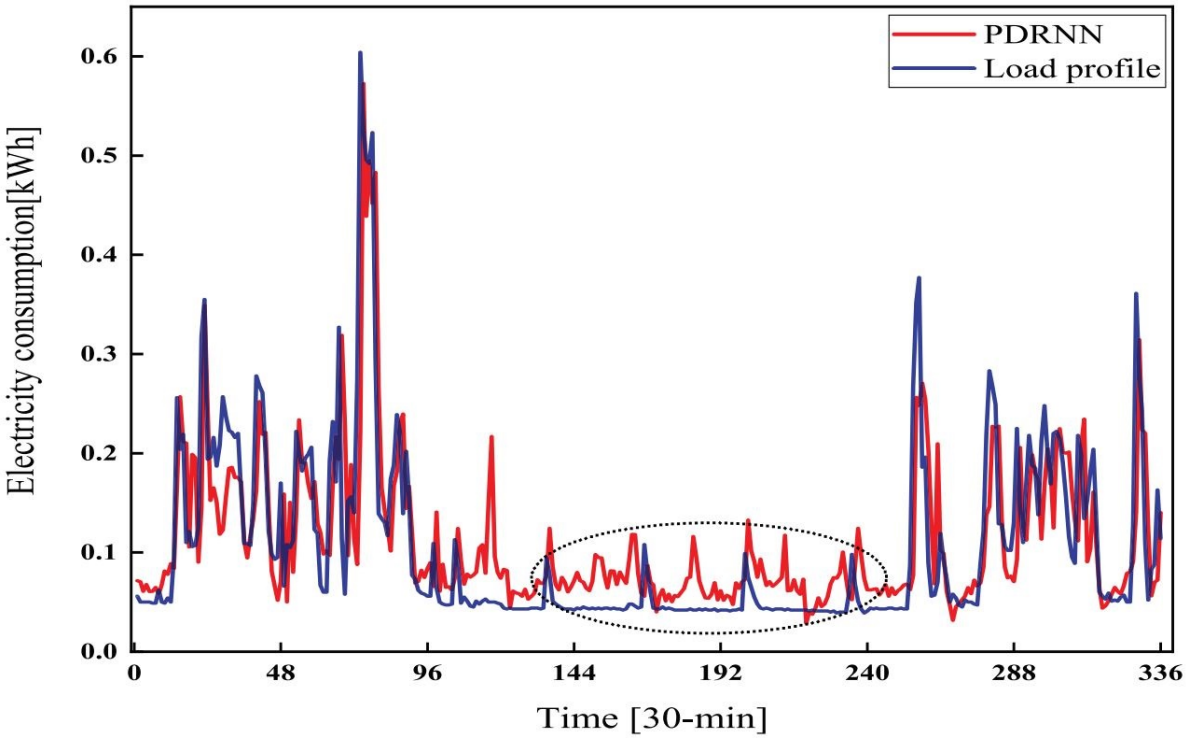


(b)

Figure B.2: Continued on next page.



(c)



(d)

Figure B.2: Bigger version of plots in Figure 8.3, also Figure 8 in [80].

BIBLIOGRAPHY

[1] R. Rossi and K. Hirama, "Characterizing Big Data Management," *CoRR*, vol. abs/2201.05929, 2022, Accessed on: 23.11.2022. [Online]. Available: https://arxiv.org/abs/2201.05929

[2] W. Li, Y. Chai, F. Khan, S. R. U. Jan, S. Verma, V. G. Menon, X. Li *et al.*, "A Comprehensive Survey on Machine Learning-Based Big Data Analytics for IoT-Enabled Smart Healthcare System," *Mobile Networks and Applications*, vol. 26, no. 1, pp. 234–252, 2021.

[3] F. Cappa, R. Oriani, E. Peruffo, and I. McCarthy, "Big Data for Creating and Capturing Value in the Digitalized Environment: Unpacking the Effects of Volume, Variety, and Veracity on Firm Performance," *Journal of Product Innovation Management*, vol. 38, no. 1, pp. 49–67, 2021.

[4] Y. Hajjaji, W. Boulila, I. R. Farah, I. Romdhani, and A. Hussain, "Big Data and IoT-Based Applications in Smart Environments: A Systematic Review," *Computer Science Review*, vol. 39, 2021, Article no. 100318.

[5] W. Haoxiang, S. Smys *et al.*, "Big Data Analysis and Perturbation Using Data Mining Algorithm," *Journal of Soft Computing Paradigm*, vol. 3, no. 01, pp. 19–28, 2021.

[6] D. Lazos, A. B. Sproul, and M. Kay, "Optimisation of Energy Management in Commercial Buildings with Weather Forecasting Inputs: A Review," *Renewable and Sustainable Energy Reviews*, vol. 39, pp. 587–603, 2014.

[7] A. R. Khan, A. Mahmood, A. Safdar, Z. A. Khan, and N. A. Khan, "Load Forecasting, Dynamic Pricing and DSM in Smart Grid: A Review," *Renewable and Sustainable Energy Reviews*, vol. 54, pp. 1311–1322, 2016.

[8] S. K. Sharma and X. Wang, "Live Data Analytics with Collaborative Edge and Cloud Processing in Wireless IoT Networks," *IEEE Access*, vol. 5, pp. 4621–4635, 2017.

[9] M. A. Mat Daut, M. Y. Hassan, H. Abdullah, H. A. Rahman, M. P. Abdullah, and F. Hussin, "Building Electrical Energy Consumption Forecasting Analysis Using Conventional and

Artificial Intelligence Methods: A Review," *Renewable and Sustainable Energy Reviews*, vol. 70, pp. 1108–1118, 2017.

[10] R. J. Hyndman and A. B. Koehler, "Another look at measures of forecast accuracy," *International Journal of Forecasting*, vol. 22, no. 4, pp. 679–688, 2006.

[11] J. Li, "Assessing the accuracy of predictive models for numerical data: Not r nor r2, why not? then what?" *PloS one*, vol. 12, no. 8, 2017, Article no. e0183250.

[12] H. B. Akyol, C. Preist, and D. Schien, "Avoiding Overconfidence in Predictions of Residential Energy Demand Through Identification of the Persistence Forecast Effect," *IEEE Transactions on Smart Grid*, vol. 14, no. 1, pp. 228–238, 2023.

[13] B. L. Bowerman and R. T. O'Connell, *Time Series Forecasting: Unified Concepts and Computer Implementation*, 2nd ed. Boston, United States: Duxbury Press, 1987, ISBN: 978-0871500700.

[14] G. Janacek and L. Swift, *Time Series: Forecasting, Simulation, Applications*. New York, United States: E. Horwood, 1993, ISBN: 978-0139184598.

[15] J. D. Hamilton, *Time Series Analysis*. New Jersey, United States: Princeton University Press, 1994, ISBN: 978-0691042893.

[16] C. Chatfield, *Time-Series Forecasting*. New York, United States: Chapman and Hall/CRC, 2000, ISBN: 978-1584880639.

[17] S. Bisgaard and M. Kulahci, *Time Series Analysis and Forecasting by Example*. New Jersey, United States: Wiley, 2011, ISBN: 978-0470540640.

