DATA-DRIVEN APPROACHES TO FORECASTING IN ENERGY SYSTEMS: WEATHER-INDUCED OUTAGE FORECASTING, NET LOAD FORECASTING, AND SOLAR ESTIMATION

by

Vinayak Sharma

A dissertation submitted to the faculty of
The University of North Carolina at Charlotte
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in
Electrical Engineering

Charlotte

2024

Approved by:

____

Dr. Valentina Cecchi

____

Dr. Badrul Chowdhury

____

Dr. Linquan Bai

____

Dr. Umit Cali

____

Dr. Carlos Orozco

ABSTRACT

VINAYAK SHARMA. Data-Driven Approaches to Forecasting in Energy Systems:
Weather-Induced Outage Forecasting, Net Load Forecasting, and Solar Estimation.
(Under the direction of DR. VALENTINA CECCHI)

In recent years, the global energy sector has been undergoing a significant transformation, characterized by an increasing shift towards data-driven operations and the widespread adoption of renewable energy such as solar photovoltaics (PV). This transition is largely motivated by the urgent need to address climate change and the realization of the potential that large-scale data collection and analysis hold for enhancing energy efficiency and sustainability. As the energy landscape becomes more complex and interconnected, the role of sophisticated energy forecasting techniques has grown in importance. These techniques are crucial for managing the variability and uncertainty inherent in renewable energy sources, such as wind and solar power, which are subject to fluctuations in weather and environmental conditions. Moreover, the integration of big data analytics into energy systems facilitates more accurate and timely predictions, thereby enabling more effective planning, operation, and maintenance of energy infrastructure. This dissertation introduces novel, data-driven methodologies to address key challenges in energy forecasting: predicting weather-induced power outages, net load forecasting, and accurately estimating solar PV penetration.

In the first part of the study, a methodology to forecast weather-related power distribution outages one day ahead on an hourly basis is presented. A solution to address the data imbalance issue is proposed, where only a small portion of the data represents the hours impacted by outages, in the form of a weighted logistic regression model. Data imbalance is a key modeling challenge for small and rural electric utilities. The weights for outage and non-outage hours are determined by the reciprocals of their corresponding number of hours. To demonstrate the effectiveness of the

proposed model, two case studies using data from a small electric utility company in the United States are presented. One case study analyses the weather-related outages aggregated up to the city level. The other case study is based on the distribution substation level, which has rarely been tackled in the outage prediction literature. Compared with two variants of ordinary logistic regression with equal weights, the proposed model shows superior performance in terms of geometric mean.

The dissertation then explores net load forecasting in the context of increasing behind-the-meter (BTM) solar PV system adoption. This adoption introduces complexities to grid management, especially concerning net load-the difference between demand and PV generation. The intermittent nature of PV generation, influenced by weather and time, adds to net load volatility, posing challenges to grid reliability. This dissertation presents a review of state-of-the-art net load forecasting with a focus on forecasting approaches, techniques, explanatory variables, and the impact of PV penetration on net load forecasting. Additionally, the study conducts a critical analysis of existing literature to identify gaps in the field of net load forecasting and PV integration. To address some of these gaps, a benchmark net load forecasting model is proposed. The proposed model uses publicly available data from ISO New England. Through the case study, it is demonstrated that the proposed net load forecasting model outperforms the current benchmark load forecast model significantly in terms of forecasting accuracy, as measured by Mean Absolute Percentage Error. Moreover, the case study also demonstrates the effectiveness of the proposed model over a range of PV penetration, which is an important consideration as the use of solar energy continues to grow.

Furthermore, the dissertation addresses two critical questions regarding PV integration: (1) How much PV is there in the system?; (2) Which meters have BTM PV? To address the challenge of estimating PV penetration in systems, existing supervised and unsupervised methods are reviewed, which reveal common limitations,

especially when PV installation information is limited or completely unavailable. To overcome these challenges, a regression-based approach is developed by leveraging the difference in performance in the benchmark load and net load forecasting models in forecasting net load. The proposed framework is deployed for real-world data from an ISO and a medium-sized in the United States. The results validate the effectiveness of the proposed method in accurately estimating PV penetration levels, even without explicit PV installation data, using only historical load data.

The final part of the study focuses on identifying meters with BTM PV installations. Again by, leveraging the performance disparities between load forecasting models and net load forecasting models, a methodology is devised to differentiate meters with and without PV installations. The effectiveness of the proposed frameworks is confirmed using an empirical case study at a medium-sized US utility with meter-level load data meters. The results illustrate that accurate identification of meters with PV installations was achieved while maintaining a low rate of false identifications. This methodology provides valuable insights for utilities, empowering them to comprehend the adoption and impact of distributed solar energy within their service territories.

Overall, this study contributes significantly to the field of energy system forecasting by developing data-driven models that enhance the understanding and management of weather-induced outages, net load variability, and solar PV integration. These advancements enable utilities to make informed decisions for grid planning, capacity management, and service customization, paving the way for more resilient and efficient energy systems.

DEDICATION

To my family.

ACKNOWLEDGEMENTS

Reflecting on the journey towards my Ph.D., I am deeply thankful for the unwavering support and encouragement I've received during both the challenging and triumphant moments. My heartfelt appreciation goes out to all who have inspired, guided, and supported me along this path.

I would like to specifically acknowledge Dr. Valentina Cecchi for the constant support she has provided throughout my journey. She has been a source of guidance and support anytime there was a hiccup in my journey. Thank you so much, Dr.Cecchi.

I also wish to express my deep appreciation to my committee members- Dr. Badrul Chowdhury, Dr. Linquan Bai, Dr. Umit Cali, and Dr. Carlos Orozco, for their support and guidance.

During my master's studies, I was introduced to the field of energy forecasting by Dr. Umit Cali while working on a project. I got particularly interested in the field as it gave me a chance to work on problems where there is an opportunity to improve every day as forecasts can never be perfect. I made energy forecasting the topic of my Master's thesis and subsequently enrolled in a Ph.D. program to further work on energy forecasting.

I want to acknowledge Dr. Tao Hong for his valuable contributions to my research work on outage forecasting, net load forecasting, and solar estimation and detection.

A huge thank you to my friend Bhav Sardana for his unwavering support throughout my Ph.D. journey. His willingness to listen, brainstorm ideas, and offer constant encouragement was invaluable.

Finally, this would not have been possible without the ndless love, support, and belief of my parents Ashwani Sharma and Suman Sharma, my brother Kartikaye Sharma, and my wife Pranita Vashisth. This achievement would not have been possible without you.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

1.1 Overview

In November 2021, the most significant United Nations Framework Convention on Climate Change (UNFCCC) Conference of the Parties (COP 26) took place since the historic Paris Agreement was enacted in 2015. At this pivotal COP 26 meeting, there was a surge in the number of countries declaring their ambitious long-term objectives to reach net-zero greenhouse gas emissions within the forthcoming decades (Bouckaert et al., 2021). Achieving this ambitious target requires aggressive decarbonization measures that will fundamentally transform how energy is generated, distributed, and consumed globally, and significantly reshape the power sector. In this context, the vast amounts of data generated from various energy sources, consumption patterns, and environmental factors become invaluable. Data-driven methods, leveraging the latest advancements in big data analytics, machine learning, and AI, are adept at dissecting this complex web of data to produce accurate and reliable energy forecasts. These forecasts are critical in managing the balance between the intermittent nature of renewable energy sources and the fluctuating energy demands of a modern economy. By accurately predicting future demand patterns and potential supply gaps, these methods enable energy providers and policymakers to make informed decisions about infrastructure investments and resource allocation. This proactive approach is essential for integrating renewable energy sources effectively, ensuring a stable power supply, and mitigating overgeneration and outages.

As modern energy systems grow more complex and integrate increasing amounts of renewable energy, there is a pressing need to improve the accuracy and operational efficiency of energy forecasts. This dissertation utilizes advanced data-driven modeling techniques to meet these challenges, offering practical solutions that utilities can implement to enhance grid management and reliability. The focus is on developing and applying sophisticated forecasting models that excel in real-world settings. To methodically address these issues, the research is organized around three main questions. Each question focuses on a distinct aspect of energy forecasting: improving power outage prediction accuracy, optimizing net load forecasts, and identifying photovoltaic installations within operational frameworks. These questions aim to bridge current gaps in academic research and contribute practical improvements to energy system operations

Main Research Question: How can advanced data-driven modeling techniques be applied to enhance the operational efficiency and accuracy of energy forecasts in real-world scenarios?

Additionally, the main research question was divided into the following sub-research questions to better address the problem and in order to come up with a road map for the thesis.

Research Question 1: What are the key factors influencing the accuracy of weather-related power outage forecasts, and how can the forecasting framework be optimized to address data imbalance in these predictions?

- This question focuses on applying advanced modeling techniques, such as weighted logistic regression models to refine the accuracy and reliability of power outage forecasts. The aim is to optimize these models for practical use by utilities, enhancing their response strategies and operational planning.

Research Question 2: How can a benchmark framework be developed for net load forecasting to address the gaps in literature while effectively accommodating different scenarios of PV penetration?

- The objective here is to develop a practical, robust benchmark model for net load forecasting that can be adopted by energy providers to stabilize and optimize grid operations across various energy systems, considering the impact of PV integration on net load dynamics. This model will address real-world variations in renewable energy inputs and provide a standardized approach for managing these dynamics.

Research Question 3: How can PV penetration be estimated, and behind-the-meter (BTM) PV installations be accurately identified using net load data?

- This question explores practical techniques for using net load data to estimate PV penetration and detect BTM PV installations. The focus is on refining these methods to provide utilities with actionable insights into the distribution of solar resources, aiding in effective grid management and planning.

These research questions aim to directly address the practical challenges of operationalizing energy forecasting within the modern power grid. By focusing on applied methodologies and their implementation in actual grid scenarios, the dissertation seeks to contribute valuable, operational solutions that enhance the adaptability, efficiency, and resilience of energy systems in the face of increasing renewable integration and complex demand dynamics.

This chapter provides an overview of the evolving field of energy forecasting, emphasizing how the complexity of energy forecasting has expanded with the emergence of issues like outages, renewable energy integration, and net load management. This evolution corresponds with the increasing availability of data collected by the utilities and advancements in data analytics. Additionally, it delves into the basics of outage management and net load considerations, which form the foundational motivation for this research. The chapter aims to underscore the significant contributions of this work in addressing these contemporary challenges in energy forecasting.

- Background and motivation

- Fundamentals of outage management
- Fundamentals of net load
- Summary of the contributions
- Dissertation organization

1.2 Background and Motivation

Energy forecasting has been a critical business problem since the inception of the electric power industry. Over the years, the realm of energy forecasting has undergone a significant evolution, expanding from its initial focus on load forecasting to encompass a variety of complex components such as renewable energy and outage forecasting. This evolution reflects the changing dynamics and increasing complexity of the energy sector.

Initially, energy forecasting was predominantly centered around load forecasting with many notable papers published in the field throughout the years (Hong, 2010; Hong et al., 2016; Gross and Galiana, 1987; Park et al., 1991b). This entailed predicting the electricity demand of a region, which was crucial for ensuring that power generation met consumer needs without overloading the grid. Load forecasting has historically been the backbone of energy management, aiding utilities in planning and operating their systems efficiently. Over the years, numerous methodologies and models have been developed and refined, incorporating advancements in statistical analysis and computational power.

As environmental concerns and technological advancements spurred the growth of renewable energy sources like solar and wind power, renewable generation forecasting emerged as a new frontier in energy forecasting. This development addressed the need to predict the variable and intermittent output from these renewable sources, which is crucial for integrating them effectively into the power grid. As a result, the last two decades saw an influx of papers on renewable generation forecasting (Costa et al., 2008; Yang et al., 2020a; Kuzlu et al., 2020; Sharma et al., 2018).

With the advancement of Advanced Metering Infrastructure (AMI) and sophisticated data analytic techniques, another subfield of energy forecasting, outage forecasting emerged as an essential component of energy forecasting (Yu et al., 2015). This subfield focuses on predicting power outages, which are critical for maintaining grid reliability and reducing the impact of disruptions. Outage forecasting signifies a shift from traditionally reactive management of power outages to a more proactive and anticipatory approach. Diverging from conventional energy forecasting domains like load or renewable energy prediction, outage forecasting grapples with the inherently unpredictable nature of power outages in electrical grids. This aspect of energy forecasting has gained attention as a critical area of research due to the increasing complexity and demands of modern power systems.

The outage forecasting problem shares many similar characteristics with load forecasting, especially in the fact that both are driven by weather-related variables such as temperature, wind speed, humidity, etc (Black et al., 2018; Xie et al., 2016a). While temperature is a key driver of electricity demand (Wang et al., 2016), enabling the development of accurate temperature-based forecasting models for load, finding such a strong correlation for outage prediction is challenging (Hong and Hofmann, 2022). This notable difference highlights the unique challenges and the paramount importance of outage forecasting in the sphere of energy forecasting, underlining its critical role in modern energy management and grid stability.

In the most recent decade, the added effects of high renewable integration in the grid started to impact and change the traditional load. This introduced another subfield to the large energy forecasting problem, in the form of net load forecasting. This involves predicting the difference between the total electricity demand and the available renewable generation, taking into account the variable nature of renewable energy sources. In 2015, in the 10-year ahead forecast of energy forecasting, Hong et al. (2016) identified behind-the-meter (BTM) PV estimation and net load forecasting as future research topics in the field of energy forecasting. These were again highlighted by Hong et al. (2020) as emerging topics with few notable works in the field.

In order to accurately predict net load, it is crucial to have a thorough understanding of both load and renewable energy forecasting. The field of load forecasting is well established, with a large number of articles published each year. The majority of these papers focus on point forecasting at high and medium voltage levels (Wang et al., 2019). However, there is growing interest in probabilistic load forecasting, as noted in a review by Hong and Fan (2016). Weather variables, such as temperature (Hong, 2010; Wang et al., 2016), humidity (Xie et al., 2016b), and wind speed (Xie and Hong, 2017), are commonly included in load forecasting models, as weather significantly impacts electricity demand. Researchers have also worked to identify the most appropriate weather stations to use to obtain weather data (Hong et al., 2015; Sobhani et al., 2019). Various techniques have been employed for load forecasting, such as artificial neural networks (Park et al., 1991a; Hippert et al., 2001), multiple linear regression (Papalexopoulos and Hesterberg, 1990; Hong, 2010), semiparametric additive models (Fan and Hyndman, 2011), and fuzzy regression (Hong and Wang, 2014). Over the years, many notable load forecasting reviews have also been done (Gross and Galiana, 1987; Hong and Fan, 2016; Hong et al., 2019).

As compared to load forecasting, renewable forecasting is a relatively newer field. Among the renewable sources of energy, wind, and PV are amongst the most popular ones. Both PV and wind power can supply energy to the transmission system through large centralized plants. In such cases, for the day-ahead market, the system operators consider not only the day-ahead load forecasts but also day-ahead renewable forecasts to forecast net load (Yang et al., 2022). However, with PV, there can also be small-scale distributed generation units in the form of BTM PV, connected to the distribution system, something that is not a concern with wind energy. At the consumer level, PV adoption is a very appealing investment opportunity owing to tax benefits, low initial cost, ease of installation, and quick return on investment. As a consequence, a significant share of solar adoption occurs at the customer level. This shift in energy production toward customer-level generation has resulted in changes in the operation and planning of power systems, particularly in load forecasting during critical times of the day. Furthermore, most utilities only have access to the accumulated net load, making it even more difficult to observe BTM PV separately from the load. In large quantities, BTM PV systems can significantly alter the shape of regional net load profiles and pose balancing and reliability challenges.

The field of solar forecasting has seen an influx of papers in the past decade, however, until 2018, it was still considered an immature field compared to load and wind energy (Hong et al., 2016). Yang et al. (2022) highlighted the major challenges and limitations that occur when approaching the solar forecasting problem in the electrical engineering way, similar to the load forecasting problem. Many notable reviews have been published on solar forecasting, focusing on the challenges, forecast horizon, approaches, and techniques. One of the first major reviews on solar forecasting was presented by Inman et al. (2013), followed by notable review by Antonanzas et al. (2016), Raza et al. (2016), and more recently by Yang et al. (2022). Literature on PV forecasting is mainly focused on forecasting the output of individual PV units, with very little attention toward large-scale unobserved BTM PV. More recently, Erdener et al. (2022) highlighted the very same issue and presented a review of BTM solar forecasting. The authors classified the forecasting methods for BTM PV into bottom-up and top-down approaches, based on the availability of information on the PV system. The bottom-up approach is used when information regarding all individual BTM PV installations is known. In this case, each individual PV system can be forecasted and aggregated to get the total BTM PV forecast for the region. The top-down approach is used when either information regarding the BTM PV is partially available or completely unavailable. In partial availability, often the approach of upscaling from a subset of representative sites is used to estimate the region's total PV (Killinger et al., 2018). More often, in cases with little to no information on individual photovoltaic installations, estimation techniques have been applied to estimate the aggregated capacity of installed BTM PV (Zhang and Grijalva, 2016a).

1.3 Fundamentals of Outage Management

Outage management is a critical component in the utility industry, focused on efficiently restoring power and maintaining grid stability during and after disruptions. Over the years, the strategies and technologies used in outage management have evolved significantly.

Traditionally, outage management was a reactive and manual process (Scott, 1990; Benner et al., 2017). Utilities depended largely on customer reports for outage detection, using paper maps and rudimentary computer systems for tracking and coordinating repair efforts (Nielsen, 2002). This approach often resulted in longer restoration times and less efficient use of resources, as the information about the extent and location of outages was not always accurate or timely. The primary tools were basic communication systems like telephones and radios, and the decision-making was largely based on the experience and intuition of the utility staff. Grid infrastructure was less complex, and renewable energy sources were minimal, making grid management somewhat simpler but less flexible.

Today, the advent of smart grid technologies has transformed outage management into a more efficient and responsive process. AMI, a key component of smart grid systems, allows for communication with millions of Smart Meters, providing near-real-time data throughout the utility's service territory (Pathak, 2016). This development enables utilities to detect outages immediately, without relying solely on customer reports. The data from these systems is integrated into sophisticated Outage Management Systems (OMS), which offer detailed insights into the status of the grid (Zhou et al., 2016).

Modern OMS platforms also use Geographic Information Systems (GIS) to visualize outages, predict their impacts, and optimize the dispatch of repair crews. They can also integrate weather data and predictive analytics to anticipate outages before they occur, enabling utilities to mobilize resources preemptively. Moreover, improved communication networks, including Customer Information Systems (CIS), enhance coordination among repair crews, utility departments, and external agencies (Pathak, 2016). The use of social media and mobile applications for instantaneous customer communication regarding outage updates and restoration processes has enhanced not only operational efficiency but also consumer satisfaction and trust.

Furthermore, the inclusion of distributed energy resources such as solar and wind power, coupled with battery storage, has introduced new complexities and prospects in managing outages. These innovations enable more resilient grid structures, like microgrids, capable of isolating specific regions from extensive grid disruptions.

1.3.1 Outage Forecasting

The most significant advancement in outage management is the ability to forecast outages. This approach adds a new dimension to traditional outage management by predicting potential disruptions before they occur. This aspect entails predicting various dimensions of power outages, which includes not only their frequency (Kankanala et al., 2013; Sharma et al., 2023) but also the expected duration (Nateghi et al., 2011, 2014) and the potential number of customers impacted (Soares et al., 2021). These predictive insights are integral to the evaluation of reliability indices, which measure the effect of outage frequency and duration on both the system and its consumers. Key distribution reliability indices include the System Average Interruption Frequency Index (SAIFI), the Customer Average Interruption Frequency Index

(CAIFI), the System Average Interruption Duration Index (SAIDI), and the Customer Average Interruption Duration Index (CAIDI). Each of these indices provides crucial metrics for assessing and enhancing the reliability and efficiency of power distribution systems. Data analytics plays a pivotal role in this, enabling the analysis of large datasets, including historical outage data and weather patterns, to detect anomalies and predict possible outages.

Weather conditions have a substantial influence on the reliability of electric power systems (Lawanson et al., 2021). Various weather elements such as wind speed and gusts, temperature fluctuations, precipitation types, and severe ice storms are often primary contributors to power outages. For example, hurricane-level winds can directly damage infrastructure by snapping power lines and breaking poles or indirectly cause damage by hurling trees and debris into power lines. Additionally, extreme weather conditions can greatly hinder the speed and efficiency of power restoration efforts (Caswell et al., 2011).

Despite the largely uncontrollable nature of these natural forces, their impact on electric utilities can be substantially reduced through thoughtful design, meticulous planning, and strategic management. This involves a detailed analysis of historical weather data to understand the correlation between weather events and power outages (Lawanson et al., 2021; Sharma et al., 2023). Accurate weather forecasting plays a crucial role in helping utilities predict and prepare for weather-related outages.

If a utility can accurately predict which weather conditions are likely to cause outages and pinpoint their locations, it can better prioritize its resources. This foresight allows for strategic long-term resilience enhancements and effective short-term resource distribution, including the arrangement of mutual aid. Implementing such measures leads to improvements in both the reliability of power supply and the efficiency of restoration processes in the aftermath of weather-related disruptions (Campbell and Lowry, 2012).

Outage management, as it stands today, is a sophisticated blend of technology, predictive analytics, and strategic planning. The shift from a reactive to a proactive approach, primarily through outage forecasting, represents a significant advancement in the field. By leveraging AMI data, GIS, and advanced data analytics, utilities can now anticipate and prepare for potential outages more effectively. This proactive stance is crucial in dealing with the inherent unpredictability of weather-related disruptions and in enhancing the overall resilience of power systems.

1.4 Fundamentals of Net Load

Solar PV has experienced the most substantial growth among various renewable energy sources. In 2020 alone, the adoption of solar PV power increased by 107 GW, according to the International Energy Agency (IEA) (IEA, 2020). IEA estimates that by 2024, global photovoltaic capacity will reach 530 GW (IEA, 2020). The widespread availability, security, and high energy generation potential in most countries make solar energy adoption on a large scale feasible. At the individual level, solar energy presents an attractive investment opportunity due to its tax benefits, low initial cost, ease of installation, and quick pay-back periods. As a result, a significant portion of solar power adoption occurs through rooftop BTM PV systems installed at the consumer level. According to the report by IEA, residential solar PV capacity is expected to reach 143 GW by 2024 (IEA, 2019).

The adoption of distributed solar energy systems is heavily influenced by remuneration schemes for renewable energy. These schemes can be categorized into three main types, as stated in the IEA report (IEA, 2019):

1. Buy-all, Sell-all: In this scheme, the utility purchases all of the power generated by the PV owners and meets their demand. PV owners act as small-scale power plants.

2. Net metering: In this type of scheme, the PV owners consume the power gen-

erated by their PV system and purchase any additional power needed from the utility. If there is an excess of power generated, it is sold back to the utility for credits.

3. Real-time self-consumption: This scheme is similar to net metering, except that instead of receiving credits for excess power generated, the PV owners are paid for the power at a specific rate.

The majority of BTM PV growth is likely to be accounted for through buy-all, sell-all, or net metering schemes. Net metering is expected to be more popular of the two since it is the more profitable strategy in terms of yearly savings. Since net metering provides credits on future bills, if PV generation is sufficient to satisfy annual consumption, the electricity bill is covered by credits, while PV owners only pay annual fixed charges.

As distributed PV continues to develop, it has the potential to meet a substantial portion of the electricity needs of a region. Nevertheless, as the penetration of PV increases, there may be concerns about its impact on the stability and functionality of the power grid, which could hinder its future growth.

1.4.1 Invisible BTM PV

Large solar farms often independently metered. However, for small-scale solar adoption, a majority of the adoption happens behind-the-meter. The cumulative influence of these BTM PV installations can be considerable on the distribution network, taking into account the unique specifications of each PV system, particularly its installed capacity. It is, therefore, crucial for operators to understand the installed capacity of BTM solar PV systems.

However, due to the infrastructural constraints, BTM PV generation cannot be monitored by the utilities independently of the demand. Furthermore, often, even their basic capacity information is unknown. Consequently, this small-scale solar adoption becomes virtually "invisible" to power system operators, due to their power generation not being consistently monitored in real-time at the grid operations level.

The accelerated growth of invisible BTM PV brings new challenges to power system operators. With high PV penetration, invisible solar generation could drastically alter the net load, creating complexities for power system operations. Knowing how much PV penetration is there in the system, in form of the PV capacity or which meters have PV installations could be very crucial for the system operators. Therefore, a crucial component of BTM PV integration is to incorporate a capacity and specification estimation procedure to determine the state of the network at any given moment.

1.4.2 Load Shape Impact

A large number of customers with BTM PV can significantly alter load profiles. Additionally, PV generation is influenced by meteorological factors, such as cloud cover and solar irradiance, which are subject to change. Unlike stable and reliable sources of generation, such as coal and nuclear power plants, PV-generated electricity is intermittent and variable, causing the net load to be erratic and unstable. This results in a new load curve with significant ramping during the morning and evening hours, as well as sudden fluctuations during the day. Although ramping has been an issue system operators have been dealing with for years, ramping caused by human behavior had predictable seasonal patterns. However, the ramping seen in net load arises from variable PV generation, presenting a new forecasting challenge.

A report was published by NREL in 2008 (Denholm et al., 2008), aiming to investigate the preparation required for the future widespread incorporation of PV generation in the electricity grid. The report analyzed the impact of integrating higher levels of PV generation into the existing grid infrastructure and identified a distinct change in the shape of the electric load profile that conventional power plants typically meet. Later in 2013, California Independent System Operator (CAISO) released projections of impacts of increased PV on net load on its system and observed similar changes in its load shape, coining the term "duck chart" or "duck curve" due to the visual resemblance of the graph to a duck's silhouette.

Figure 1.1 shows a typical duck curve, depicting the net load curves for a typical day with various levels of PV penetration, based on data from ISO New England. The morning hours, starting around 4 a.m., witness the first ramp in the upward direction of the "duck's tail" as people begin their day. The second ramp, in the downward direction, occurs after the sun rises around 7 a.m., where on-line conventional power plants are substituted by solar generation resources, leading to the formation of the "belly" of the duck. As the sun sets, which usually starts around 4-5 p.m., the supply from solar generation resources ends, requiring the system operator to deploy resources to meet the third and most significant daily ramp, which forms the arch of the duck's neck. Following this steep upward ramp, demand on the system decreases into the evening hours, leading the system operator to scale back or shut down power generation to meet the final downward ramp.

Moreover, PV generation can lead to overgeneration on clear days when PV generation is high, resulting in reverse power flow in the feeder and the grid. This reverse flow can damage the utility grid and most systems are not designed to accommodate this change. Additionally, the variable generation from PV may make it difficult to keep voltage within the permitted range, and fluctuations associated with PV generation have the potential to damage voltage regulators. These challenges associated with large-scale PV integration will only grow in the face of more aggressive government goals aimed at 100% renewable energy generation. Finding solutions to these challenges by making the grid more flexible to accommodate these changes is a critical problem to address.

1.4.3 Dealing with the Duck Curve: Net Load Forecasting

According to a report by NREL (Denholm et al., 2015), there are two potential methods to deal with the duck curve and facilitate a greater integration of PV gener-

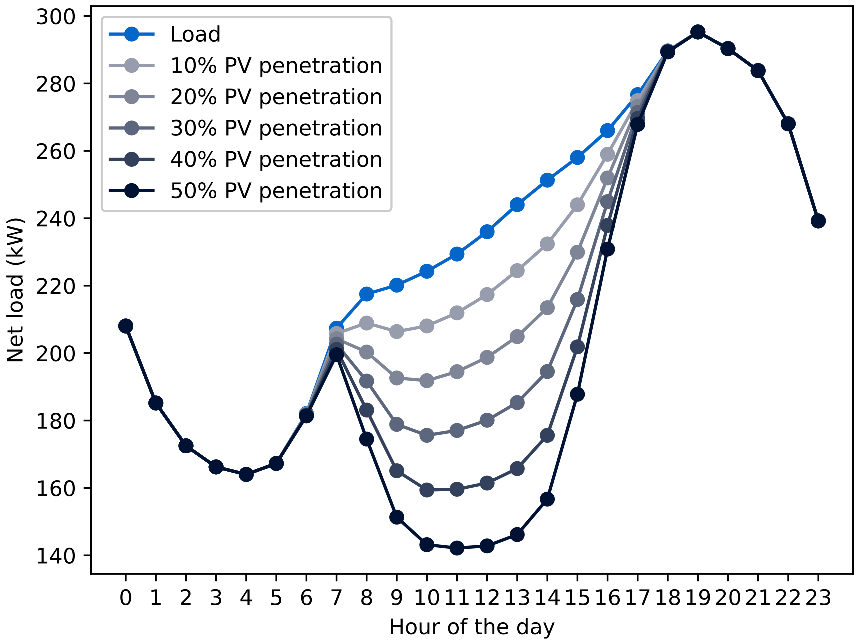


Figure 1.1: Net load curve with varying levels of PV penetration resulting in a duck curve

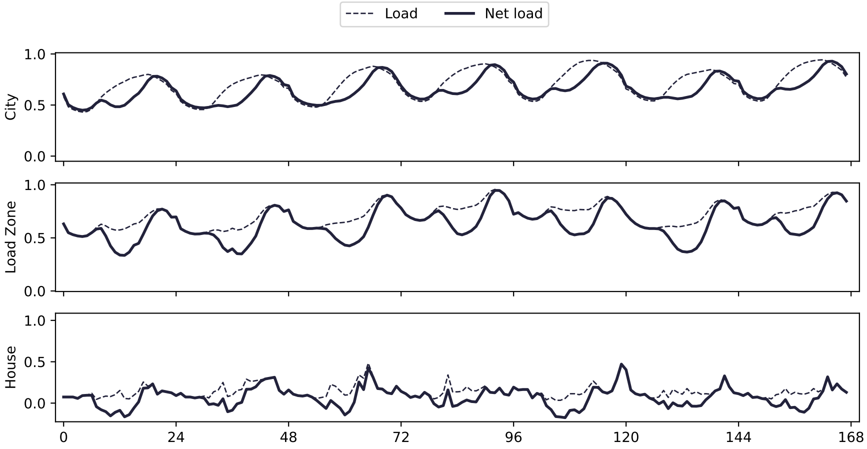


Figure 1.2: Normalized hourly net load profiles with 30% PV for a week for three types of net loads

ation into the electricity grid. The first approach is to "fatten" the duck, increasing its belly by enhancing the flexibility of the power system. This can be achieved by modifying operational practices to allow more frequent power plant cycling, starts, and stops, among other measures. The second approach is to "flatten" the duck, reducing its belly by shifting supply and demand to allow PV generation to fulfill parts of the load that are typically not met during the day. Flattening the duck requires the deployment of energy storage or demand response techniques, both of which are already being implemented in various regions throughout the United States. Both of the aforementioned solutions to address changing load and mitigating the challenges associated with the duck curve can benefit tremendously from more accurate load forecasts. By having a better understanding of future energy needs, system operators can make informed decisions about power generation and distribution, specifically during peak usage periods, enabling them to manage their resources more effectively.

Since load is driven by temperature, many load forecasting models use temperature and its cross effects to forecast load. However, with increasing PV penetration in the grid, load is no longer governed only by temperature. During the hours when the Sun is out, it is governed by a combination of temperature, irradiance, and cloud cover and, for higher penetration levels, the load is completely governed by irradiance and cloud cover. As a result, load forecasting models must be updated to reflect this change. Load forecasting models are no longer effective in regions with significant PV penetration.

Figure 1.2 displays the net load profiles for three distinct types of loads: a city, a load zone, and a residential house, sourced from a medium-sized utility company in the United States. Each net load profile contains 30% PV, with the load shown in a dotted line. The load of a house is highly volatile, and when BTM PV is added, the net load becomes even more volatile. The net load at the load zone level is less volatile than at the house level, but still volatile compared to the city-level net load.

This is because, at higher levels, the load and net load profiles get smoother due to aggregation. It is challenging to develop highly accurate forecasts for lower levels of net loads.

Furthermore, in a study by Razavi et al. (2020), the authors present a comprehensive comparison between the load and net load forecasts at a single-household level and at a low-aggregated level. In general, it has been observed that the error in the net load forecasts is greater than the error in the load forecasts. The difference between net load forecasts and load forecasts is roughly 30% at the household level, while the difference between the two is even more significant at the aggregate level.

1.5 Summary of Contributions

The contributions of this dissertation are closely aligned with these research questions highlighted in Section 1.1, offering practical solutions and advancements in the field of energy forecasting:

1. A day-ahead weather-related outage forecasting model is developed, addressing data imbalances through a weighted logistic regression model with easily interpretable weights. This model is tested in two distinct scenarios: city-level outages and distribution substation outages, the latter of which has received limited attention in academic research. This contribution directly addresses the first research question by applying advanced modeling techniques to enhance the accuracy and reliability of outage forecasts, optimizing these models for practical utility operations.

2. Based on the comprehensive literature review and identified gaps, a benchmark model for net load forecasting is proposed. This model serves as a foundation for future research and aims to standardize approaches in net load forecasting, especially in scenarios involving various levels of PV penetration. This contribution responds to the second research question by developing a practical, robust

model that can be universally applied to stabilize and optimize grid operations across different energy systems.

3. A novel methodology for estimating PV penetration is developed, analyzing the discrepancy between load and net load forecasting models. This approach is capable of distinguishing between load and PV generation even in systems with minimal PV penetration, offering a significant enhancement to the accuracy of PV estimation in the grid. We also address the challenge of detecting behind-the-meter (BTM) PV installations. A method is developed that leverages disparities between the load and net load forecasting models to identify meters with PV installations, thus providing utilities with actionable insights into distributed solar energy adoption.

These contributions methodically address the specified research question, addressing specific aspects of energy forecasting challenges in modern power grids. Each contribution not only answers a specific research question but also enhances practical understanding and operational efficiency in real-world energy system management, showcasing the dissertation's direct impact on improving grid reliability and integrating renewable energy sources effectively.

1.6 Dissertation Organization

The organization of this dissertation is presented in Figure 1.3. An overview of the literature on net load forecasting and PV estimation is presented in Chapter 2. In this study, the net load forecasting papers are analyzed based on their forecasting approach, forecasting technique, and the explanatory variables employed. Furthermore, studies on the impact of PV penetration on net load are also reviewed. The background information needed for the work in this dissertation is presented in Chapter **??**. This includes an overview of multiple linear regression, recency effect, forecast evaluation, and sliding simulation. Chapter 3 presents a day ahead weather-related outage forecasting model. Chapter 4 proposes a benchmark model for net load forecasting. Chapter 5 presents the proposed PV estimation methodology addressing estimating PV penetration in the system and BTM PV detection. The dissertation is concluded in Chapter 6 with a summary of the contributions and an overview of the future vision.