[18] G. E. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, *Time Series Analysis: Forecasting and Control*, 5th ed. New Jersey, United States: John Wiley & Sons, 2015, ISBN: 978-1118675021.

[19] R. J. Hyndman and G. Athanasopoulos, *Forecasting: Principles and Practice*, 2nd ed. Melbourne, Australia: OTexts, 2018, ISBN: 978-0987507112.

[20] R. Adhikari and R. K. Agrawal, "An Introductory Study on Time Series Modeling and Forecasting," *CoRR*, vol. abs/1302.6613, 2013, Accessed on: 16.05.2022. [Online]. Available: http://arxiv.org/abs/1302.6613

[21] J. Grandell, "Time series analysis," *Lecture Notes (KTH, Sweden, 2000). http://www.math.kth.se/matstat/gru/sf2943/ts.pdf*, 1998, Accessed on: 22.12.2022.

[22] R. Prado and M. West, *Time Series Modelling, Inference and Forecasting*. London, United Kingdom: Chapman and Hall/CRC, 2010, ISBN: 978-1420093360.

[23] A. S. Weigend, *Time Series Prediction: Forecasting The Future and Understanding The Past*. London, United Kingdom: Routledge, 2019, ISBN: 978-0429972270.

[24] W. Palma, *Time Series Analysis*. New Jersey, United States: John Wiley & Sons, 2016, ISBN: 978-1118634325.

[25] R. H. Shumway and D. S. Stoffer, *Time Series: A Data Analysis Approach Using R*. London, United Kingdom: Routledge, 2019, ISBN: 978-0367221096.

[26] G. Kirchgässner, J. Wolters, and U. Hassler, *Introduction to Modern Time Series Analysis (Springer Texts in Business and Economics)*, 2nd ed. Berlin, Germany: Springer, 2013, ISBN: 978-3642334351.

[27] P. Diggle, *Time Series: A Biostatistical Introduction*. New York, United States: Oxford University Press, 1990, ISBN: 978-0198522263.

[28] W. W. Wei, "Time Series Analysis," in *The Oxford Handbook of Quantitative Methods in Psychology: Vol. 2: Statistical Analysis*. New York, United States: Oxford University Press, 2013, ISBN: 978-0199934898.

[29] C. Chatfield and X. Haipeng, *The Analysis of Time Series: An Introduction with R*, 7th ed. Chapman and Hall/CRC, 2019, ISBN: 978-1138066137.

[30] J. D. Cryer and K.-S. Chan, *Time Series Analysis: With Applications in R*, 2nd ed. New York, United States: Springer, 2008, ISBN: 978-0387759593.

[31] D. Pena, G. C. Tiao, and R. S. Tsay, *A Course in Time Series Analysis*. New York, United States: Wiley, 2011, ISBN: 978-0471361640.

[32] P. Brabban, "Daily Minimum Temperatures in Melbourne," [Online], https://www.kaggle.com/datasets/paulbrabban/daily-minimum-temperatures-in-melbourne, Accessed on: 10.10.2022.

[33] D. Andreazzini, "International Airline Passengers," [Online], https://www.kaggle.com/datasets/andreazzini/international-airline-passengers, Accessed on: 10.10.2022.

[34] P. J. Brockwell and R. A. Davis, *Introduction to Time Series and Forecasting*, 2nd ed. Berlin, Germany: Springer, 2002, ISBN: 978-0387216577.

[35] D. S. G. Pollock, R. C. Green, and T. Nguyen, *Handbook of Time Series Analysis, Signal Processing, and Dynamics*. Elsevier, 1999, ISBN: 978-0080507873.

[36] S. Ben Taieb, "Machine Learning Strategies for Multi-Step-Ahead Time Series Forecasting," Ph.D. dissertation, Université Libre de Bruxelles, Belgium, 2014.

[37] C. Chatfield, A. B. Koehler, J. K. Ord, and R. D. Snyder, "A New Look at Models for Exponential Smoothing," *Journal of the Royal Statistical Society. Series D (The Statistician)*, vol. 50, no. 2, pp. 147–159, 2001.

[38] F. Chollet, *Deep Learning with Python*, 2nd ed. New York, United States: Manning Publications Inc., 2021, ISBN: 978-1638350095.

[39] N. Sanders, *Forecasting Fundamentals*, 4th ed. New York, United States: Business Expert Press, 2016, ISBN: 978-1606498712.

[40] T. Ahmad, H. Chen, Y. Guo, and J. Wang, "A Comprehensive Overview on the Data Driven and Large Scale Based Approaches for Forecasting of Building Energy Demand: A Review," *Energy and Buildings*, vol. 165, pp. 301–320, 2018.

[41] P. Mesarić and S. Krajcar, "Home Demand Side Management Integrated with Electric Vehicles and Renewable Energy Sources," *Energy and Buildings*, vol. 108, pp. 1–9, 2015.

[42] K. B. Debnath and M. Mourshed, "Forecasting Methods in Energy Planning Models," *Renewable and Sustainable Energy Reviews*, vol. 88, pp. 297–325, 2018.

[43] P. Lusis, K. R. Khalilpour, L. Andrew, and A. Liebman, "Short-Term Residential Load Forecasting: Impact of Calendar Effects and Forecast Granularity," *Applied Energy*, vol. 205, pp. 654–669, 2017.

[44] R. Bonetto and M. Rossi, "Machine Learning Approaches to Energy Consumption Forecasting in Households," *CoRR*, vol. abs/1706.09648, 2017, Accessed on: 01.04.2019. [Online]. Available: http://arxiv.org/abs/1706.09648

[45] M. Q. Raza and A. Khosravi, "A Review on Artificial Intelligence Based Load Demand Forecasting Techniques for Smart Grid and Buildings," *Renewable and Sustainable Energy Reviews*, vol. 50, pp. 1352–1372, 2015.

[46] A. Rahman, V. Srikumar, and A. D. Smith, "Predicting Electricity Consumption for Commercial and Residential Buildings Using Deep Recurrent Neural Networks," *Applied Energy*, vol. 212, pp. 372–385, 2018.

[47] T. Hong and S. Fan, "Probabilistic Electric Load Forecasting: A Tutorial Review," *International Journal of Forecasting*, vol. 32, no. 3, pp. 914–938, 2016.

[48] P. Cichosz, *Data Mining Algorithms: Explained Using R.* New Jersey, United States: Wiley, 2014, ISBN: 978-1118332580.

[49] A. V. Hill, *The Encyclopedia of Operations Management*. New Jersey, United States: FT Press, 2012, ISBN: 978-0132883702.

[50] V. Cerqueira, L. Torgo, J. Smailović, and I. Mozetič, "A Comparative Study of Performance Estimation Methods for Time Series Forecasting," in *Proceedings of IEEE International Conference on Data Science and Advanced Analytics*, 2017, pp. 529–538.

[51] C. Bergmeir, R. J. Hyndman, and B. Koo, "A Note on the Validity of Cross-Validation for Evaluating Autoregressive Time Series Prediction," *Computational Statistics Data Analysis*, vol. 120, pp. 70–83, 2018.