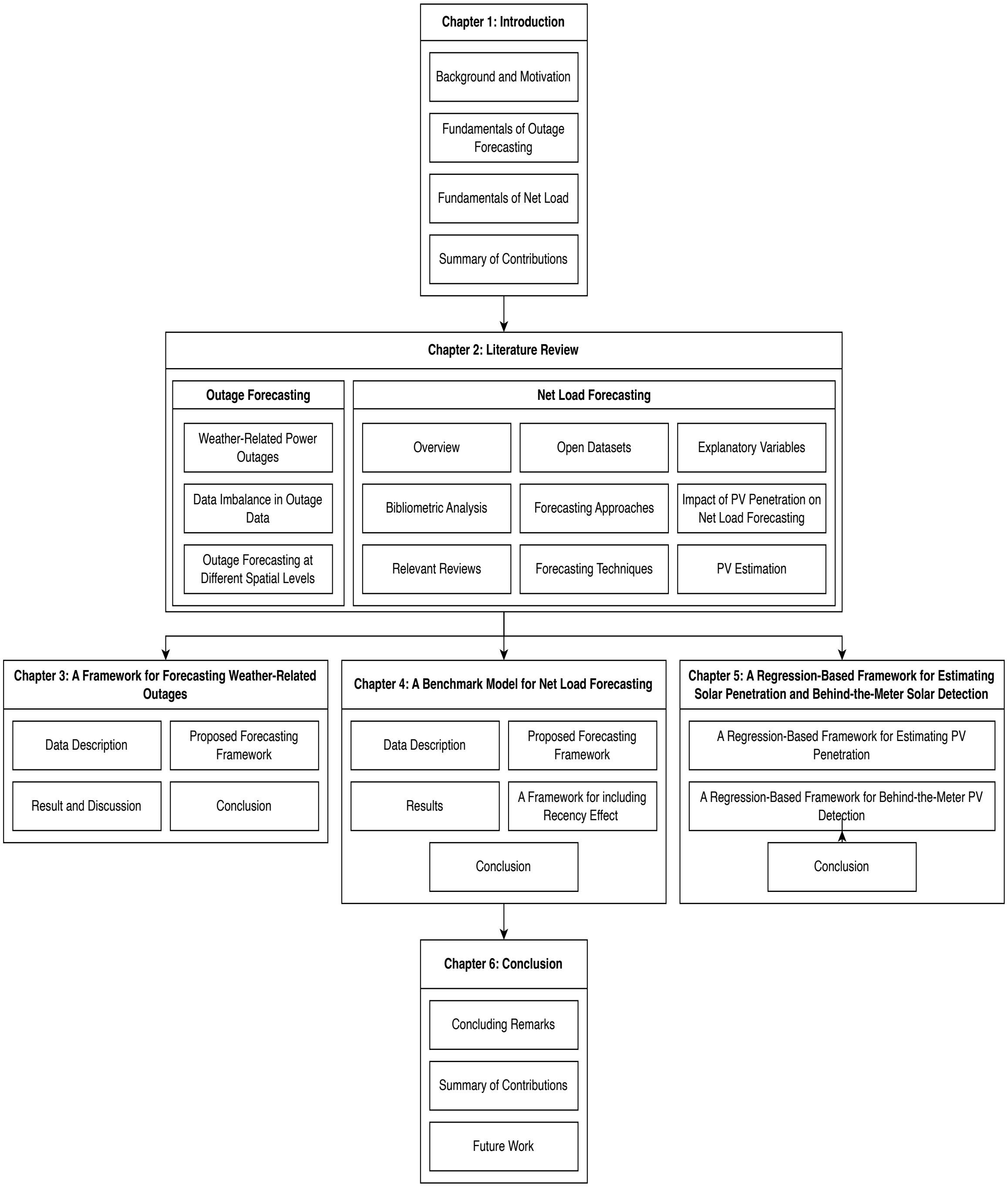

Chapter 1: Introduction
Background and Motivation
Fundamentals of Outage Forecasting
Fundamentals of Net Load
Summary of Contributions
Chapter 2: Literature Review
Outage Forecasting
Weather-Related Power Outages
Data Imbalance in Outage Data
Outage Forecasting at Different Spatial Levels
Net Load Forecasting
Overview
Open Datasets
Explanatory Variables
Bibliometric Analysis
Forecasting Approaches
Impact of PV Penetration on Net Load Forecasting
Relevant Reviews
Forecasting Techniques
PV Estimation
Chapter 3: A Framework for Forecasting Weather-Related Outages
Data Description
Proposed Forecasting Framework
Result and Discussion
Conclusion
Chapter 4: A Benchmark Model for Net Load Forecasting
Data Description
Proposed Forecasting Framework
Results
A Framework for including Recency Effect
Conclusion
Chapter 5: A Regression-Based Framework for Estimating Solar Penetration and Behind-the-Meter Solar Detection
A Regression-Based Framework for Estimating PV Penetration
A Regression-Based Framework for Behind-the-Meter PV Detection
Conclusion
Chapter 6: Conclusion
Concluding Remarks
Summary of Contributions
Future Work

Figure 1.3: Organization of dissertation

CHAPTER 2: LITERATURE REVIEW

2.1 Overview

This chapter aims to provide a comprehensive overview of the current state-of-the-art in outage and net load forecasting. Various methodologies, their applications in different contexts, and the challenges that remain in achieving accurate and reliable forecasts are discussed in reference to outage and net load forecasting. The following are discussed in this chapter:

- A review of the state-of-the-art of outage forecasting
- A review of the state-of-the-art of net load forecasting

2.2 A Review of the State-of-the-Art of Outage Forecasting

Outage forecasting is a critical aspect of power system management, involving the prediction and analysis of power outages. The importance of this field has grown significantly with the increasing complexity of electrical grids and the need for efficient and reliable energy distribution. Research in outage forecasting has expanded significantly in recent years, driven by the availability of large datasets, advances in computational techniques, and the integration of renewable energy sources into the power grid. Notable papers in the field are discussed in Ferreira et al. (2021). Additionally, Hong and Hofmann (2022) offered an in-depth review of the progress made in the field over the last twenty years, highlighting several important publications in this area. As mentioned in Section 3, many aspects of outages can be predicted, which includes the frequency of outages (Kankanala et al., 2013; Sharma et al., 2023), duration of outages (Nateghi et al., 2011, 2014), and the number of customers impacted by outages (Soares et al., 2021). However, most studies have focused on predicting the frequency of outages or the number of outage events. Additionally, it is noted that over 70% of power outages in the US are directly or indirectly attributable to weather-related causes (Kezunovic et al., 2017), resulting in a large majority of research focusing on forecasting outages due to weather events. This review delves into outage forecasting, with a particular emphasis on weather-related electrical disruptions, the challenges associated with imbalance in outage data, and the prediction of outages across different geographic scales.

Earlier work in this field was mainly focused on root cause classification, fault diagnosis, and fault detection. Numerous studies have delved into the effects of different factors on the duration of outages, including the cause of the outage, actions undertaken by repair teams, and temporal factors like month of the year, day of the week, and time of day. The primary objective was to assist distribution operators in rapidly gaining an understanding of what had occurred in order to hasten the process of restoring power after an outage. The literature on fault diagnosis and outage prediction has a significant number of studies (Xu and Chow, 2005, 2006; Dehbozorgi et al., 2020; Xu et al., 2007a; Cai and Chow, 2011; Xu et al., 2007b; Gui et al., 2009).

However, electrical power outages can be caused by a multitude of factors, each varying in frequency and impact (Bashkari et al., 2020). Natural disasters like earthquakes and floods pose significant risks by disrupting power supply systems. Beyond natural causes, equipment failure is a frequent cause of outages; as power grids age, transformers, substations, and other components can fail, leading to outages (Ward, 2013). Human error and wildlife interference, such as animals coming into contact with electrical equipment, can also disrupt power supply. Additionally, increased demand on the power grid, particularly during extreme weather conditions, can lead to overloads and subsequent outages (Ekisheva et al., 2021). Another factor is planned outages, which are necessary for maintenance and upgrades to the power system. Cybersecurity threats have emerged as a modern concern, where malicious attacks on power grid software systems can lead to significant disruptions (Sheng et al., 2011). Each of these causes poses unique challenges and requires tailored strategies for mitigation and management to ensure a reliable power supply. It is also possible that the cause of outage may be unknown (Power et al., 2014). As a consequence, more recent studies have focused on outages due to single causes such as specific weather events (Black et al., 2018; Doostan and Chowdhury, 2020), vegetation (Doostan et al., 2020), etc.

2.2.1 Weather-Related Power Outages

The intricate relationship between meteorological conditions and electrical power outages is a subject of increasing importance in the context of modern utility management. Severe weather events, including storms, hurricanes, snow, ice, and extreme temperatures, pose significant risks to the stability and efficiency of power supply systems. For instance, storms and hurricanes can physically damage power lines and infrastructure, while snow and ice can lead to line breakage and equipment failure. Additionally, extreme temperatures strain the grid by increasing electricity demand, as seen with air conditioning loads during heatwaves or heating requirements during cold spells (Boretti, 2024). Floods, resulting from heavy rains or storm surges, also present a considerable threat by damaging substations and electrical installations. These weather-induced outages not only cause immediate inconvenience but also lead to broader socioeconomic impacts such as substantial economic losses, public health and safety concerns, and impaired communication and emergency response (Bhattacharyya et al., 2021). Therefore, addressing these challenges is crucial, necessitating a proactive approach encompassing grid modernization, predictive analysis, and enhanced weather forecasting, supported by robust policy and regulatory frameworks.

Among the outage forecasting literature, researchers have mostly concentrated on outages caused by weather. Regression-based approaches have been widely proposed in outage forecasting literature. The authors in Black et al. (2018) presented a multiple linear regression model with 37 variables extracted from weather variables such as daily mean, maximum, and minimum of wind speed, speed gusts, wind chill, relative humidity, temperature, precipitation, etc to forecast daily System Average Interruption Duration Index (SAIDI) values. In Zhou et al. (2006), the authors proposed a Poisson regression model and a Bayesian network model to forecast the failure rates of overhead distribution lines. The regression model took into account input variables such as the natural log of lightning strikes, speed gusts, and their interaction, whereas the Bayesian model took into account only the daily values of the natural log of lightning strikes and speed gusts. Both models produced results that were comparable to one another.

With the advancement in machine learning-based forecasting techniques, these techniques are also widely utilized in outage forecasting, just as they are in other areas of energy forecasting. In Kankanala et al. (2013), the authors proposed an ensemble learning method based on a boosting algorithm to forecast wind and lightning-related outages, using total daily lightning strikes and daily wind speed as inputs. Major weather events such as storms and hurricanes can cause widespread outages, lasting several days. As a consequence of this, there have been a significant number of noteworthy studies have looked into forecasting power outages caused by particular weather events, such as storms (He et al., 2017; Zhu et al., 2007), hurricanes (Nateghi et al., 2014; Eskandarpour and Khodaei, 2017) and lightning (Doostan and Chowdhury, 2020). The authors in He et al. (2017) proposed a methodology to predict the number of outages caused due to storms on the distribution network using quantile regression forests (QRF) and a Bayesian additive regression trees (BART) model. A support vector machine (SVM) based classifier was proposed by Eskandarpour and Khodaei (2017), to predict the outages in power grid components post hurricanes. The proposed SVM-based model is also compared with and outperforms a simple logistic regression-based model. The authors in Das et al. (2021) proposed a deep neural network ensemble model to forecast total daily outages caused by wind and lightning using the log values of total daily lightning strikes and daily maximum speed gusts as inputs for the model. An important poiont to note is that even though machine learning-based models are being presented more frequently in literature, these solutions continue to be more difficult to implement in an operational setting due to their lack of interpretability and extensive parameter tweaking requirements.

2.2.2 Data Imbalance in Outage Data

A major challenge when analyzing segments of large utility networks more closely, or when attempting to forecast outages at small and rural utilities in general, is managing the data imbalance caused by the varying hourly frequency of outages. Data imbalance arises when the population of one class of data is considerably higher than the population of the other. In other words, the number of hours not impacted by outages can be significantly more than the number of hours impacted by outages and vice versa. Such an imbalance in the data can lead to outage predictions that underestimate outage events and overestimate non-outage events. Outage forecasting is often formulated as a classification problem. An event classification approach that forecasts power outages by subdividing the training dataset into specific subgroups was proposed to address the imbalance issue (Yang et al., 2020b) Another popular way to address the issue of imbalance in the data is to synthetically balance the data. To do this, researchers frequently devise novel strategies to oversample the minority population. The synthetic minority over-sampling approach (SMOTE), which was proposed by Chawla et al. (2002), is a technique that received a significant amount of attention. SMOTE is often used for the forecasting of outages, such as outages caused by thunderstorms (Kabir et al., 2019a) and typhoons (Hou et al., 2021).

Logistic regression is a frequently used technique for classification (Kleinbaum et al., 2002), and a good candidate technique for outage forecasting. However, ordinary logistic regression models are not designed to cope with imbalanced data. Therefore, they must be tweaked in order to be effectively used to forecast outages. In Xu and Chow (2005), a logistic regression-based model for power distribution fault cause identification for animal or tree-related faults is proposed. The threshold value that is used to convert the probabilities predicted by the logistic regression model to class labels is tuned in order to find an optimal threshold value to compensate for the imbalance in the data. In Doostan and Chowdhury (2020), a weighted logistic regression model for lightning-related outage forecasts is presented. The weights of the model are fine-tuned by a process of trial and error to compensate for the imbalance in the data.

2.2.3 Outage Forecasting at Different Spatial Levels

Forecasting outages at different spatial levels within the distribution network can provide utilities with better situational awareness and help make better operational decisions. Outage forecasting, however, presents unique challenges when addressed at varying spatial levels-regional, zonal, and local. At the regional level, forecasts must account for broad meteorological patterns and diverse infrastructure conditions, predicting outages a complex task that often relies on sophisticated statistical models and large-scale data analysis. Moving to the zonal level, the focus shifts to localized weather events and specific operational practices of the area, necessitating a more granular approach to data collection and analysis. The local level poses its distinct set of challenges, as predictions must be incredibly precise, often down to the level of individual transformers or lines, integrating data on local weather conditions, equipment age, and consumer behavior patterns. This multi-tiered approach to forecasting requires not only a deep understanding of the various factors influencing outages at each spatial level but also an adeptness in employing a range of forecasting methodologies, from traditional statistical methods to advanced machine learning techniques, each tailored to the unique demands of the spatial context.

Forecasting outages at a higher granular level, for example, at the substation level can provide valuable insights to utilities, but only a few substation-level studies have been reported in the literature. A major reason for this is that at higher granular levels (and lower voltage), the data becomes more imbalanced, which makes the challenge of outage forecasting exceedingly challenging. In addition, because of the higher degree of unpredictability that occurs at lower voltages, outages become even less predictable. The various challenges associated with forecasting outages at the substation level are highlighted in Doostan and Chowdhury (2020) and Doostan et al. (2020). As a result, among the outage forecasting literature, only a few papers take into account forecasting outages at the substation level (He et al., 2017; Hou et al., 2021). A methodology to forecast storm-related outages at different spatial levels in the distribution network is proposed in He et al. (2017). The authors in Hou et al. (2021) presented a two-stage methodology to predict the outage area for typhoon-related outages up to 1km*1km grid cells.

2.2.4 Research Gaps

The literature on outage forecasting provides a comprehensive overview of the current advancements in the field. Despite its extensive research, outage forecasting, a subset of energy forecasting, shares both commonalities and unique challenges with its broader domain. This section offers a concise overview of key areas within outage forecasting that demand further exploration:

- Integration of Advanced Meteorological Prediction Models: Enhancing outage forecasting accuracy is feasible through the integration of advanced meteorological models. These models, capable of predicting extreme weather events like hurricanes, storms, and wildfires, can significantly improve forecasting precision.

- Application of Machine Learning and Artificial Intelligence: The use of artificial intelligence (AI) and machine learning (ML) can refine the predictive capabilities of outage models. Analyzing historical data on outages, weather conditions, and grid performance through AI and ML, including technologies like artificial neural networks (ANN) (Xu and Chow, 2005, 2006), decision trees (Dehbozorgi et al., 2020), fuzzy systems (Xu et al., 2007a; Cai and Chow, 2011), and artificial immune systems (Xu et al., 2007b; Gui et al., 2009), shows promise. However, operational implementation remains challenging due to these methods' complexity and the need for extensive customization.

- Benchmarking and Standardization of OMS Data: Unlike other areas such as state estimation and load forecasting, which have benefited from standardized data sets, outage forecasting suffers from a lack of accessible, standardized data (Hong and Hofmann, 2022). Creating benchmark datasets through collaboration among utilities, governmental bodies, and researchers is crucial for advancing outage prediction research and enhancing grid reliability planning. .
- Big Data Analytics in Outage Prediction: Employing big data analytics to examine information from smart meters, sensors, and IoT devices can uncover grid vulnerabilities and forecast potential outages. Additionally, exploring how grid digitization, through smart grids and advanced metering infrastructure, can improve outage detection, response times, and overall grid management is vital.
- Impact of Renewable Energy Integration: Investigating the effects of incorporating renewable energy sources into the grid on outage rates and developing predictive models for grids with significant renewable energy contributions is an essential direction for future research.
- Cybersecurity and Data Integrity in Outage Management Systems (OMS): As highlighted by Hong and Hofmann (2022), the lack of reported data integrity attacks on OMS highlights a significant research gap. Given the critical role of OMS in distribution operations and their vulnerability to data attacks, a

focused research agenda on simulating data attacks, anomaly detection, mitigation strategies for data contamination, and improving the robustness of reliability operations is imperative.

In summary, the field of outage forecasting is vital for enhancing the resilience and reliability of power systems in the face of increasing demand and complexity. The growing body of research in this area reflects its importance, offering insights and tools to predict and manage power outages more effectively. As technology and data analytics continue to advance, the scope and accuracy of outage forecasting are expected to improve, contributing significantly to the stability of electrical grids worldwide.

2.3 A Review of the State-of-the-Art of Net Load Forecasting

A literature review on net load forecasting is included in this section. Publications are evaluated with an emphasis on assessing the methodologies used and explanatory factors used in net load modeling in order to give a holistic overview of the literature on net load forecasting. Each paper is reviewed in its scope with a focus on the forecasting approach, technique, and the explanatory variables used. Furthermore, a section on methodologies for PV estimation is also presented.

2.3.1 Bibliometric Analysis

A bibliometric study is frequently used to examine the research trends in a certain topic. By analyzing the current research data in the subject, a bibliometrics study offers insights into new trends, research topics, and evolutionary subtleties of the field. In this Section, we present a biliometric analysis on net load forecasting that was conducted on 1 July 2023, using a well-known and respected database, Web of Science (WoS). The following query was used in WOS:

TS=(("net load forecasting" OR "net demand forecasting" OR "net load prediction" OR "net demand prediction" OR "forecasting net load" OR "net electricity prediction" OR "net electricity prediction") AND ("power" OR "rooftop" OR "PV" OR " electricity"))

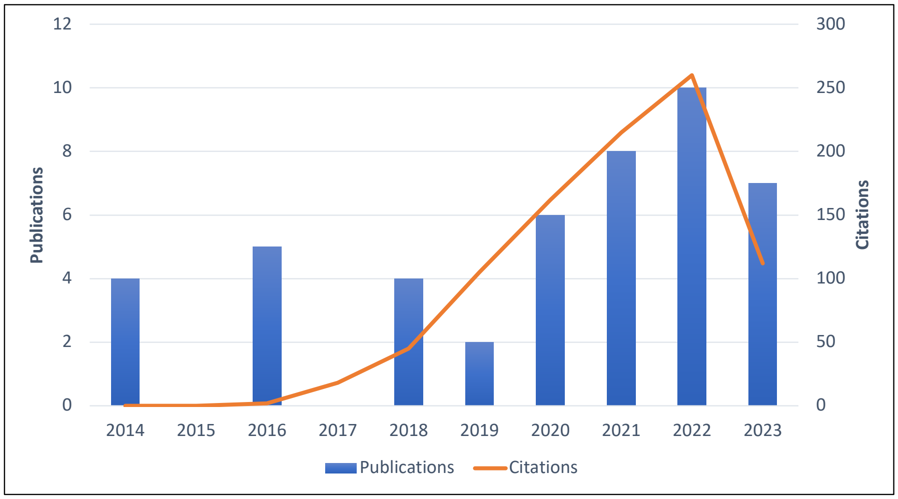


Figure 2.1: Number of journal articles on net load forecasting based on Web of Science query (2014 to 2023)

Figure 2.1 shows the number of publications indexed by WoS on net load forecasting. WoS refers to papers presented in conferences as proceeding papers, and papers published in journals as articles. As can be observed, the first publications on net load forecasting started appearing in 2014. Much like other fields, first conference papers started appearing, and later they were converted to journal articles. From 2014 until 2023, the field has seen an increasing trend in the number of publications per year, as well as the total citations. This demonstrates that net load forecasting is still a developing topic that attracts increasing interest every year. An important point to note here is that WoS indexed articles represent only the journals and conferences that are indexed in WoS. Although WoS databases are continuously being updated to include more conferences and journals, the number is still less than the true number of publications in the field.

2.3.2 Relevant Reviews

While the literature on energy forecasting is extensive, with thousands of papers published and reviews of the literature covering various aspects of load forecasting, none of these have focused on net load forecasting. In 2018, Van der Meer et al. (2018b) highlighted the need for a review of net load forecasting, but due to the

Table 2.1: Summaries of several open net load data sets

Name	**Description**	**Frequency**	**Length**	**References**
ISONE (ISONE, 2023)	8 load zones BTM PV Normalized BTM PV	1 hr	01/01/2014 - 12/31/2022	
PecanStreet (Street, 2015)	24 households Weather data	1 hr	01/01/2014 - 12/31/2016	
Ausgrid Residents (Ratnam et al., 2017)	300 households PV data	30 min	07/01/2010 - 06/30/2013	Sun et al. (2019) Van Der Meer et al. (2018)
Open Power System data (Data Package Household Data, 2020)	100 Households Weather data	1 min	12/11/2014 - 05/01/2019	Alipour et al. (2020)
Umass Smart SunDance data set (UMass, 2017)	100 buildings Weather data	1 hr	1/1/2015 - 1/1/2016	
Solar Home System Jharkhand India (Mehra, 2016)	3 households Weather data	1 min	06/25/2015 - 01/31/2016	

insufficient number of research publications on the topic, they opted for a separate review on probabilistic solar and load forecasting. Since the publication of Van der Meer et al. (2018b), some noteworthy papers on net load forecasting have emerged, although the volume of literature on this subject still remains relatively low. In their recent review on behind-the-meter solar forecasting, Erdener et al. (2022) highlighted the significance of net load forecasting for unit commitment and economic dispatch at the transmission scale. The authors also examined some noteworthy papers in the field of net load forecasting, however the review mainly addressed BTM PV forecasting. Currently, in our knowledge, there is no literature review present on net load forecasting. This paper attempts to provide a comprehensive overview of the advances made in the field of net load forecasting. The publications are evaluated with an emphasis on assessing the methodologies utilized and explanatory factors used in net load modeling in order to give a holistic overview of the literature on net load forecasting. Each paper is reviewed in its scope with a focus on the forecasting approaches, techniques, and the explanatory variables used.

2.3.3 Open Data sets

Due to net load being a more recent phenomenon, most power companies have less less history of net load data. In addition, many power providers do not publicly disclose net load data due to concerns about data privacy and other related topics. Table 2.1 provides a summary of some of the freely accessible data sets that can be used to conduct studies on net load forecasting. Some of the open source data sets are as follows:

- **Independent System Operator New England (ISONE):** ISONE is an independent regional transmission organization responsible for operating the power grid and wholesale electricity markets covering the six states in the north-eastern region of the United States, divided into 8 load zones. The load data for each of the load zones is freely available on ISONE's website. Starting 2023, along with hourly load data, ISONE also released estimates of total hourly production from behind-the-meter PV installations for each of their eight wholesale load zones as well as the total system BTM generation. Along with the BTM PV power production data, ISONE also released the normalized BTM PV data, which is the ratio of estimated hourly BTM PV power production relative to the total BTM PV installed capacity for each zone.

- **PecanStreet:** The data was compiled as part of the Pecan Street Demonstration and comprises 3 years' worth of hourly data for 24 residences with rooftop PV systems having an install capacity ranging from 2.9 to 8.8 kW. The households are located in Austin, Texas, US with the data is made available via Pecan Street Dataport from January 2014 until December 2016 (Street, 2015).

- **Ausgrid Residents:** The data was compiled by Ausgrid, a power distributor located on the east coast of Australia. The data contains half-hourly gross meter data comprising rooftop PV generation for 300 households located within the service territory of Ausgrid's electricity network over 3 years (Ratnam et al., 2017).

- **Open Power System Data, Household dataset:** The data contains measured solar power generation as well as electricity consumption down to the single device consumption for several small businesses and residential households.

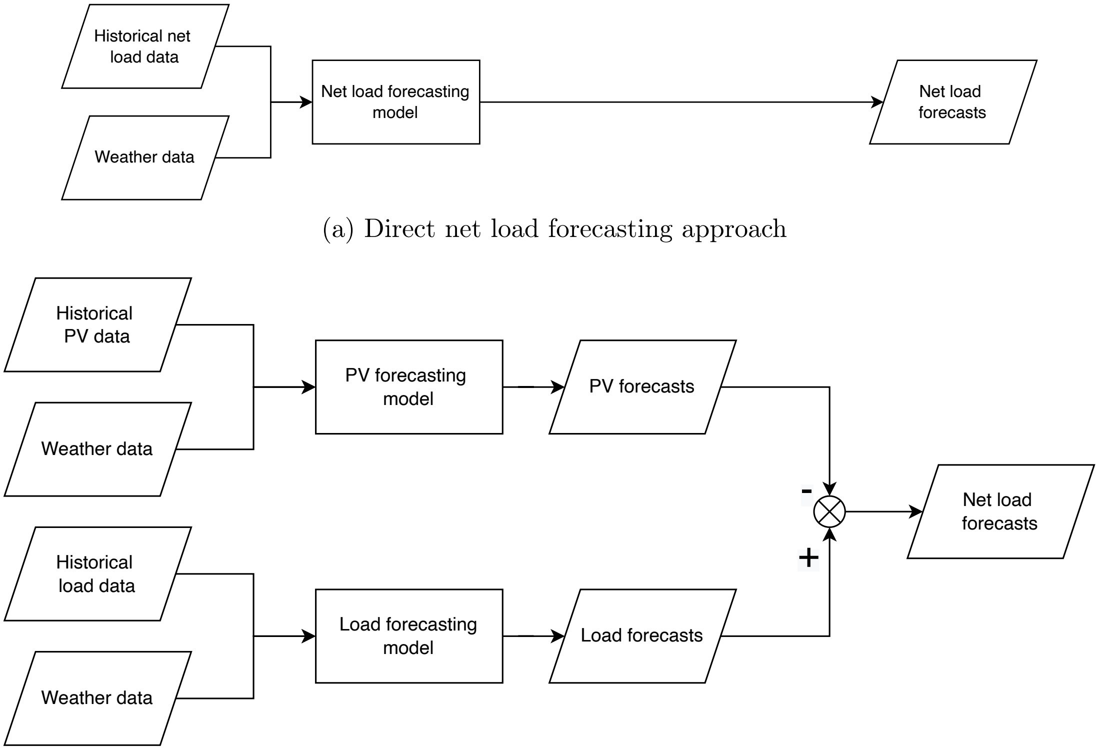
The data is available at a 1 minute resolution for more than 4 years (US Department of Commerce, 2020).

- **UMass Smart, SunDance dataset:**The data comprises of hourly net meter, solar generation and corresponding weather readings from a weather station for 100 buildings in North America over the course of one year (**?**).
- **Solar Home System Jharkhand India:** The data contains around 6 months electricity consumption and solar generation data for 3 households located in Jharkhand, India. The data is available at a 1 minute resolution (**?**).

2.3.4 Forecasting Approaches

Net load forecasting approaches can be classified into two broad categories, namely, direct and indirect net load forecasting approaches. In the direct approach, historical net load and corresponding weather variables are taken as inputs to estimate the parameters of the forecasting model. Once the parameters are estimated, the weather forecast is used to generate net load forecasts. In the indirect approach, load and renewable generation are forecasted individually and then combined to get a net load forecast. Figure 2.2 shows the direct and indirect net load forecasting approaches.

In the realm of net load forecasting, the indirect approach or the additive approach is proposed due to its potential to leverage existing methodologies and potential advantages. This approach primarily involves forecasting the individual components of net load separately, such as renewable energy generation and the total load, before subsequently deriving the net load by integrating these forecasts. The rationale behind this method is the belief that forecasting individual components can harness more specific models tailored to the unique characteristics and patterns of each component, potentially leading to enhanced accuracy Wang et al. (2018b). However, it is imperative to note that the indirect approach also introduces the challenge of accumulating errors from multiple forecasts when they are combined Ruiz-Abellón et al.



(a) Direct net load forecasting approach

(b) Indirect net load forecasting approach

Figure 2.2: Net load forecasting approaches

(2020). Sreekumar et al. (2018) presented an indirect approach for very short-term interval net load forecasting for a system with wind and solar generation. The authors proposed a modified grey model to forecast load, solar, and wind generation individually. Net load forecasts were then obtained by subtracting the total renewable generation forecasts from the load forecasts. Ruiz-Abellón et al. (2020) also proposed an indirect methodology to forecast the day-ahead net load for demand response strategies for prosumers. The authors observed that the majority of errors in net load forecasts came from PV forecasts rather than load forecasts. The authors addressed this issue by proposing a very short-term adjusted PV forecast, inspired by the demand-correcting procedures used by the ISOs during power events. An indirect net load forecasting model is also proposed in Kaur et al. (2016) with separate models for forecasting load and PV generation. In most cases, however, load and renewable generation data are not available individually. In such cases, authors have proposed techniques to first estimate the output power or capacity of individual BTM solar PV panels, in order to separate out the load and renewable generation data from net load. For instance, Wang et al. (2018b) presented a methodology to first extract the BTM PV's capacity using a correlation analysis and grid search. Subsequently, load and PV forecasting models were used to forecast each time series individually. In Stratman et al. (2023a), a disaggregation algorithm is applied to separate load and PV generation from net load. Then, an LSTM model is used to forecast load and PV separately using the historical disaggregated load and PV, respectively. The additional step of disaggregation or capacity estimation, however, also adds additional uncertainty to the forecasting framework.

Unlike the indirect approach, the direct approach forecasts the net load directly by considering the integrated effect of both renewable generation and load. This approach offers the advantage of inherently accounting for the interdependencies between renewable generation and load patterns, thereby potentially improving the accuracy of the forecast. Moreover, by eliminating the need for separate forecasting models, the direct approach simplifies the forecasting process, making it more computationally efficient. Chu et al. (2017) proposed a direct approach for short-term net load forecasting for four feeders in San Diego, CA. Mei et al. (2019) also proposed a direct net load forecasting approach to forecast ultra short-term net load. Kobylinski et al. (2020) proposed a direct net load forecasting model to forecast the total net load for a neighborhood with 75 single-family houses. Kaur et al. (2013) presented a comparison of multiple direct net load forecasting techniques for 1 h and 15 min horizon to analyze the impact of on site solar generation. The performance of the foresting models was compared for a case with no PV penetration and a case with high PV penetration. A detailed discussion of this work on the impact of solar penetration on net load forecasting models is presented in Section 2.3.7.

Some papers also present a comparison of the direct and indirect approaches. Kaur et al. (2016), presented a comparison between direct and indirect approaches to forecasting net load. For the direct net load forecasting model, solar forecasts were used as the primary input to train the model. It was observed that the direct approach outperformed the indirect approach. The direct approach proposed in this study, however, is not a fully independent net load forecast because it includes solar forecasts as an input to train the net load model. The proposed direct solution is constrained due to the additional computation required to forecast solar generation separately and the unavailability of solar generation data in real-world scenarios. Furthermore, the added error from the solar forecast is included to the net load forecast in this method. Pierro et al. (2020) presented a study on probabilistic net load forecasting to manage day-ahead reserves. The paper presented two direct approaches, namely persistence and seasonal autoregressive integrated moving average (SARIMA) and two indirect approaches, namely artificial neural network (ANN) and smart persistence. The forecasting models were developed based on previous studies done by the same authors in Pierro et al. (2017). The work presented in Alipour et al. (2020) also compared net load forecasts obtained from a direct and an indirect approach for short-term and medium-term net load forecasting. The authors observed that the indirect approach yields better results than the direct approach for mid-term net load forecast. Van Der Meer et al. (2018) compared direct and indirect probabilistic net load forecasting models. It was found that the two approaches have their own advantages and disadvantages in terms of prediction interval (PI) and accuracy. The authors concluded, "Overall, selecting the best strategy depends mainly whether one prefers higher informativeness of PIs, or higher coverage probability."

The decision to use one over the other is generally governed by the availability of data. In the case when only the net load data is available, the only option is to use the direct approach while when load and PV data are available independently, either the direct or the indirect approach can be used. Based on the studies reviewed, the two approaches have their own advantages and disadvantages.

2.3.5 Forecasting Techniques

A number of forecasting techniques have been applied to forecast net load. Generally, these can be classified into two broad categories: statistical techniques and artificial intelligence (AI) techniques. Statistical methods, such as autoregressive integrated moving averages (ARIMA) or exponential smoothing, are based on statistical properties of the data. They are used to detect patterns and trends and to make forecasts based on those. These methods are easy to understand, interpret, and implement, and they often work well when the underlying system follows a linear trend or pattern. AI techniques, on the other hand, such as artificial neural networks (ANN), support vector machines (SVM), or random forests, are capable of capturing complex, non-linear relationships in data. They can handle high-dimensional inputs and are more adaptable to changes in the underlying system. They also have the potential to incorporate multiple influencing factors, such as weather variables, into the forecasting model. However, the downside of these models is that they can be more computationally intensive and harder to interpret, often referred to as "black-box" models due to their lack of transparency in how inputs are transformed into outputs. The forecasting techniques used in net load forecasting literature are summarized in Table 2.2. Based on the techniques used in the papers reviewed for this study, we classify the techniques into three categories, namely statistical, neural network, and others.

2.3.5.1 Statistical techniques

A 24 hour persistence model along with a seasonal autoregressive integrated moving average (SARIMA) model have been proposed in Pierro et al. (2020) for day ahead net load forecasting. Kaur et al. (2013) presented five time series techniques for 15 min and 1 h ahead net load forecasting. The techniques included a persistence model,

Table 2.2: Summary of the forecasting techniques used in net load forecasting literature

Forecasting Techniques	**Reference**
Statistical	Kaur et al. (2016), Kaur et al. (2013), Chu et al. (2017), Kaur et al. (2013), Pierro et al. (2020)
Neural Networks	Chu et al. (2017), Alipour et al. (2020), Kobylinski et al. (2020), Mei et al. (2019), Pierro et al. (2020), Razavi et al. (2020), Sepasi et al. (2017), Stratigakos et al. (2021), Sun et al. (2019), Wang et al. (2018b), Stratman et al. (2023b),Zhang et al. (2023)
Other	Chu et al. (2017), Kaur et al. (2016), Kaur et al. (2013), Sreekumar et al. (2018), Van Der Meer et al. (2018)

two smart persistence models, one with the assumption that the difference in terms of trend persists and other with the assumption that ratio between the time series and its trend remains unchanged. Furthermore, an autoregressive (AR) model and an ARMA model was also presented.

2.3.5.2 Neural Network Based Techniques

Artificial neural network (ANN) models are a category of artificial intelligence models inspired by the structure of the human brain. ANNs have been extensively used in energy forecasting since the past four decades due to their ability model complex relationships and incorporate multiple influencing factors, adapting to changing patterns over time. A typical neural network includes an input layer, several hidden layers, and an output layer. The input layer assimilates raw data such as historical net load data, weather variables, and time-of-day indicators. This data is then processed through the hidden layers via weighted connections and activation functions introducing non-linearity, with the output layer ultimately generating the forecasted net load. Despite the challenges posed by their need for careful design, parameter tuning, and less interpretable results compared to linear models, neural networks have become increasingly popular due to their considerable potential for accuracy, and have been used in a variety of applications, including net load forecasting. As a result of their popularity, a variety of neural network architectures have been developed over time, each associated with a particular type of neural network layer.

Chu et al. (2017) presented an ANN-based model with a Bayesian regularization process with Levenberg-Marquardt optimization to forecast net load for four utility-scale feeders with different levels of solar integration. In Sepasi et al. (2017), a complex-valued neural network (CVNN) was proposed. CVNNs are similar to ANNs with the difference that the inputs, outputs, and parameters like as weights and thresholds, are complex numbers in the CVNN. Kobylinski et al. (2020) proposed a multilayer perceptron (MLP) model with one hidden layer consisting of 16 hidden neurons for net load forecasting.

Although there is no clear cutoff point for when a neural network becomes a deep neural network, it is generally accepted that as a model's complexity rises with numerous hidden layers, it is referred to as a deep neural network. Alipour et al. (2020) presented a deep learning model that combined an auto-encoder neural network with a cascade neural network layer to forecast net load. The role of the auto-encoder is to extract important features and for dimension reduction. Generally, auto-encoders are used in image classification tasks, paired with a classifier layer. However, the authors paired the auto-encoder neural network with a cascade neural network model which is similar to a feed-forward neural network except that the input layer is linked directly to all subsequent layers. This resulted in a hybrid model that was able to take a number of features as input and outperform other machine learning models that it was compared to. The authors further fine-tuned the neural network model by varying its architecture. Zhou et al. (2021) proposed a deep belief network approach to forecast multi-energy net loads for prosumers in a local energy system. The prosumers were first aggregated using a k-means clustering algorithm and then a separate forecasting model was run for each cluster.