[52] J. P. Donate, P. Cortez, G. G. Sánchez, and A. S. de Miguel, "Time Series Forecasting Using A Weighted Cross-Validation Evolutionary Artificial Neural Network Ensemble," *Neurocomputing*, vol. 109, pp. 27–32, 2013, new trends on Soft Computing Models in Industrial and Environmental Applications.

[53] C. Bergmeir and J. M. Benítez, "On the use of Cross-Validation for Time Series Predictor Evaluation," *Information Sciences*, vol. 191, pp. 192–213, 2012, data Mining for Software Trustworthiness.

[54] Q. Li, M. Yan, and J. Xu, "Optimizing Convolutional Neural Network Performance by Mitigating Underfitting and Overfitting," in *Proceedings of the 19th International Conference on Computer and Information Science*, 2021, pp. 126–131.

[55] K. Kouvaris, J. Clune, L. Kounios, M. Brede, and R. A. Watson, "How Evolution Learns to Generalise: Principles of under-fitting, over-fitting and induction in the evolution of developmental organisation," *CoRR*, vol. abs/1508.06854, 2015, Accessed on: 26.09.2022. [Online]. Available: http://arxiv.org/abs/1508.06854

[56] W. M. Van Der Aalst, V. Rubin, H. M. Verbeek, B. F. Van Dongen, E. Kindler, and C. W. Günther, "Process Mining: A Two-Step Approach to Balance Between Underfitting and Overfitting," *Software and Systems Modeling*, vol. 9, no. 1, pp. 87–111, 2010.

[57] H. Zhang, L. Zhang, and Y. Jiang, "Overfitting and Underfitting Analysis for Deep Learning Based End-to-end Communication Systems," in *Proceedings of the 11th International Conference on Wireless Communications and Signal Processing*, 2019, pp. 1–6.

[58] D. Bashir, G. D. Montañez, S. Sehra, P. S. Segura, and J. Lauw, "An Information-Theoretic Perspective on Overfitting and Underfitting," in *Proceeding of AI 2020: Advances in Artificial Intelligence*, M. Gallagher, N. Moustafa, and E. Lakshika, Eds. Cham, Switzerland: Springer International Publishing, 2020, pp. 347–358.

[59] E. Enes, "Determining Overfitting and Underfitting in Generative Adversarial Networks Using Fréchet Distance," *Turkish Journal of Electrical Engineering and Computer Sciences*, vol. 29, no. 3, pp. 1524–1538, 2021.

[60] P. Baheti, "What is Overfitting in Deep Learning [+10 Ways to Avoid It]," [Online], 2022, https://www.v7labs.com/blog/overfitting, Accessed on: 22.12.2022.

[61] J. A. Bullinaria, "Bias and Variance, Under-Fitting and Over-Fitting," *Neural Computation: Lecture 9*, 2015.

[62] A. Conway, K. Macpherson, and J. Brown, "Delayed Time Series Predictions with Neural Networks," *Neurocomputing*, vol. 18, no. 1, pp. 81–89, 1998.

[63] M. Lange, "On the Uncertainty of Wind Power Predictions - Analysis of the Forecast Accuracy and Statistical Distribution of Errors," *Journal of Solar Energy Engineering*, vol. 127, no. 2, pp. 177–184, 2005.

[64] Y. Fujimoto, Y. Takahashi, and Y. Hayashi, "Alerting to Rare Large-Scale Ramp Events in Wind Power Generation," *IEEE Transactions on Sustainable Energy*, vol. 10, no. 1, pp. 55–65, 2019.

[65] P. Dixit, S. Londhe, and Y. Dandawate, "Removing Prediction Lag in Wave Height Forecasting Using Neuro - Wavelet Modeling Technique," *Ocean Engineering*, vol. 93, pp. 74–83, 2015.

[66] L. L. Takacs, "A Two-Step Scheme for the Advection Equation with Minimized Dissipation and Dispersion Errors," *Monthly Weather Review*, vol. 113, no. 6, pp. 1050–1065, 1985.

[67] D. Hou, E. Kalnay, and K. K. Droegemeier, "Objective Verification of the SAMEX '98 Ensemble Forecasts," *Monthly Weather Review*, vol. 129, no. 1, pp. 73–91, 2001.

[68] M. Cui, B. M. Hodge, J. Zhang, D. Ke, A. Florita, and Y. Sun, "Solar Power Ramp Events Detection Using an Optimized Swinging Door Algorithm," in *Proceeding of the ASME International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, vol. 2A: 41st Design Automation Conference, Boston, Massachusetts, USA, 2015, Article no. V02AT03A027.

[69] W. Zhu, L. Zhang, M. Yang, and B. Wang, "Solar Power Ramp Event Forewarning with Limited Historical Observations," *IEEE Transactions on Industry Applications*, vol. 55, no. 6, pp. 5621–5630, 2019.

[70] M. Abuella and B. Chowdhury, "Forecasting of Solar Power Ramp Events: A Post-Processing Approach," *Renewable Energy*, vol. 133, pp. 1380–1392, 2019.

[71] C. Ferreira, J. Gama, L. Matias, A. Botterud, and J. Wang, "A Survey on Wind Power Ramp Forecasting," Argonne National Laboratory, Tech. Rep., 2011.

[72] T. Ouyang, X. Zha, L. Qin *et al.*, "A Survey of Wind Power Ramp Forecasting," *Energy and Power Engineering*, vol. 5, no. 04, pp. 368–372, 2013.

[73] C. Gallego-Castillo, A. Cuerva-Tejero, and O. Lopez-Garcia, "A Review on The Recent History of Wind Power Ramp Forecasting," *Renewable and Sustainable Energy Reviews*, vol. 52, pp. 1148–1157, 2015.

[74] L. Vallance, B. Charbonnier, N. Paul, S. Dubost, and P. Blanc, "Towards a Standardized Procedure to Assess Solar Forecast Accuracy: A New Ramp and Time Alignment Metric," *Solar Energy*, vol. 150, pp. 408–422, 2017.

[75] N. J. de Vos and T. H. M. Rientjes, "Constraints of Artificial Neural Networks for Rainfall-Runoff Modelling: Trade-Offs in Hydrological State Representation and Model Evaluation," *Hydrology and Earth System Sciences*, vol. 9, no. 1/2, pp. 111–126, 2005.

[76] G. H. Erharter and T. Marcher, "On the Pointlessness of Machine Learning Based Time Delayed Prediction of TBM Operational Data," *Automation in Construction*, vol. 121, 2021, Article no. 103443.

[77] S. Haben, J. Ward, D. Vukadinovic Greetham, C. Singleton, and P. Grindrod, "A New Error Measure for Forecasts of Household-Level, High Resolution Electrical Energy Consumption," *International Journal of Forecasting*, vol. 30, no. 2, pp. 246–256, 2014.

[78] M. Voss, "Permutation-Based Residential Short-term Load Forecasting in the Context of Energy Management Optimization Objectives," in *Proceeding of the 11th ACM International Conference on Future Energy Systems*, 2020, pp. 231–236.