LSTM (Long Short-Term Memory) is a Recurrent Neural Network (RNN) based architecture that is widely used in time series forecasting. The LSTM rectifies some of the issues that the recurrent neural networks suffer from: short-memory and vanishing gradient (Hochreiter and Schmidhuber, 1997). This makes LSTM a particularly appealing model, as summarized in Table 2.2, by the increasing number of publications that propose an LSTM-based model for net load forecasting. Stratigakos et al. (2021) proposed a hybrid model that combines singular spectrum analysis (SSA) based decomposition with LSTM. To tune the hyperparameters of the model a random search algorithm is implemented. The authors also compared the proposed model with other popular techniques such as ARIMA, SVM, LSTM and MLP. Sun et al. (2019), proposed a Bayesian deep LSTM network (BDLSTM) model, that combined the architecture of the LSTM model and introduced a prior distribution on the weights and parameters of the LSTM model and inferring the posterior distribution. The motivation to combine Bayesian and the LSTM models came from the distinct properties of the two models. Bayesian model is inherently a probabilistic model that allows the LSTM model to represent uncertainty, while the LSTM model is known to have the ability to capture long-term dependencies in the data. The proposed model was applied to forecast net load for residential customer. The customers were first grouped into two groups, namely, customers with visible PV and customers with invisible PV. Using hierarchical clustering, these two categories were clustered based on their load profiles. A separate BDLSTM model was then built for each individual cluster. The probabilistic forecast of each individual cluster was then aggregated to get the net load at the aggregated level. Razavi et al. (2020) also proposed a version of LSTM: multi-input single-output (MISO) LSTM model for load and net load forecasting. The proposed model takes the historical net load of individual households as input to forecast either household or low-aggregated net load. Compared to a traditional LSTM based model, the MISO LSTM model performed better at the low-aggregate level, however it did not show any improvement at the household level.

2.3.5.3 Other techniques

Other techniques such as support vector machines (SVM) (Chu et al., 2017; Kaur et al., 2016), k-nearest neighbour (kNN) (Kaur et al., 2013) have also been used by authors, but mainly to compare their proposed forecast model's performance. Probabilistic net load forecasts using a dynamic Gaussian process (GP) based timeseries forecasting technique was proposed in Van Der Meer et al. (2018). In Van der Meer et al. (2018), the authors build upon their previous work in Van Der Meer et al. (2018) and improved the GP model with by finding the optimal window size to train the model, making it a good model even with limited amount of training data. A quantile regression (QR) model is also proposed in this work.

In the literature on net load forecasting, a variety of techniques have been employed, including time series methods, ANN, SVM, and others. A growing trend towards the use of machine learning techniques, similar to other energy forecasting fields, is also evident in the field of net load forecasting, as corroborated by the findings in Table 2.2. A significant portion of the published papers propose models that fall under the neural network family, encompassing everything from MLP to LSTM models. However, a major drawback of these techniques is their lack of interpretability. Most machine learning approaches are still regarded as 'black-box' models, which may lead utilities to prefer statistical models for their transparency and interpretability. Despite their widespread use in load forecasting, regression-based techniques are rarely employed in net load forecasting.

As highlighted by Hong and Fan (2016), the notion of a universally "best" technique in forecasting is indeed more myth than reality, largely due to the diverse nature of forecasting problems that can differ greatly in scope, scale, and complexity. Each forecasting technique has its own strengths and weaknesses, and the effectiveness of a particular method can vary depending on the characteristics of the specific forecasting problem at hand. For instance, net load forecasting involves predicting the net electricity consumption, which is influenced by a multitude of factors including weather conditions, time of day, day of the week, economic conditions, and consumer behavior. Different forecasting techniques capture these elements in varying ways, and the effectiveness of a particular method may depend heavily on the specific characteristics of the data, the time frame of the forecast, and the particular needs of the user, such as the trade-off between accuracy and interpretability. A technique that excels in short-term forecasting might not be suitable for long-term predictions. Likewise, a complex neural network might capture intricate patterns in one dataset but may overfit and perform poorly in another context. Furthermore, the success of a forecasting technique doesn't only depend on the choice of the model but also on proper preprocessing of the data, feature selection, parameter tuning, and error analysis. Even the best model can provide inaccurate forecasts if not implemented and tuned correctly. The best technique can also change over time as new data becomes available, and the underlying system changes.

2.3.6 Explanatory Variables

Weather has a significant influence on load, PV generation, and, as a direct consequence, net load. As a result, several meteorological factors and their effects on net load forecasting models have been investigated. Some of the most widely studied weather variables are temperature, solar irradiance, wind speed, humidity, electricity pricing, and cloud cover.

The use of weather variables for net load forecasting is dependent on a number of parameters, including historical data availability, meteorological conditions, and so on. It is critical to carefully identify weather factors that will increase the forecasting models' performance while having no detrimental influence on the forecast. As a result, researchers have been exploring for new ways to harness meteorological data in order to improve net load estimates. One such approach is presented in Alipour et al. (2020). The authors begin with meteorological variables such as temperature, solar radiation, wind speed and their lags. Additionally, electricity price of the current and previous day have also been considered. A discrete wavelet transformation of these variables is taken to decompose the variables into low frequency and high frequency components. Subsequently an autoencoder neural network is implemented with the aim of reducing the dimensions of the input features. Stratigakos et al. (2021) also explore temperature and solar radiation as explanatory variables along with class variables for holiday and weekdays to forecast day-ahead hourly net load.

The authors in Chu et al. (2017) presented three strategies to enhance the net load foresting models and in doing so, present some very interesting analysis. (1) the net load timeseries was detrended into two components: a low-frequency daily trend that represents everyday human activity, and a high-frequency component that represents variations coming from PV generation. It was observed that the error was reduced in models trained on detrended data. (2) to capture substantial variability throughout the day, the authors train separate models for day and night. The findings, however, do not support this hypothesis as there not much improvement observed by training separate models for day and night. (3) the authors investigated the influence of including cloud cover data as an exogenous input to net load forecasting models. Cloud cover data was collected using two sky imagers stationed near two feeders. The improvement in the performance of the forecasts for the feeder with high PV penetration was higher, although the overall improvement was low. The authors concluded that "the sky-imaging techniques are expected to noticeably enhance the daytime performance of data-driven forecasts for feeders with high solar penetration levels". Sepasi et al. (2017) proposed a similar day model, in which the forecasted net load is a combination of the net load of a few previous similar days. To further make sure that the model does not lose out on the recent data, an additional model with a moving window of the 20 most recent data points was also developed. The final forecast was calculated using a weighted average of the two models. The authors in Sun et al. (2019) proposed an approach for forecasting net load for residential consumers. The authors add a "Feature Construction Stage" as part of their forecasting methodology, with the goal of identifying highly contributing explanatory variables for the forecasting model. However, the authors end up using raw features as input owing to the capability of the proposed Bayesian deep neural network model to effectively handle raw data without the need of careful feature selection. The raw input features for the model included the historical net load, lags of net load and calendar variables such as hour of the day, day of the week and month of the year. Wang et al. (2018b) compared the indirect approach with two direct approach based models, one with temperature as the input and the other with temperature and irradiance as the inputs. The model with temperature and irradiance outperforms the model with only temperature by a significant margin.

Kobylinski et al. (2020) presented an approach to select the best features for the proposed ANN model to forecast the net load for a neighborhood. Additionally, daily and yearly cycles are included in the model using sine and cosine functions. A class variable representing single day holidays and two class variables representing the first and second day of two day holidays were also been included in the model. Lags of net load were included in the model by optimally selecting the number of lags based on the minimum value of the autocorrelation function. The authors also introduced GHI to capture the PV generation. Calendar variables such as day of the week were also included in the model. However, the categorical variables were not correctly included in the model, for example, seven class variables representing each day of the week were included in the model. When encoding categorical variables like "weekday" in a model, one category should be taken as a reference and be left out. This means that for the "weekday" variable, only six dummy variables should be included in the model. $n-1$ features can entirely describe a categorical variable with n categories.

In the studies examined in this paper, all the studies that utilized GHI did it in

Table 2.3: Summary of the explanatory variables used in net load forecasting literature

Explanatory Variable	**References**
Temperature	Alipour et al. (2020), Sreekumar et al. (2018), Stratigakos et al. (2021), Wang et al. (2018b), Stratman et al. (2023b), Zhang et al. (2023)
Irradiance	Alipour et al. (2020), Kobylinski et al. (2020), Stratigakos et al. (2021), Wang et al. (2018b), Stratman et al. (2023b), Zhang et al. (2023)
Recency information	Kaur et al. (2016), Sun et al. (2019), Zhang et al. (2023), Kaur et al. (2013),Kobylinski et al. (2020), Van Der Meer et al. (2018)
Calendar Variables	Kobylinski et al. (2020), Razavi et al. (2020), Sun et al. (2019), Zhang et al. (2023)
Cloud Cover	Chu et al. (2017), Zhang et al. (2023)

its original form. These investigations, however, did not delve into exploring interactions, polynomials, or other transformations of GHI data to potentially uncover nuanced relationships or hidden patterns. The existing literature, while acknowledging the importance of GHI, does not seem to fully explore potential enhancements in the predictive power of forecasting models that might be achieved through transformations or interactions of GHI data. For example, Stratman et al. (2023a) proposed a direct net load forecasting model using only GHI as the input, based on a correlation analysis that showed that GHI had a higher correlation with net load than temperature. The authors failed to further explore models that incorporated both GHI and temperature. Therefore, it could be beneficial for future research to investigate these aspects, as they may further improve the accuracy and reliability of solar generation and net load forecasts.

A few studies also present univariate models that do not include any exogenous variable and rely only on the historical net load timeseries. Van Der Meer et al. (2018), only use endogenous explanatory variables in their model. The authors select from different set of inputs, comprising of different combinations of time based lags of net load. A univariate model with only lag variables of net load is presented in Stratigakos et al. (2021).

A summary of the explanatory variables frequently employed in net load forecasting literature is presented in Table 2.3. Foremost, temperature and irradiance are the most important variables, consistently used across multiple studies. This highlights the pivotal role of temperature in driving electricity demand and the significance of irradiance, especially GHI, in driving PV output. The inclusion of recency information in many studies indicates the importance of recent trends and patterns in predicting future net loads. This suggests that net load is not just influenced by external factors like weather but also by its own historical values, emphasizing the time-series nature of the data. The use of calendar variables in several studies highlights the influence of cyclical patterns, such as weekends, holidays, and seasons, on net load. Such variables can capture the variations in human activity and energy consumption patterns that are tied to specific days or times of the year. While cloud cover is not as frequently cited as temperature or irradiance, its inclusion in some studies points to its potential significance, especially in areas with high solar penetration. Cloud cover can directly impact solar PV output, making it a crucial factor in certain geographies or for specific forecasting models.

2.3.7 Impact of PV Penetration on Net Load Forecasting

The impact of large-scale integration of PV into the electricity grid is a key area of research since it has a direct impact on net load. A modest amount of PV penetration might not be able to offset load enough to draw attention, whereas higher PV penetration levels might significantly impact load.

Wang et al. (2018b) study the impact of PV penetration on net load at the independent system operator (ISO) level in their research. PV penetration of the system is varied from 0% to 20% by manually adding PV to the system. It is observed that the performance of the models degrades significantly as the PV penetration in the system increases. In an analysis presented in Shaker et al. (2016b), the authors found that on a daily and weekly level, the net load can be eight times more volatile than load, making it much harder to forecast. The study concludes, “Thus, with the increasing penetration of renewables, predicting the net load for operation planning purposes would become more challenging.". Chu et al. (2017) compared the net load forecasts for four feeders with different levels of PV penetration. They also found that the forecasting error for feeders with higher PV penetration was significantly higher. Another analysis on the impact of PV penetration on the net load forecasting accuracy is presented in Pierro et al. (2020). Six levels of PV penetration from no PV to 45% have been compared by analyzing their impact on day-ahead reserve requirement. The results show that as the PV penetration grows, the uncertainty in following reserve requirement also increases, again highlighting the difficulty in forecasting net load with high PV penetration.

Kaur et al. (2013) dig deeper into understanding what drives the net load forecasting error by analyzing the impact of variability in PV generation along with PV penetration. To investigate the impact of PV penetration, two cases are presented, one with onsite PV and one without it. The overall forecast error is found to be higher for the case with onsite PV generation compared to the case without PV generation. To investigate the impact of PV variability, days with variable PV variability were compared. It was observed that a low forecast error was observed for sunny and overcast days regardless of the solar penetration level. However, on cloudy days, a high forecast error was observed even for medium solar penetration, indicating that forecast error is impacted more by solar variability than by solar penetration. However, as noted in Van der Meer et al. (2018a), the main reason for these findings could be that the analysis is carried out on a single PV farm, thus ignoring the smoothing effect.

For BTM PV, Sun et al. (2019) examine a different facet of PV integration relevant to residential customers. Most of these residential customers have invisible BTM PV, and only a small fraction of them have visible PV generation, due to the requirement of installing a separate meter for demand and PV generation. To analyze the impact of visibility of BTM PV generation, the authors compare their proposed model for varying degrees of visibility of PV. They observe that as PV visibility increases, net load forecasting results improve. As a result, it can be stated visibility into the BTM PV generation can improve net load forecasting performance. To achieve 100 percent visibility would require the installation of separate meters to measure PV generation. There is a trade-off between the accuracy of the forecast and the cost of installing meters for each customer.

The authors in Van der Meer et al. (2018) also investigate the impact of PV penetration on net load forecasts. To do this, the authors vary the PV penetration from 10% to 100% and observe the outcome of the net load forecasting model. However, the findings are contradictory to what has been observed in previous studies. The authors observe a positive impact on performance as PV penetration increases. The authors attribute this explanation to the differences in the scale of data included in all of these studies. Other studies deal with data in the MW range, whereas the data in this study is of a few hundred kW. Since this study deals with local distribution grid data, the smoothing effect of PV power is more pronounced than in other larger-scale systems. This argument, however, cannot be generalized, and additional research on the topic is required.

2.3.8 Forecast Evaluation

Forecast evaluation is the process of assessing the accuracy and effectiveness of a forecasting model. This is done by comparing the forecasted values against the actual values observed during the period for which the forecast was made. Forecast evaluation helps in identifying the strengths and weaknesses of a forecasting model, and provides insights on how to improve it. It's an important part of the forecasting process as it not only guides the model selection, tuning, and refinement, but also assists stakeholders in understanding the performance of the forecasting system, its reliability, and its potential impact on decision-making processes. Definitions and discussion on multiple error metrics are extensively reviewed in (Zhang et al., 2015).

Unlike in traditional load forecasting where the Mean Absolute Percentage Error (MAPE) is a widely accepted and predominantly used metric both in academia and businesses, the net load forecasting community is yet to agree around a singular, universally accepted metric. This can be attributed to the intricate nature of net load forecasting, with its intricate challenges that arise due to the variability and unpredictability of renewable energy sources.

A review of the literature on net load forecasting reveals a wide variety of metrics used by researchers. Some of the commonly used metrics include MAPE, Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE). These metrics provide different perspectives on the forecast errors. For instance, MAE offers insights into the average magnitude of the errors, while RMSE gives more weight to larger errors, making it more sensitive to outliers. On the other hand, MAPE, a relative metric, offers a percentage error, making it convenient for comparisons across different scales. Other metrics cited in the studies include normalized RMSE (nRMSE) and the Mean Bias Error (MBE). While nRMSE offers a scale-normalized version of the RMSE, catering to datasets of different scales, the MBE captures the systematic bias in predictions, providing insights into whether the model consistently overestimates or underestimates the actual values. Furthermore, metrics like the R-squared and forecast Skill have also been used. For studies that look into probabilistic net load forecasting, metrics such as the prediction interval coverage probability (PICP), prediction interval normalized average width (PINAW), continuous ranked probability score (CRPS) and pinball score have been proposed.

One metric worthy of special attention is the MAPE. In the context of net load forecasting, MAPE is one of the most frequently used error metric, reflecting its significance in load forecasting and ease of interpretation. While MAPE offers several advantages, it has a significant limitation, especially for net load with high renewable

Table 2.4: Summary of the error metrics used in net load forecasting literature

Reference	**MAE**	**MAPE**	**RMSE/nRMSE**	**MBE**	**R-Squared**
Alipour et al. (2020)	x	x	x		
Chu et al. (2017)		x	x	x	
Kaur et al. (2016)	x	x	x	x	x
Kaur et al. (2013)	x		x	x	x
Kobylinski et al. (2020)	x	x			
Mei et al. (2019)		x	x		
Pierro et al. (2020)		x	x		
Razavi et al. (2020)		x			
Sreekumar et al. (2018)	x	x			
Van Der Meer et al. (2018)	x	x	x		
Sepasi et al. (2017)	x		x		
Stratigakos et al. (2021)		x			
Sun et al. (2019)	x	x			
Wang et al. (2018b)		x	x		
Stratman et al. (2023b)	x		x		
Zhang et al. (2023)	x	x			

penetration. MAPE is defined as:

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{A_t - F_t}{A_t}\right| \tag{2.1}$$

where A_t is the actual value and F_t is the forecasted value at time t. Since the definition of MAPE involves division by the actual value, it can be undefined or overly sensitive when the actual value approaches zero. In the context of net load, which accounts for both demand and renewable energy generation like solar and wind, there could be times when the net load is zero or close to it, such as during periods of high renewable generation and low demand. Razavi et al. (2020) address this limitation by using Mean Arctangent Absolute Percentage Error (MAAPE). MAAPE retains the advantages of MAPE but is more robust against outliers and is more accommodating when dealing with values near zero. The arctangent transformation ensures that the error remains bounded, thus providing a more reliable representation of the forecasting performance. Its adoption indicates a search for more robust and resilient metrics in the face of the unique challenges presented by net load forecasting. It is also important to note that for lower renewable penetration levels, the known drawbacks of MAPE, such as challenges associated with small or zero denominators, may not be particularly pertinent.

The trend of presenting multiple error metrics in research studies is driven by a pragmatic need. It allows researchers to overcome the inherent limitations of singular metrics and offers a deeper, multi-dimensional evaluation of forecasting models. By presenting multiple metrics, researchers can convey a comprehensive understanding of a model's capabilities, strengths, and potential areas that warrant further refinement. However, while this multi-metric approach can be insightful, it also introduces challenges. The lack of a standardized metric complicates the task of model selection. When different studies propose different models based on different metrics, it makes it challenging for practitioners seeking the most effective forecasting tool. This also impedes direct comparison between studies, making it a challenging task to gauge relative advancements in the field.

Table 2.4 summarizes studies that have used different metrics for evaluating net load forecasting models. The table highlights that different studies use different combinations of error metrics for evaluation. For instance, Kaur et al. (2016) used all listed error metrics, while Stratigakos et al. (2021) only used MAPE. This table serves as a quick reference for researchers to understand which error metrics have been commonly used in the net load forecasting literature and may help in the selection of appropriate error metrics for their own studies. Furthermore, the use of multiple error metrics highlights the absence of a universally accepted evaluation metric in net load forecasting and the field's evolving nature. While metrics like MAPE, MAE, and RMSE dominate the space, the diversity of metrics suggests that researchers are actively seeking the most fitting measure that balances ease of interpretation with robustness against anomalies inherent to net load data. The lack of a unanimous standard highlights the importance for researchers to clearly articulate their choice of metric, ensuring that model evaluations remain transparent and comparisons between studies are contextualized appropriately. Bridging this gap is of utmost importance to further advancing the field, streamlining comparisons across studies, and ultimately enhancing the efficacy of net load forecasting methodologies.

2.3.9 PV Estimation

Integration of a large number of distributed PV systems into the grid presents numerous challenges in terms of proper planning, management, and operations. Many of these challenges are discussed in Section 1.4.2. A major issue remains the metering of PV systems. For larger-scale PV systems, these sites are metered separately, and there is a detailed record of these sites including location, capacity, generation, etc. However, for small-scale PV systems, these systems are typically installed behind the meter and are not metered individually. Thus, the PV generation from these systems is not available individually, and only the net load is available. As residential PV adoption continues to grow rapidly, utilities with high PV penetration are starting to see some of its effects, such as overvoltage and reverse power flow. To effectively tackle these emerging challenges, utilities must acquire a comprehensive understanding of the integrated PV systems, with a particular focus on behind-the-meter adoption.

Although the research community has made considerable efforts to investigate the impacts and associated risks of PV integration into distribution systems, the identification and estimation of residential PV systems remains a relatively new research area that has garnered significant interest in recent years. This section aims to fill this knowledge gap by providing a review of existing literature.

In recent years a popular approach to estimate BTM PV is using satellite images. This method leverages the advancements in satellite imaging technology and computer vision techniques to identify and map PV installations on a large scale. Deep learning models, which have made significant advances in the past decade, are often used in this approach. These models have the ability to automatically learn patterns and features from images without requiring extensive manual feature engineering. Malof et al. (2016) present a Random Forest Classifier based model to detect PV arrays. In Shen et al. (2022), the authors proposed a deep learning model, which they called U-net to detect PV arrays from complex scenes with an accuracy of 99.79%

To develop an effective deep learning model, a substantial collection of image samples is crucial. Deep learning algorithms rely on a large number of labeled samples during the training process to achieve robust generalization capabilities. One such data set was presented by Bradbury et al. (2016). The data contains manually labeled data points from four cities in California. Yu et al. (2018) presented a deep learning based approach called DeepSolar to detect solar sites in the United States. The authors released a publicly available data set containing the locations and sizes of solar installations in the United States for more than 1.4 million installations.

It is important to note that while this approach can provide valuable insights into the distributed PV capacity, it has limitations. Factors such as image quality, cloud cover, tree shadows, or nearby structures can affect the accuracy of the detection. Moreover, satellite imagery may not capture small-scale or rooftop PV installations with high precision. Therefore, ground-based surveys and other data sources are often used in conjunction with satellite imagery to improve accuracy and provide a more comprehensive assessment of PV capacity. In this work, we focus on data-driven approaches for detection and estimation of BTM PV.

Many studies use the changes in PV generation under varying weather circumstances as a means to identify PV systems and estimate their capacity (Li et al., 2019; Zhang and Grijalva, 2016b; Wang et al., 2018a; Zhang and Grijalva, 2016b; Chen and Irwin, 2017). In Li et al. (2019), the authors proposed a methodology that leverages the difference in PV generation under different weather conditions, such as an overcast day versus a sunny day. Two days with significantly differing weather conditions are selected, and the difference between their daily net load curves is used to extract suitable characteristics, which are then used as inputs for an ensemble SVR model to estimate the capacity. The estimated capacity is then multiplied by the output of a known local PV system to obtain the PV output for the individual meter. The proposed approach is tested on data from 300 households, showing promising results. However, the proposed methodology assumes that the actual PV generation output for a small set of meters would be available, which might not be the case in a practical scenario. Zhang and Grijalva (2016b) proposed a three-step methodology for the detection and estimation of residential PV sites. To detect PV sites, the authors present a change point detection model to identify abrupt changes in load behavior due to PV. Then, to verify if the change-point was caused by PV installation, the difference between the load profiles before and after the change-point is then to a local PV system, and the Spearman's rank correlation coefficient is used to verify the identified PV systems. The size of the PV system is estimated using days when there is no cloud cover, i.e., clear sky days by comparing the difference in the typical load profiles of before and after the change-point detection with the local PV profile and estimating the size of the installed PV. The results show significantly better performance in estimating the size of PV when cloud coverage information is included in the model than when it is not. Moreover, the proposed framework assumes load behavior to be unchanged before and after PV installation, which is an impractical assumption. Wang et al. (2018a) present a two-step methodology for the detection and estimation of the capacity of BTM PV. For detecting the meters with PV, a support vector classifier (SVC) is proposed. The SVC model is trained using input features extracted from typical net load profiles for four different weather conditions. The output is a binary label that indicates whether or not a PV installation is present. In addition, a bootstrap support vector regression (SVR) model is proposed to estimate the capacities of the detected PV systems. The proposed approach is implemented on data from 183 households with rooftop PV systems. When compared with the methodology presented in Zhang and Grijalva (2016b), their proposed method displayed superior performance. In Chen and Irwin (2017), the authors proposed a black-box model called SunDance to disaggregate a building's PV generation from net load. The SunDance model includes two modules, first is a clear sky model that uses the historical net load data to model the maximum clear sky PV generation and a second model that maps the effect of weather on the clear sky solar irradiance. The SunDance model is evaluated in 100 buildings, showing results comparable to those of other supervised approaches that use historical PV data for training. The major advantage offered by the SunDance model is that it only requires the location and a minimal amount of historical net meter data.

Several research studies have focused their attention on predicting the aggregated PV generation from invisible solar sites. Shaker et al. (2015) proposed a scaling-up approach to estimate the total generation from invisible PV sites using the measured PV generation data from a small number of representative sites. The authors propose a hybrid K-means approach and a Principal Component Analysis (PCA) based approach to select a set of solar PV sites that represent total generation. Subsequently, a regression-based approach is proposed to map the total power generation of all PV sites using the data from the subset of PV sites selected from the previous step. Expanding on the work in Shaker et al. (2015), Shaker et al. (2016a) proposed another methodology to estimate the aggregated PV generation of invisible sites using a subset of known PV sites. First, the region is divided into smaller subregions and a subset of PV sites within the region is selected using the approach proposed in Shaker et al. (2015). Subsequently, for each subregion the normalized PV generation variation from one site to the other is represented in the form of fuzzy numbers. Finally, a fuzzy number representing uncertainty related to aggregated photovoltaic generation is modeled using the real-time generation of the representative sites and the known capacity of invisible sites. Compared to the work in Shaker et al. (2015), the updated methodology shows inferior performance, however since it does not require historical aggregated PV generation information and can also captures the uncertainties associated with PV generation. Both the works assume access to PV generation data from multiple sites as well as the PV capacity information. Shaker et al. (2020) proposed a framework to estimate the generation from aggregated BTM PV sites using a small subset of known PV sites. A Fuzzy Arithmetic Wavelet Neural Network (FAWNN) model is trained using historical PV generation data from a small subset of representative PV sites identified using the approach presented in Shaker et al. (2015) and numerical weather prediction data to estimate the aggregated BTM PV generation. Mason et al. (2020) proposed a deep neural network model to estimate the PV size, tilt, and azimuth angle for BTM PV installations using historical net load data and a set of data points with known PV size, tilt, and azimuth specifications.

The majority of studies examined in this work rely on data collected from a select group of PV sites for which data is readily available. However, in a practical scenario, this data might not be readily available with the utility. Contrarily, very few studies have solely relied on metered net load data, without necessitating any additional data inputs. In Kabir et al. (2019b), the authors proposed a combination of a physical PV model and a Hidden Markov regression model for load modeling using the 15-minute net load data from the customer. The proposed model shows superior performance compared to the SunDance model proposed in Chen and Irwin (2017). In Stainsby et al. (2020), the authors proposed a methodology to estimate BTM PV generation using data from before the PV was installed. The proposed framework assumes that the customer behavior remains the same after installing PV, which may not be the case. In some cases the customers may change their behavior after installing PV, often termed as 'solar rebound effect' (Qiu et al., 2019)

2.3.10 Research Gaps

Net load forecasting, being a subdomain of energy forecasting, shares some of the traditional research problems that remain unresolved, many of which have been highlighted by Hong et al. (2016). Additionally, there are some distinct research problems that are unique to net load forecasting. In this section, we will outline a few areas of net load forecasting that require further development:

- **Direct vs Indirect approach:** A number of the studies examined in this review address the net load forecasting problem through the indirect approach, which involves breaking down the problem into separate load and PV forecasting tasks. However, in practice, utilities typically have access only to net metering data and lack individual demand and PV generation data. This limitation restricts the applicability of the indirect approach to situations where PV generation and feeder demand are monitored separately. As a result, there is a need to focus more on developing direct net load forecasting methods.

- **Benchmark Data sets:** Over the last several years, global energy forecasting competitions (GEFCOM) have supplied the energy forecasting community with open source benchmark datasets for electricity demand, pricing, solar and wind forecasting, as well as corresponding meteorological data. Since then, several articles have used these data sets to develop and evaluate and compare forecasting algorithms. While there is an availability of open source datasets for net load forecasting, the utilization of these resources in existing literature has been surprisingly limited. Among the few papers that have employed these open datasets, there exists a distinct lack of comparative analysis between different studies using the same dataset. This absence of cross-study comparison has led to a fragmented understanding of the results and their implications. Not only does this create challenges in discerning the relative strengths and weaknesses of different forecasting methods, but it also inhibits the development of standardized benchmarks and best practices within the field. The failure to leverage shared datasets for comparative analysis represents a missed opportunity to foster collaboration, transparency, and consensus in net load forecasting research.

- **Benchmark net load forecasting model:** One of the noticeable gaps in the field of net load forecasting is the absence of a benchmark model. A benchmark model serves as a standard against which other models are evaluated, providing a consistent framework for comparison and validation. The lack of a benchmark model contributes to challenges in assessing the relative effectiveness of different forecasting approaches, as comparisons are often constrained to specific datasets or conditions, limiting the generalizability of findings. The development of a benchmark model for net load forecasting is, therefore, a critical need, as it would facilitate more transparent and consistent evaluations, drive methodological advancements, and enhance the overall coherence and progression of the field.

- **Reproducible research:** Similar to load forecasting, empirical studies are crucial for net load forecasting. While numerous published studies claim to offer novel solutions, not all of them are technically reliable as they lack reproducibility. Researchers must address this issue, by opting to conduct empirical investigations utilizing publicly accessible data and furnish adequate details for others to reproduce their work.

- **Error measures for net load forecasting:** A prominent gap in the field of net load forecasting is the lack of specifically tailored error metrics. Many existing studies employ general error metrics, such as MAPE, RMSE, or MAE. While these metrics are useful and widely adopted, they may not capture certain specific nuances and complexities inherent in net load forecasting. For instance, net load values can be very close to zero for hours with high PV generation, leading to MAPE getting infinite values. The current suite of error metrics also does not provide a standardized approach to compare and evaluate the performance of various forecasting models across different scenarios and datasets. Therefore, there is a need for the development of more refined and context-specific error metrics for net load forecasting that can facilitate a more detailed understanding of model performance and help achieve greater consistency and comparability in evaluations across the field.

- **Leveraging from load and solar forecasting literature:** As net load consists of components of load and renewable energy, it could be beneficial to leverage from the vast literature on load and renewable forecasting to build net load forecasting models. Drawing insights from both load and solar forecasting literature holds considerable potential for enhancing net load forecasting methods. Load forecasting literature provides crucial knowledge about traditional energy demand dynamics, which are impacted by factors such as time, weather, and socio-economic variables. On the other hand, solar forecasting literature offers valuable information about solar power production, which is largely driven by weather variables, particularly GHI. Leveraging from the rich literature these domains can allow us to better understand the complex interplay between consumption and renewable generation and develop more accurate and robust net load forecasting models can be developed.

The results of this study indicate that net load forecasting is a rapidly expanding field, with the number of published articles increasing annually. As PV penetration levels increase, its impact on load becomes more significant, resulting in forecasts from traditional load forecasting models becoming less reliable. Furthermore, with longer available histories of net load data, more researchers are starting to look into the net load forecasting problem.

There are mainly two forecasting approaches used in net load forecasting; direct and indirect. The decision to use one over the other is generally governed by the availability of data. In the case when only the net load data is available, the only option is to use the direct approach while when load and PV data are available independently, either the direct or the indirect approach can be used. Based on the studies reviewed, the two approaches have their own advantages and disadvantages.

In regards to forecasting techniques, the literature review in this report covers a wide variety of techniques used to forecast net load including time series techniques, ANN, SVM, etc. Much like in many other applications, the continuing buzz towards using machine learning techniques is also seen in net load forecasting literature. The large majority of authors devoted their research towards developing and applying machine learning techniques for net load forecasting. The lack of interpretability of most machine learning approaches, on the other hand, has remained a significant shortcoming of these techniques. Machine learning models are still considered as 'black-box' models. Utilities may prefer statistical model due to their transparency and interpretability. Regression-based techniques are rarely used, despite their popularity in load forecasting.

The impact of weather on load, PV generation, and consequently, net load is substantial. As a result, numerous meteorological factors and their effects on net load forecasting models have been examined. Temperature, solar radiation, wind speed, humidity, and cloud cover are some of the most commonly studied weather variables. When considering weather variables, most studies also utilize a variable selection approach to meticulously identify weather factors that can improve the performance of forecasting models.

Finally, the impact of PV penetration on net load forecasting was also observed by many studies. It can be concluded that the impact of PV penetration on net load forecasting is significant. Many studies observed that as the PV penetration in the system increases, it becomes more challenging to forecast net load.

Detection and capacity estimation of PV systems, especially BTM PV installations, presents significant challenges to grid management. While recent research efforts have made strides toward understanding and managing these systems, accurate detection, identification, and capacity estimation of BTM PV installations remain critical areas of ongoing research. Several strategies have been adopted to address these challenges. These include using satellite imagery and machine learning techniques to identify and map PV installations, studying changes in PV generation under varying weather conditions to estimate capacity, and leveraging data from a small subset of representative PV sites to estimate aggregated generation.

However, each approach has its limitations and assumptions. Satellite imagery methods may suffer from image quality issues and may not accurately capture small-scale installations. Weather-based approaches often require the actual PV generation output for a set of meters, which may not be readily available. Approaches focusing on the aggregated generation rely heavily on access to generation data from multiple sites and PV capacity information, which may not always be feasible.

In general, an efficient, robust, and practical approach to detect and estimate BTM PV systems remains a key research priority. The development of novel methods that only rely on readily available data, such as metered net load data, could offer promising solutions to these challenges.

During our review of the existing literature on net load forecasting, we have identified several gaps and avenues for future research. In the following section, we aim to address some of these gaps by introducing a benchmark model for net load forecasting. Our proposed model is developed as a direct model and is based on freely available data sets, which we use as case studies. To develop the model, we have leveraged previous work on load forecasting by using a benchmark load forecasting model (Vanilla model) as the foundation for the benchmark net load forecasting model. Additionally, we showcase the performance of our proposed model for net load forecasting under varying levels of PV penetration, thereby highlighting the impact of PV penetration on net load forecast performance.

Furthermore, we also present a regression-based framework to estimate the PV penetration or the PV capacity in the system using only the historical net load information. We also present a framework to detect meters with BTM PV installations using only the meter-level net load information.

CHAPTER 3: A FRAMEWORK FOR FORECASTING WEATHER-RELATED POWER OUTAGES

[1]

In this chapter, we explore a practical approach to forecasting hourly day-ahead weather-related power outages. We examine the key factors that influence the accuracy of these forecasts and discuss how a weighted logistic regression model can be optimized to address the challenge of data imbalance in outage predictions. This model is applied and evaluated in practical scenarios, specifically focusing on city-level outages and those at the distribution substation level, the latter of which has not been extensively studied in the academic literature.

3.1 Overview of Contributions

Power outages significantly disrupt safety, economic activities, and daily life, necessitating reliable forecasts, especially in the face of severe weather events like hurricanes, tornadoes, snowstorms, and floods. The ability to accurately forecast weather-related outages is crucial for enabling utility companies and governments to make timely preparations that mitigate the adverse effects on infrastructure and communities.

This chapter introduces a methodology for predicting weather-related outages, emphasizing the role of various weather variables in modeling these events. Additionally, it tackles the prevalent issue of data imbalance through the development of a weighted logistic regression model that incorporates interpretable weights. This model is compared against standard logistic regression and an optimized threshold

[1] This chapter is based on the following paper: Sharma, V., Hong, T., Cecchi, V., Hofmann, A., Lee, J. Y. (2023). Forecasting weather related power outages using weighted logistic regression. IET Smart Grid, 6(5), 470-479 (2023). doi: https://doi.org/10.1049/stg2.12109

logistic regression model to demonstrate its efficacy.

The main contributions of this work are as follows:

1. The study presents a reproducible framework for forecasting weather-related power outages on an hourly basis one day ahead. This practical framework is designed to enhance the utility's response capabilities by providing timely and accurate outage predictions.

2. To address data imbalances commonly found in outage data, the proposed framework introduces an innovative approach employing weighted logistic regression. This model includes straightforward yet robust interpretative weights derived directly from the dataset's imbalance, providing a novel solution to a common issue in outage forecasting.

3. A methodical variable selection process that identifies the most impactful weather-related variables for the forecasting model is presented. This contribution enhances the model's accuracy and reliability by focusing on the most significant predictors.

4. The model is applied and tested through two practical case studies: one at the city level and another at the distribution substation level, an area previously less explored in research. These case studies validate the model's effectiveness and adaptability in practical settings, showcasing its broad applicability.

5. Finally, the proposed model is evaluated against standard logistic regression and logistic regression with optimized thresholds. This comparison is crucial for demonstrating the enhanced performance and reliability of the prposed approach in predicting weather-related outages.

These contributions collectively advance the field of energy forecasting and emphasize the applied nature of the work, focusing on developing operational tools and insights that can be directly utilized by utilities to manage and mitigate the impact of weather-related power outages more effectively.

The remainder of the chapter is organized as follows: Section 3.2 describes the outage data used in this study, including exploratory data analysis in Section 3.2.1 and the data preprocessing techniques applied to the data in Section 3.2.2. The proposed forecasting framework is described in Section 3.3. In Section 3.4, the results of the proposed model on out-of-sample test sets are presented. The study is concluded in Section 3.5 with a brief discussion of future research.

3.2 Data Description

The data used for this work comes from a small electric utility located in the United States of America. Often, but not always, electric utilities have software packages known as outage management systems (OMS) that take data from various grid-connected devices, such as smart meters, and use it to record outage events and facilitate restoration through the dispatch of crew or automated switching schemes. OMS assists grid operators by monitoring and displaying outage events and additional information related to those events. OMS is typically integrated with other systems such as Supervisory Control and Data Acquisition (SCADA), Customer Information System (CIS), Geographic Information System (GIS), and Advanced Metering Infrastructure (AMI) (Hong and Hofmann, 2022). In addition, the data gathered by the OMS is essential for establishing reliability indices, which in turn assists engineers in determining which improvements should be prioritized in order to increase the overall reliability of the system.

The OMS data used in this study includes information on outage incidents from January 1, 2013, to December 31, 2019. The data contains critical information on outage events, including the number of customers who were impacted, the cause of the outage, and the amount of time it took to restore service, amongst other things. Additionally, hourly weather data from a corresponding weather station was also available for the study. The weather data used in this work is the actual weather, making this an ex-post forecasting problem.

3.2.1 Exploratory Data Analysis

This section presents an exploratory analysis of the outage data. OMS data may be utilized to extract useful information, to have valuable insights from the data, and to take preventive measures to avoid outages. One of the most critical pieces of information gathered from OMS data is the root cause of the outage. The OMS data that was used in this analysis includes specifics on outages that occurred within the service area of the utility. For each outage event, the utility assigns a cause of the outage. These causes are broadly based on recommendations that are listed in IEEE 1782 (*IEEE Guide for Collecting, Categorizing, and Utilizing Information Related to Electric Power Distribution Interruption Events*, 2014), which is a guide that provides recommendations for collecting, classifying, and utilizing information related to power outages. The causes are organized hierarchically, with four levels. The first degree of causation describes whether or not the power outage was pre-planned, labeling it as unscheduled or scheduled. In the second level, the causes from the first level are broken down further into ten subcategories, which are as follows:

1. Public
2. Natural
3. Equipment
4. Power Supply
5. Utility Human Error
6. Non-Utility Construction
7. Non-Customer Requests

8. Service
9. Maintenance and Repairs
10. Unknown

Once the outage is categorized into one of the major ten categories, the outage is assigned more precise subcategories based on the specific information gathered. The third level of causation differentiates the reasons for outages even further from cause level two. Natural causes, for example, are broken down even further into outages caused by weather, wildlife, and vegetation. At level four, the causes of power outages are broken down into more specific categories. For instance, the causes that are associated with the weather may be further assigned categories such as storm-related, wind-related, ice-related, heat-related outages, and so on. This level of detailed information about the causes of power outages is helpful in determining the reliability of the distribution system for the purpose of presenting it to the general public and commercial customers. It also assists engineers in prioritizing work that improves reliability, such as concentrated tree trimming, upgrades to overhead lines, and so on.

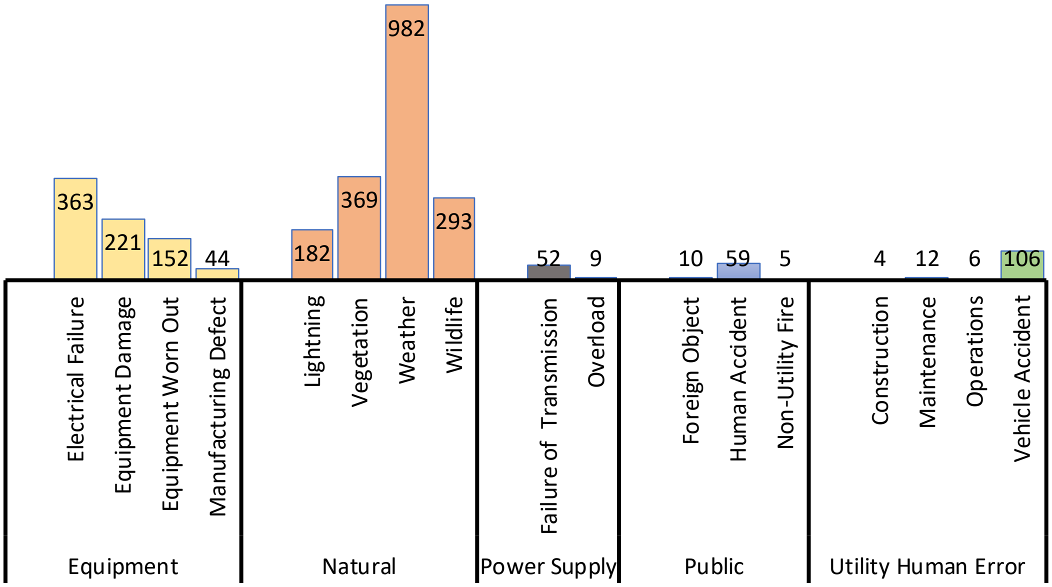


Figure 3.1: Number of unscheduled outages by different root cause levels

Figure 3.1 shows the total number of unscheduled outages by outage cause. It can be observed that most unscheduled outages occur due to natural causes. The primary reason for this is the unpredictable nature of natural events and their impact on the grid. Under natural causes, most number of outages are caused by weather-related events, followed by vegetation, wildlife, and lightning. Lightning strikes could also be classified as weather-related incidents, which would make the number of power outages caused by the weather even higher. After outages caused by natural causes, the next most outages are caused by equipment failure followed by human error, power supply-related and public-related outages. Since the largest number of outages are caused by natural causes, more specifically by weather-related events, this study focuses on forecasting weather-related outages.

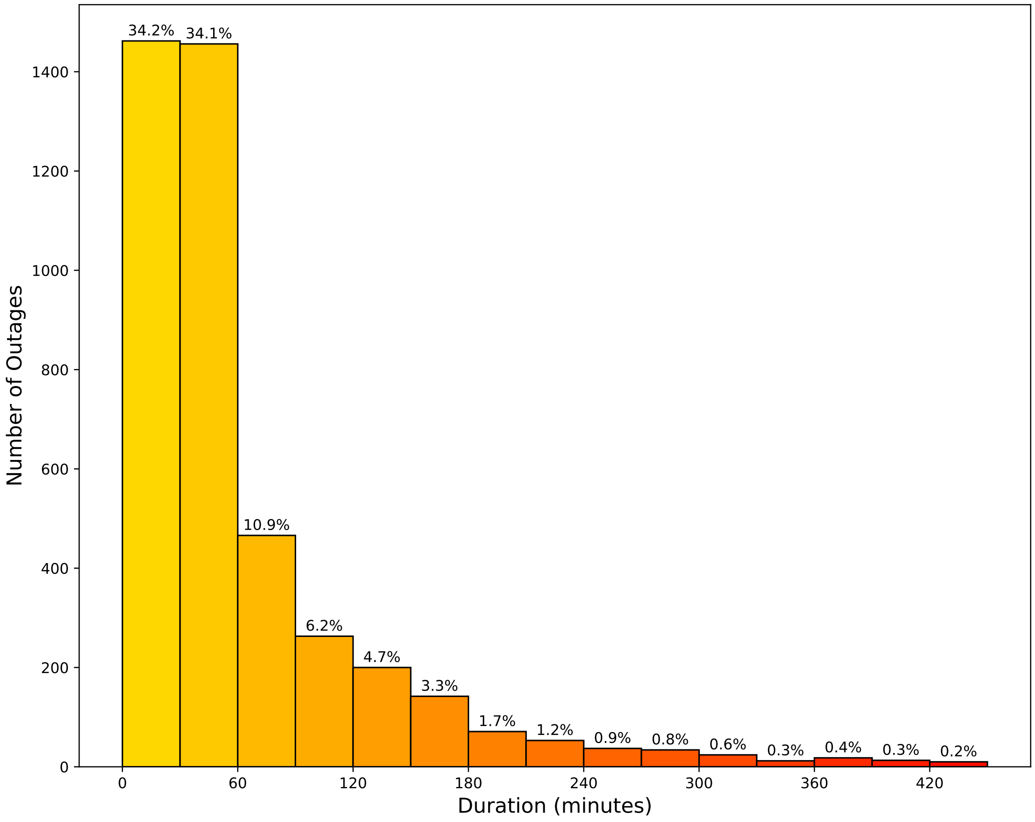


Figure 3.2: Distribution of power outage duration (mins) at the city level

Figure 3.2 shows the distribution of the duration of outages for each outage event.

The figure indicates a distribution that is right-skewed, given the number of outages decreases as the duration increases. Most outages occur within shorter durations, as evidenced by higher percentages at the lower end of the duration scale (34.2% and 34.1% at the beginning). There's a clear decrease in frequency as the duration increases, showcasing fewer long-term outages. This distribution might reflect the utility company's efficiency in addressing most outages quickly, while longer outages, less frequent, could indicate more complex issues or severe conditions requiring extended resolution times.

Forecasting outages at the substation level can offer several advantages to utilities, including the ability to optimize the deployment of repair crews and the strategic allocation of resources, leading to quicker response times and minimized downtime. This granular level of forecasting could also support making more informed decision-making for infrastructure investment, ensuring that enhancements in reliability are equitably distributed across the network. However, this approach is not without challenges. The inherent unpredictability of outages, coupled with their relatively lower frequency at the substation level, results in highly imbalanced data, making accurate predictions difficult. Moreover, the complexity of factors influencing outages at this level, including varying equipment types and local environmental conditions, adds to the forecasting challenges. For this case study, we select five substations within the service area of the utility. Table 3.1 shows the summary of the outage events for different levels. It can be observed that the average outage duration lasts between 62 minutes to around 78 minutes. The challenges associated with forecasting outages at the substation level can also be quantified by the low number of outages at the substation level.

Figure 3.3 shows the distribution of outage duration at the substation level for the five selected substations. Similar to the city level, it can be observed that the distribution of outage duration for each substation is right-skewed. Meaning that most outages occur for a shorter duration.

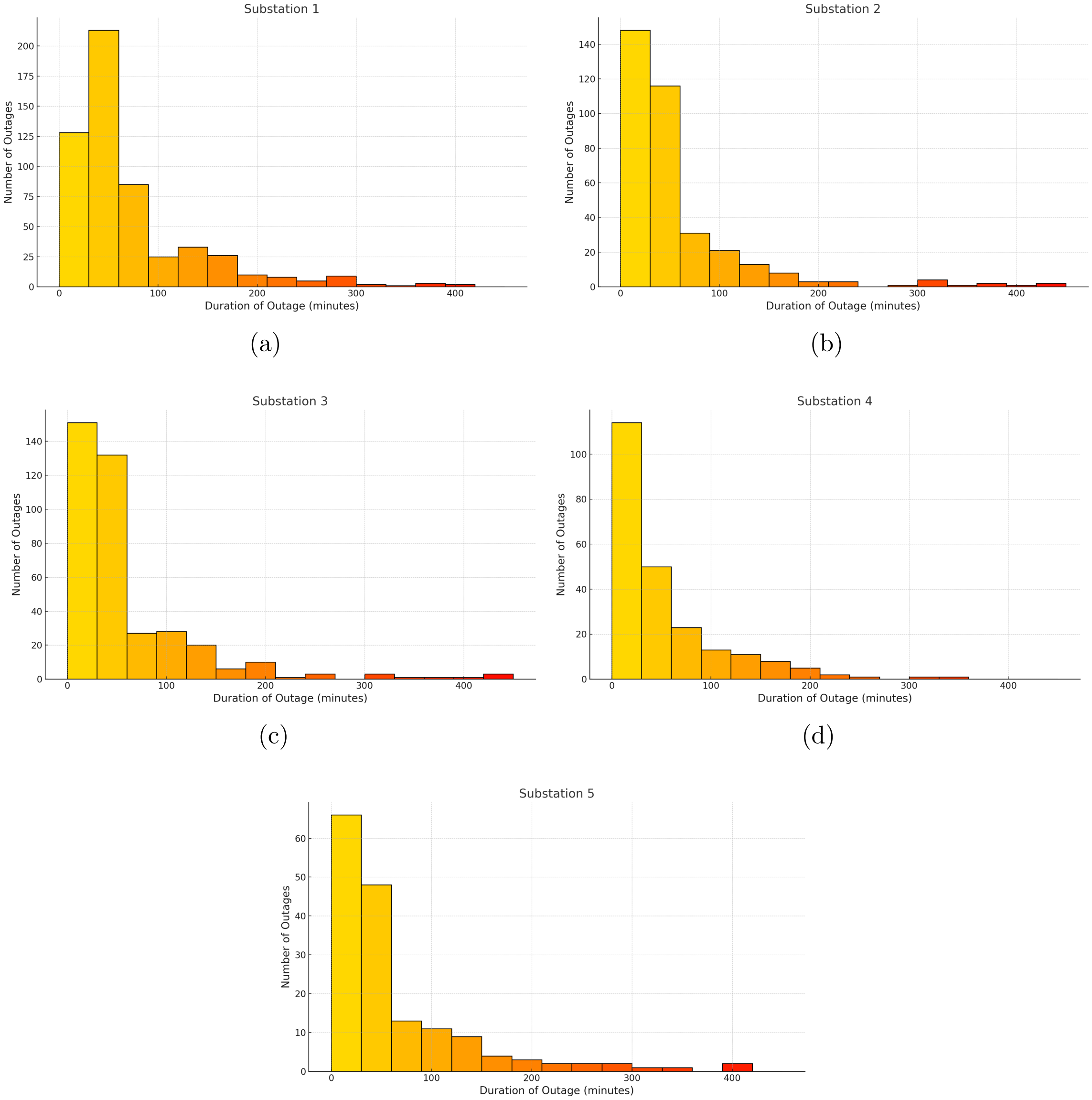
Table 3.1: Summary of outage events for different levels

Level	**Number of Outages**	**Avg. Outage Duration (min)**
City	463	75.47
Substation 1	78	78.85
Substation 2	94	62.06
Substation 3	60	68.66
Substation 4	37	77.78
Substation 5	13	75.38

3.2.2 Outage Data Preprocessing

The OMS data provided by the utility offers information on an individual basis regarding each outage incident. Before we can begin the process of developing the model that will allow us to predict weather-related outages on an hourly basis, the data that comes from the OMS as well as the data that corresponds to the weather need to be transformed. The OMS data, in its most basic form, consists of a list of outage occurrences and the duration of time each outage event lasted. The first step is to transform this into a continuous timeseries. In order to do this, an hourly time series is generated for the corresponding years. Following this, we will need to label the training data in order to develop a classification model that will predict which hours may experience outages. To do this the hours during which there is an outage are labelled as 1, and the hours during which there is no outage are labelled as 0. Following that, we only examine outages caused by natural events that are connected to the weather, i.e. the cause of the outage is weather, vegetation, or lightning. Finally, we get a timeseries where each hour provides information on whether or not there was a power outage in that hour, as well as the corresponding meteorological conditions during that hour.

Cross-validation and model selection are important steps in model building that help mitigate the overfitting and underfitting of models (Arlot and Celisse, 2010). A



(a)

(b)

(c)

(d)

(e)

Figure 3.3: Distribution of outage duration (mins) at the substation level

rolling window-based model selection strategy is utilized in this study. The models are trained using three years of data and are validated over one year. We train the forecasting model with the most recent three years of data and forecast weather-related outages for the day ahead. Then, we move forward one day and re-estimate the model with the latest three years of data and forecast the next day. This process is repeated each day for the validation year. The results of the validation years are averaged using simple averaging, and the model with the best average results is selected. Models are validated using data from 2016, 2017, and 2018, with 2019 serving as an out-of-sample test set. Figure 3.4 shows the model selection strategy deployed in this work.

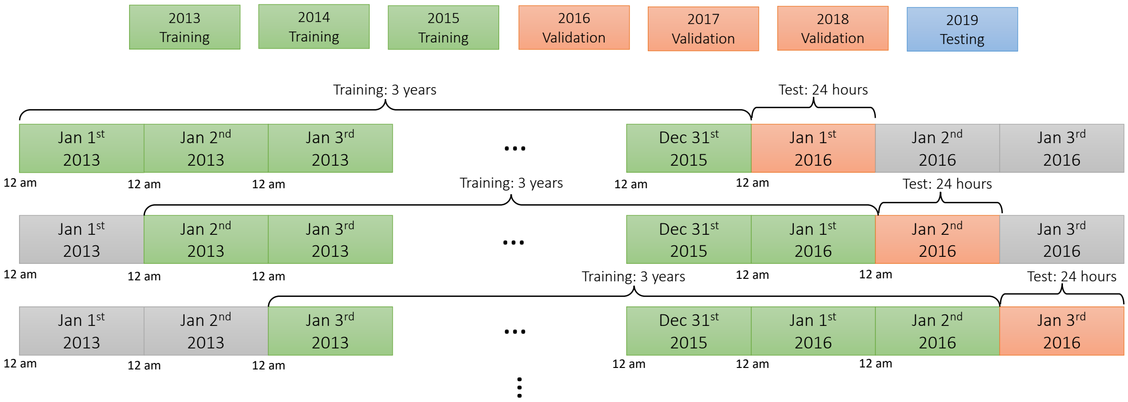


Figure 3.4: Model selection strategy

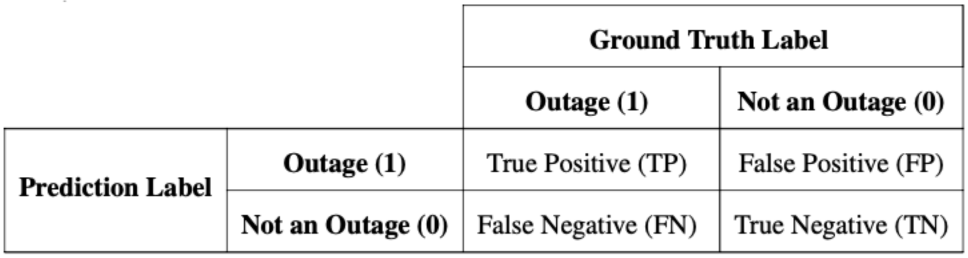
3.3 Proposed Forecasting Framework

In this section, we outline the process of developing an effective outage forecasting solution. We begin by discussing model performance measures, then forecasting techniques, and delve into feature selection methods. Finally, we present a recommended outage forecasting model based on our findings. Each section contributes to enhancing the accuracy and reliability of the forecasting model, crucial for ensuring the resilience of electrical grids.

3.3.1 Model Performance Measures

An important component in the process of model building is selecting an adequate measure to evaluate the performance of the model. This is especially important when dealing with imbalanced data. In classification analysis, a confusion matrix is typically used to evaluate a classifier as it shows a comprehensive overview of the classification model's performance. A typical confusion matrix is shown in Table 3.5, with the columns corresponding to the actual classifications, and the rows corresponding to the classifier's predictions. In Table 3.5, True Positive (TP) is the number of cases for which the positive cases were correctly classified as positive. False Negative (FN) is the number of cases that were actually positive but were incorrectly classified as negatives. False Positive (FP) is the number of cases that were actually negative and are incorrectly identified as positive cases and True Negative (TN) is the number of negative cases that are correctly classified as negative. When modeling unbalanced data, it is common practice to think of the minority class as the "positive class," and the majority class as the "negative class." The confusion matrix can also be used to derive other complex error measurements, such as classification accuracy, F-score, etc. However, when working with data that is not evenly distributed, certain error measures that are commonly utilized might generate conclusions that are not accurate. One example of such an error measure is classification accuracy, given as:

$$\text{Accuracy } = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.1}$$



		Ground Truth Label	
		Outage (1)	**Not an Outage (0)**
Prediction Label	**Outage (1)**	True Positive (TP)	False Positive (FP)
	Not an Outage (0)	False Negative (FN)	True Negative (TN)

Figure 3.5: Confusion matrix for a binary classifier

In a 1:1000 imbalanced dataset, for example, there are 1000 data points in the majority class for 1 data point in the minority class. If a model predicts the majority class for all data points, the model's classification accuracy will be around 99%. Looking at the classification accuracy, we could be persuaded to assume that the model did well when, in reality, it failed to anticipate the minority class. Because of this, classification accuracy is not a useful error statistic to use when working with unbalanced data.

$$\text{Precision } = \frac{TP}{TP + FP} \tag{3.2}$$

$$\text{Recall or True positive rate } (\text{Acc}^{+}) \ = \frac{TP}{TP + FN} \tag{3.3}$$

$$\text{True negative rate } \left(Acc^{-}\right) = \frac{TN}{TN + FP} \tag{3.4}$$

In the context of hourly outage forecasting, key metrics such as precision, recall or true positive rate (TPR or Acc^{+}), and true negative rate (TNR or Acc^{-})) are essential for evaluating the accuracy of prediction models. Precision measures the proportion of correctly predicted outages among all predicted outages, indicating the model's ability to minimize false alarms. Recall, or the true positive rate, assesses the proportion of actual outages that were correctly identified, reflecting the model's effectiveness in capturing all potential disruptions. The true negative rate, on the other hand, indicates the proportion of non-outage situations correctly identified, ensuring that normal operations are not mistakenly flagged as outages.

G-mean or geometric mean provides a single metric that balances the TPR of the model (its ability to detect outages) with its TNR (its ability to recognize non-outage situations). This balance is crucial because focusing solely on maximizing recall might lead to many false alarms (low precision), while maximizing precision could result in missed outages (low recall). Thus, G-mean serves as a comprehensive measure, ensuring that the forecasting model performs well across both dimensions, which is vital for maintaining system reliability and operational efficiency. G-mean evaluates the model's performance across both classes and ensures that the model does not over-fit the majority class and under-fit the minority class. Due to the imbalanced nature of power outage data and the requirement that the minority class of data be well predicted, G-mean is an optimal error metric to use. Hence, in this work, we use G-mean to evaluate the performance of the outage forecasting models. G-mean is defined as:

$$\text{G-mean} = \sqrt{Acc^{+} \cdot Acc^{-}} \tag{3.5}$$

3.3.2 Forecasting Technique

The initial step in creating a forecasting solution involves choosing an appropriate forecasting technique. There are several established methods for classification available. It's important to select the most suitable method considering practical factors such as ease of implementation, computational requirements, and accuracy.

One commonly used method is logistic regression, which falls under generalized linear models. Logistic regression is particularly useful for binary classification tasks, where it estimates the probability of an event happening (King and Zeng, 2001; Kleinbaum et al., 2002). Unlike linear regression, logistic regression transforms its output using a logistic function to predict the probability of the default class. We begin by defining the hypothesis, to approximate y as a function of x:

$$h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_n x_n = \theta^T x \tag{3.6}$$

where θ_i's represent the parameters of the model. Since in logistic regression, we know that $y \in 0, 1$, i.e. y can either be 0 or 1, we can redefine the hypothesis $h_\theta(x)$ in the form of a sigmoid function or a logistic function, given as:

$$h_\theta(x) = \frac{1}{1 + e^{-\theta^T x}} \tag{3.7}$$

Using m training samples, the parameters of the model are estimated via maximum log-likelihood, given as:

$$J(\theta) = \sum_{i=1}^{m} y^{(i)} \log h\left(x^{(i)}\right) + \left(1 - y^{(i)}\right) \log\left(1 - h\left(x^{(i)}\right)\right) \tag{3.8}$$

In the context of outage forecasting, logistic regression is used to predict the likelihood of the default class (outage occurrence), which is bounded between 0 and 1. The dependent variable in this model is the occurrence of an outage, while the independent variables encompass both quantitative factors, such as the month and season, and qualitative factors, including weather conditions and their interactions, that may influence outage occurrences. A critical aspect of this model is the threshold value (T), which establishes the decision boundary for categorizing predicted probabilities. Probabilities that meet or exceed (T) are classified as indicating an outage, whereas lower probabilities suggest no outage. Although the default threshold is often set at 0.5, adjustments can be made to better align with the specific needs of outage forecasting, thereby enhancing the precision of the logistic regression model.

$$\begin{aligned} h_\theta(x) \geq T \rightarrow y = 1 \\ h_\theta(x) < T \rightarrow y = 0 \end{aligned} \tag{3.9}$$

A major challenge encountered in the defined classification problem is the existence of imbalanced distributed classes. Logistic regression is a powerful tool for binary classification problems. However, as shown in Equation 3.8, the default weights allocated to the two classes are equal, implying that the model assigns equal weightage to the two classes. This means that when there is an imbalance in the data, the model will underestimate the probabilities for the minority class (King and Zeng, 2001). For instance, in an electricity grid, outages might be significantly less frequent than non-outages, representing an imbalance. To overcome this issue, the training function used to fit the logistic regression needs to be modified in order to accommodate the imbalance in the data. This adjustment involves controlling how the logistic regression coefficients are updated during training by assigning different weights to each class. Specifically, a lower weight is assigned to the dominant class (non-outages) and a higher weight to the minority class (outages), aiming to balance the relative importance of the two classes. This strategy penalizes the model less for errors in the majority class and more for errors in the minority class. The outcome is a weighted logistic regression model, with a weighted maximum-likelihood estimator to manage this imbalance effectively (Manski and Lerman, 1977). The adjusted cost function for this model is tailored to better account for the specific challenges presented by the imbalance in outage data.

$$J(\theta) = \sum_{i=1}^{m} \omega_0 * y^{(i)} \log h\left(x^{(i)}\right) + \omega_1 * \left(1 - y^{(i)}\right) \log\left(1 - h\left(x^{(i)}\right)\right) \tag{3.10}$$

ω_0 and ω_1 are the weights corresponding to class 0 and 1 respectively. In this work, we propose to calculate the weights ω_0 and ω_1 as the reciprocals of the number of data samples of the respective class. Hence, the class with more data samples, i.e., the non-outage class, will have a lower weight and the class with fewer samples, i.e., the outage class, will have a higher weight. The weights are calculated based on the number of outage and non-outage hours in the training set.

The proposed weights, ω_0 and ω_1 can be calculated as:

$$\omega_0 = \frac{1}{\text{Number of non-outage hours}} \tag{3.11}$$

$$\omega_1 = \frac{1}{\text{Number of outage hours}} \tag{3.12}$$

3.3.3 Feature Selection

Feature selection in forecasting is a critical step in the modeling process where a subset of input variables that are most relevant to predicting the target variable are identified and selected. Since the outages we are predicting are driven by weather, including meteorological information can help build accurate outage forecasting models. However, it is important to select the optimal set of variables in order to maximize the accuracy of the model while also taking into account the ability of the model to generalize, prevent overfitting, and avoid complexity. In this particular case study, the process of feature selection begins with the establishment of eight models, from M_1 to M_8, each with varying sets of inputs.

To assess the performance of the models, we employ a three-fold cross-validation method, where the data from 2012 to 2018 is divided into three parts. For each part, three years of data is used to estimate the model's parameters, with one year serving as validation data for post-sample fit. In each model, we add meteorological variables and their cross-effects one at a time and evaluate whether or not the addition of these factors improves the perfoamnce of the model. Then, we select the terms that exhibit a positive influence on the model. The model equations and the performance metrics of models M_1 to M_8 are listed in Table 3.3 and Table 3.4, respectively.

In our study, we evaluate several metrics for each model: precision, recall, ACC$^+$, ACC$^-$, and notably, the G-mean. These metrics provide detailed insights into the model's ability to classify each category accurately. Precision measures the proportion of correctly predicted positive observations to the total predicted positives. Recall assesses the effectiveness of a model in identifying positive cases. ACC$^+$, the true positive rate, evaluates the model's accuracy in identifying positive instances, while ACC$^-$, the true negative rate, evaluates the accuracy in identifying negative instances. In scenarios of imbalanced classification, ACC$^+$ may hold more relevance than ACC$^-$. Nevertheless, for a comprehensive analysis, it's crucial to consider both precision and recall or ACC$^+$ and ACC$^-$ together. The G-Mean is a combined metric that integrates both ACC$^+$ and ACC$^-$ to provide a balanced overview of the model's performance.

Outage forecasting and load forecasting are quite comparable in many respects. They are often driven by similar qualitative and quantitative variables. Since temperature (Xie and Hong, 2016), relative humidity (Xie et al., 2016a) and wind speed (Xie and Hong, 2017) variables are known to be used in load forecasting, we test them for outage forecasting as well. We start with model M_1, consisting of temperature as the only variable. Next, in model M_2 we use relative humidity as the input variable. We use humidity because periods of high humidity can often lead to rainfall, which has a greater physical impact on electricity lines and may result in more outages. For model M_3, we use wind speed gust as the input variable. Speed gusts are rapid bursts of high-speed winds that last for only a brief period of time. Sudden gusts of high wind speed have the potential to inflict significant damage, which may result in power outages. In addition, strong wind speeds can be the underlying reason for other factors, such as trees toppling onto electricity lines, poles breaking, and other similar occurrences. We use model M3 as the base model and build on it by including more variables.

To further explore speed gust, we introduce to model M_4, the daily mean and the daily maximum of speed gusts, along with the wind speed gust. The addition of the daily average and maximum speed gust could provide the model with additional information regarding the variations in speed gust that take place during the day. Here $f(SG_t)$ is defined as:

$$f\left(SG_t\right) = \alpha_1 SG_t + \alpha_2 \overline{SG}_{t,d} + \alpha_3 SG_{max\,t,d} \tag{3.13}$$

Where SG_t is the value of the speed gust at the t^{th} hour. $\overline{SG}_{t,d}$ is the average value of speed gust on the d^{th} day defined as:

$$\overline{SG}_{t,d} = \frac{1}{24} \sum_{t=24d-23}^{24d} SG_t \tag{3.14}$$

$SG_{max\,t,d}$ is the maximum value of speed gust on the d^{th} day defined as:

$$SG_{max\,t,d} = \max_{24d-23 \leq t \leq 24d} SG_t \tag{3.15}$$

According to the guidelines provided by the National weather service (US Department of Commerce, 2020), wind speed and speed gusts can be classified into a number of distinct categories based on the speeds in miles per hour (mph). Table 3.2 shows the classification of wind speed/speed gust into six distinct categories ranging from light to strong and damaging winds. These categories are included as class variables in model M_5, indicated by dummy variable S_t. In model M_6, we include the interaction of speed gust, its daily mean, and its daily maximum with S_t. We also investigate precipitation and snowfall in models M_7 and M_8 respectively. Precipitation and snowfall, combined with strong wind speeds, can cause trees to fall or electricity lines to shatter, resulting in outages.

Table 3.2: Wind speed/speed gust classification

Wind Speed (mph)	**Descriptive Term**
0-5	Light/ Light and variable wind
5-20	None
15-25	Breezy/ Blustery
20-30	Windy
30-40	Very Windy
40 or greater	Strong/ Dangerous/ Damaging/ High

The addition of precipitation to model M_7 further improves the model's performance. The inclusion of snowfall in model M_8 shows a very minor improvement in the overall performance of the model, specifically in the sixth decimal place. Snowstorms are the primary cause of power outages in many areas during the winter season. As a result, we recommend that if the location is known to receive significant snowfall, incorporating snowfall can potentially help improve the accuracy of the forecasts. Model M_8 may still be utilized in regions with little to no snowfall as including the snowfall variable has no negative effect on the model's overall performance. Furthermore, since the objective of this work is to present a generalized model, we include snowfall in the recommended model.

Examining the performance metrics for eight models, M_1 through M_8, we observe a high level of precision across the board, with values ranging narrowly between 0.9669 and 0.9721. This suggests that all models are quite adept at correctly identifying positive instances. However, when it comes to recall, there is a notable divergence among the models. Models M_1 and M_2 exhibit significantly lower recall values at approximately 0.436. This shows that temperature and humidity alone are not ideal in modeling outages. Inclusion of speed-gust in model M_3 stands out with the highest recall of 0.746, indicating its superior capability in detecting positive cases.

Moving on to ACC$^+$, which represents the accuracy with which models predict positive cases, we again see M_3-M_8 show high performance. In contrast, the ACC$^-$ values, indicating each model's accuracy in predicting negative cases, are inversely related, with models M_3-M_8 achieving lower ACC$^-$ scores. This suggests a trade-off in performance where models with higher recall and ACC^+ may have lower accuracy for negative cases.

The G-mean metric, which combines both positive and negative accuracy to provide a balanced measure of performance, shows that Models M_6, M_7, and M_8 are the most balanced, each achieving a G-mean of approximately 0.6007 or 0.6008. Despite not having the highest individual scores in recall or ACC^+, these models offer a more evenly distributed performance across classes. Models M_4 and M_5 display consistent scores across all metrics, indicating a stable performance regardless of the class being predicted. Overall, while M_3 excels in identifying positive cases, Models M_6, M_7, and M_8 present the most balanced performance, which may be preferred in scenarios where treating both classes equally is essential. The additive contribution from incorporating S_t in model M_5 on its own is minor, when we include it together with its interaction with speed gust, its daily mean, and its daily maximum in model M_6, we observe an improvement over model M_5, resulting from the related cross effects. The addition of precipitation to model M_7 further improves the model's performance. Furthermore, the inclusion of snowfall in model M_8 shows a very minor improvement in the overall performance of the model, specifically in the sixth decimal place. Snowstorms are the primary cause of power outages in many areas during the winter season. As a result, we recommend that if the location is known to receive significant snowfall, incorporating snowfall can potentially help improve the accuracy of the forecasts. Model M_8 may still be utilized in regions with little to no snowfall as including the snowfall variable has no negative effect on the model's overall performance. Furthermore, since the objective of this work is to present a generalized model, we include snowfall in the recommended model.

Table 3.3: Model inputs for M_1-M_8

Model	Model Equation
M1	$y_t = \beta_0 + \beta_1 Temperature_t$
M2	$y_t = \beta_0 + \beta_1 Humidity_t$
M3	$y_t = \beta_0 + \beta_1 SG_t$
M4	$y_t = \beta_0 + f(SG_t) \quad 2$
M5	$y_t = \beta_0 + f(SG_t) + \beta_1 S_t$
M6	$y_t = \beta_0 + f(SG_t) + \beta_1 S_t + f(SG_t) \times S_t$
M7	$y_t = \beta_0 + f(SG_t) + \beta_1 S_t + f(SG_t) \times S_t + \beta_2 Precipitation_t$
M8	$y_t = \beta_0 + f(SG_t) + \beta_1 S_t + f(SG_t) \times S_t + \beta_2 Precipitation_t + \beta_3 SnowFall_t$

Table 3.4: Performance of models M_1-M_8 for the validation set

Model	Precision	Recall	ACC$^+$	ACC$^-$	G-mean
M1	0.9721	0.438	0.432	0.774	0.5433
M2	0.9669	0.436	0.431	0.659	0.5515
M3	0.9689	0.746	0.751	0.48	0.5936
M4	0.9697	0.703	0.706	0.541	0.6002
M5	0.9697	0.703	0.706	0.541	0.6002
M6	0.9699	0.723	0.727	0.533	0.6007
M7	0.9699	0.724	0.727	0.533	0.6008
M8	0.9699	0.724	0.727	0.533	0.6008

3.3.4 Recommended Outage-Forecasting Model

Based on the improvement in accuracy and the model complexity, we recommend the variables defined in model $M8$ as the outage forecasting model.

$$y_t = \beta_0 + f\left(SG_t\right) + \beta_1 S_t + f\left(SG_t\right) \times S_t + \beta_2 \text{ Precipitation }_t + \beta_3 \text{ SnowFall }_t$$

3.4 Results and Discussion

3.4.1 Out-Of-Sample Test

In order to assess how well the model performs, the proposed model (M_8) is implemented on an out-of-sample test set. The data for the year 2019 is withheld from the model selection process and is used as the out-of-sample test set. To further evaluate the performance of the proposed weighted logistic regression model, it is compared to two other models. The first model is a simple logistic regression model, while the second is a logistic regression model with an optimized threshold value (T). The logistic regression model with an optimized threshold value has earlier been proposed in (Xu and Chow, 2006, 2005) for power distribution fault cause identification with imbalanced data. Tuning the threshold value to an optimum value rather than using the fixed value of 0.5 can compensate for the imbalance in the data (Xu and Chow, 2006). To obtain the optimal value for threshold we perform a grid search by exhaustively considering values ranging from 0.01 to 0.99 with a step size of 0.01, and then we select the value that produces the highest G-mean score in the validation set. The optimal threshold value obtained from the validation set is then used as the threshold value for the out-of-sample test set.