[79] T. Zufferey, A. Lepouze, and G. Hug, "Inadequacy of Standard Algorithms and Metrics for Short-Term Load Forecasts in Low-Voltage Grids," in *Proceedings of the IEEE Milan PowerTech*. IEEE, 2019, pp. 1–6.

[80] L. Jiang, X. Wang, W. Li, L. Wang, X. Yin, and L. Jia, "Hybrid Multitask Multi-Information Fusion Deep Learning for Household Short-Term Load Forecasting," *IEEE Transactions on Smart Grid*, vol. 12, no. 6, pp. 5362–5372, 2021.

[81] I. K. Nti, M. Teimeh, A. F. Adekoya, and O. Nyarko-boateng, "Forecasting Electricity Consumption of Residential Users Based on Lifestyle Data Using Artificial Neural Networks," *ICTACT Journal on Soft Computing*, vol. 10, no. 3, pp. 2107–2116, 2020.

[82] C. Tian, C. Li, G. Zhang, and Y. Lv, "Data Driven Parallel Prediction of Building Energy Consumption Using Generative Adversarial Nets," *Energy and Buildings*, vol. 186, pp. 230–243, 2019.

[83] D. L. Marino, K. Amarasinghe, and M. Manic, "Building Energy Load Forecasting using Deep Neural Networks," in *Proceedings of IECON 42nd Annual Conference of the IEEE Industrial Electronics Society*. Florence, Italy: IEEE, 2016, pp. 7046–7051.

[84] J. Xiao, Y. Li, L. Xie, D. Liu, and J. Huang, "A Hybrid Model Based on Selective Ensemble for Energy Consumption Forecasting in China," *Energy*, vol. 159, pp. 534–546, 2018.

[85] L. Fan, J. Li, and X. P. Zhang, "Load Prediction Methods Using Machine Learning for Home Energy Management Systems Based on Human Behavior Patterns Recognition," *CSEE Journal of Power and Energy Systems*, vol. 6, no. 3, pp. 563–571, 2020.

[86] F. Lin, S. J. Liu, H. C. Chao, and J. S. Pan, "Short-Term Household Load Forecasting Model Based on Variational Mode Decomposition and Gated Recurrent Unit with Attention Mechanism," *Journal of Network Intelligence*, vol. 6, no. 1, pp. 143–153, 2021.

[87] X. Guo, Y. Gao, Y. Li, D. Zheng, and D. Shan, "Short-Term Household Load Forecasting Based on Long- and Short-Term Time-Series Network," *Energy Reports*, vol. 7, pp. 58–64, 2021, iCPE 2020-The International Conference on Power Engineering.

[88] F. He, J. Zhou, Z. kai Feng, G. Liu, and Y. Yang, "A Hybrid Short-Term Load Forecasting Model Based on Variational Mode Decomposition and Long Short-Term Memory Networks Considering Relevant Factors with Bayesian Optimization algorithm," *Applied Energy*, vol. 237, pp. 103–116, 2019.

[89] W. Kong, Z. Y. Dong, Y. Jia, D. J. Hill, Y. Xu, and Y. Zhang, "Short-Term Residential Load Forecasting Based on LSTM Recurrent Neural Network," *IEEE Transactions on Smart Grid*, vol. 10, no. 1, pp. 841–851, 2019.

[90] M. Alhussein, K. Aurangzeb, and S. I. Haider, "Hybrid CNN-LSTM Model for Short-Term Individual Household Load Forecasting," *IEEE Access*, vol. 8, pp. 180 544–180 557, 2020.

[91] W. Kong, Z. Y. Dong, D. J. Hill, F. Luo, and Y. Xu, "Short-Term Residential Load Forecasting Based on Resident Behaviour Learning," *IEEE Transactions on Power Systems*, vol. 33, no. 1, pp. 1087–1088, 2018.

[92] Z. Zheng, H. Chen, and X. Luo, "A Kalman Filter-Based Bottom-up Approach for Household Short-Term Load Forecast," *Applied Energy*, vol. 250, pp. 882–894, 2019.

[93] A. Estebsari and R. Rajabi, "Single Residential Load Forecasting Using Deep Learning and Image Encoding Techniques," *Electronics*, vol. 9, no. 1, pp. 68–85, 2020.

[94] S. Wang, X. Deng, H. Chen, Q. Shi, and D. Xu, "A Bottom-up Short-Term Residential Load Forecasting Approach Based on Appliance Characteristic Analysis and Multi-Task Learning," *Electric Power Systems Research*, vol. 196, 2021, Article no. 107233.

[95] N. M. Ibrahim, A. I. Megahed, and N. H. Abbasy, "Short-Term Individual Household Load Forecasting Framework Using LSTM Deep Learning Approach," in *Proceeding of the*

5th International Symposium on Multidisciplinary Studies and Innovative Technologies, 2021, pp. 257–262.

[96] H. Shi, M. Xu, and R. Li, "Deep Learning for Household Load Forecasting – A Novel Pooling Deep RNN," *IEEE Transactions on Smart Grid*, vol. 9, no. 5, 2018.

[97] H. Zhong, J. Wang, H. Jia, Y. Mu, and S. Lv, "Vector Field-Based Support Vector Regression for Building Energy Consumption prediction," *Applied Energy*, vol. 242, pp. 403–414, 2019.

[98] S. Seyedzadeh, F. Pour Rahimian, P. Rastogi, and I. Glesk, "Tuning Machine Learning Models for Prediction of Building Energy Loads," *Sustainable Cities and Society*, vol. 47, 2019, Article no. 101484.

[99] K. Li, X. Xie, W. Xue, X. Dai, X. Chen, and X. Yang, "A Hybrid Teaching-Learning Artificial Neural Network for Building Electrical Energy Consumption Prediction," *Energy and Buildings*, vol. 174, pp. 323–334, 2018.

[100] K. Gajowniczek and T. Zabkowski, "Electricity Forecasting on the Individual Household Level Enhanced Based on Activity Patterns," *PLoS ONE*, vol. 12, no. 4, pp. 1–26, 2017.

[101] N. Andriopoulos, A. Magklaras, A. Birbas, A. Papalexopoulos, C. Valouxis, S. Daskalaki, M. Birbas, E. Housos, and G. P. Papaioannou, "Short Term Electric Load Forecasting Based on Data Transformation and Statistical Machine Learning," *Applied Sciences*, vol. 11, no. 1, pp. 1–22, 2021.

[102] A. Yang, W. Li, and X. Yang, "Short-Term Electricity Load Forecasting Based on Feature Selection and Least Squares Support Vector Machines," *Knowledge-Based Systems*, vol. 163, pp. 159–173, 2019.

[103] J. Zhang, Y.-M. Wei, D. Li, Z. Tan, and J. Zhou, "Short Term Electricity Load Forecasting Using a Hybrid Model," *Energy*, vol. 158, pp. 774–781, 2018.

[104] L. Li, C. J. Meinrenken, V. Modi, and P. J. Culligan, "Short-Term Apartment-Level Load Forecasting Using a Modified Neural Network with Selected Auto-Regressive Features," *Applied Energy*, vol. 287, 2021, Article no. 116509.