The G-mean scores for the three models on the out-of-sample test set are presented in Table 3.5. From the results, it is clear that the simple logistic regression model performs the worst, with the lowest G-mean score. Since the simple logistic regression model was designed with the assumption of balanced data, it gives equal weightage to learning both classes. However, when the data is imbalanced, the number of data points from the majority class is significantly more than the minority class. As a result, the model learns just the features of the majority class while disregarding the characteristics of the minority class, i.e., the hours when there was an outage. This is seen in the low G-mean score achieved by the simple logistic regression model. The logistic regression model with an optimized threshold performs much better than the simple logistic regression model. This demonstrates how the additional step of fine-tuning the threshold value is able to cope with data imbalance and improve performance over the simple logistic regression model. The proposed weighted logistic regression model outperforms the other two models in terms of G-mean score, demonstrating the efficacy of the proposed methodology. The proposed reciprocal weighting method efficiently addresses the imbalance in the data. This enables the model to learn the characteristics of the minority class on power with the majority class. Fur-

Table 3.5: G-mean of the proposed model on the out-of-sample-test-set at the utility level

Model	**G-mean**
Simple Logistic Regression Model	0.2415
Logistic Regression Model w/ Optimized Threshold	0.7197
Proposed Weighted Logistic Regression Model	0.7328

thermore, the proposed model is also computationally less expensive since it does away with the need for additional threshold optimization stages. Overall, the proposed weighted logistic regression model using the variables indicated in model M_8 outperforms its counterparts and demonstrates that it is a reliable model to forecast weather-related outages.

3.4.2 Applying The Proposed Model To Substation Level

The proposed model was developed and implemented to predict outages on a utility-wide level. However, the ability to forecast outages at a more granular level, such as the substation level, provides additional value to utilities seeking to better position crews and ensure equitable system-wide investment in reliability in light of the rising frequency of extreme weather conditions. Forecasting outages at the substation level with minimal data can be very challenging due to the high degree of outage event stochasticity. Additionally, there are fewer power outages at the substation than there are at the utility level, which results in more imbalanced data. Many challenges associated with forecasting outages at the substation level are highlighted in Doostan and Chowdhury (2020). For this case study, we select five substations within the service area of the utility implement the proposed model M_8, and observe its performance. The results of this analysis are presented in Table 3.6.

At the substation level, the proposed weighted logistic regression model outperforms both the simple logistic regression model and the logistic regression model with an optimized threshold value. These findings are consistent with those that were observed at the city level. This is particularly noteworthy considering the intricate

Table 3.6: G-mean of the proposed model on the out-of-sample-test-set at the substation level

Model	Substation 1	Substation 2	Substation 3	Substation 4	Substation 5
Simple Logistic Regression Model	0.2	0.2439	0.199	0.2132	0
Logistic Regression Model w/ Optimized Threshold	0.7233	0.7083	0.5854	0.4763	0.4079
Proposed Weighted Logistic Regression Model	0.7429	0.715	0.6933	0.6259	0.5396

nature of the problem. Among the five substations, substation 1 shows better G-mean performance compared to that at the utility level. One explanation for this could be that substation 1 consists of the highest number of outage events among the selected substations and has the highest average outage duration making it more predictable. For substation 5, the standard logistic regression model gives a G-mean of 0, as the model fails to predict any outages due to an extremely high imbalance at the substation. However, the proposed weighted logistic regression model can compensate for the imbalance in the data. Overall, the proposed weighted logistic regression model with the parameters described in model M_8 can be used by utilities to forecast weather-related outages not only at the city-wide level but also at the substation level.

3.5 Conclusion

Outages have a significant impact on both individual customers and power companies, frequently resulting in severe economic and social disruptions. Effectively forecasting weather-related outages can help power system planners and operators improve reliability and resilience. Accurate forecasting models may be used to evaluate which areas or utility systems are more vulnerable to new and more extreme weather patterns, offering another aspect for prioritizing resilience-related maintenance and upgrades. Similarly, such models may be used to forecast when expected weather would have the most impact on power outages, assisting in finding the best moments to engage mutual aid and bring in outside resources for assistance. A high-quality location-specific outage prediction model can be used to improve the equity of reliability investments across a utility system by identifying excessive impacts on under-served communities far enough in advance that reliability and resiliency investments can be made accordingly.

This work presents a weighted logistic regression model to forecast weather-related power outages one day ahead on an hourly basis. We address the inherent data imbalance issue by proposing a weighted logistic regression model and allocating different weights for the outage and non-outage classes based on the reciprocals of their respective number of hours. The best-performing inputs for the forecasting model are chosen using a variable selection technique. The proposed model is used to forecast weather-related outages aggregated from the distribution substation level up to the city level. The out-of-sample tests showed that for both cases, the proposed model outperforms a simple logistic regression model and a logistic regression model with an optimized threshold value. Using the methodology described in this work as a foundation, researchers may expand the investigations to several directions, such as robustness of models, and forecasting with high-resolution data.

This work also opens up various new research avenues for future researchers. The current work offers an ex-post forecasting model. In future studies, ex-ante forecasts can be explored. An area of interest for future analytical research may be to investigate how lagged and moving averages of meteorological data can enhance outage forecasting models, similar to what is proposed in Wang et al. (2016) for load forecasting. Additionally, the climatic conditions in a certain place may differ from those in the surrounding areas. Choosing the best weather stations for each location has the potential to give more accurate and relevant weather information, as well as improve the forecast model's performance. A weather station selection process, as proposed by Hong et al. (2015), can be used to complement the current work in future research to enhance the performance of the models. In some instances, based on real-world occurrences, certain historical outage incidents could be misclassified. An examination of the robustness of the model in the face of varying degrees of misclassification might be an intriguing issue for researchers to investigate in further studies. Furthermore, the impact of data integrity attacks on OMS systems is highlighted and discussed in Hong and Hofmann (2022). A future direction of work could also be to analyze the robustness of outage forecasting models under data integrity attacks and develop more robust and reliable models. Another interesting area of future work can also be to forecast the duration of the outages. Furthermore, the current work can serve as a strong foundation for future research into forecasting weather-related outages not only on an hourly but also on a sub-hourly basis.

CHAPTER 4: A BENCHMARK MODEL FOR NET LOAD FORECASTING

This chapter outlines a framework for creating a benchmark model for net load forecasting. The framework employs multiple linear regression and includes an extensive feature engineering process to choose the best features for the model. Additionally, the framework takes into account different scenarios with varying levels of solar penetration.

4.1 Overview of Contributions

During the review of the existing literature on net load forecasting, we have identified several gaps and avenues for future research. In this chapter, we aim to address some of these gaps by introducing a benchmark model for net load forecasting, which is notably absent in current studies. Our proposed model is developed as a direct model and is based on a publicly available data set, which we use as a case study. To develop the model, we have leveraged previous work on load forecasting by using a benchmark load forecasting model as the foundation for the benchmark net load forecasting model. We extend this model with a systematic approach to including GHI variables, resulting in the recommendation of a benchmark model for net load forecasting with a group of GHI and temperature variables. Additionally, we showcase the performance of our proposed model for net load forecasting under varying levels of PV penetration, thereby highlighting the impact of PV penetration on net load forecast performance.

The principal contributions of this study are summarized as follows:

1. The study presents a benchmarking framework for net load forecasting, which establishes a standard reference for future research in this area. This model

fills a notable gap in the literature, providing a foundation that enhances both theoretical and applied research in net load forecasting.

2. By incorporating GHI and temperature variables into an established load forecasting model, we significantly enhance its relevance and utility for net load forecasting. This approach introduces a novel way of integrating critical weather variables, improving the accuracy and practical applicability of the model for electric utilities.

3. The modelâs performance is thoroughly assessed across different levels of PV penetration, demonstrating its robustness and adaptability. This evaluation offers valuable insights into the modelâs utility in various scenarios of renewable energy integration, highlighting its practical benefits for energy system management.

4. The use of a publicly available dataset for developing and validating the model ensures that our research is reproducible and accessible. This practice supports transparency and encourages further investigation by other researchers, strengthening the reliability and practical relevance of our findings.

These contributions significantly advance the field of energy forecasting by providing electric utilities with a more effective and practical approach to net load forecasting. The integration of environmental factors and the thorough testing across various scenarios highlight the innovative and applied nature of our work, setting a new standard for future research in energy system management.

The remainder of the chapter is organized as follows: Section 4.2 describes the data used in this study, including data preprocessing and exploratory data analysis. The proposed forecasting framework is described in Section 4.3. In Section 4.3.5, the results of the proposed model on out-of-sample test sets are presented. Section 4.4 presents a framework for including recency effect on the benchmark net load forecasting model. The study is concluded in Section 4.5 with a brief discussion of future research.

4.2 Data Description

The data used for this work comes from three major Independent System Operators (ISO) in the United States, namely the Independent System Operator of New England (ISONE), the Electric Reliability Council of Texas (ERCOT), and the California Independent System Operator (CAISO). The data covers a substantial period of time and the three IOSs are located in different parts of the United States, providing a comprehensive overview of electricity demand and supply dynamics across these three regions. Figure 4.1 shows the three IOSs used in the case study. Table 4.2 shows the summary of the data from the three utilities. For ease of presentation, the analysis from ISONE is presented in the main text of the dissertation, while the analysis from CAISO and ERCOT are presented in A and B, respectively.

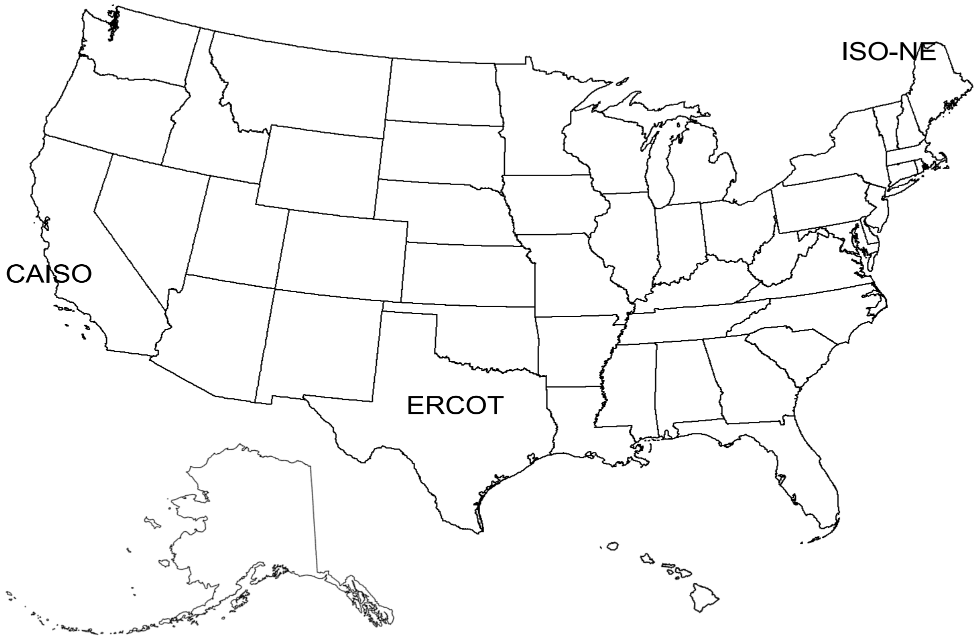
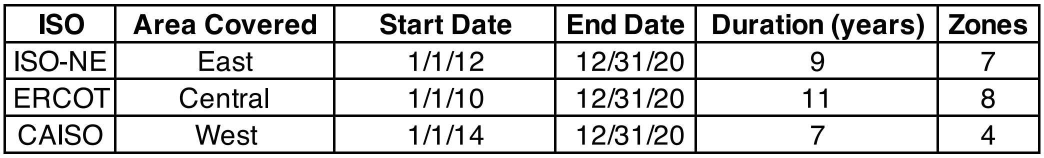


Figure 4.1: Map showing the locations of the three IOSs used in this study

Table 4.1: Summary of the data from the three case studies



ISO	Area Covered	Start Date	End Date	Duration (years)	Zones
ISO-NE	East	1/1/12	12/31/20	9	7
ERCOT	Central	1/1/10	12/31/20	11	8
CAISO	West	1/1/14	12/31/20	7	4

4.2.1 ISO New England (ISONE)

ISONE is an independent regional transmission organization responsible for operating the power grid and wholesale electricity markets covering the six states in the northeastern region of the United States, namely Connecticut (CT), Maine (ME), Massachusetts (MA), New Hampshire (NH), Rhode Island (RI), and Vermont (VT). ISONE divides its territory into three levels, namely top, middle and bottom level. The top level consists of the aggregated system load for ISONE. The middle level consists of six states covered by ISONE, and the bottom level consists of eight load zones based on the geographic location and characteristics of the electricity load in each zone. The five states form their own load zone, while additionally, Massachusetts is divided into three load zones, namely, NEMASS, SEMASS and WCMASS. The load data for each of the load zones is freely available on ISONE's website. Starting 2023, along with hourly load data, ISONE also released estimates of total hourly production from behind-the-meter PV installations for each of their eight wholesale load zones as well as the total system BTM generation. Along with the BTM PV power production data, ISONE also released the normalized BTM PV data, which is the ratio of estimated hourly BTM PV power production relative to the total installed capacity of BTM PV for each zone. The weather data including temperature and GHI were obtained from the National Solar Radiation Database (NSRDB), which is a freely available database of solar radiation and related weather data for locations across the United States, developed and maintained by the National Renewable Energy Laboratory (NREL).

To obtain the weather data for each load zone, we use the data from the nearest major airport. For the total system load of ISONE, the average weather data from all the eight load zones is used. Table 4.2 presents the summary statistics of the load, PV and weather data for 2014-2019. The total data for this study ranges from 1st January 2014 to 31st December 2019.

Table 4.2: Summary data on load, temperature, and GHI for ISONE load zones (2014-2019)

Zone	Weather Station	Lat/Lon	Load (MW)		Temperature (°F)		GHI (W/m^2)	
			Mean	STD.	Mean	STD.	Mean	STD.
NEMA	KBOS	42.36, -71.01	2815	563	49.97	49.37	154.90	241.68
VT	KBTV	44.47, -73.15	623	111	45.07	52.94	154.90	241.68
NH	KCON	43.2, -71.5	1316	263	47.20	51.54	164.08	247.29
ME	KPWM	43.64, -70.3	1306	207	47.25	48.86	165.33	250.03
RI	KPVD	41.72, -71.43	913	204	51.01	49.43	168.28	253.27
SEMA	KPVD	41.72, -71.43	1652	385	51.01	49.43	168.28	253.27
CT	KBDL	41.94, -72.68	3391	760	50.06	51.46	166.49	251.38
WCMA	KORH	42.27, -71.87	1911	370	47.20	51.40	166.49	251.38
ISONE	N/A	N/A	13928	2788	48.25	50.57	170.05	254.87

For this work, we use ISONE as the primary case study to conduct initial experiments and validate our proposed model. We use the total system load from 2014-2018 for validation and model selection, while we use the total system load and the data from 2019 for all of ISONE's load zones for out-of-sample testing.

4.2.2 Data Pre-processing

As part of pre-processing, the data is adjusted for daylight savings time (DST). At the beginning of the DST, we fill in the missing hour (3:00 AM) by taking the average of the values at 2:00 AM and HE 4:00 AM. At the end of the DST, we take the average of the duplicate values at HE 1:00 AM.

As reviewed in Section 2.3.7, as the penetration of PV increases, forecasting the net load becomes more challenging due to the added variability of PV generation. For this work we define PV penetration as the ratio of annual peak PV generation to annual peak load. To develop a robust benchmark net load forecasting model, we scale the normalized BTM PV to obtain net load time series with varying PV penetration levels. The following equation is used to calculate the net load.

$$Net\ load = Load - \alpha * Normalized\, BTM\ PV \tag{4.1}$$

The value of α is varied such that we get net load time series with 5%, 10%, 15%, 20%, 25%, and 30% PV penetration, where PV penetration is defined as:

$$PV penetration = \frac{Annual\, peak\ PV\ generation}{Annual\, peak\ load} \tag{4.2}$$

4.2.3 Exploratory Data Analytics

Figure 4.2 and Figure 4.3 show the typical seasonal patterns of net load, temperature, and GHI. The seasonal cycles of temperature and GHI are quite similar, with higher values in summer and lower values in winter. This can be attributed to the fact that the months of May through August receive a substantial amount of sunlight, causing the temperature to increase significantly during this period. On the contrary, the winter months receive less GHI from the Sun, resulting in colder temperatures. This seasonal fluctuation is also evident in load and net load, with the former being influenced by temperature and the latter being influenced by temperature and PV generation, which in turn depends on the GHI.

Figure 4.2 displays the average daily GHI by month, highlighting a discernible diurnal pattern. The GHI peaks around noon, with the highest values between 8 am and 4 pm. Additionally, GHI levels in the summer months (April-August) are considerably higher than those received during winter months (November-January), while transition months (February and October) experience intermediate GHI levels. In Figure 4.3, we observe the average daily load and net load plots by month. The load curves for winter months demonstrate a morning and an evening peak, while summer months exhibit a solitary, yet sizable, evening peak. On the contrary, the

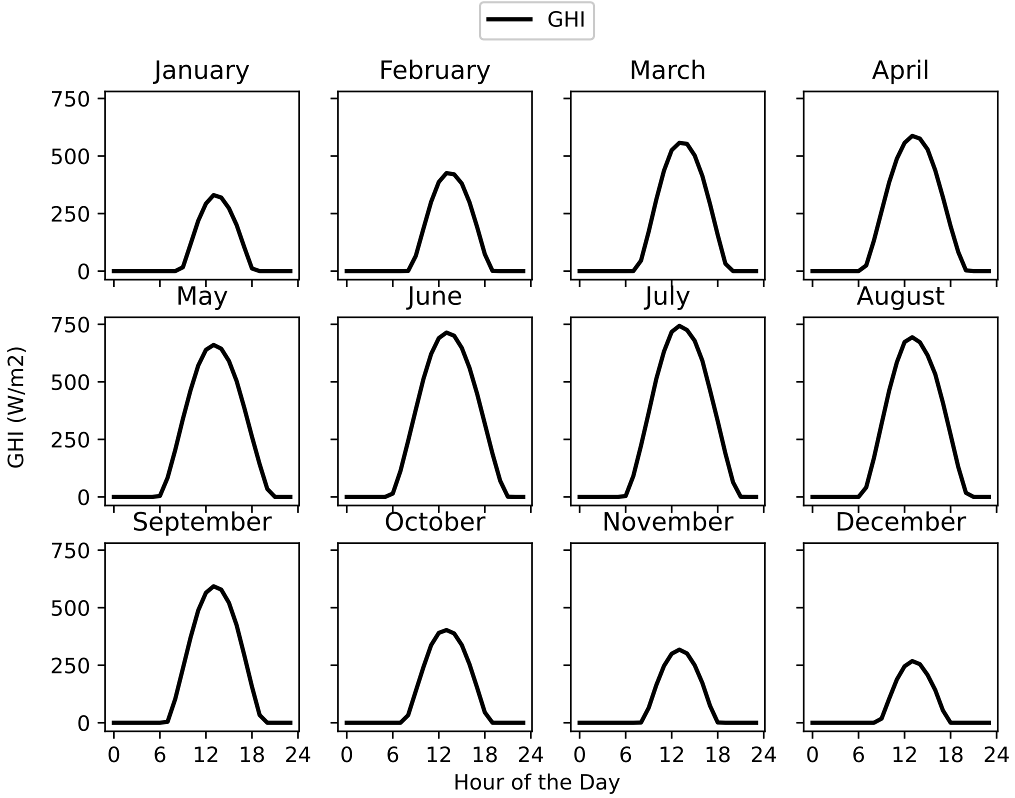


Figure 4.2: Average hourly GHI profiles by month of the year

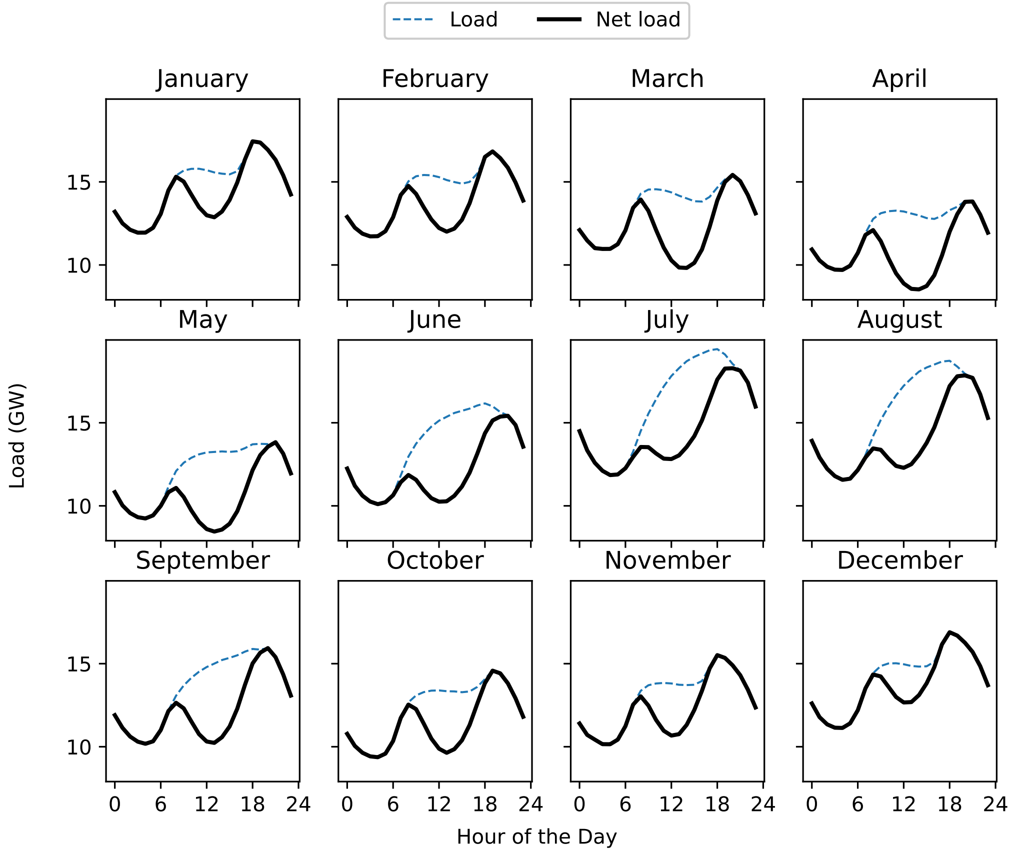


Figure 4.3: Average hourly load and net load profiles by month of the year

net load curves show a divergent daily pattern. During the early morning and late evening hours, the net load resembles the load; however, as PV generation increases with sunrise, the net load decreases, resulting in rapid ramping. This daily reduction in net load is regulated by PV generation which, in turn, is governed by GHI.

In conclusion, the observed seasonality in the net load presents an opportunity for further investigation to enhance the development of net load forecasting models. It is crucial to conduct additional research on the relationship between GHI and net load to develop more efficient and effective net load forecasting models.

4.3 Proposed Forecasting Framework

In this section, we propose a framework for net load forecasting, crucial for optimizing power system operations with rising renewable energy integration. We discuss the evaluation metrics used to assess model performance, delve into the application of the Multiple Linear Regression technique, and emphasize the significance of feature selection, particularly focusing on weather variables, specifically solar irradiance. Furthermore, we compare the proposed model's performance against a load benchmark to showcase its efficacy in improving net load forecast accuracy, especially in contexts characterized by substantial solar power integration.

4.3.1 Model Performance Measures

To evaluate the performance of a forecast, we need to compare the forecast values with the actual values and analyze the error in forecast. The forecast error tells us what portion of the data remained unpredictable. Forecast error can be written as:

$$Error_t = Actual_t - Prediction_t \tag{4.3}$$

Forecast error may be represented in a variety of ways to obtain additional forecast accuracy measures.

Scale-dependent measures summarize errors at the scale of the data. Due to being scale-dependent, these error measures are used to compare different forecasting models on the same data, and should not be used to compare forecasts across different data sets. A popular scale-dependent error measure is the root mean square error (RMSE). A common way of utilizing scale-dependent measures is to normalize them for example, using Normalized RMSE (nRMSE) instead of RMSE. nRMSE is defined as:

$$\text{nRMSE} = \frac{\sqrt{\frac{\sum_{i=1}^{N}(Error_i)^2}{N}}}{A_{\max} - A_{\min}} \tag{4.4}$$

The percentage error measures or scale-independent measures are obtained by dividing the error by the actual value, i.e. $\frac{Actual_t - Prediction_t}{Actual_t}$. One of the most commonly used percentage measures is mean absolute percentage error (MAPE), given as:

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{Error_t}{Actual_t}\right| \tag{4.5}$$

Because of being scale-free, MAPE can be used to compare the forecasts from different forecasting models on the same data as well as across different data sets. Furthermore, being a percentage measure makes MAPE easily interpretable without having to understand the scale of the data. However, there can be some downsides to using MAPE. MAPE can be infinite or undefined in cases when the $Actual_t = 0$. Additionally, MAPE puts a heavier penalty on negative errors as compared to positive errors.

4.3.2 Forecasting Technique

Multiple Linear Regression is a statistical technique that formulates a relationship between independent variables and a dependent variable and uses this relationship to predict the dependent variable. A comprehensive overview of MLR can be found in Rawlings et al. (1998).

A general regression model is given as:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{k-1} X_{i,k-1} + e_i \tag{4.6}$$

where $\beta_0 \cdots \beta_{k-1}$ are the regression parameters to be estimated, $X_{i1} \cdots X_{i,k-1}$ are the known variables and e_i is the error term. The error has constant variance and zero mean. In this work, the dependent variable is hourly net load and the independent variables include quantitative, qualitative, and interaction variables. Temperature is an example of a quantitative variable. The demand for electricity might increase as the temperature rises as people tend to turn on their air conditioners, fans, etc. for cooling. Furthermore, the temperature might not have a linear relationship with the load. As a result, a higher-level polynomial of temperature can be included in the model as a quantitative variable.

Qualitative variables or class variables such as month of the year, day of the week, hour of the day, etc can also be included in the model. These qualitative variables are included in the model as dummy variables. Values of 0 and 1 can indicate the class of a quantitative variable. Interaction terms, which are the multiplication of two or more independent variables can also be included in the model when one predictor variable depends on some other predictor variable.

4.3.3 Feature Selection

Weather is a key driving factor of net load, both in terms of electricity demand as well as PV generation. Temperature is known to impact electricity demand and is the most commonly used weather variable in load forecasting models. On the other hand, irradiance affects the generation of solar power. Solar irradiance is defined as the amount of sunlight falling on a square meter of area per second Garner and Dunbar (2008). There are three types of solar irradiance variables, namely, Direct Normal Irradiance (DNI), Diffuse Horizontal Irradiance (DHI), and Global Horizontal

Irradiance (GHI).

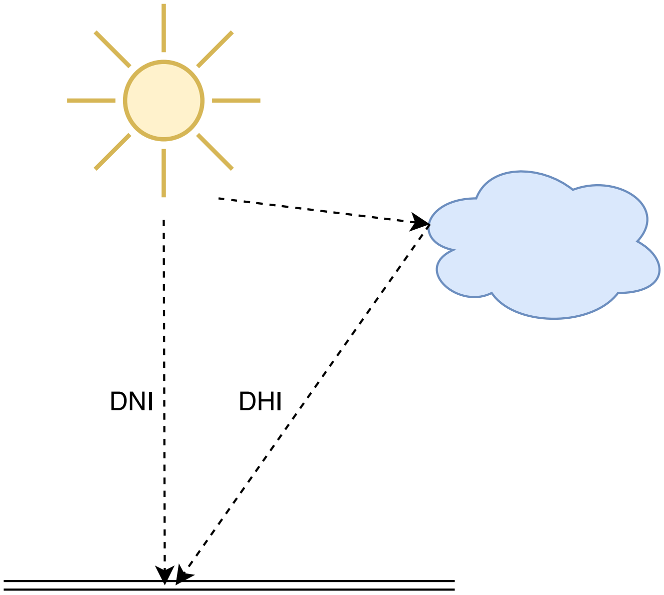


Figure 4.4: Visual representation of DNI and DHI.

DNI is the amount of direct solar radiation falling on an object or a surface per unit area normally Mousavi Maleki et al. (2017). In other words, DNI is the light falling directly perpendicular from the Sun on any surface. This is represented in Figure 4.4. The rays falling in a straight line directly from the Sun to the incident surface represent DNI.

DHI is the amount of solar radiation hitting an object or a surface per unit area not directly from the Sun. In other words, it is the amount of radiation incident on a surface after light has been emitted Mousavi Maleki et al. (2017). This is shown in Figure 4.4 as the irradiance that falls on the surface after scattering from the clouds.

GHI or total radiation is a combination of DHI and DNI Mousavi Maleki et al. (2017). It is the total amount of solar radiation incident on an object or a surface. GHI can be defined as:

$$\mathrm{GHI} = \mathrm{DHI} + \mathrm{DNI} \cdot \cos(\mathrm{Z}) \tag{4.7}$$

where Z is the zenith angle. GHI is the most important type of irradiance as it is a combination of all others and is responsible for PV generation. There are many instruments available to measure GHI. The most common of these instruments is the pyranometer. In some cases when GHI cannot be measured, it can also be calculated form DNI and DHI. There are significantly more weather stations that measure GHI as compared to DHI and DNI, thus we use GHI to make the work easily and more widely adoptable (Vignola, 2012).

High GHI levels typically correlate with increased solar energy production, subsequently reducing the net load on the grid. Accurately modeling these dependencies is essential for precise net load forecasting, as fluctuations in temperature and GHI can lead to significant variations in energy demand and renewable generation.

To achieve this objective, we begin with Tao's Vanilla model (refered to as the Vanilla model subsequently), which is a benchmark model for load forecasting and has been widely adopted in the field of load forecasting. The Vanilla model is a highly accepted and cited MLR-based model load forecasting model, which was first introduced in Hong (2010). This model has been adopted by many scholars in their work (Høverstad et al., 2015; Black and Henson, 2014), and was also used as the benchmark model in GEFCom2012 (Hong et al., 2014). The Vanilla model is defined as:

$$\begin{aligned} y_t =&\beta_0 + \beta_1 \text{ Trend }_t + \beta_2 H_t + \beta_3 W_t + \beta_4 W_t H_t + \beta_5 M_t + f\left(T_t\right) + \beta_6 T_t + \beta_7 T_t^2 + \beta_8 T_t^3 \\ &+ \beta_9 T_T M_t + \beta_{10} T_t^2 M_t + \beta_{11} T_t^3 M_t + \beta_{12} T_k H_t + \beta_{13} T_t^2 H_t + \beta_{14} T_t^3 H_t \end{aligned} \tag{4.8}$$

where y_t stands for the forecasted load at time t, estimated based on the independent variables on the right side of the equation. These independent variables consist of a combination of quantitative and qualitative. $Trend_t$ signifies a linearly ascending variable that models the linear upward trend of load within the data history. T_t denotes the temperature at time t. A third-order polynomial of temperature is utilized to capture the non-linear relation between temperature and load. H_t is the class variable with 24 levels representing the 24 hours of the day. W_t is the class variable with 7 levels for the day of the week, representing the variation in load with the day of the week. M_t is the class variable for the month of the year with 12 levels, representing the monthly variation in load. There is an interaction between temperature terms with the hours in a day and the months of the year. Likewise, the hours of a day also interact with the days of the week.

The vanilla model serves as a solid foundation for building a benchmark net load forecasting model, denoted as $M0$. To investigate the influence of GHI and determine the most effective way to incorporate it, we develop twelve different models from $M1$ to $M12$ by incrementally adding GHI and its interactions to $M0$. In model $M1$, we add GHI to the base model. In model $M2$, we take model $M1$ and add the interaction of GHI with the month of the year. In model $M3$, we add the interaction of GHI with the hour of the day to the previous model. In model $M4$, we add the interaction of GHI with temperature.

In subsequent models ($M5$ and $M6$), we include the interaction of GHI and temperature with the hour of the day and the month of the year, respectively. For models $M7 - M9$ and $M10 - M12$, we incorporate the same interactions as $M4 - M6$, but with the second and third-order polynomial of temperature.

Initial experiments were carried out to investigate the impact of distinct GHI terms and their corresponding cross effects on the aggregated system load of ISONE. We use a sliding simulation validation approach. From the available data (2014-2019), we use three years of data (2014-2016) for parameter estimation, with one year (2017) as the validation data for post-sample fit. Then we move one year ahead and repeat the process. The simple average of the MAPE values from the two validation years (2017 and 2018) is used for variable selection. MAPE values are reported up to two decimal places in accordance with standard reporting conventions.

Table 4.3 reports the average MAPEs of the different models. Notably, incorporat-

Table 4.3: MAPE values (in %) of the twelve benchmarking model candidates for the validation data (2017 & 2018) for the aggregated system load of ISONE

Model	**Input Parameters**	**Parameters Calculated**	**PV Penetration Level**						
			No PV	**5%**	**10%**	**15%**	**20%**	**25%**	**30%**
M0	Vanilla Model	285	4.50	4.94	5.51	6.17	6.94	7.84	8.91
M1	M0 + GHI	286	4.25	4.42	4.64	4.92	5.25	5.64	6.11
M2	M1 + $GHI \cdot H$	309	4.21	4.37	4.58	4.85	5.16	5.55	6.01
M3	M2 + $GHI \cdot M$	320	4.20	4.37	4.58	4.85	5.18	5.56	6.02
M4	M3 + $GHI \cdot T$	321	4.21	4.37	4.59	4.86	5.18	5.57	6.03
M5	M4 + $GHI \cdot T \cdot H$	344	4.23	4.40	4.62	4.88	5.20	5.58	6.04
M6	M5 + $GHI \cdot T \cdot M$	355	4.15	4.31	4.53	4.79	5.11	5.49	5.94
M7	M6 + $GHI \cdot T^2$	356	4.14	4.31	4.52	4.79	5.11	5.49	5.94
M8	M7 + $GHI \cdot T^2 \cdot H$	379	4.13	4.29	4.51	4.78	5.11	5.49	5.95
M9	M8 + $GHI \cdot T^2 \cdot M$	390	4.13	4.30	4.52	4.79	5.11	5.50	5.96
M10	M9 + $GHI \cdot T^3$	391	4.14	4.31	4.53	4.80	5.13	5.52	5.99
M11	M10 + $GHI \cdot T^3 \cdot H$	414	4.14	4.31	4.54	4.81	5.15	5.54	6.02
M12	M11 + $GHI \cdot T^3 \cdot M$	425	4.13	4.30	4.52	4.80	5.13	5.53	6.01

ing the GHI and its interactions into the models leads to a significant improvement in the accuracy of the forecast. Green indicates a lower MAPE value, i.e. better performance, and red indicates a higher MAPE value, i.e. worse performance. Even the addition of a single variable, GHI, leads to a marked decrease in the MAPE from the Vanilla model, indicating a positive impact on the model performance. Furthermore, the improvement in MAPE offered by introducing GHI-related variables increases substantially as the PV penetration increases.

The results show that there is a gradual improvement in MAPE from model $M1$ to model $M6$, beyond which the improvement levels off. Among the models, $M6 - M9$ exhibit promising results as benchmark net load forecasting models. Comparing the four models, it is observed that models $M7$ to $M9$ exhibit slightly better performance for lower levels of PV penetration than model $M6$, although their performance at higher PV penetration levels is comparable. Out of the four models, model $M6$ requires the least number of parameters to be estimated while also performing on par with the other models. These results are in line with other experiments performed using other data sets. Therefore, based on the improvement in accuracy and the model complexity, we recommend the variables defined in model $M6$ as the benchmark net load forecasting model. An additional benefit of model $M6$ is its easy-to-remember pattern of variables, which is similar to Tao's Vanilla model. This feature enhances the model's usability and suitability as a benchmark model for net load forecasting.

4.3.4 Recommended Benchmark Net Load Forecasting Model

The benchmark model includes the following:

1. Quantitative variables: $Trend$, $Temp$, $Temp^2$, $Temp^3$, GHI
2. Class variables: $Month$, $Weekday$, $Hour$
3. Interaction effects: $Weekday \times Hour$, $Temp \times Hour$, $Temp^2 \times Hour$, $Temp^3 \times Hour$, $Temp \times Month$, $Temp^2 \times Month$, $Temp^3 \times Month$, $GHI \times Hour$, $GHI \times Month$, $GHI \times Temp$, $GHI \times Temp \times Hour$, $GHI \times Temp \times Month$

The cross sign represents the interaction effect (also called as cross effects). The recommended benchmark model can defined as the following:

$$\begin{aligned} y_t = &\beta_0 + \beta_1 \text{Trend}_t + \beta_2 H_t + \beta_3 M_t + \beta_4 W_t + \beta_5 W_t H_t \\ &+ \beta_6 T_t + \beta_7 T_t^2 + \beta_8 T_t^3 + \beta_9 T_t H_t + \beta_{10} T_t^2 H_t \\ &+ \beta_{11} T_t^3 H_t + \beta_{12} T_t M_t + \beta_{13} T_t^2 M_t + \beta_{14} T_t^3 M_t \\ &+ \beta_{15} GHI_t + \beta_{16} GHI_t H_t + \beta_{17} GHI_t M_t \\ &+ \beta_{18} GHI_t T_t + \beta_{19} GHI_t T_t H_t + \beta_{20} GHI_t T_t M_t \end{aligned}$$

4.3.5 Results

To demonstrate the effectiveness of the proposed model, we conducted an out-of-sample for a total of ten time series from ISONE. These include the eight load zones of ISONE at the lowermost level, the combined load of three zones situated in Massachusetts at the intermediary level, and the aggregated system load at the top level, which encompasses the summation of the aforementioned eight zones. The data from the year 2019 was excluded from the parameter estimation and model selection process and reserved as a hold-out set for evaluating the model's performance. Additionally, the performance of the Vanilla model was also presented to facilitate a comparative analysis between the two benchmark models.

Table 4.4 shows the results for the out-of-sample test performance for all load zones and the total system load for ISONE. Across all load zones, the proposed model outperforms the Vanilla model. Compared to the benchmark model, the relative reduction to MAPE values when using the proposed model ranges from around 17% for net load with 5% PV penetration to around 45% for net load with 30% PV penetration. Additionally, the proposed model outperforms the Vanilla model by approximately 10% in the case without PV in the system, i.e. when net load equals the load. This improvement may be due to the additional information added to the model in the form of irradiance.

The actual and predicted values for a typical summer and winter day with 30% PV penetration are presented in Figure 4.5 and 4.6, respectively. We also include the demand for reference for visualization proposes. In particular, during the hours with sunlight, the predicted values from the proposed model follow the net load more closely, while the Vanilla model struggles to forecast the drop in demand due to PV generation. In contrast, during hours without sunlight, the forecasts of the two models appear similar, indicating that both models effectively capture demand patterns in the absence of PV generation. The proposed model demonstrated a considerable improvement in MAPE, reducing it from 4.87% to 3.06% for a typical summer day, and from 5.14% to 3.01% for a typical winter day compared to the Vanilla model, representing a relative improvement of approximately 37% and 41%, respectively.

The improved forecast accuracy with the proposed model highlights the importance of updating the previous load forecasting model and developing net load forecasting models that are designed to accommodate renewable generation.

Table 4.4: MAPE values (in %) of the Vanilla model vs the proposed model on the out-of-sample test data (2019) for ISONE

Zone	**Model**	**0% PV**	**5% PV**	**10% PV**	**15% PV**	**20% PV**	**25% PV**	**30% PV**
VT	Vanilla Model	7.83	8.65	9.67	10.93	12.54	14.68	17.85
	Proposed Model	6.93	7.20	7.52	7.92	8.44	9.14	10.24
NH	Vanilla Model	3.95	4.42	5.12	5.99	7.02	8.24	9.68
	Proposed Model	3.76	3.83	3.92	4.02	4.14	4.29	4.47
ME	Vanilla Model	3.43	3.86	4.48	5.25	6.15	7.18	8.38
	Proposed Model	3.26	3.31	3.37	3.44	3.53	3.63	3.75
RI	Vanilla Model	4.69	5.37	6.38	7.66	9.24	11.22	13.83
	Proposed Model	4.44	4.55	4.68	4.84	5.04	5.30	5.64
CT	Vanilla Model	5.32	6.01	7.00	8.26	9.81	11.76	14.36
	Proposed Model	5.00	5.14	5.30	5.49	5.74	6.05	6.48
SEMASS	Vanilla Model	5.55	6.41	7.57	9.07	11.01	13.65	17.81
	Proposed Model	5.00	5.14	5.32	5.54	5.82	6.21	6.84
NEMASS	Vanilla Model	4.48	5.17	6.11	7.27	8.67	10.36	12.50
	Proposed Model	4.18	4.28	4.39	4.53	4.71	4.92	5.20
WCMASS	Vanilla Model	5.69	6.44	7.42	8.61	10.06	11.84	14.12
	Proposed Model	5.21	5.36	5.54	5.76	6.02	6.36	6.80
MASS	Vanilla Model	4.88	5.54	6.33	7.26	8.36	9.69	11.35
	Proposed Model	4.22	4.46	4.75	5.11	5.53	6.04	6.69
ISONE	Vanilla Model	4.43	5.02	5.73	6.56	7.54	8.70	10.10
	Proposed Model	3.85	4.06	4.32	4.63	5.00	5.45	6.00

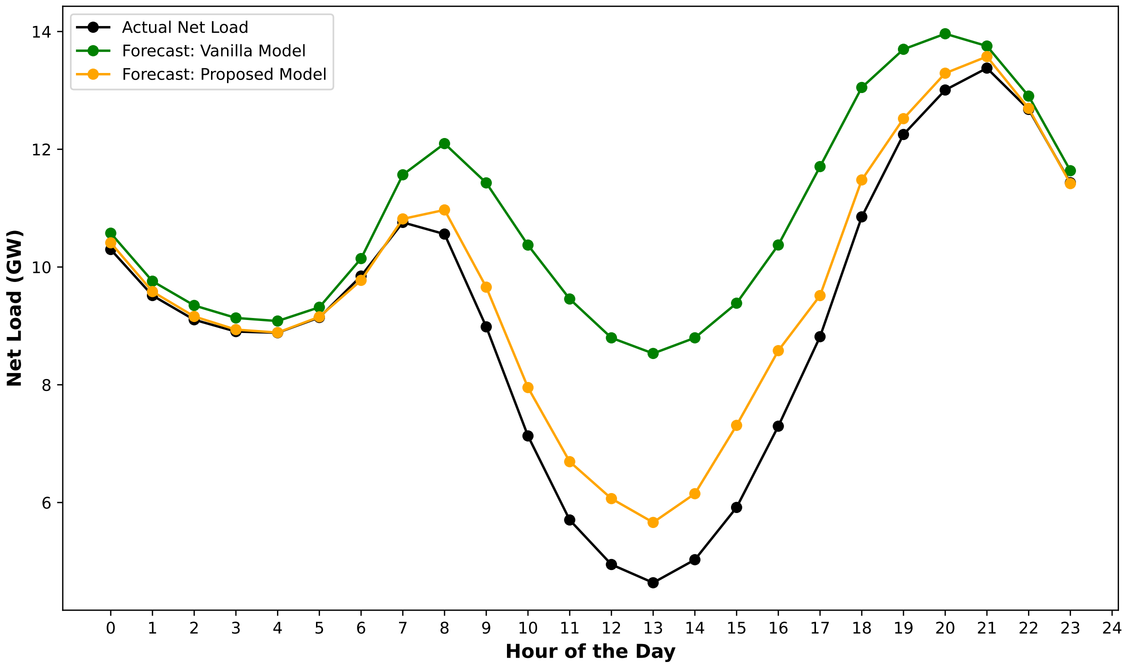


Figure 4.5: Actual vs. prediction of a summer day in the test year (2019)

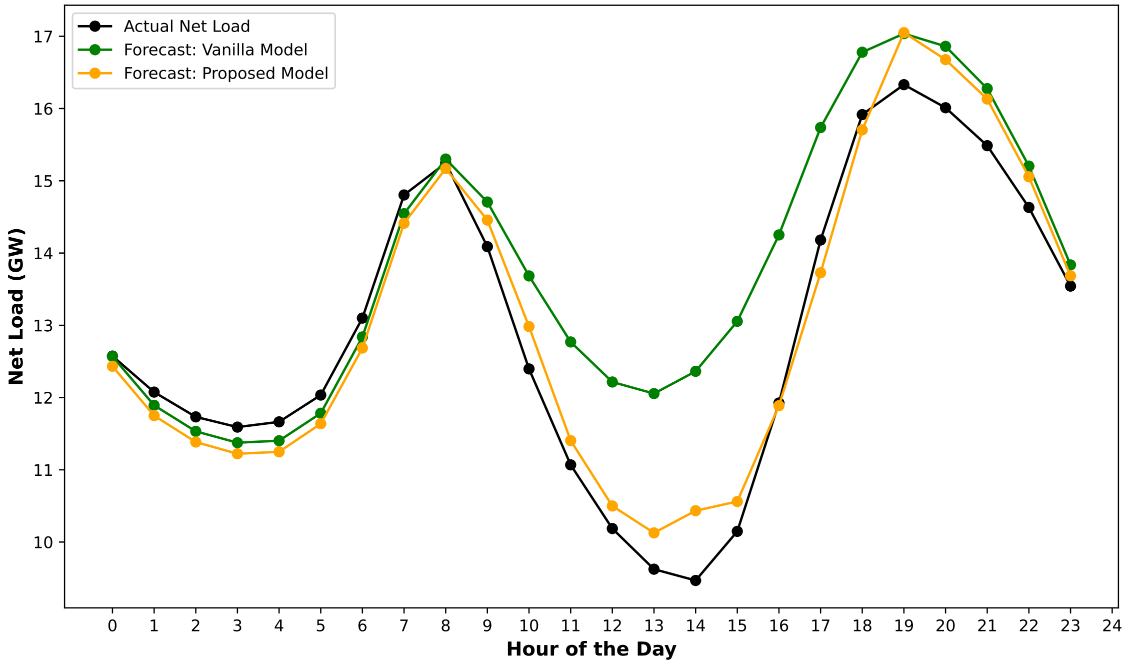


Figure 4.6: Actual vs. prediction of a winter day in the test year (2019)

This study further expands upon the previous research by including case studies from CAISO and ERCOT. The results of the proposed model and the Vanilla model for varying levels of PV penetration for CAISO and ERCOT are presented in Table A.2 and Table B.2, respectively in Appendix A.

The findings from ISONE, CAISO, and ERCOT strongly indicate that the proposed net load forecasting model surpasses the Vanilla model in performance. As a result, it is recommended to consider this model as the benchmark model for forecasting net load.

4.4 A Framework For Including Recency Effect To The Proposed Benchmark Net Load Forecasting Model

The term "recency effect" refers to the phenomena in which the most recent values have an impact on current and future values. In terms of load forecasting, this entails incorporating historical weather variables into the model in order to forecast current values. This was originally introduced by Hong (2010), and subsequently Wang et al. (2016) proposed a methodology to include the recency effect in the short-term load forecast. Recency effect can be added to the model by including lags and moving averages of weather variables. For the vanilla model, this means including lags and a moving average of temperature. The Vanilla model, defined in 4.8, extended to include the recency effect, can be written as:

$$\begin{aligned} y_t = \beta_0 + &\beta_1 \operatorname{Trend}_t + \beta_2 H_t + \beta_3 W_t + \beta_4 M_t + \beta_5 H_t W_t + f(T_t) \\ &+ \sum f\left(\hat{T}_{t,d}\right) + \sum f(T_{t-h}) \end{aligned} \tag{4.9}$$

where T_{k-l} is the lag of temperature of the lth hour and $\sum_d f(\tilde{T}_{k,m})$ is the moving average (MA) of the temperature of the mth day.

Wang et al. (2016) propose a methodology to select the best pair of lag and MA of temperature to add to Tao's Vanilla model. In this work, we follow the framework provided by Wang et al. (2016) and extend it to apply it to net load forecasting. Since the benchmark net load forecasting model consists of GHI in addition to temperature, we extend the recency effect to include GHI.

To do this, we add the recency effect in two parts. In the first part, we find the best lag-MA pair for temperature. For this, we vary the lags from 0 to 24 hours and the MA from 0 to 7 days. This gives us the impact of the recency effect of temperature on the net load forecasting model. In addition to temperature, the lag and MA of the temperature are also added to the net load forecasting model. We do not make any changes to GHI.

Once we get the best lag-MA pair for temperature, we fix that and explore the recency effect for GHI. In this step, we follow the same steps as in Wang et al. (2016) but for GHI. We vary the lags from 0 to 24 hours and the MA from 0 to 7 days and select the lag-MA pair that yields the best accuracy.

The proposed benchmark model ($M6$) and the benchmarking process provide a generic framework to develop a net load forecasting model. However, this model can be further improved by adding some more information, specific to the use case. We also add recency effect to the Vanilla model to compare the model with the proposed model. Net load with 30% PV penetration is used for this analysis. The data from 2017 and 2018 is used as the validation data to select the lg-MA pairs and the data from 2019 is used as the out-of-sample test set.

Table 4.5 shows the heatmap of the MAPE's of the Vanilla model with recency effect on the validation data. A greener shade indicates better performance and a redder shade indicates worse performance. The best (d, h) pair is highlighted in bold. It can be observed that including additional temperature information in the form of lag and MA significantly improves the models. The best model is observed for (1, 9).

Since the proposed net load forecasting model includes temperature and GHI as the two weather variables, we need to include recency effect for both the variables. We begin by first adding recency effect for only the temperature variables, keeping the GHI-related variables as is in the model. Table 4.6 shows the heatmap of the MAPE's by adding recency effect for temperature on the proposed net load model. Similar to the Vanilla model, adding recency effect for temperature significantly improves the model. The best model is observed for (1, 5).

Keeping the best lag-MA pair for temperature fixed in the model, the next step is to include recency effect for GHI. The heatmap of the MAPE'S after including recency effect for GHI to the net load model is presented in Table 4.7. We observe that by including additional information for GHI further improves the model. The best model is observed for (0, 6).

Table 4.5: Heatmap of the MAPE values (in %) for recency effect modeling for the Vanilla model on the validation data (years 2017 and 2018)

MA Lag	0	1	2	3	4	5	6
0	8.91	7.85	8.30	8.64	8.84	8.94	8.99
1	8.39	7.37	7.88	8.21	8.39	8.46	8.47
2	8.29	7.34	7.85	8.15	8.31	8.37	8.38
3	8.20	7.34	7.82	8.09	8.24	8.29	8.29
4	8.07	7.32	7.76	8.00	8.13	8.17	8.16
5	7.90	7.27	7.67	7.88	7.99	8.02	8.01
6	7.73	7.22	7.58	7.77	7.84	7.86	7.84
7	7.58	7.19	7.49	7.67	7.73	7.72	7.71
8	7.47	7.17	7.43	7.59	7.64	7.62	7.61
9	7.40	**7.17**	7.39	7.55	7.58	7.56	7.54
10	7.39	7.17	7.38	7.53	7.57	7.54	7.51
11	7.43	7.20	7.40	7.56	7.60	7.58	7.55
12	7.52	7.26	7.47	7.63	7.68	7.65	7.64
13	7.62	7.33	7.56	7.73	7.77	7.75	7.75
14	7.74	7.38	7.66	7.83	7.88	7.87	7.87
15	7.86	7.38	7.78	7.95	8.00	7.99	7.99
16	7.98	7.34	7.90	8.06	8.12	8.12	8.13
17	8.10	7.30	8.01	8.18	8.23	8.25	8.26
18	8.22	7.29	8.10	8.29	8.35	8.37	8.40
19	8.33	7.32	8.17	8.39	8.45	8.49	8.52
20	8.42	7.36	8.22	8.48	8.55	8.59	8.62
21	8.50	7.40	8.24	8.55	8.63	8.67	8.71
22	8.56	7.43	8.24	8.58	8.67	8.72	8.77
23	8.59	7.45	8.20	8.60	8.69	8.75	8.79

Table 4.6: Heatmap of the MAPE values (in %) for recency effect modeling for temperature for the proposed model on the validation data (years 2017 and 2018)

MA Lag	0	1	2	3	4	5	6
0	5.94	5.29	5.55	5.79	5.98	6.05	6.10
1	5.63	5.13	5.38	5.59	5.76	5.82	5.86
2	5.52	5.10	5.34	5.53	5.68	5.74	5.77
3	5.42	5.08	5.31	5.48	5.61	5.66	5.69
4	5.34	5.07	5.28	5.44	5.56	5.60	5.62
5	5.26	**5.07**	5.27	5.40	5.51	5.54	5.55
6	5.21	5.07	5.26	5.39	5.48	5.49	5.50
7	5.18	5.09	5.27	5.39	5.47	5.47	5.47
8	5.17	5.11	5.28	5.40	5.47	5.46	5.45
9	5.18	5.13	5.30	5.41	5.48	5.47	5.45
10	5.21	5.14	5.32	5.43	5.49	5.48	5.47
11	5.25	5.14	5.34	5.44	5.51	5.50	5.50
12	5.29	5.14	5.36	5.47	5.54	5.53	5.53
13	5.33	5.14	5.38	5.50	5.56	5.56	5.56
14	5.36	5.13	5.40	5.52	5.59	5.58	5.58
15	5.39	5.12	5.41	5.55	5.61	5.61	5.61
16	5.42	5.12	5.43	5.58	5.64	5.65	5.64
17	5.47	5.12	5.44	5.60	5.67	5.68	5.68
18	5.50	5.13	5.46	5.64	5.71	5.72	5.72
19	5.53	5.14	5.49	5.67	5.74	5.75	5.76
20	5.54	5.18	5.52	5.70	5.77	5.78	5.79
21	5.56	5.22	5.55	5.74	5.81	5.82	5.83
22	5.59	5.27	5.58	5.78	5.85	5.85	5.86
23	5.61	5.32	5.62	5.82	5.88	5.87	5.89

Table 4.7: Heatmap of the MAPE values (in %) for recency effect modeling for GHI for the proposed model on the validation data (years 2017 and 2018)

MA Lag	0	1	2	3	4	5	6
0	5.07	5.26	5.40	5.31	5.24	5.21	5.23
1	5.01	5.21	5.33	5.23	5.16	5.14	5.16
2	5.02	5.23	5.34	5.24	5.17	5.15	5.17
3	5.02	5.25	5.36	5.26	5.18	5.16	5.19
4	5.02	5.26	5.37	5.26	5.19	5.17	5.19
5	5.01	5.27	5.38	5.28	5.21	5.17	5.20
6	**5.01**	5.28	5.39	5.29	5.22	5.18	5.21
7	5.01	5.29	5.40	5.30	5.23	5.19	5.22
8	5.01	5.28	5.40	5.31	5.23	5.20	5.23
9	5.01	5.26	5.39	5.30	5.23	5.19	5.23
10	5.03	5.24	5.38	5.29	5.23	5.19	5.22
11	5.04	5.23	5.38	5.30	5.24	5.21	5.23
12	5.07	5.25	5.39	5.31	5.26	5.22	5.25
13	5.09	5.26	5.40	5.33	5.28	5.24	5.26
14	5.12	5.27	5.42	5.34	5.29	5.25	5.28
15	5.13	5.27	5.43	5.36	5.30	5.27	5.29
16	5.15	5.27	5.44	5.36	5.31	5.27	5.30
17	5.17	5.27	5.43	5.36	5.30	5.27	5.30
18	5.18	5.27	5.41	5.35	5.29	5.26	5.29
19	5.18	5.26	5.39	5.33	5.28	5.25	5.28
20	5.22	5.26	5.39	5.34	5.30	5.28	5.32
21	5.29	5.31	5.42	5.39	5.36	5.34	5.38
22	5.37	5.40	5.48	5.45	5.42	5.41	5.45
23	5.43	5.45	5.52	5.49	5.47	5.46	5.50

4.4.1 Results

The (d,h) pairs from the validation data are used to evaluate the performance of the models in the out-of-sample test set. We customize the model for each load zone of ISONE using the validation data. The Table 4.8 shows the MAPE values on the out-of-sample test set with 30% PV penetration. To present a comprehensive analysis, we compare the vanilla model and the proposed net load mode - with and without recency effect.

On average, the MAPE values are lower for the recency effect models than for the base models. The Vanilla model incorporating the recency effect shows significant improvement over the base Vanilla model. On average, across all load zones, the model performance improves by approximately 16%. However, the real value comes from adding recency effect to the proposed net load model.

Moreover, the best results are achieved when temperature and GHI recency are included within the proposed net load model. The net load model with the recency effect outperforms the Vanilla model by about 41% and shows an approximate 8% improvement over the base net load model, thus emphasizing the value of integrating the recency effect into net load forecasting.

Table 4.8: MAPE values (in %) of the Vanilla model vs the proposed model with recency effect modeling on the out-of-sample test data (2019) for ISONE

Model	**Vanilla**	**Vanilla Model**	**Proposed**	**Proposed Model**
Zone	**Model**	**w/ Recency**	**Model**	**w/ Recency**
VT	17.16	14.85	11.87	11.55
NH	8.68	7.09	5.82	5.03
ME	7.37	6.28	4.78	4.42
RI	11.06	9.31	6.89	6.30
CT	11.96	10.43	8.11	7.22
SEMASS	14.40	12.14	8.78	7.99
NEMASS	10.28	8.78	6.37	6.02
WCMASS	12.21	10.17	8.22	7.69
MASS	11.35	9.42	6.69	6.37
ISONE	10.10	8.23	6.00	5.28

Figure 4.7 shows the percentage improvement of the vanilla model with recency effect, the proposed model, and the proposed model with recency effect over the base vanilla model. The results are for a system with 30% PV penetration. It can be noted that the proposed benchmark model shows an improvement between 30%-40% over the vanilla model. Furthermore, the proposed model with the recency effect added shows an even greater improvement between 32% - 47%. The biggest improvement comes from aggregated zones such as MASS and ISONE. This is because, at the aggregated level, the load becomes more predictable.

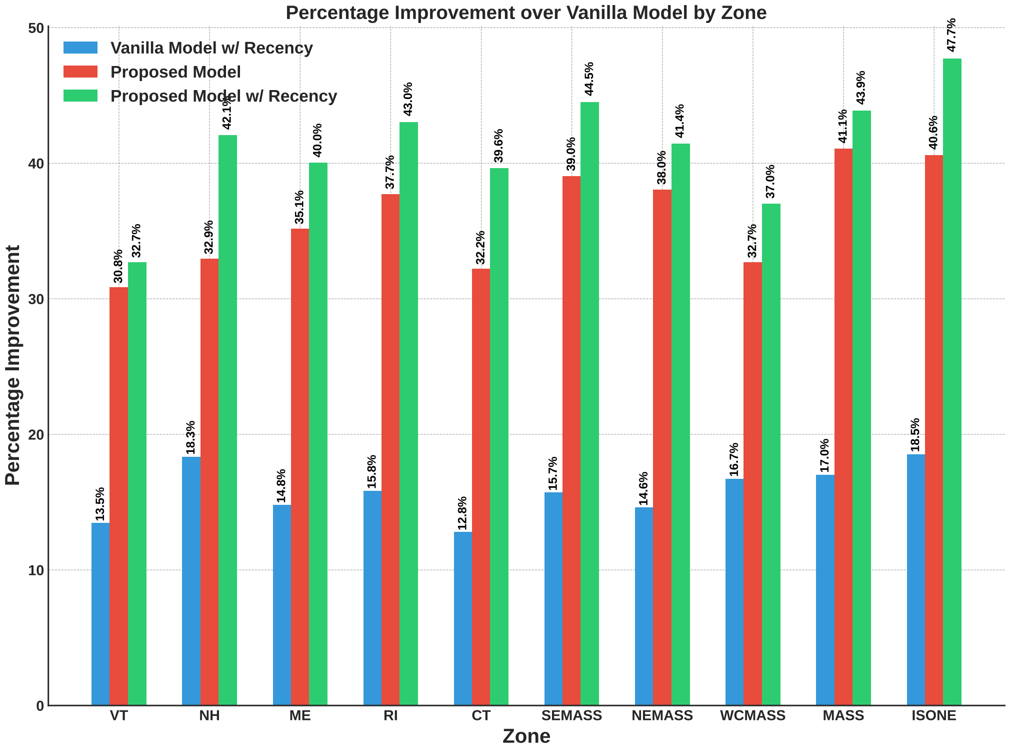


Figure 4.7: Percentage improvement over the Vanilla model for the test year (2019) with 30% PV penetration

Since the goal of this work is to present a benchamrk model for net load forecasting, we also present the results of the prposed model using additional case studies from CAISO and ERCOT. The data description and the results of the proposed modeling framework for CAISO and ERCOT are presented in Appendix A and Appendix B, respectively.

4.5 Conclusion

In this study, we propose a benchmark net load forecasting model. The proposed model leverages existing load forecasting techniques and extends them to the domain of net load forecasting. Additionally, we showcase the performance of our proposed model for net load forecasting under varying levels of PV penetration, thereby highlighting the impact of PV penetration on net load forecast performance. The results indicate that the benchmark model outperforms the existing benchmark model for load forecasting and provides a robust foundation for future research in this area.

Furthermore, we also extend the proposed model by adding recency effect to it. Adding recency effect to the proposed model shows an improvement of over 40% compared to the base net load benchmark model.

The insights derived from the study are of significant practical relevance to the planning and management of utilities, as precise net load forecasting is pivotal for efficiently incorporating renewable energy sources into the power grid.

The proposed work on net load forecasting opens various avenues for future research. The currently proposed benchmark net load forecasting model serves as a foundational model with scope for further improvement. The weather variables considered in this research are limited to temperature and GHI. The inclusion of additional weather variables such as cloud cover, wind speed, and humidity, among others, could potentially enhance the model's performance. The model could also benefit from the utilization of data from multiple weather stations, adopting the weather station selection methodology proposed by Hong et al. (2015). Currently, the model uses multiple linear regression for forecasting. Exploring other methods, such as machine learning techniques like neural networks or gradient-boosted trees, could be valuable for predicting net load.

CHAPTER 5: A REGRESSION BASED FRAMEWORK FOR ESTIMATING PV PENETRATION AND BTM PV DETECTION

The rising adoption of renewable energy sources, especially solar PV systems, has revolutionized the dynamics of electricity generation and distribution. As our world shifts towards a more sustainable, low-carbon future, understanding the incorporation and impact of PV systems on the electrical grid has become of utmost importance. This paradigm shift presents utilities with two main challenges: "What is the extent of PV in the system?" and "Which meters have behind-the-meter PV installations?" Addressing these questions is vital for utilities and energy providers, as it fosters efficient grid management, resource distribution, and the development of informed policies.

PV penetration estimation is the process of quantifying the amount of PV capacity that is integrated into the electricity system. The presence of distributed PV generation within a power system is indicated by the PV penetration or PV capacity information embedded in the corresponding net load time series. With PV generation data readily available, the estimation of capacity would be relatively straightforward. However, in reality, data for PV generation and load often exists in an aggregate form as net load information, with separate data usually inaccessible. This lack of segregated data makes it difficult to distinguish between a decrease in electricity consumption and an increase in PV output power when analyzing a single net load curve. Accurate estimation, therefore, is vital to allow utilities to assess solar energy's contribution and its impact on grid dynamics, load profiles, and the overall system's performance.

Another important aspect of PV integration is the detection of BTM-PV installations, where PV systems are installed by individual consumers on their premises. Identifying meters with behind-the-meter PV installations is crucial for utilities to gain insight into the particular customer adoption and impact of distributed solar energy. It allows for targeted strategies for customer engagement, grid management, and the optimization of distributed energy resources. However, distinguishing between meters with and without PV installations based solely on load or net load data is a challenging task.

5.1 Overview of Contributions

In this chapter, we tackle key challenges in the energy sector by developing methodologies to estimate PV penetration and detect BTM PV installations. Our approach utilizes differences between load and net load forecasting models to extract critical insights, enabling precise quantification and identification of PV systems.

The main contributions of this study are as follows:

1. A novel methodology for estimating PV penetration is presented. The proposed approach capitalizes on the disparities between existing load and net load forecasts. This method provides electric utilities with a refined tool for using historical load data to gauge PV penetration, addressing the critical question of "What is the extent of PV in the system?", enhancing their ability to manage and integrate solar resources effectively.

2. Next, another critical question "Which meters have behind-the-meter PV installations?" is addressed. A novel framework that identifies meters with BTM PV installations, addressing the challenge of detecting unreported or invisible PV installations is presented. By analyzing the discrepancies in forecast performance between load and net load models, the proposed framework accurately identifies meters with PV installations, providing crucial data to utilities on the distribution and adoption of distributed solar energy.

3. The effectiveness and accuracy of the proposed methodologies are demonstrated through real-world case studies. Utilizing actual utility data, the proposed frameworks validate the practical applicability of the proposed frameworks. These case studies serve as tangible evidence of the methodologies' potential for real-world implementation by utilities.

These contributions significantly improve the practical tools available to electric utilities, enabling more precise management and integration of solar energy sources. By providing effective methods to accurately gauge PV penetration and identify BTM installations, the work presented in this section assists utilities in optimizing grid performance and planning for future energy needs, ultimately supporting the transition to more sustainable energy systems.

The remainder of the chapter is structured as follows: This chapter presents two main contributions. Firstly, a regression-based framework for estimating PV penetration in the system is detailed in Section 5.2. Secondly, a regression-based framework for BTM PV detection is presented in Section 5.3. In the first study, the data utilized is outlined in Section 5.2.1, followed by the presentation of the proposed metric for measuring PV penetration, the framework for estimating PV penetration, and the corresponding results in Section 5.2.2. For the second study, the data is described in Section 5.3.1, and the proposed framework along with the results are presented in Section 5.3.2. The chapter concludes with a summary in Section 5.4.

5.2 A Regression-Based Framework for Estimating PV Penetration in the System

This Section presents a robust framework for estimating PV penetration in a system. The proposed framework utilizes the previously established load forecasting model, namely Tao's Vanilla model, and the proposed benchmark model for net load forecasting presented in Section 4.3.

5.2.1 Data Description

Datasets from two sources are used in this study. The dataset from ISO New England is used to formulate the study and validate the proposed framework. The second dataset from a mid-sized utility in the US is used to test adn evaluate the framework proposed in the study.

5.2.1.1 ISO New England (ISONE)

The data used in this work comes from ISONE. We use the total system load of ISONE as well as the normalized and aggregated BTM PV data from ISONE. The details of the data are presented in Section 4.2.1. Figure 5.1 shows the hourly load, aggregated BTM PV and normalized BTM PV data for the total system load of ISONE.

Since ISONE provides the load, normalized PV and the aggregated BTM PV data separately, we can use the load and the aggregated BTM PV data to generated the net load time series, using Equation 5.1:

$$Net\,load = Load - Aggregated\,BTM\ PV \tag{5.1}$$

5.2.1.2 Aggregated Data from a Mid-Sized US Utility

The data set includes hourly load data and corresponding weather data spanning over a three-year period from January 1, 2016, to December 31, 2018. The hourly aggregate PV generation was available for the utility the while the normalized BTM PV data was not recorded available in this case. Although, the aggregated PV generation data was available, in most cases this data is not available with the utilities. In such scenarios, we would need to rely on simulation PV generation for the location.

To better replicate real-world conditions and evaluate the effectiveness of our proposed approach in cases where simulated PV data is required, we incorporated sim-

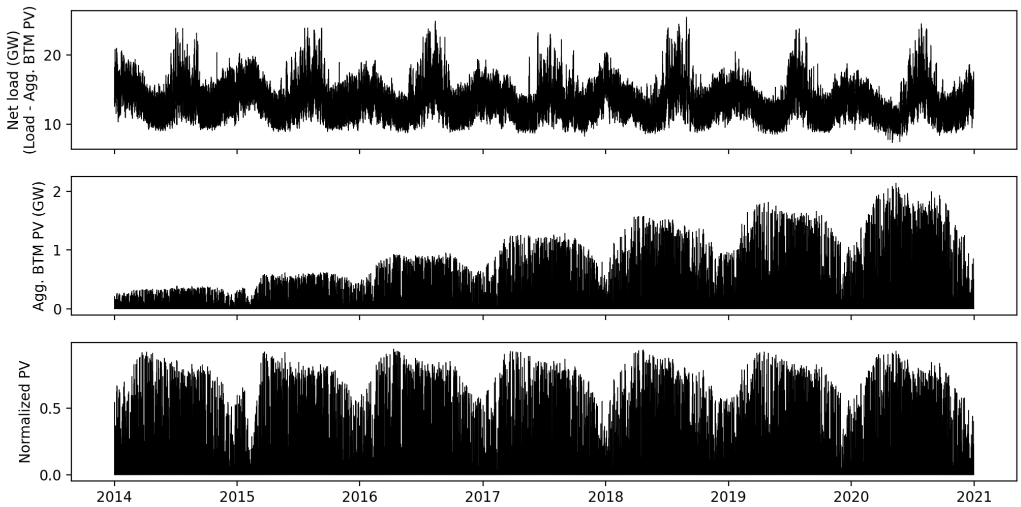


Figure 5.1: Load, aggregated BTM PV and normalized BTM PV time series from ISONE (2014 to 2021)

ulated PV generation data as part of our analysis.

To generate simulated PV generation data, we use the System Advisor Model (SAM), which uses real weather data to simulate PV generation, based on precise PV models, taking into account various weather variables such as solar irradiance, wind speed and temperature data Gilman et al. (2018). A study by Freeman et al. (2014) examined the performance of SAM and validated its accuracy by comparing its simulation results with actual measurements. The results indicate that the average normalized hourly root mean square error between the SAM simulations and the actual measurements was approximately 4%. Given the high level of precision associated with the simulated data obtained through SAM, it was deemed appropriate for use in simulating PV generation in this study.

Based on the high level of precision associated with the simulated data generated via SAM, we consider it appropriate to use the simulated PV generation data for our case study. Figure 5.2 shows the hourly load and simulated PV generation for the case study. The results of our analysis using the simulated data are presented in subsequent sections.

5.2.1.3 Proposed Metric for Quantifying PV Penetration: Mean Daily PV Share (MDPVS)

Before we can begin to estimate PV penetration in the system, we need to first define what PV penetration is. The term "PV penetration" lacks a widely accepted definition in the literature, and various definitions are utilized by researchers. For example, Oliver and Perfumo (2015) defines PV penetration as the percentage of customers in a feeder who have PV systems installed. However, most researchers use some ratio of PV and load to calculate PV penetration, which is often some form of peak PV and load, for example PV penetration is defined as the ratio of peak PV active power to the peak load active power (ul Abideen et al., 2019), or as the ratio of total peak PV real power to peak load apparent power (Hoke et al., 2012a,b), or the ratio of total peak PV real power to peak load real power (Kordkheili et al., 2014). Cheng et al. (2015) define PV penetration as the ratio of the total PV nameplate capacity to the annual peak load of the circuit.

Defining photovoltaic PV levels solely based on the intersection of peak PV generation and peak load may lead to an inaccurate representation of actual PV penetration levels. This approach neglects other factors that affect PV generation, such as geographical location and load profile. Load patterns can experience sudden spikes due to weather events such as heat waves. Additionally, PV systems can have varying levels of generation depending on factors such as weather conditions and system design. Therefore, defining PV penetration solely on the basis of peak PV generation and peak load may not accurately capture the full range of PV system output over time. As a result, relying on peak PV generation and peak load to define PV penetration levels may not provide an accurate representation of the extent to which PV has penetrated the electrical grid.

While peak PV generation and peak load can provide a useful snapshot of PV penetration, a more comprehensive approach to measuring PV penetration is required which would involve considering the total amount of PV energy generated over a given period and comparing it to the total amount of energy consumed on the grid during that same period. As a result, we develop a new PV penetration statistic known as mean daily PV share (MDPVS), defined as:

$$MDPVS = \frac{1}{n}\sum_{t=1}^{n}\left(\frac{\sum_{h=1}^{24} PV_h \times 100}{\sum_{h=1}^{24} \text{Load}_h}\right) \tag{5.2}$$

where,n is the number of days, $\sum_{h=1}^{24} PV_h$ is the total daily PV generation and $\sum_{h=1}^{24} Load_h$ is the total daily demand.