[105] Z. Wang, Y. Wang, R. Zeng, R. S. Srinivasan, and S. Ahrentzen, "Random Forest Based Hourly Building Energy Prediction," *Energy and Buildings*, vol. 171, pp. 11–25, 2018.

[106] M. Imani and H. Ghassemian, "Residential Load Forecasting Using Wavelet and Collaborative Representation Transforms," *Applied Energy*, vol. 253, 2019, Article no. 113505.

[107] S. Naji, A. Keivani, S. Shamshirband, U. J. Alengaram, M. Z. Jumaat, Z. Mansor, and M. Lee, "Estimating Building Energy Consumption Using Extreme Learning Machine Method," *Energy*, vol. 97, pp. 506–516, 2016.

[108] T.-Y. Kim and S.-B. Cho, "Predicting Residential Energy Consumption using CNN-LSTM Neural Networks," *Energy*, vol. 182, pp. 72–81, 2019.

[109] J.-S. Chou and D.-S. Tran, "Forecasting Energy Consumption Time Series Using Machine Learning Techniques Based on Usage Patterns of Residential Householders," *Energy*, vol. 165, pp. 709–726, 2018.

[110] D. W. van der Meer, J. Widén, and J. Munkhammar, "Review on Probabilistic Forecasting of Photovoltaic Power Production and Electricity Consumption," *Renewable and Sustainable Energy Reviews*, vol. 81, pp. 1484–1512, 2018.

[111] Y. T. Chae, R. Horesh, Y. Hwang, and Y. M. Lee, "Artificial Neural Network Model for Forecasting Sub-Hourly Electricity Usage in Commercial Buildings," *Energy and Buildings*, vol. 111, pp. 184–194, 2016.

[112] X. Liu, Z. Zhang, and Z. Song, "A Comparative Study of The Data-Driven Day-Ahead Hourly Provincial Load Forecasting Methods: From Classical Data Mining to Deep Learning," *Renewable and Sustainable Energy Reviews*, vol. 119, 2020, Article no. 109632.

[113] C. Deb, F. Zhang, J. Yang, S. E. Lee, and K. W. Shah, "A Review on Time Series Forecasting Techniques for Building Energy Consumption," *Renewable and Sustainable Energy Reviews*, vol. 74, pp. 902–924, 2017.

[114] M. Rowe, T. Yunusov, S. Haben, C. Singleton, W. Holderbaum, and B. Potter, "A Peak Reduction Scheduling Algorithm for Storage Devices on the Low Voltage Network," *IEEE Transactions on Smart Grid*, vol. 5, no. 4, pp. 2115–2124, 2014.

[115] M. Rowe, T. Yunusov, S. Haben, W. Holderbaum, and B. Potter, "The Real-Time Optimisation of DNO Owned Storage Devices on the LV Network for Peak Reduction," *Energies*, vol. 7, pp. 3537–3560, 2014.

[116] K. Muralitharan, R. Sakthivel, and Y. Shi, "Multiobjective Optimization Technique for Demand Side Management with Load Balancing Approach in Smart Grid," *Neurocomputing*, vol. 177, pp. 110–119, 2016.

[117] A. Ensslen, P. Ringler, L. Dörr, P. Jochem, F. Zimmermann, and W. Fichtner, "Incentivizing Smart Charging: Modeling Charging Tariffs for Electric Vehicles in German and French Electricity Markets," *Energy Research Social Science*, vol. 42, pp. 112–126, 2018.

[118] M. Kühnbach, J. Stute, and A.-L. Klingler, "Impacts of Avalanche Effects of Price-Optimized Electric Vehicle Charging - Does Demand Response Make It Worse?" *Energy Strategy Reviews*, vol. 34, 2021, Article no. 100608.

[119] D. Bunn and E. D. Farmer, "Comparative Models for Electrical Load Forecasting," *John Wiley and Sons*, 1985.

[120] R. Chen, C. S. Lai, C. Zhong, K. Pan, W. W. Ng, Z. Li, and L. L. Lai, "MultiCycleNet: Multiple Cycles Self-Boosted Neural Network for Short-Term Electric Household Load Forecasting," *Sustainable Cities and Society*, vol. 76, 2022, Article no. 103484.

[121] C. Fan, Y. Li, L. Yi, L. Xiao, X. Qu, and Z. Ai, "Multi-Objective LSTM Ensemble Model for Household Short-Term Load Forecasting," *Memetic Computing*, vol. 14, pp. 115–132, 2022.

[122] S. Haben, G. Giasemidis, F. Ziel, and S. Arora, "Short Term Load Forecasting and the Effect of Temperature at the Low Voltage Level," *International Journal of Forecasting*, vol. 35, pp. 1469–1484, 2019.

[123] W. El-Baz and P. Tzscheutschler, "Short-Term Smart Learning Electrical Load Prediction Algorithm for Home Energy Management Systems," *Applied Energy*, vol. 147, pp. 10–19, 2015.

[124] F. Chollet, "Keras," [Online], 2015, https://github.com/fchollet/keras, Accessed on: 13.10.2019.

[125] R. Al-Rfou, G. Alain, A. Almahairi, C. Angermueller, D. Bahdanau *et al.*, "Theano: A Python Framework for Fast Computation of Mathematical Expressions," *CoRR*, vol. arXiv.1605.02688, 2016, Accessed on: 13.10.1990. [Online]. Available: https://doi.org/10.48550/arXiv.1605.02688

[126] Australian Government (data.gov.au), "Smart-Grid Smart-City Customer Trial Data," [Online], https://data.gov.au/data/dataset/smart-grid-smart-city-customer-trial-data, Accessed on: 03.03.2020.

[127] N. Somu, G. R. M R, and K. Ramamritham, "A Hybrid Model for Building Energy Consumption Forecasting Using Long Short Term Memory Networks," *Applied Energy*, vol. 261, 2020, Article no. 114131.

[128] G. Zhang, C. Tian, C. Li, J. J. Zhang, and W. Zuo, "Accurate Forecasting of Building Energy Consumption via a Novel Ensembled Deep Learning Method Considering the Cyclic Feature," *Energy*, vol. 201, 2020, Article no. 117531.

[129] M. Bourdeau, X. qiang Zhai, E. Nefzaoui, X. Guo, and P. Chatellier, "Modeling and Forecasting Building Energy Consumption: A Review of Data-Driven Techniques," *Sustainable Cities and Society*, vol. 48, 2019, Article no. 101533.

[130] C. Heghedus, A. Chakravorty, and C. Rong, "Energy Informatics Applicability; Machine Learning and Deep Learning," in *Proceedings of IEEE/ACIS International Conference on Big Data, Cloud Computing, Data Science and Engineering*. Yonago, Japan: IEEE, 2018, pp. 97–101.

[131] M. Ghofrani, M. Ghayekhloo, A. Arabali, and A. Ghayekhloo, "A Hybrid Short-Term Load Forecasting with a New Input Selection Framework," *Energy*, vol. 81, pp. 777–786, 2015.

[132] G. F. Fan, Y. H. Guo, J. M. Zheng, and W. C. Hong, "A Generalized Regression Model Based on Hybrid Empirical Mode Decomposition and Support Vector Regression with Back-Propagation Neural Network for Mid-Short-Term Load Forecasting," *Journal of Forecasting*, vol. 39, pp. 737–756, 2020.