MDPVS defines PV penetration as the average of the total daily PV generation divided by the total daily load. MDPVS hence considers the daily demand and how much of it was covered by PV generation. This approach provides a more complete picture of the overall impact of PV systems on the grid.

5.2.2 Proposed Framework

Figure 4.5 and Figure 4.6 (presented in Section 4.3) show the net load forecast curves derived from both the load and net load forecasting models. We can observe a clear distinction between the two forecasts compared to the actual net load, with the forecast from the net load forecasting model aligning more closely with the actual net load. Furthermore, Table 4.4 (presented in Section 4.3) shows the difference in performance between the two models. As the PV penetration increases, the difference between the two models also increases. This observation inspires us that we can utilize the discrepancy between two forecasting models to extract useful information for capacity estimation. By comparing the forecasts generated by these two models, the relationship between the amount of PV in the system referred to as the MDPVS, can be established.

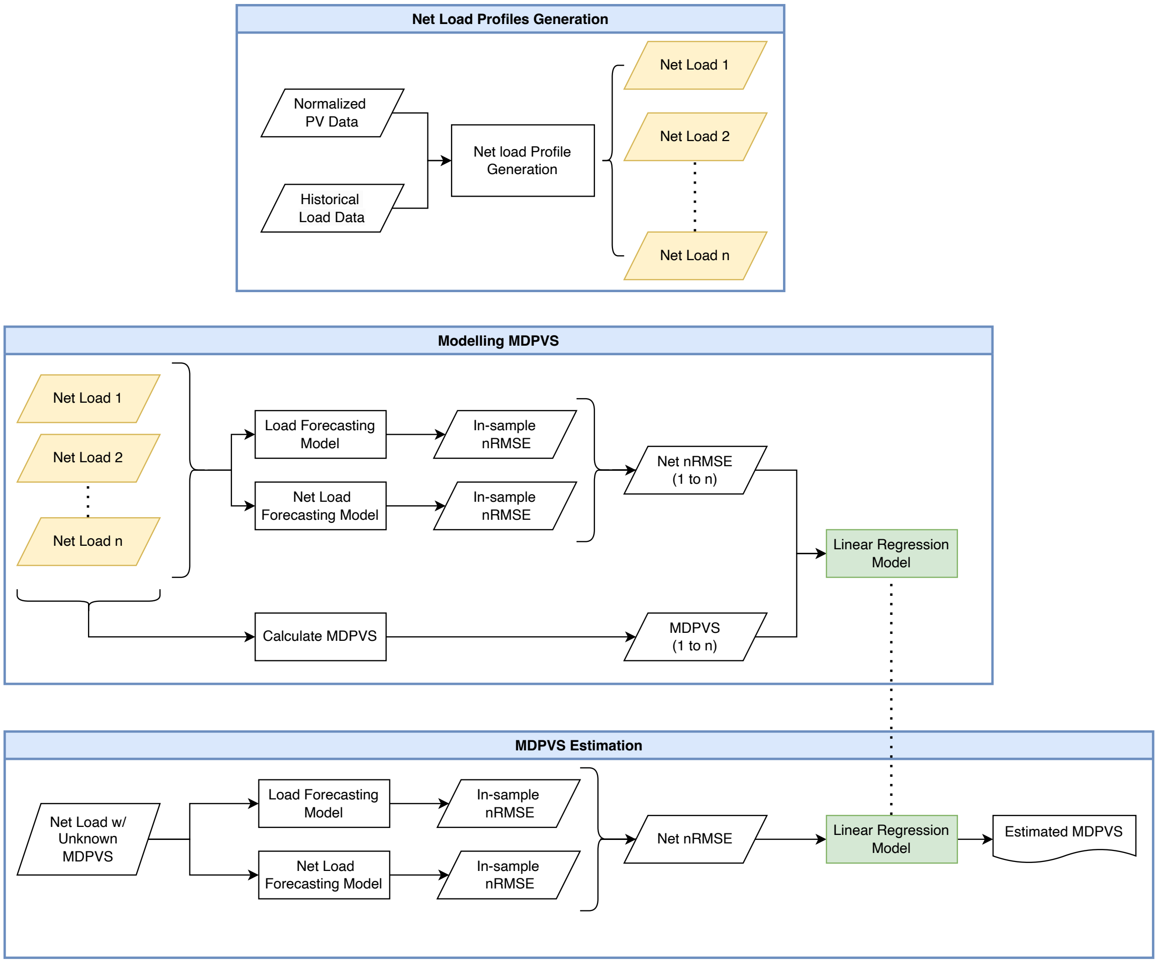


Figure 5.2: High-level workflow for the PV estimation model

The high-level workflow of the proposed methodology is illustrated in Figure 5.2. This three-step framework aims to estimate the MDPVS by utilizing historical load and PV generation data. The process initiates with the generation of synthetic net load profiles. Subsequently, it involves modeling the relationship between these profiles and MDPVS. The final step applies these models to estimate MDPVS for unknown cases. The steps of the framework are outlined as follows:"

1. **Net load profile generation**

   The initial phase of the proposed research framework involves generating multiple net load profiles from available load data. This process starts by using historical load time series data, which ideally predates the adoption of PV systems and is commonly available from utility companies. Subsequently, we use normalized PV generation time series for the specified location to create net load profiles. In instances where this data is not available from the utility, tools like the National Renewable Energy Laboratory's (NREL) System Advisor Model (SAM) can be used to simulate PV generation time series. SAM is a freely accessible program that provides techno-economic modeling and can simulate PV systems with high accuracy across different U.S. locations.

With the historical load data and normalized PV time series at hand, we proceed to generate synthetic net load profiles (for example, 200 profiles). This is done by iteratively varying the amount of PV added to the load, where the $ScalingFactor$ is adjusted to produce diverse net load profiles with varying MDPVS, as per the equation:

$$Net\,load = Load - ScalingFactor * Aggregated\,BTM\,PV \tag{5.3}$$

An example of these profiles is illustrated in Figure 5.3, showcasing net load profiles with different MDPVS levels.

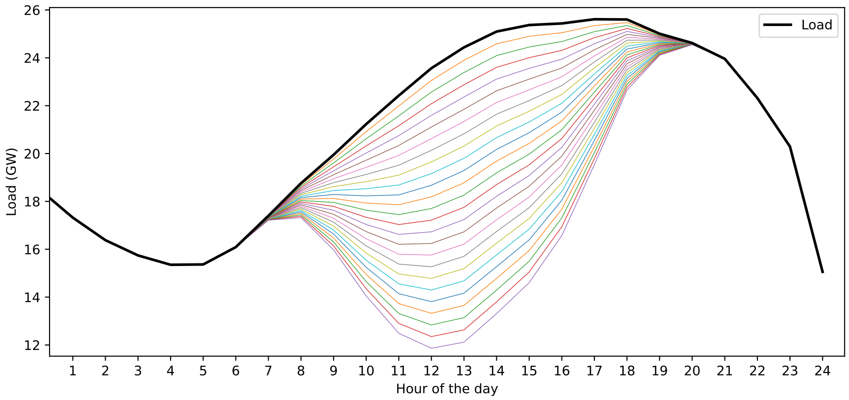


Figure 5.3: Net load profiles with varying levels of MDPVS

2. **Modeling**

In the second part of the proposed methodology, we seek to establish a relationship between the net error from the benchmark load and net load forecasting models and the MDPVS in the system. To accomplish this, we apply Tao's Vanilla load benchmark model (defined in Equation 4.8) and the net load benchmark model (defined in Equation 4.3.4) to each synthetic net load time series and calculate the in-sample normalized Root Mean Square Error (nRMSE) values. This provides us with a load nRMSE value and a net load nRMSE value for each net load time series.

We then calculate the net nRMSE values by subtracting the nRMSE value of the load forecasting model from that of the net load forecasting model, as defined in Equation 5.4. The assumption here is that for net load data, the net load forecasts are more accurate than the load forecasts. We aim to leverage this discrepancy in the forecasts of the two models, represented in the net nRMSE.

$$Net : nRMSE = nRMSE_{\text{Load forecasting model}} - nRMSE_{\text{Net load forecasting model}} \tag{5.4}$$

Subsequently, we determine the MDPVS for each net load series. This analysis generates a collection of data points, each characterized by distinct net nRMSE and MDPVS values. To visualize and analyze the relationship between these two variables, we plot the net nRMSE against the corresponding MDPVS values in a scatter plot, as depicted in Figure 5.4.

An examination of the scatter plot reveals a discernible relationship between the net nRMSE and the MDPVS. Based on this observed relationship, we propose developing a linear regression model, referred to as Model B. The aim of Model B is to predict the MDPVS or PV penetration level of a net load with unknown PV penetration. The parameters of Model B are estimated by training it on the net nRMSE and MDPVS data points obtained from the various net load timeseries, thus enabling us to establish a predictive framework based on our findings.

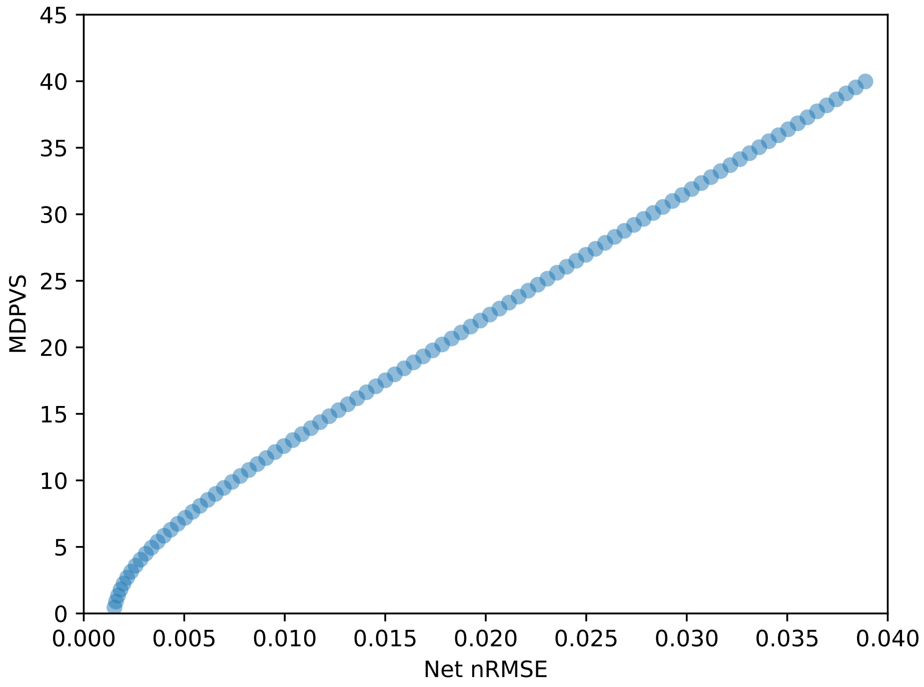


Figure 5.4: Scatterplot of net nRMSE vs. MDPVS based on the training data (2017)

3. **PV estimation**

The final phase utilizes the regression model developed in the previous step to estimate MDPVS for any given year. For a net load with an unknown PV penetration, we calculate the net nRMSE using the load and the net load models. The fitted parameters from Model B are then used to estimate the MDPVS for the target year, providing a reliable estimation of PV penetration.

Through this structured methodology, we establish a robust framework for accurately estimating PV penetration across various settings, contributing to enhanced grid management and planning in the context of increasing renewable energy integration.

5.2.2.1 Validation Results from ISONE

In this section, we implemented the proposed framework for estimating MDPVS using the total system load data from ISONE for the years 2018 and 2019.

As the first step (net load profile generation), we use the historical load data and BTM PV data from 2017 to generated 200 synthetic net load profiles. This was achieved by incrementally adjusting the scaling factor by 0.005, resulting in 200 data points for net nRMSE and corresponding MDPVS values.

Next, in the modeling step, we use the data points generated in the previous step, each with different net nRMSE and corresponding MDPVS values to establish a net load estimating model. In order to model the relationship between the net nRMSE and MDPVS, we test three different linear regression models ($B1 - B3$) with different sets of input variables. The models and the model equations are defined in Table 5.1. In the first model $B1$, we use net nRMSE as the input to estimate the MDPVS. In model $B2$, we take the square root of net nRMSE and use that to model MDPVS. Finally, in model $B3$, we use the log of net nRMSE to estimate the MDPVS.

Table 5.1: Tested MDPVS estimation models

Model	**Model Equation**
B1	$MDPVS = \beta_0 + \beta_1 Net\, nRMSE$
B2	$MDPVS = \beta_0 + \beta_1 \sqrt{Net\, nRMSE}$
B3	$MDPVS = \beta_0 + \beta_1 \log(Net\, nRMSE)$

Table 5.2: MDPVS estimation results for ISONE for validation years (2018 & 2019)

Year	**Actual (%)**	**B1 (%)**	**B2 (%)**	**B3 (%)**
2018	1.67	1.59	0.98	0.63
2019	2.04	2.65	2.66	3.01
Avg. Absolute Error		0.34	0.65	1.00

Finally, in the PV estimation step, using the regression model developed in Step 2, we estimated the MDPVS for 2018 and 2019. We performed an in-sample fit and computed the net nRMSE for both years using our load and net load forecasting models. . To do this, we do an in-sample fit and calculate the net nRMSE using the load and the net load model for 2018 and 2019. We then use the calculated net nRMSE values form 2018 and 2019 and the fitted linear regression model to estimate the MDPVS.

Table 5.2 shows the actual and estimated PV penetration as measured by MDPVS, for two test years (2018 and 2019) and the average absolute error value, compared to the actual MDPVS for the three models. We can observe that model $B1$ that uses the net nRMSE as the input variable outperforms the other two models in estimating the MDPVS in the system for both the years, with an average absolute error of around 0.34%. For 2018, the model slightly underestimated the actual MDPVS, with an estimated value of 1.59% compared to the actual value of 1.67%. However, the difference between the actual and estimated values is less than 5% of the actual value. For 2019, the model slightly overestimated the actual PV penetration, with an estimated value of 2.65% compared to the actual value of 2.04%. This suggests that the model shows high performance in estimating the average PV penetration for the two years. Hence, we recommend using model $B1$ to estimate the MDPVS in the system. The recommended model can be defined as:

$$MDPVS = \beta_0 + \beta_1 NetnRMSE \tag{5.5}$$

5.2.2.2 Results from Aggregated Data from a Mid-Sized US Utility

In order to further test the validity of our proposed framework, we conduct an additional case study using the aggregated data from a mid-sized utility in the United States. This case study represents a scenario in which actual PV generation data is not available, which is often the case in the real world. As detailed in Section 5.2.1.2 we use simulated PV generation data from NREL's SAM tool to obtain the PV data, which we then use to estimate a linear regression model. We use the data from 2016 to generate the time series and test the performance of the model on the 2017 and 2018 data.

Table 5.3: MDPVS estimation results for ISONE for test years (2017 & 2018) for a medium-sized US utility

Year	**Actual MDPVS (%)**	**Estimated MDPVS (%)**
2017	11.01	11.59
2018	12.69	11.19

The actual and estimated MDPVS values using model $B1$ for the two test years are presented in Table 5.3. The results indicate that the proposed model accurately estimates the MDPVS, consistent with our previous findings for the ISONE case study. On average, the proposed model estimates approximately 11.85% MDPVS, while the actual MDPVS is 11.39% for the two years. These findings provide strong evidence of the efficacy of our proposed methodology in estimating MDPVS in a system, even in cases where actual BTM PV data is unavailable.

The proposed framework has demonstrated consistent accuracy across different scenarios, including both a regional transmission organization like ISO New England and a mid-sized U.S. utility. This consistency is particularly notable given the challenges associated with the unavailability of actual PV generation data, a common obstacle in renewable energy studies. The utilization of NRELâs SAM tool for generating simulated data exemplifies an innovative approach to overcoming data limitations, enhancing the framework's applicability in real-world settings where actual data may not be accessible. The success of this framework not only contributes to the academic field of energy system modeling but also holds significant implications for utility managers and policymakers in planning and optimizing the integration of PV systems into the grid.

5.3 A Regression-Based Framework for Behind-the-Meter PV Detection

The detection of BTM PV installations is a significant research challenge with practical implications for utilities and energy providers. The ability to identify meters equipped with PV systems can offer valuable insights into solar energy penetration and facilitate efficient grid management. In this context, this section presents a methodological framework that leverages previously established benchmark load and net load forecasting models to distinguish meters with BTM PV installations.

5.3.1 Data Description

The meter-level data set used in work comes from a medium-sized utility in the US. The electricity consumption data set contains hourly load values for 800 residential meters. The data set includes hourly load data and corresponding weather data spanning a three-year period from January 1, 2016, to December 31, 2018. To generate the simulated PV generation data, we again make use of NREL's SAM tool.

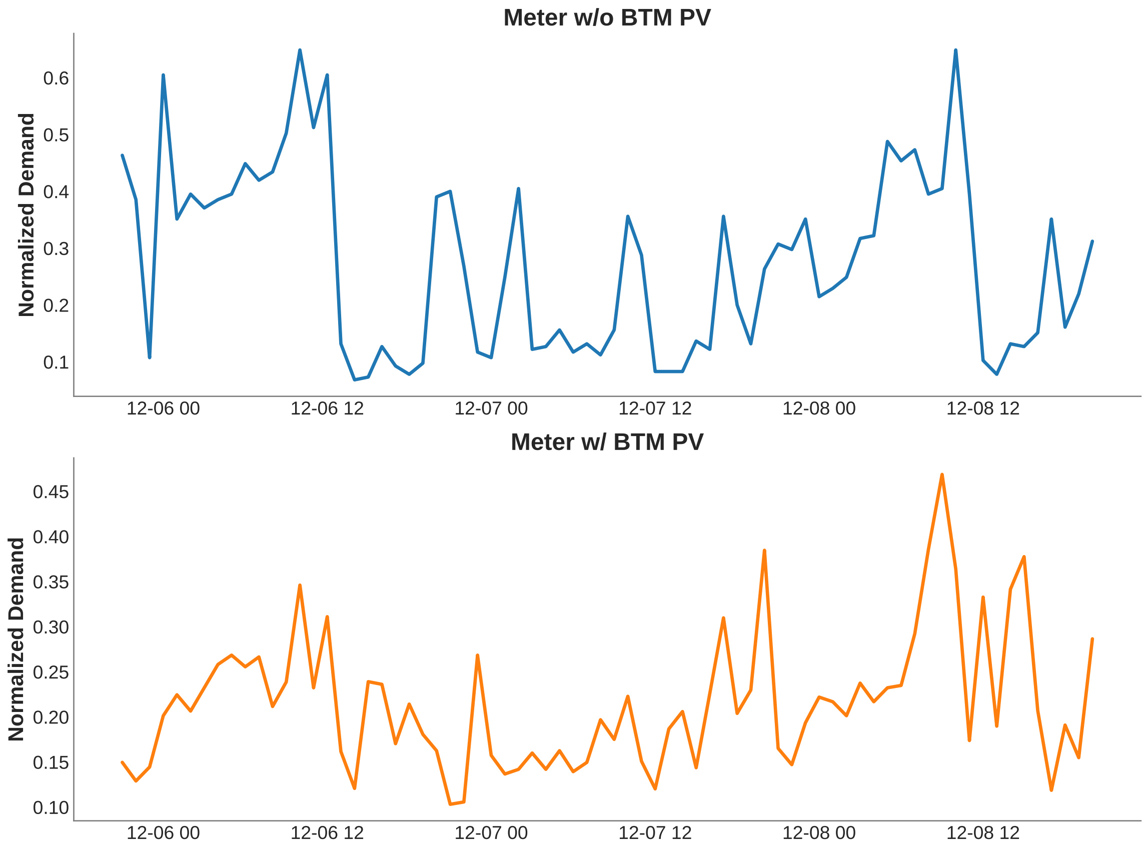


Figure 5.5: Meter-level load time series for a week for two meters

Figure 5.5 shows the meter-level load and net load time series for two meters for a week. The time series in the upper plot shows the load for a meter without rooftop PV, while the lower plot shows the net load for a meter with PV installed. Unlike the load profiles at the high voltage levels, the load profiles at the meter level are more susceptible to fluctuations and customer behaviors. Hence, just by visually comparing the two plots, it is very hard to tell which meters have BTM PV and which do not.

5.3.2 Proposed Framework

The primary objective of this section is to address a central research question: "Given only the meter-level data, can we identify the meters that have behind-the-meter PV installations?" To address this question, we once again propose a comparative analysis approach that utilizes two forecasting models, namely, a load forecasting model (Vanilla model) and a net load forecasting model (proposed benchmark net load forecasting model, introduced earlier in Section 4.3). By examining the disparities between the forecasts generated by these models, the proposed framework aims to identify meters with BTM PV installations.

Based on the empirical case studies detailed in Section 4.3.5 we observe a noticeable disparity between the performance of two forecasting models. The net load forecasting model demonstrated superior performance over the load forecasting model, specifically when PV was integrated into the system, that is, for net load. Consequently, the proposed methodology is built on a comparative analysis of the predictions generated by these two models, with the aim of identifying meters that have BTM PV installations.

The central idea is that in-sample forecasts from both the load and net load forecasting models should exhibit similar characteristics for meters that do not have BTM PV installations, thereby yielding a lower net nRMSE value. Conversely, for meters with BTM PV installations, the net load forecasting model would yield a lower nRMSE value in comparison to the load forecasting model, subsequently resulting in a greater net nRMSE value. Hence, by evaluating the net nRMSE of the meters in question, it should be possible to identify meters with BTM PV installations

A high-level workflow of the proposed approach is illustrated in Figure 5.6. We begin by employing one year of load/net load data for each meter. Subsequently, we apply Tao's Vanilla load benchmark model (as formulated in Equation 4.8) and the net load benchmark model (as defined in Equation 4.3.4), on a per-meter basis. Subsequently, the in-sample nRMSE values are computed for each meter. Following this, the net nRMSE values are calculated by subtracting the nRMSE of the load forecasting model from the nRMSE of the net load forecasting model, as detailed in Equation 5.4. We then analyze the net nRMSE values. Based on a threshold value, we classify the meters as 'with PV' and 'without PV'. If the threshold value is more than the net nRMSE for the meter, it is classified as 'without PV' and if the net nRMSE for the meter is greater than the threshold, we classify it as 'with PV'. The selection of an optimal threshold value is critical, as it influences the classification accuracy. Furthermore, this threshold may need adjustment based on new data sets. A method to calculate the optimal threshold value is proposed by performing a sensitivity analysis, ensuring accurate classification based on the observed net nRMSE values.

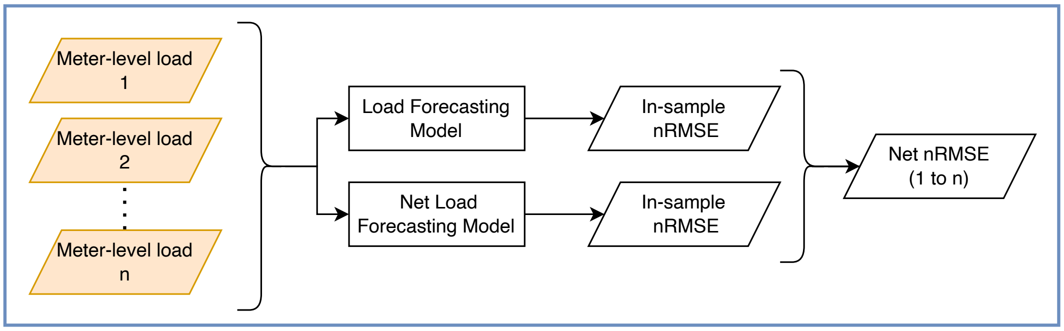


Figure 5.6: High-level workflow for the BTM PV detection model

5.3.2.1 Sensitivity Analysis for Threshold

Correctly identifying the threshold value is a crucial step in the classification of a meter as 'with PV' or 'without PV'. To do that, we take a sample of known meters from the area and use them to calculate the optimal threshold value. We assume that for these meters, the information regarding which meters have BTM installed is known. Here we consider a total of 100 such meters, out of which 50 random meters have BTM PV installed and 50 do not.

We begin by calculating the in-sample nRMSEs using the load model and the net load model. Next, we calculate the net nRMSE for all 100 meters. Next, we plot and analyze the distribution of the net nRMSE values. Figure 5.7 shows the histogram of the net nRMSE values of the 100 meters. It can be observed that the net nRMSE values vary between 0 and 0.008, with a large number of the meters having a relatively lower net nRMSE value. This gives us an indication of the possible threshold values that can be used to separate the meters.

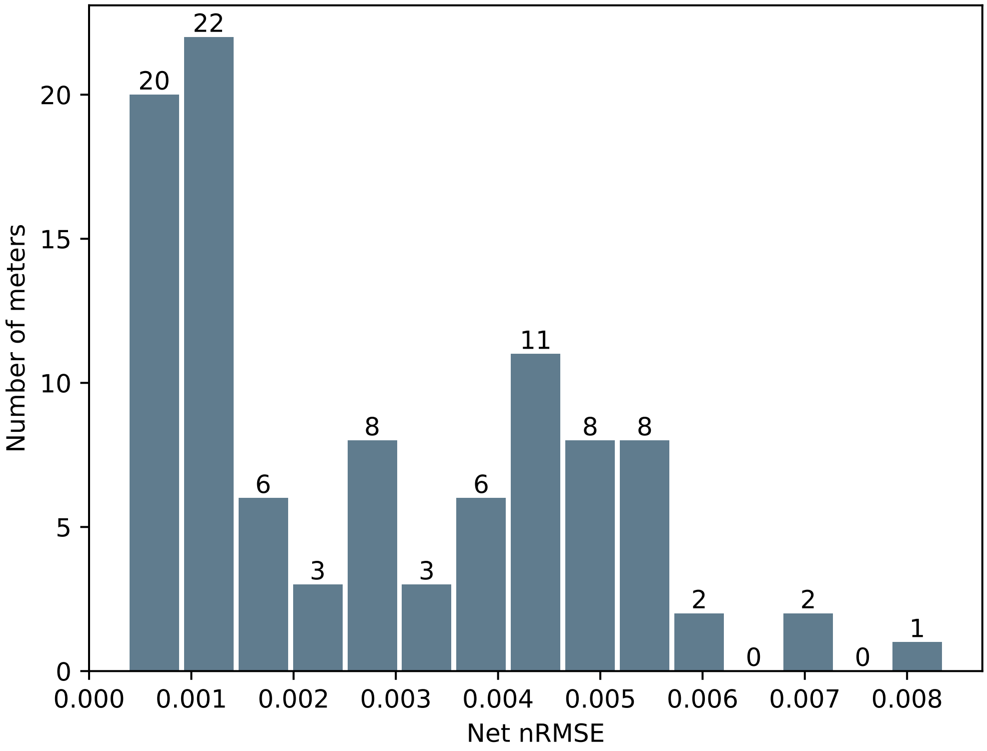


Figure 5.7: Histogram of the net nRMSE values for 100 meters

To do this, we select a range of potential threshold values based on the distribution of the net nRMSE values. In this case, we vary the threshold from 0.0010 to 0.0050 with an increment of 0.0005. This will vary with each new dataset.

Based on each threshold value, we can assign labels to each meter, identifying them as 'with PV' or 'without PV'. We can then evaluate the performance of each threshold value by comparing the label assigned to the actual label. A common method of assessing classification problems is to look at the confusion matrix. A confusion matrix provides a summary the number of correct and incorrect predictions with count values and broken down by each class. However, in this case it can be confusing to look at each element of the confusion matrix in order to decide the optimal threshold value. Hence, in this case we summarize the values of the confusion matrix by calculating the accuracy score. The accuracy score offers insights into the ratio of correct predictions. This metric is widely used due to its simplicity in calculation and interpretation, along with its ability to quantify the model's performance using a single figure. The accuracy is computed as:

$$\text{Accuracy} \;=\; \frac{\text{TN} \;+\; \text{TP}}{\text{TP} \;+\; \text{FP} \;+\; \text{TN} \;+\; \text{FN}} * 100 \tag{5.6}$$

Here, TP or True Positives signifies correctly classified samples from the positive class. TN or True Negatives represent correctly classified samples from the negative class. FP or False Positives describe samples wrongly classified as belonging to the positive class while they belong to the negative class. FN or False Negatives stand for samples wrongly classified as belonging to the negative class while they actually belong to the positive class.

Table 5.4 shows the accuracy values for the threshold values tested for the 100 known meters. It can be observed that the value of 0.0020 gives the highest accuracy among the threshold values tested. This indicates that using 0.0020 as the threshold value, out of the 100 predictions made by the model, 94 were correctly identified, and 6 were incorrectly identified. This shows promising results considering that the model was provided only the historical load information.

Table 5.4: Accuracy values (in %) for the threshold values tested for 100 known meters

Threshold	**Accuracy (%)**
0.0010	69.00
0.0015	89.00
0.0020	94.00
0.0025	91.00
0.0030	87.00
0.0035	86.00
0.0040	83.00
0.0045	73.00
0.0050	68.00

5.3.3 Results from Meter-Level Data from a Mid-Sized US Utility

In order to put our proposed methodology to the test, we employ meter-level data derived from a mid-sized US utility, as elaborated in Section 5.3.1. We use a threshold value of 0.002, obtained from the set of known PV meters. While actual PV generation data was not available for this data set, we used simulated PV generation data from NREL's SAM tool to simulate the PV generation profiles. As none of the residential units in the data set had BTM PV installations, we combined the residential load measurements with the simulated PV generation data to create various customer net load profiles.

In order to add PV to the meters, we first need to calculate the size of the PV installation for each meter. The sizing of the PV capacity that homeowners typically opt for can be influenced by a range of factors, including their previous electricity usage, their target bill offset, and the available space on their rooftops. While some homeowners may opt for a system sized to offset their entire annual consumption, others may prefer systems that offset their consumption during peak hours. As a result, there is no straightforward heuristic for sizing a PV system.

$$\text{PV Size (kW)} = \frac{\text{Total Annual Consumption (kWh)/365}}{\text{Average Daily PSH (h)}\eta_{\text{system}}} \tag{5.7}$$

In the context of the present study, a simple method based on the general location of the data is used to determine the size of the PV system, detailed in Equation 5.7. The first step involves dividing the total annual kilowatt-hour (kWh) consumption by 365, to get the average daily consumption in kWh. This average daily consumption is then divided by the average number of daily peak sun hours (PSH) experienced in the given location multiplied by the overall system efficiency (η_{system}). For simplicity, the system efficiency is assumed to be 1. The resulting value provides an estimate of the required size of the array in kilowatts needed to offset the annual energy consumption. For the location considered in this study, the average daily PSH was found to be 5 hours.

We evaluate the performance of the proposed methodology on the 800 meters for the year 2019. We add PV to randomly selected meters ranging from 100 meters to 600 meters meters. This approach allows us to test and validate the effectiveness of the proposed methodology in a diverse range of meters with PV installations.

The results of the analysis are summarized in Table 5.5. The table provides key findings for 800 meters, including the number of meters with PV installations, and the percentages of correctly identified meters with PV (True Positives) and without PV (True Negative), as well as falsely identified meters (False Positives + False Negatives).

The results demonstrate high accuracy in detecting meters with PV installations, with percentages ranging from 88% to 91%. This high accuracy empowers utilities to precisely identify households with PV systems, providing valuable insights into the penetration of solar energy within their service areas. Furthermore, the methodology consistently exhibits high accuracy in correctly identifying meters without PV installations, with percentages ranging from 95% to 98%. This accuracy is crucial for

Table 5.5: BTM PV detection results for 800 meters from a medium-sized utility in the US for the test year (2019)

Meters w/ PV	**Meters w/PV Correctly Identified (%)**	**Meters wo/ PV Correctly Identified (%)**	**Meters Falsely Identified (%)**
100	91	95	6
200	89	96	6
300	89	96	7
400	88	98	7
500	89	98	8
600	90	98	8

utilities to avoid misidentifying non-PV households as PV-equipped. Additionally, a very low percentage of the meters were falsely identified, ranging from 6% to 8%.

These findings confirm the efficacy of the proposed methodology in differentiating meters with BTM PV systems from those without. By leveraging the disparities between load forecasting models and net load forecasting models, the methodology achieves high accuracy in identifying PV installations. The comprehensive analysis of a diverse range of meter configurations strengthens the robustness of the methodology's performance evaluation.

5.4 Conclusion

The research presented in this study focuses on the estimation of PV penetration in electricity systems and the identification of meters with BTM PV installations. The aim was to develop robust methodologies that can provide accurate information on the integration of PV systems into the grid and allow utilities to effectively manage renewable energy resources.

The initial part of the study addresses the challenge of estimating PV penetration in systems where this information may not be available to the utility. Existing supervised and unsupervised methods are reviewed, revealing their limitations, particularly when information on PV installations is unknown or limited. To overcome these challenges, a regression-based approach is proposed based on existing forecasting models. By learning the relationship between the discrepancy between the load and the net load forecasting models and the corresponding PV penetration on synthetic net load profiles, we develop a regression-based methodology for estimating PV penetration, which is successfully applied to real-world scenarios. The results demonstrate the efficacy of the proposed approach in accurately estimating PV penetration levels, even in the absence of explicit PV installation information, using only historical load data.

The second part of the study addresses the identification of meters with BTM PV installations. Leveraging the performance disparities between load forecasting models and net load forecasting models, a methodology is devised to differentiate meters with and without PV installations. The results show that an accurate identification of meters with PV installations was achieved while maintaining a low rate of false identifications. This methodology provides valuable information for utilities, enabling them to understand the adoption and impact of distributed solar energy within their service areas.

CHAPTER 6: CONCLUSION

6.1 Overview

This chapter summarizes the work done in the dissertation, the key contributions, and the prospects for future research.

6.2 Concluding Remarks

The work presented in this dissertation addresses the pivotal shift towards data-centric operations and the adoption of renewable energy sources, notably solar photovoltaics (PV) in the modern power system by proposing data-driven approaches to tackle significant forecasting challenges: predicting weather-related power outages, forecasting net load, and estimating solar PV penetration accurately.

The study presents a forecasting methodology for predicting day ahead weather-related power outages on an hourly basis. We address the inherent data imbalance issue by proposing a weighted logistic regression model and allocating different weights for the outage and non-outage classes based on the reciprocals of their respective number of hours. The best-performing inputs for the forecasting model are chosen using a variable selection technique. The proposed model is used to forecast weather-related outages aggregated from the distribution substation level up to the city level. The out-of-sample tests showed that for both cases, the proposed model outperforms a simple logistic regression model and a logistic regression model with an optimized threshold.

Secondly, this dissertation presents an in-depth study of net load forecasting, beginning with an extensive review of the existing literature in the field. Various forecasting approaches, techniques, explanatory variables, the impact of PV penetration on net load forecasting, and PV estimation methods are discussed. The study's findings highlight a notable rise in the number of related research published annually, marking this as a growing field of study. Furthermore, with an increasing amount of historical net load data available, more researchers are starting to look into the net load forecasting problem. As part of the review of the existing literature on net load forecasting, several gaps and potential avenues for future research were also identified, including a lack of reproducible research, a lack of a benchmark net load forecasting model, and a lack of use of benchmark open-source datasets.

This dissertation aims to address these gaps by presenting an MLR-based benchmark model for net load forecasting. The proposed model is a benchmark net load forecasting model that is interpretable, reproducible, and leverages existing load forecasting techniques and extends them to the domain of net load forecasting. Furthermore, the proposed model is developed using open-source data for easy reproducibility. Additionally, we showcase the performance of our proposed model for net load forecasting under varying levels of PV penetration, thereby highlighting the impact of PV penetration on net load forecast performance. The proposed solutions offer practical value in field operations. Through empirical case studies on three ISOs in the US, the results show that the net load forecast accuracy can be significantly improved by the proposed model compared to the existing load forecasting model.

The second part of the study focuses on addressing two important questions: "What is the amount of PV penetration in the system?" and "Which meters have behind-the-meter PV installations?". Most of the existing literature builds and tests PV capacity estimation models and BTM PV identification models using limited PV generation information in some form or the other. However, the robustness of these models under operational settings has rarely been studied, where information on historical PV generation is rarely available independent of the net load. This dissertation addresses this gap by presenting an approach that takes advantage of existing models for load and net load forecasting, without the need for additional PV information. The proposed approach uses the difference in performance between the two models to develop a regression-based approach to estimate PV penetration in the system. Through empirical case studies on data from an ISO and a medium-sized utility in the US, the results demonstrate that high accuracy in estimating PV penetration levels can be achieved using only historical load data. Furthermore, the approach is easy to implement and offers practical values in field operations.