[133] Z. Liu, X. Sun, S. Wang, M. Pan, Y. Zhang, and Z. Ji, "Midterm Power Load Forecasting Model Based on Kernel Principal Component Analysis and Back Propagation Neural Network with Particle Swarm Optimization," *Big Data*, vol. 7, no. 2, pp. 130–138, 2019.

[134] L. Zajmi, F. Y. Ahmed, and A. A. Jaharadak, "Concepts, methods, and performances of particle swarm optimization, backpropagation, and neural networks," *Applied Computational Intelligence and Soft Computing*, vol. 2018, 2018, Article no. 9547212.

[135] S. Masum, Y. Liu, and J. Chiverton, "Multi-step time series forecasting of electric load using machine learning models," in *Proceedings of International Conference on Artificial Intelligence and Soft Computing*. Cham, Switzerland: Springer, 2018, pp. 148–159.

[136] M. Cai, M. Pipattanasomporn, and S. Rahman, "Day-Ahead Building-Level Load Forecasts Using Deep Learning vs. Traditional Time-series Techniques," *Applied Energy*, vol. 236, pp. 1078–1088, 2019.

[137] J. Zhang and S. Qu, "Optimization of Backpropagation Neural Network Under the Adaptive Genetic Algorithm," *Complexity*, vol. 2021, 2021, Article no. 1718234.

[138] Y. Wei, X. Zhang, Y. Shi, L. Xia, S. Pan, J. Wu, M. Han, and X. Zhao, "A Review of Data-Driven Approaches for Prediction and Classification of Building Energy Consumption," *Renewable and Sustainable Energy Reviews*, vol. 82, pp. 1027–1047, 2018.

[139] G. Chen, "A Gentle Tutorial of Recurrent Neural Network with Error Backpropagation," *CoRR*, 2018, Accessed on: 20.05.2023. [Online]. Available: https://arxiv.org/abs/1610.02583

[140] H. Sak, A. W. Senior, and F. Beaufays, "Long Short-Term Memory Recurrent Neural Network Architectures for Large Scale Acoustic Modeling," in *Proceedings of the 15th Annual Conference of International Speech Communication Association (INTERSPEECH)*, Singapore, 2014, pp. 338–342.

[141] A. Graves, A. R. Mohamed, and G. Hinton, "Speech Recognition with Deep Recurrent Neural Networks," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, no. 6. Vancouver, BC, Canada: IEEE, 2013, pp. 6645–6649.

[142] F. M. Bianchi, E. Maiorino, M. C. Kampffmeyer, A. Rizzi, and R. Jenssen, "An Overview and Comparative Analysis of Recurrent Neural Networks for Short Term Load Forecasting," *CoRR*, vol. abs/1705.04378, 2017, Accessed on: 02.06.2022. [Online]. Available: http://arxiv.org/abs/1705.04378

[143] B. Nepal, M. Yamaha, A. Yokoe, and T. Yamaji, "Electricity Load Forecasting Using Clustering and ARIMA Model for Energy Management in Buildings," *Japan Architectural Review*, vol. 3, no. 1, pp. 62–76, 2020.

[144] V. Cohen-Addad, V. Kanade, F. Mallmann-Trenn, and C. Mathieu, "Hierarchical Clustering: Objective Functions and Algorithms," *Journal of the ACM*, vol. 66, no. 4, pp. 1–42, 2019.

[145] Y. Zhao and G. Karypis, "Evaluation of Hierarchical Clustering Algorithms for Document Datasets," in *Proceedings of the 11th International Conference on Information and Knowledge Management*, McLean, Virginia, 2002, pp. 515–524.

[146] N. H. An and D. T. Anh, "Comparison of Strategies for Multi-step-Ahead Prediction of Time Series Using Neural Network," in *Proceeding of International Conference on Advanced Computing and Applications*. IEEE, 2015, pp. 142–149.

[147] F. Teng, J. Teng, L. Qiao, S. Du, and T. Li, "A Multi-Step Forecasting Model of Online Car-Hailing Demand," *Information Sciences*, vol. 587, pp. 572–586, 2022.

[148] F. J. Chang, Y. M. Chiang, and L. C. Chang, "Multi-Step-Ahead Neural Networks for Flood Forecasting," *Hydrological Sciences Journal*, vol. 52, no. 1, pp. 114–130, 2007.

[149] P. Du, J. Wang, W. Yang, and T. Niu, "Multi-Step Ahead Forecasting in Electrical Power System Using a Hybrid Forecasting System," *Renewable Energy*, vol. 122, pp. 533–550, 2018.

[150] H. Cheng, P.-n. Tan, J. Gao, and J. Scripps, "Multistep-Ahead Time Series Prediction," in *Proceeding of Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Berlin, Heidelberg: Springer, 2006, pp. 765–766.

[151] S. Ben Taieb, G. Bontempi, A. F. Atiya, and A. Sorjamaa, "A review and comparison of strategies for multi-step ahead time series forecasting based on the nn5 forecasting competition," *Expert Systems with Applications*, vol. 39, no. 8, pp. 7067–7083, 2012.

[152] J. Wang, Y. Song, F. Liu, and R. Hou, "Analysis and Application of Forecasting Models in Wind Power Integration: A Review of Multi-Step-Ahead Wind Speed Forecasting Models," *Renewable and Sustainable Energy Reviews*, vol. 60, pp. 960–981, 2016.

[153] A. Ahmed and M. Khalid, "Multi-Step Ahead Wind Forecasting Using Nonlinear Autoregressive Neural Networks," *Energy Procedia*, vol. 134, pp. 192–204, 2017.

[154] R. Das, A. I. Middya, and S. Roy, "High Granular and Short Term Time Series Forecasting of $PM_{2.5}$ Air Pollutant - A Comparative Review," *Artificial Intelligence Review*, vol. 55, pp. 1253–1287, 2022.

[155] R. J. Murphy and J. Parsons, "Finer Granularity Means Better Data: A Crowdsourcing Lab Experiment," in *Proceedings of the 2nd Crowd Science Workshop: Trust, Ethics, and Excellence in Crowdsourced Data Management at Scale*, vol. 2932, Copenhagen, Denmark, 2021, pp. 72–86.

[156] S. Makonin, "AMPds2: The Almanac of Minutely Power dataset (Version 2)," 2016, Accessed on: 24.12.2018. [Online]. Available: https://doi.org/10.7910/DVN/FIE0S4

[157] S. Makonin, B. Ellert, I. V. Bajić, and F. Popowich, "Electricity, Water, and Natural gas Consumption of a Residential House in Canada from 2012 to 2014," *Scientific data*, vol. 3, no. 1, 2016, Article no. 160037.

[158] R. J. Hyndman *et al.*, "Another Look at Forecast-Accuracy Metrics for Intermittent Demand," *Foresight: The International Journal of Applied Forecasting*, vol. 4, no. 4, pp. 43–46, 2006.

[159] A. Jierula, S. Wang, T.-M. Oh, and P. Wang, "Study on Accuracy Metrics for Evaluating the Predictions of Damage Locations in Deep Piles Using Artificial Neural Networks with Acoustic Emission Data," *Applied Sciences*, vol. 11, no. 5, 2021, Article no. 2314.

[160] F. Zhang, C. Deb, S. E. Lee, J. Yang, and K. W. Shah, "Time Series Forecasting for Building Energy Consumption Using Weighted Support Vector Regression with Differential Evolution Optimization Technique," *Energy and Buildings*, vol. 126, pp. 94–103, 2016.

[161] Q. Li, Q. Meng, J. Cai, H. Yoshino, and A. Mochida, "Applying Support Vector Machine to Predict Hourly Cooling Load in the Building," *Applied Energy*, vol. 86, no. 10, pp. 2249–2256, 2009.

[162] V. Vapnik, *The Nature of Statistical Learning Theory*, 2nd ed. Berlin, Germany: Springer, 1999, ISBN: 978-0387987804.

[163] W.-C. Hong, "Electric Load Forecasting by Support Vector Model," *Applied Mathematical Modelling*, vol. 33, no. 5, pp. 2444–2454, 2009.

[164] T. Ackermann and L. Söder, "Wind Energy Technology and Current Status: A Review," *Renewable and Sustainable Energy Reviews*, vol. 4, no. 4, pp. 315–374, 2000.

[165] M.-D. Liu, L. Ding, and Y.-L. Bai, "Application of Hybrid Model Based on Empirical Mode Decomposition, Novel Recurrent Neural Networks and the ARIMA to Wind Speed Prediction," *Energy Conversion and Management*, vol. 233, 2021, Article no. 113917.

[166] Max Planck Institute for Biogeochemistry, "Jena Climate Dataset," [Online], https://www.kaggle.com/datasets/mnassrib/jena-climate, Accessed on: 24.02.2022.

[167] W. Yu, D. An, D. Griffith, Q. Yang, and G. Xu, "On Statistical Modeling and Forecasting of Energy Usage in Smart Grid," in *Proceedings of Conference on Research in Adaptive and Convergent Systems*, Towson, MD, USA, 2014, pp. 12–17.

[168] M. Robnik-Šikonja and I. Kononenko, "Theoretical and Empirical Analysis of ReliefF and RReliefF," *Machine Learning*, vol. 53, no. 1, pp. 23–69, 2003.

[169] E. Skomski, J.-Y. Lee, W. Kim, V. Chandan, S. Katipamula, and B. Hutchinson, "Sequence-to-Sequence Neural Networks for Short-Term Electrical Load Forecasting in Commercial Office Buildings," *Energy and Buildings*, vol. 226, 2020, Article no. 110350.

[170] G. Chitalia, M. Pipattanasomporn, V. Garg, and S. Rahman, "Robust Short-Term Electrical Load Forecasting Framework for Commercial Buildings Using Deep Recurrent Neural Networks," *Applied Energy*, vol. 278, 2020, Article no. 115410.

[171] A. Taïk and S. Cherkaoui, "Electrical Load Forecasting Using Edge Computing and Federated Learning," in *Proceedings of IEEE International Conference on Communications*, 2020, pp. 1–6.

[172] D.-C. Wu, B. Bahrami Asl, A. Razban, and J. Chen, "Air Compressor Load Forecasting Using Artificial Neural Network," *Expert Systems with Applications*, vol. 168, 2021.

[173] M. Sajjad, Z. A. Khan, A. Ullah, T. Hussain, W. Ullah, M. Y. Lee, and S. W. Baik, "A Novel CNN-GRU-Based Hybrid Approach for Short-Term Residential Load Forecasting," *IEEE Access*, vol. 8, pp. 143 759–143 768, 2020.

[174] W.-Q. Li and L. Chang, "A Combination Model with Variable Weight Optimization for Short-Term Electrical Load Forecasting," *Energy*, vol. 164, pp. 575–593, 2018.

[175] L. Xu, C. Li, X. Xie, and G. Zhang, "Long-Short-Term Memory Network Based Hybrid Model for Short-Term Electrical Load Forecasting," *Information*, vol. 9, no. 7, 2018, Article no. 165.

[176] M. Imani, "Fuzzy-Based Weighting Long Short-Term Memory Network for Demand Forecasting," *The Journal of Supercomputing*, vol. 79, pp. 435–460, 2023.

[177] B. Saha, K. F. Ahmed, S. Saha, and M. T. Islam, "Short-Term Electrical Load Forecasting Via Deep Learning Algorithms to Mitigate the Impact of Covid-19 Pandemic on Power Demand," in *Proceedings of International Conference on Automation, Control and Mechatronics for Industry 4.0*, 2021, pp. 1–6.

[178] Z. A. Khan, T. Hussain, A. Ullah, S. Rho, M. Lee, and S. W. Baik, "Towards Efficient Electricity Forecasting in Residential and Commercial Buildings: A Novel Hybrid CNN with a LSTM-AE based Framework," *Sensors*, vol. 20, no. 5, 2020, Article no. 1399.

[179] S. Bouktif, A. Fiaz, A. Ouni, and M. A. Serhani, "Optimal Deep Learning LSTM Model for Electric Load Forecasting using Feature Selection and Genetic Algorithm: Comparison with Machine Learning Approaches," *Energies*, vol. 11, no. 7, 2018, Article no. 1636.

[180] S. Singh and A. Yassine, "Big Data Mining of Energy Time Series for Behavioral Analytics and Energy Consumption Forecasting," *Energies*, vol. 11, no. 2, 2018, Article no. 452.

[181] X. Li, J. Wen, and A. Malkawi, "An Operation Optimization and Decision Framework for a Building Cluster with Distributed Energy Systems," *Applied Energy*, vol. 178, pp. 98–109, 2016.

[182] W. Mai, C. Y. Chung, T. Wu, and H. Huang, "Electric Load Forecasting for Large Office Building Based on Radial Basis Function Neural Network," in *Proceedings of IEEE PES General Meeting, Conference Exposition*, 2014, pp. 1–5.

[183] C. Li, M. Tang, G. Zhang, R. Wang, and C. Tian, "A Hybrid Short-Term Building Electrical Load Forecasting Model Combining the Periodic Pattern, Fuzzy System, and Wavelet Transform," *International Journal of Fuzzy Systems*, vol. 22, no. 1, pp. 156–171, 2020.

[184] H. Eskandari, M. Imani, and M. P. Moghaddam, "Convolutional and Recurrent Neural Network Based Model for Short-Term Load Forecasting," *Electric Power Systems Research*, vol. 195, 2021, Article no. 107173.

[185] X. Xu and W. Ren, "Random Fourier Feature Kernel Recursive Maximum Mixture Correntropy Algorithm for Online Time Series Prediction," *ISA Transactions*, vol. 126, pp. 370–376, 2022.

[186] J. Wang, X. Li, J. Li, Q. Sun, and H. Wang, "NGCU: A New RNN Model for Time-Series Data Prediction," *Big Data Research*, vol. 27, 2022, Article no. 100296.

[187] R. Solgi, H. A. Loáiciga, and M. Kram, "Long Short-Term Memory Neural Network (LSTM-NN) for Aquifer Level Time Series Forecasting Using In-Situ Piezometric Observations," *Journal of Hydrology*, vol. 601, 2021, Article no. 126800.

[188] S. Karasu and A. Altan, "Crude Oil Time Series Prediction Model Based on LSTM Network With Chaotic Henry Gas Solubility Optimization," *Energy*, vol. 242, 2022, Article no. 122964.

[189] M. Iqbal, M. S. Iqbal, F. H. Jaskani, K. Iqbal, and A. Hassan, "Time-Series Prediction of Cryptocurrency Market Using Machine Learning Techniques," *EAI Endorsed Transactions on Creative Technologies*, vol. 8, no. 28, 2021.

[190] H. Nasiri and M. M. Ebadzadeh, "MFRFNN: Multi-Functional Recurrent Fuzzy Neural Network for Chaotic Time Series Prediction," *Neurocomputing*, vol. 507, pp. 292–310, 2022.

[191] Z. I. Erzurum Cicek and Z. Kamisli Ozturk, "Optimizing the Artificial Neural Network Parameters Using a Biased Random Key Genetic Algorithm for Time Series Forecasting," *Applied Soft Computing*, vol. 102, 2021, Article no. 107091.

[192] Z. Hajirahimi, M. Khashei, and S. Etemadi, "A Novel Class of Reliability-Based Parallel Hybridization (RPH) Models for Time Series Forecasting," *Chaos, Solitons Fractals*, vol. 156, 2022, Article no. 111880.

[193] S. Zhang, Y. Chen, W. Zhang, and R. Feng, "A Novel Ensemble Deep Learning Model with Dynamic Error Correction and Multi-Objective Ensemble Pruning for Time Series Forecasting," *Information Sciences*, vol. 544, pp. 427–445, 2021.

[194] S. Suradhaniwar, S. Kar, S. S. Durbha, and A. Jagarlapudi, "Time Series Forecasting of Univariate Agrometeorological Data: A Comparative Performance Evaluation via One-Step and Multi-Step Ahead Forecasting Strategies," *Sensors*, vol. 21, no. 7, 2021, Article no. 2430.

[195] Z.-K. Feng, P.-F. Shi, T. Yang, W.-J. Niu, J.-Z. Zhou, and C.-T. Cheng, "Parallel Cooperation Search Algorithm and Artificial Intelligence Method for Streamflow Time Series Forecasting," *Journal of Hydrology*, vol. 606, 2022, Article no. 127434.

[196] F. Ding and C. Luo, "Interpretable Cognitive Learning with Spatial Attention for High-Volatility Time Series Prediction," *Applied Soft Computing*, vol. 117, 2022, Article no. 108447.

[197] L. Ruan, Y. Bai, S. Li, S. He, and L. Xiao, "Workload Time Series Prediction in Storage Systems: a deep Learning Based Approach," *Cluster Computing*, pp. 1–11, 2021.

[198] R. Chandra, S. Goyal, and R. Gupta, "Evaluation of Deep Learning Models for Multi-Step Ahead Time Series Prediction," *IEEE Access*, vol. 9, pp. 83 105–83 123, 2021.

[199] H. Abbasimehr and R. Paki, "Improving Time Series Forecasting Using LSTM and Attention Models," *Journal of Ambient Intelligence and Humanized Computing*, vol. 13, no. 1, pp. 673–691, 2022.

[200] S. Du, T. Li, Y. Yang, and S.-J. Horng, "Multivariate Time Series Forecasting via Attention-Based Encoder-Decoder Framework," *Neurocomputing*, vol. 388, pp. 269–279, 2020.

[201] Z. Shen, Y. Zhang, J. Lu, J. Xu, and G. Xiao, "A Novel Time Series Forecasting Model with Deep Learning," *Neurocomputing*, vol. 396, pp. 302–313, 2020.

[202] D. S. de O. Santos Júnior, J. F. de Oliveira, and P. S. de Mattos Neto, "An Intelligent Hybridization of ARIMA with Machine Learning Models for Time Series Forecasting," *Knowledge-Based Systems*, vol. 175, pp. 72–86, 2019.

[203] J. Chen, G.-Q. Zeng, W. Zhou, W. Du, and K.-D. Lu, "Wind Speed Forecasting Using Nonlinear-Learning Ensemble of Deep Learning Time Series Prediction and Extremal Optimization," *Energy Conversion and Management*, vol. 165, pp. 681–695, 2018.

[204] D. Hussain, T. Hussain, A. A. Khan, S. A. A. Naqvi, and A. Jamil, "A Deep Learning Approach for Hydrological Time-Series Prediction: A Case Study of Gilgit River Basin," *Earth Science Informatics*, vol. 13, no. 3, pp. 915–927, 2020.

[205] A. Sagheer and M. Kotb, "Unsupervised Pre-Training of a Deep LSTM-Based Stacked Autoencoder for Multivariate Time Series Forecasting Problems," *Scientific reports*, vol. 9, no. 1, pp. 1–16, 2019.

[206] A. Y. Barrera-Animas, L. O. Oyedele, M. Bilal, T. D. Akinosho, J. M. D. Delgado, and L. A. Akanbi, "Rainfall Prediction: A Comparative Analysis of Modern Machine Learning Algorithms for Time-Series Forecasting," *Machine Learning with Applications*, vol. 7, 2022, Article no. 100204.

[207] A. F. Faisal, A. Rahman, M. T. M. Habib, A. H. Siddique, M. Hasan, and M. M. Khan, "Neural Networks Based Multivariate Time Series Forecasting of Solar Radiation Using Meteorological Data of Different Cities of Bangladesh," *Results in Engineering*, vol. 13, 2022, Article no. 100365.

[208] N. F. Sopelsa Neto, S. F. Stefenon, L. H. Meyer, R. G. Ovejero, and V. R. Q. Leithardt, "Fault Prediction Based on Leakage Current in Contaminated Insulators Using Enhanced Time Series Forecasting Models," *Sensors*, vol. 22, no. 16, 2022, Article no. 6121.

[209] K. P. Amber, M. W. Aslam, A. Mahmood, A. Kousar, M. Y. Younis, B. Akbar, G. Q. Chaudhary, and S. K. Hussain, "Energy Consumption Forecasting for University Sector Buildings," *Energies*, vol. 10, no. 10, 2017, Article no. 1579.

[210] R.-G. Cirstea, B. Yang, C. Guo, T. Kieu, and S. Pan, "Towards Spatio- Temporal Aware Traffic Time Series Forecasting," in *Proceeding of IEEE 38th International Conference on Data Engineering (ICDE)*, 2022, pp. 2900–2913.

[211] T. Wang, Z. Li, X. Geng, B. Jin, and L. Xu, "Time Series Prediction of Sea Surface Temperature Based on an Adaptive Graph Learning Neural Model," *Future Internet*, vol. 14, no. 6, 2022, Article no. 171.

[212] L. Shen and Y. Wang, "TCCT: Tightly-Coupled Convolutional Transformer on Time Series Forecasting," *Neurocomputing*, vol. 480, pp. 131–145, 2022.