This dissertation also presents a methodology to identify the meters with BTM PV installation using net load data. The proposed framework builds on the central idea that the in-sample forecasts from the load and net load forecasting models should be similar for meters without PV, while for the meters with PV, the forecasts from the net load forecasting model should be more accurate than the forecasts from the load forecasting model. Hence, by evaluating the disparities between the load and net load forecasting models, it should be possible to identify meters with BTM PV installations. Practical testing of this approach, using meter-level data from a mid-sized US utility, with varying numbers of PV installations, demonstrated a consistently high detection rate of meters with and without PV. The low percentage of falsely identified meters further underscores the effectiveness of this approach. This research contributes valuable insights for utilities and energy providers, aiding in grid management and understanding the penetration of solar energy within their service areas.

A major point kept in mind in the work presented in this dissertation is the ease of implementation, interpretability, and reproducibility of the work. First and foremost, the regression-based techniques have been proposed in this work since regression analysis is interpretable and easy to implement in most tools available in the market. Secondly, as much as possible the data used in this dissertation is real-world data that is openly available. Thirdly, the frameworks proposed in this dissertation are explained in detail making them easily reproducible by fellow researchers.

In this dissertation, a strong emphasis is placed on making sure that the methods used are easy to apply, clear to understand, and simple to reproduce. The focus on regression-based techniques is a key aspect of this, as these methods are well-known for their interpretability and ease of use with widely available industry tools. Emphasizing practicality, the research primarily uses real-world data that is publicly available. This choice enhances the real-life applicability and relevance of the study. Additionally, the frameworks developed in this research are thoroughly explained, ensuring that other researchers can easily understand and reproduce the results obtained in this work. This approach highlights the dissertation's contribution toward setting a benchmark in the field for research that is practical, clear, and reproducible in the area of energy forecasting.

6.3 Summary of Contributions

In this work, we conducted a comprehensive literature review on outage forecasting and net load forecasting, areas vital for improving grid management and stability. This included an in-depth analysis of state-of-the-art outage forecasting methods, particularly focusing on weather-related forecasting techniques, and a thorough examination of net load forecasting approaches, techniques, variables, and error metrics. Special emphasis was placed on understanding the effects of photovoltaic (PV) system penetration on net load forecasting, a crucial area impacting grid management and stability. The review identified gaps and provided directions for framing research questions for this dissertation.

The key contributions of this study are as follows:

1. We developed a day-ahead weather-related outage forecasting model that addresses data imbalances using a weighted logistic regression approach. This model was rigorously tested against traditional logistic regression methods and logistic regression with optimized thresholds in two distinct case studies: one

at the city level and another at the distribution substation levelâan area less explored in academic research. The new model significantly outperformed traditional methods, demonstrating improvements of 100% at the city level and 117% at the substation level compared to simple logistic regression, and 2% and 12% improvements respectively compared to the optimized threshold model. This enhancement in forecasting accuracy is crucial for utility companies in improving response strategies and operational planning during outage events.

2. A benchmarking framework for net load forecasting was proposed, developed from insights gathered in the literature review and designed to address identified gaps in existing methods. The proposed model serves as a new benchmark, contributing substantially to the development and improvement of forecasting practices. It demonstrated an average improvement of 50% in accuracy for net load forecasts with 30% PV penetration compared to existing models, proving its efficacy in handling renewable integration.

3. Addressing critical questions about the extent of PV penetration and the identification of meters with BTM PV installations, we developed a framework that leverages differences between load and net load model performances. The framework developed for estimating PV penetration achieved estimates with an absolute error under 0.5%, while the detection framework for identifying meters with BTM PV installations achieved approximately 90% accuracy. This capability provides utilities with essential insights into the distribution of distributed solar energy, enhancing their ability to manage and plan grid operations effectively.

These contributions significantly advance the practical application of forecasting methodologies, offering electric utilities robust tools for enhancing grid reliability and accommodating the growing integration of renewable energy sources. By improving the accuracy and applicability of these forecasts, this work aids utilities in making informed decisions that support sustainable and efficient energy systems.

6.4 Future Work

This section discusses some ideas to extend the work presented in this dissertation as well as its limitations.

- Regarding weather-related outage forecasting, the current work offers an ex-post outage forecasting model. In future studies, ex-ante forecasts can be explored. An area of interest for future analytical research may also be to investigate how lagged and moving averages of meteorological data can enhance outage forecasting models.

- Additionally, the climatic conditions in a certain place may differ from those in the surrounding areas. Choosing the best weather stations for each location has the potential to give more accurate and relevant weather information, as well as improve the forecast model's performance. A weather station selection process may be adopted to complement the current work in future research to enhance the performance of the models. In some instances, based on real-world occurrences, certain historical outage incidents could be misclassified.

- The currently proposed benchmark net load forecasting model serves as a foundational model with scope for further improvement. The weather variables considered in this research are limited to temperature and GHI. The inclusion of additional weather variables such as cloud cover, wind speed, and humidity, among others, could potentially enhance the model's performance. The model could also benefit from the utilization of data from multiple weather stations.

- Finally, concerning PV capacity estimation, the current method employs a year's worth of data to train the estimation model. Future studies might evaluate the sensitivity related to the duration of training history necessary for the model.

REFERENCES

Alipour, M., Aghaei, J., Norouzi, M., Niknam, T., Hashemi, S., and Lehtonen, M. (2020). A novel electrical net-load forecasting model based on deep neural networks and wavelet transform integration. *Energy*, 205:118106.

Antonanzas, J., Osorio, N., Escobar, R., Urraca, R., Martinez-de Pison, F. J., and Antonanzas-Torres, F. (2016). Review of photovoltaic power forecasting. *Solar energy*, 136:78–111.

Arlot, S. and Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistics Surveys*, 4:40–79.

Bashkari, M. S., Sami, A., and Rastegar, M. (2020). Outage cause detection in power distribution systems based on data mining. *IEEE Transactions on Industrial Informatics*, 17(1):640–649.

Benner, C. L., Peterson, R. A., and Russell, B. D. (2017). Application of dfa technology for improved reliability and operations. In *2017 IEEE Rural Electric Power Conference (REPC)*, pages 44–51. IEEE.

Bhattacharyya, A., Yoon, S., and Hastak, M. (2021). Economic impact assessment of severe weather–induced power outages in the us. *Journal of Infrastructure Systems*, 27(4):04021038.

Black, J., Hoffman, A., Hong, T., Roberts, J., and Wang, P. (2018). Weather data for energy analytics: From modeling outages and reliability indices to simulating distributed photovoltaic fleets. *IEEE Power and Energy Magazine*, 16(3):43–53.

Black, J. D. and Henson, W. L. W. (2014). Hierarchical load hindcasting using reanalysis weather. *IEEE Transactions on Smart Grid*, 5(1):447–455.

Boretti, A. (2024). Investigating the correlation between extreme temperature events and the increasing frequency of power outages in the northern territory of australia. *International Journal of Electrical Power & Energy Systems*, 155:109608.

Bouckaert, S., Pales, A. F., McGlade, C., Remme, U., Wanner, B., Varro, L., D'Ambrosio, D., and Spencer, T. (2021). Net zero by 2050: A roadmap for the global energy sector.

Bradbury, K., Saboo, R., L Johnson, T., Malof, J. M., Devarajan, A., Zhang, W., M Collins, L., and G Newell, R. (2016). Distributed solar photovoltaic array location and extent dataset for remote sensing object identification. *Scientific data*, 3(1):1–9.

Cai, Y. and Chow, M.-Y. (2011). Cause-effect modeling and spatial-temporal simulation of power distribution fault events. *IEEE Transactions on Power Systems*, 26(2):794–801.

Campbell, R. J. and Lowry, S. (2012). Weather-related power outages and electric system resiliency. Congressional Research Service, Library of Congress Washington, DC.

Caswell, H. C., Forte, V. J., Fraser, J. C., Pahwa, A., Short, T., Thatcher, M., and Werner, V. G. (2011). Weather normalization of reliability indices. *IEEE Transactions on Power Delivery*, 26(2):1273–1279.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). Smote: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357.

Chen, D. and Irwin, D. (2017). Sundance: Black-box behind-the-meter solar disaggregation. pages 45–55.

Cheng, D., Mather, B. A., Seguin, R., Hambrick, J., and Broadwater, R. P. (2015). Photovoltaic (PV) impact assessment for very high penetration levels. *IEEE Journal of photovoltaics*, 6(1):295–300.

Chu, Y., Pedro, H. T., Kaur, A., Kleissl, J., and Coimbra, C. F. (2017). Net load forecasts for solar-integrated operational grid feeders. *Solar Energy*, 158:236–246.

Costa, A., Crespo, A., Navarro, J., Lizcano, G., Madsen, H., and Feitosa, E. (2008). A review on the young history of the wind power short-term prediction. *Renewable and Sustainable Energy Reviews*, 12(6):1725–1744.

Das, S., Kankanala, P., and Pahwa, A. (2021). Outage estimation in electric power distribution systems using a neural network ensemble. *Energies*, 14(16):4797.

Data Package Household Data (2020). Open Power System Data. `https://data.open-power-system-data.org/`.

Dehbozorgi, M. R., Rastegar, M., and Dabbaghjamanesh, M. (2020). Decision tree-based classifiers for root-cause detection of equipment-related distribution power system outages. *IET Generation, Transmission & Distribution*, 14(24):5809–5815.

Denholm, P., Margolis, R., and Milford, J. (2008). Production cost modeling for high levels of photovoltaics penetration. Technical report, National Renewable Energy Lab.(NREL), Golden, CO (United States).

Denholm, P., O'Connell, M., Brinkman, G., and Jorgenson, J. (2015). Overgeneration from solar energy in california. a field guide to the duck chart. Technical report, National Renewable Energy Lab.(NREL), Golden, CO (United States).

Doostan, M. and Chowdhury, B. (2020). Predicting lightning-related outages in power distribution systems: A statistical approach. *IEEE Access*, 8:84541–84550.

Doostan, M., Sohrabi, R., and Chowdhury, B. (2020). A data-driven approach for predicting vegetation-related outages in power distribution systems. *International Transactions on Electrical Energy Systems*, 30(1):e12154.

Ekisheva, S., Rieder, R., Norris, J., Lauby, M., and Dobson, I. (2021). Impact of extreme weather on north american transmission system outages. In *2021 IEEE Power & Energy Society General Meeting (PESGM)*, pages 01–05. IEEE.

Erdener, B. C., Feng, C., Doubleday, K., Florita, A., and Hodge, B.-M. (2022). A review of behind-the-meter solar forecasting. *Renewable and Sustainable Energy Reviews*, 160:112224.

Eskandarpour, R. and Khodaei, A. (2017). Leveraging accuracy-uncertainty tradeoff in svm to achieve highly accurate outage predictions. *IEEE Transactions on Power Systems*, 33(1):1139–1141.

Fan, S. and Hyndman, R. J. (2011). Short-term load forecasting based on a semi-parametric additive model. *IEEE transactions on power systems*, 27(1):134–141.

Ferreira, V. H., CorrÃªa, R., Colombini, A., Fortes, M., Mello, F., Araújo, F., and Pereira, N. (2021). Big data analytics for spatio-temporal service orders demand forecasting in electric distribution utilities. *Energies*, 14.

Freeman, J., Whitmore, J., Blair, N., and Dobos, A. P. (2014). Validation of multiple tools for flat plate photovoltaic modeling against measured data. In *2014 IEEE 40th Photovoltaic Specialist Conference (PVSC)*, pages 1932–1937. IEEE.

Garner, R. and Dunbar, B. (2008). Solar irradiance. *Retrieved from NASA TV: https://www. nasa. gov/mission_pages/sdo/science/solar-irradiance. html (2008, January 2).*

Gilman, P., DiOrio, N. A., Freeman, J. M., Janzou, S., Dobos, A., and Ryberg, D. (2018). Sam photovoltaic model technical reference 2016 update. Technical report, National Renewable Energy Lab.(NREL), Golden, CO (United States).

Gross, G. and Galiana, F. D. (1987). Short-term load forecasting. *Proceedings of the IEEE*, 75(12):1558–1573.

Gui, M., Pahwa, A., and Das, S. (2009). Analysis of animal-related outages in overhead distribution systems with wavelet decomposition and immune systems-based neural networks. *IEEE Transactions on Power Systems*, 24(4):1765–1771.

He, J., Wanik, D. W., Hartman, B. M., Anagnostou, E. N., Astitha, M., and Frediani, M. E. (2017). Nonparametric tree-based predictive modeling of storm outages on an electric distribution network. *Risk Analysis*, 37(3):441–458.

Hippert, H. S., Pedreira, C. E., and Souza, R. C. (2001). Neural networks for short-term load forecasting: A review and evaluation. *IEEE Transactions on power systems*, 16(1):44–55.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.

Hoke, A., Butler, R., Hambrick, J., and Kroposki, B. (2012a). Maximum photovoltaic penetration levels on typical distribution feeders. Technical report, National Renewable Energy Lab.(NREL), Golden, CO (United States).

Hoke, A., Butler, R., Hambrick, J., and Kroposki, B. (2012b). Steady-state analysis of maximum photovoltaic penetration levels on typical distribution feeders. *IEEE Transactions on Sustainable Energy*, 4(2):350–357.

Hong, T. (2010). *Short term electric load forecasting*. PhD thesis.

Hong, T. and Fan, S. (2016). Probabilistic electric load forecasting: A tutorial review. *International Journal of Forecasting*, 32(3):914–938.

Hong, T. and Hofmann, A. (2022). Data integrity attacks against outage management systems. *IEEE Transactions on Engineering Management*, 69(3):765–772.

Hong, T., Pinson, P., and Fan, S. (2014). Global energy forecasting competition 2012. *International Journal of Forecasting*, 30(2):357–363.

Hong, T., Pinson, P., Fan, S., Zareipour, H., Troccoli, A., and Hyndman, R. J. (2016). Probabilistic energy forecasting: Global energy forecasting competition 2014 and beyond. *International Journal of Forecasting*, 32(3):896–913.

Hong, T., Pinson, P., Wang, Y., Weron, R., Yang, D., and Zareipour, H. (2020). Energy forecasting: A review and outlook. *IEEE Open Access Journal of Power and Energy*, 7:376–388.

Hong, T. and Wang, P. (2014). Fuzzy interaction regression for short term load forecasting. *Fuzzy optimization and decision making*, 13:91–103.

Hong, T., Wang, P., and White, L. (2015). Weather station selection for electric load forecasting. *International Journal of Forecasting*, 31(2):286–295.

Hong, T., Xie, J., and Black, J. (2019). Global energy forecasting competition 2017: Hierarchical probabilistic load forecasting. *International Journal of Forecasting*, 35(4):1389–1399.

Hou, H., Zhu, S., Geng, H., Li, M., Xie, Y., Zhu, L., and Huang, Y. (2021). Spatial distribution assessment of power outage under typhoon disasters. *International Journal of Electrical Power & Energy Systems*, 132:107169.

Høverstad, B. A., Tidemann, A., Langseth, H., and Öztürk, P. (2015). Short-term load forecasting with seasonal decomposition using evolution for parameter tuning. *IEEE Transactions on Smart Grid*, 6(4):1904–1913.

IEA (2019). Renewables 2019 - analysis.

IEA (2020). Renewables 2020 - analysis.

Inman, R. H., Pedro, H. T., and Coimbra, C. F. (2013). Solar forecasting methods for renewable energy integration. *Progress in energy and combustion science*, 39(6):535–576.

ISONE (2023). ISO New England.

Kabir, E., Guikema, S. D., and Quiring, S. M. (2019a). Predicting thunderstorm-induced power outages to support utility restoration. *IEEE Transactions on Power Systems*, 34(6):4370–4381.

Kabir, F., Yu, N., Yao, W., Yang, R., and Zhang, Y. (2019b). Estimation of behind-the-meter solar generation by integrating physical with statistical models. In *2019 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm)*, pages 1–6.

Kankanala, P., Das, S., and Pahwa, A. (2013). Adaboost $^+$: An ensemble learning approach for estimating weather-related outages in distribution systems. *IEEE Transactions on Power Systems*, 29(1):359–367.

Kaur, A., Nonnenmacher, L., and Coimbra, C. F. (2016). Net load forecasting for high renewable energy penetration grids. *Energy*, 114:1073–1084.

Kaur, A., Pedro, H. T., and Coimbra, C. F. (2013). Impact of onsite solar generation on system load demand forecast. *Energy conversion and management*, 75:701–709.

Kezunovic, M., Obradovic, Z., Dokic, T., Zhang, B., Stojanovic, J., Dehghanian, P., and Chen, P.-C. (2017). Predicting spatiotemporal impacts of weather on power systems using big data science. *Data Science and Big Data: An Environment of Computational Intelligence*, pages 265–299.

Killinger, S., Lingfors, D., Saint-Drenan, Y.-M., Moraitis, P., Van Sark, W., Taylor, J., Engerer, N. A., and Bright, J. M. (2018). On the search for representative characteristics of pv systems: Data collection and analysis of pv system azimuth, tilt, capacity, yield and shading. *Solar Energy*, 173:1087–1106.

King, G. and Zeng, L. (2001). Logistic regression in rare events data. *Political Analysis*, 9(2):137–163.

Kleinbaum, D. G., Dietz, K., Gail, M., Klein, M., and Klein, M. (2002). *Logistic regression*. Springer.

Kobylinski, P., Wierzbowski, M., and Piotrowski, K. (2020). High-resolution net load forecasting for micro-neighbourhoods with high penetration of renewable energy sources. *International Journal of Electrical Power & Energy Systems*, 117:105635.

Kordkheili, R. A., Bak-Jensen, B., Jayakrishnan, R., Mahat, P., et al. (2014). Determining maximum photovoltaic penetration in a distribution grid considering grid operation limits. In *2014 IEEE PES General Meeting| Conference & Exposition*, pages 1–5. IEEE.

Kuzlu, M., Cali, U., Sharma, V., and GÃŒler, Ã. (2020). Gaining insight into solar photovoltaic power generation forecasting utilizing explainable artificial intelligence tools. *IEEE Access*, 8:187814–187823.

Lawanson, T., Sharma, V., Cecchi, V., and Hong, T. (2021). Analysis of outage frequency and duration in distribution systems using machine learning. In *2020 52nd North American Power Symposium (NAPS)*, pages 1–6.

Li, K., Wang, F., Mi, Z., Fotuhi-Firuzabad, M., Duić, N., and Wang, T. (2019). Capacity and output power estimation approach of individual behind-the-meter distributed photovoltaic system for demand response baseline estimation. *Applied energy*, 253:113595.

Malof, J. M., Bradbury, K., Collins, L. M., and Newell, R. G. (2016). Automatic detection of solar photovoltaic arrays in high resolution aerial imagery. *Applied energy*, 183:229–240.

Manski, C. F. and Lerman, S. R. (1977). The estimation of choice probabilities from choice based samples. *Econometrica*, 45(8):1977–1988.

Mason, K., Reno, M. J., Blakely, L., Vejdan, S., and Grijalva, S. (2020). A deep neural network approach for behind-the-meter residential pv size, tilt and azimuth estimation. *Solar Energy*, 196:260–269.

Mehra, V. (2016). Solar home system jharkhand india.

Mei, F., Wu, Q., Shi, T., Lu, J., Pan, Y., and Zheng, J. (2019). An ultrashort-term net load forecasting model based on phase space reconstruction and deep neural network. *Applied Sciences*, 9(7):1487.

Mousavi Maleki, S. A., Hizam, H., and Gomes, C. (2017). Estimation of hourly, daily and monthly global solar radiation on inclined surfaces: Models re-visited. *Energies*, 10(1):134.

Nateghi, R., Guikema, S. D., and Quiring, S. M. (2011). Comparison and validation of statistical methods for predicting power outage durations in the event of hurricanes. *Risk Analysis: An International Journal*, 31(12):1897–1906.

Nateghi, R., Guikema, S. D., and Quiring, S. M. (2014). Forecasting hurricane-induced power outage durations. *Natural hazards*, 74(3):1795–1811.

Nielsen, T. (2002). Improving outage restoration efforts using rule-based prediction and advanced analysis. In *2002 IEEE Power Engineering Society Winter Meeting. Conference Proceedings (Cat. No. 02CH37309)*, volume 2, pages 866–869. IEEE.

Oliver, E. and Perfumo, C. (2015). Technical report: Load and solar modelling for the nfts feeders. *Commonwealth Sci. Ind. Res. Organisation (CSIRO), Newcastle, NSW, Australia, Tech. Rep.*

Papalexopoulos, A. and Hesterberg, T. (1990). A regression-based approach to short-term system load forecasting. *IEEE Transactions on Power Systems*, 5(4):1535–1547.

Park, D., El-Sharkawi, M., Marks, R., Atlas, L., and Damborg, M. (1991a). Electric load forecasting using an artificial neural network. *IEEE Transactions on Power Systems*, 6(2):442–449.

Park, D. C., El-Sharkawi, M., Marks, R., Atlas, L., and Damborg, M. (1991b). Electric load forecasting using an artificial neural network. *IEEE transactions on Power Systems*, 6(2):442–449.

Pathak, S. (2016). Leveraging gis mapping and smart metering for improved oms and saidi for smart city. In *2016 Saudi Arabia Smart Grid (SASG)*, pages 1–5.

Pierro, M., De Felice, M., Maggioni, E., Moser, D., Perotto, A., Spada, F., and Cornaro, C. (2017). Data-driven upscaling methods for regional photovoltaic power estimation and forecast using satellite and numerical weather prediction data. *Solar Energy*, 158:1026–1038.

Pierro, M., De Felice, M., Maggioni, E., Moser, D., Perotto, A., Spada, F., and Cornaro, C. (2020). Residual load probabilistic forecast for reserve assessment: A real case study. *Renewable Energy*, 149:508–522.

Power, I. et al. (2014). Ieee guide for collecting, categorizing, and utilizing information related to electric power distribution interruption events ieee power and energy society. Institute of Electrical and Electronics Engineers.

Qiu, Y., Kahn, M. E., and Xing, B. (2019). Quantifying the rebound effects of residential solar panel adoption. *Journal of Environmental Economics and Management*, 96:310–341.

Ratnam, E. L., Weller, S. R., Kellett, C. M., and Murray, A. T. (2017). Residential load and rooftop pv generation: an australian distribution network dataset. *International Journal of Sustainable Energy*, 36(8):787–806.

Rawlings, J. O., Pantula, S. G., and Dickey, D. A. (1998). *Applied regression analysis: a research tool*. Springer.

Raza, M. Q., Nadarajah, M., and Ekanayake, C. (2016). On recent advances in pv output power forecast. *Solar Energy*, 136:125–144.

Razavi, S. E., Arefi, A., Ledwich, G., Nourbakhsh, G., Smith, D. B., and Minakshi, M. (2020). From load to net energy forecasting: Short-term residential forecasting for the blend of load and pv behind the meter. *IEEE Access*, 8:224343–224353.

Ruiz-Abellón, M. C., Fernández-Jiménez, L. A., Guillamón, A., Falces, A., García-Garre, A., and Gabaldón, A. (2020). Integration of demand response and short-term forecasting for the management of prosumers' demand and generation. *Energies*, 13(1):11.

Scott, W. (1990). Automating the restoration of distribution services in major emergencies. *IEEE Transactions on power delivery*, 5(2):1034–1039.

Sepasi, S., Reihani, E., Howlader, A. M., Roose, L. R., and Matsuura, M. M. (2017).

Very short term load forecasting of a distribution system with high PV penetration. *Renewable energy*, 106:142–148.

Shaker, H., Manfre, D., and Zareipour, H. (2020). Forecasting the aggregated output of a large fleet of small behind-the-meter solar photovoltaic sites. *Renewable Energy*, 147:1861–1869.

Shaker, H., Zareipour, H., and Wood, D. (2015). A data-driven approach for estimating the power generation of invisible solar sites. *IEEE Transactions on Smart Grid*, 7(5):2466–2476.

Shaker, H., Zareipour, H., and Wood, D. (2016a). Estimating power generation of invisible solar sites using publicly available data. *IEEE Transactions on Smart Grid*, 7(5):2456–2465.

Shaker, H., Zareipour, H., and Wood, D. (2016b). Impacts of large-scale wind and solar power integration on california's net electrical load. *Renewable and Sustainable Energy Reviews*, 58:761–774.

Sharma, V., Cali, U., Hagenmeyer, V., Mikut, R., and Ordiano, J. Á. G. (2018). Numerical weather prediction data free solar power forecasting with neural networks. In *Proceedings of the Ninth International Conference on Future Energy Systems*, pages 604–609.

Sharma, V., Hong, T., Cecchi, V., Hofmann, A., and Lee, J. Y. (2023). Forecasting weather-related power outages using weighted logistic regression. *IET Smart Grid*, 6(5):470–479.

Shen, Y., Fan, T., Lai, G., Na, Z., Liu, H., Wang, Z., Wang, Y., Jiao, Y., Chen, X., Lou, Z., et al. (2022). Modified u-net based photovoltaic array extraction from complex scene in aerial infrared thermal imagery. *Solar Energy*, 240:90–103.

Sheng, S., Yingkun, W., Yuyi, L., Yong, L., and Yu, J. (2011). Cyber attack impact on power system blackout.

Soares, T., Gonçalves, R., Louro, M., Pinto, J. F., Santos, T., Preto, I., and Monteiro, C. (2021). Forecasting the number of outage affected clients in extreme weather conditions. In *CIRED 2021 - The 26th International Conference and Exhibition on Electricity Distribution*, volume 2021, pages 1630–1633.

Sobhani, M., Campbell, A., Sangamwar, S., Li, C., and Hong, T. (2019). Combining weather stations for electric load forecasting. *Energies*, 12(8):1510.

Sreekumar, S., Sharma, K. C., and Bhakar, R. (2018). Gumbel copula based aggregated net load forecasting for modern power systems. *IET Generation, Transmission & Distribution*, 12(19):4348–4358.

Stainsby, W., Zimmerle, D., and Duggan, G. P. (2020). A method to estimate residential pv generation from net-metered load data and system install date. *Applied energy*, 267:114895.

Stratigakos, A., Bachoumis, A., Vita, V., and Zafiropoulos, E. (2021). Short-term net load forecasting with singular spectrum analysis and lstm neural networks. *Energies*, 14(14).

Stratman, A., Hong, T., Yi, M., and Zhao, D. (2023a). Net load forecasting with disaggregated behind-the-meter pv generation. *IEEE Transactions on Industry Applications*, pages 1–11.

Stratman, A., Hong, T., Yi, M., and Zhao, D. (2023b). Net load forecasting with disaggregated behind-the-meter pv generation. *IEEE Transactions on Industry Applications*.

Street, P. (2015). Dataport: the world's largest energy data resource. *Pecan Street Inc*, pages 01–01.

Sun, M., Zhang, T., Wang, Y., Strbac, G., and Kang, C. (2019). Using bayesian deep learning to capture uncertainty for residential net load forecasting. *IEEE Transactions on Power Systems*, 35(1):188–201.

IEEE Guide for Collecting, Categorizing, and Utilizing Information Related to Electric Power Distribution Interruption Events (2014). , IEEE Std 1782-2014, pp. 1-98, 2014.

ul Abideen, M. Z., Ellabban, O., Refaat, S. S., Abu-Rub, H., and Al-Fagih, L. (2019). A novel methodology to determine the maximum pv penetration in distribution networks. In *2019 2nd International Conference on Smart Grid and Renewable Energy (SGRE)*, pages 1–6.

UMass (2017). University of Massachusetts Smart Dataset.

US Department of Commerce, N. (2020). Forecast terms. https://www.weather.gov/bgm/forecast-terms, accessed April 2020.

Van der Meer, D., Munkhammar, J., and Widén, J. (2018). Probabilistic forecasting of solar power, electricity consumption and net load: Investigating the effect of seasons, aggregation and penetration on prediction intervals. *Solar Energy*, 171:397–413.

Van Der Meer, D., Shepero, M., Svensson, A., Widén, J., and Munkhammar, J. (2018). Probabilistic forecasting of electricity consumption, photovoltaic power generation and net demand of an individual building using gaussian processes. *Applied Energy*, 213:195–207.

Van der Meer, D. W., Munkhammar, J., and Widén, J. (2018a). Probabilistic forecasting of solar power, electricity consumption and net load: Investigating the effect of seasons, aggregation and penetration on prediction intervals. *Solar Energy*, 171:397–413.

Van der Meer, D. W., Widén, J., and Munkhammar, J. (2018b). Review on probabilistic forecasting of photovoltaic power production and electricity consumption. *Renewable and Sustainable Energy Reviews*, 81:1484–1512.

Vignola, F. (2012). GHI correlations with DHI and DNI and the effects of cloudiness on one-minute data. In *ASES*.

Wang, F., Li, K., Wang, X., Jiang, L., Ren, J., Mi, Z., Shafie-khah, M., and Catalão, J. P. (2018a). A distributed pv system capacity estimation approach based on support vector machine with customer net load curve features. *Energies*, 11(7):1750.

Wang, P., Liu, B., and Hong, T. (2016). Electric load forecasting with recency effect: A big data approach. *International Journal of Forecasting*, 32(3):585–597.

Wang, Y., Chen, Q., Hong, T., and Kang, C. (2019). Review of smart meter data analytics: Applications, methodologies, and challenges. *IEEE Transactions on Smart Grid*, 10(3):3125–3148.

Wang, Y., Zhang, N., Chen, Q., Kirschen, D. S., Li, P., and Xia, Q. (2018b). Data-driven probabilistic net load forecasting with high penetration of behind-the-meter pv. *IEEE Transactions on Power Systems*, 33(3):3255–3264.

Ward, D. M. (2013). The effect of weather on grid systems and the reliability of electricity supply. *Climatic Change*, 121(1):103–113.

Xie, J., Chen, Y., Hong, T., and Laing, T. D. (2016a). Relative humidity for load forecasting models. *IEEE Transactions on Smart Grid*, 9(1):191–198.

Xie, J., Chen, Y., Hong, T., and Laing, T. D. (2016b). Relative humidity for load forecasting models. *IEEE Transactions on Smart Grid*, 9(1):191–198.

Xie, J. and Hong, T. (2016). GEFCom2014 probabilistic electric load forecasting: An integrated solution with forecast combination and residual simulation. *International Journal of Forecasting*, 32(3):1012–1016.

Xie, J. and Hong, T. (2017). Wind speed for load forecasting models. *Sustainability*, 9(5):795.

Xu, L. and Chow, M.-Y. (2005). Power distribution systems fault cause identification using logistic regression and artificial neural network. *Proceedings of the 13th International Conference on, Intelligent Systems Application to Power Systems (ISAP2005)*, pages 6–.

Xu, L. and Chow, M.-Y. (2006). A classification approach for power distribution systems fault cause identification. *IEEE Transactions on Power Systems*, 21(1):53–60.

Xu, L., Chow, M.-Y., and Taylor, L. S. (2007a). Power distribution fault cause identification with imbalanced data using the data mining-based fuzzy classification e-algorithm. *IEEE Transactions on Power Systems*, 22(1):164–171.

Xu, L., Chow, M.-Y., Timmis, J., and Taylor, L. S. (2007b). Power distribution outage cause identification with imbalanced data using artificial immune recognition system (AIRS) algorithm. *IEEE Transactions on Power Systems*, 22(1):198–204.

Yang, D., Alessandrini, S., Antonanzas, J., Antonanzas-Torres, F., Badescu, V., Beyer, H. G., Blaga, R., Boland, J., Bright, J. M., Coimbra, C. F., et al. (2020a). Verification of deterministic solar forecasts. *Solar Energy*, 210:20–37.

Yang, D., Wang, W., Gueymard, C. A., Hong, T., Kleissl, J., Huang, J., Perez, M. J., Perez, R., Bright, J. M., Xia, X., Van der Meer, D., and Peters, I. M. (2022). A review of solar forecasting, its dependence on atmospheric sciences and implications for grid integration: Towards carbon neutrality. *Renewable and Sustainable Energy Reviews*, 161:112348.

Yang, F., Watson, P., Koukoula, M., and Anagnostou, E. N. (2020b). Enhancing weather-related power outage prediction by event severity classification. *IEEE Access*, 8:60029–60042.

Yu, J., Wang, Z., Majumdar, A., and Rajagopal, R. (2018). Deepsolar: A machine learning framework to efficiently construct a solar deployment database in the united states. *Joule*, 2(12):2605–2617.

Yu, N., Shah, S., Johnson, R., Sherick, R., Hong, M., and Loparo, K. (2015). Big data analytics in power distribution systems. In *2015 IEEE power & energy society innovative smart grid technologies conference (ISGT)*, pages 1–5. IEEE.

Zhang, J., Florita, A., Hodge, B.-M., Lu, S., Hamann, H. F., Banunarayanan, V., and Brockway, A. M. (2015). A suite of metrics for assessing the performance of solar power forecasting. *Solar Energy*, 111:157–175.

Zhang, T., Zhang, X., Chau, T. K., Chow, Y., Fernando, T., Iu, H. H.-C., et al. (2023). Highly accurate peak and valley prediction short-term net load forecasting approach based on decomposition for power systems with high pv penetration. *Applied Energy*, 333:120641.

Zhang, X. and Grijalva, S. (2016a). A data-driven approach for detection and estimation of residential pv installations. *IEEE Transactions on Smart Grid*, 7(5):2477–2485.

Zhang, X. and Grijalva, S. (2016b). A data-driven approach for detection and estimation of residential pv installations. *IEEE Transactions on Smart Grid*, 7(5):2477–2485.

Zhou, B., Meng, Y., Huang, W., Wang, H., Deng, L., Huang, S., and Wei, J. (2021). Multi-energy net load forecasting for integrated local energy systems with heterogeneous prosumers. *International Journal of Electrical Power & Energy Systems*, 126:106542.

Zhou, K., Fu, C., and Yang, S. (2016). Big data driven smart energy management: From big data to big insights. *Renewable and sustainable energy reviews*, 56:215–225.

Zhou, Y., Pahwa, A., and Yang, S.-S. (2006). Modeling weather-related failures of overhead distribution lines. *IEEE Transactions on Power Systems*, 21(4):1683–1690.

Zhu, D., Cheng, D., Broadwater, R., and Scirbona, C. (2007). Storm modeling for prediction of power distribution system outages. *Electric Power Systems Research*, 77:973–979.

APPENDIX A: California Independent System Operator (CAISO)

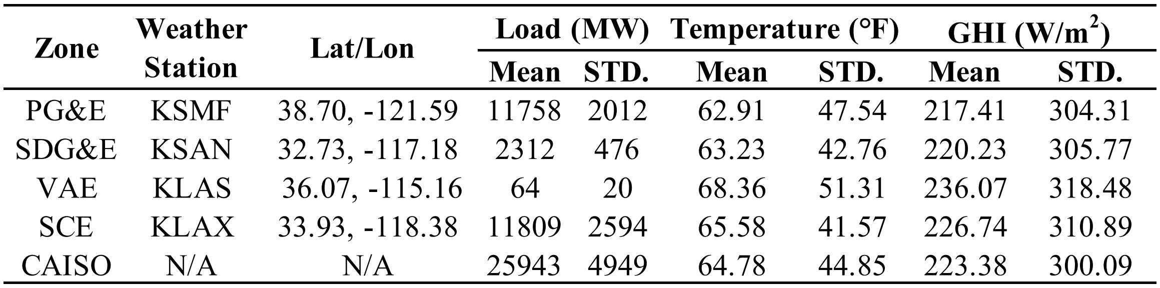
A.1 Data Description

The California Independent System Operator (CAISO) is an ISO that oversees the operation of California's bulk electric power system, transmission lines, and electricity market generated and transmitted by its member utilities.

CAISO operates in a region where the four key utilities are the Pacific Gas and Electric Company (PG&E), serving Northern and Central California; Southern California Edison (SCE), providing electricity to Southern and Central California; San Diego Gas & Electric (SDG&E), servicing San Diego and Southern Orange counties; and Valley Electric Association, Inc. (VEA). Serving around 80% of California's electricity demand and a portion of Nevada, it caters to over 30 million customers.

CAISO's data serves as an extra case study, to validate the performance of the proposed methodology. The total system load along with the load zones from CAISO is used for out-of-sample testing.

Table A.1: Summary data on load, temperature, and GHI for CAISO load zones (2016-2019)



Zone	Weather Station	Lat/Lon	Load (MW)		Temperature (°F)		GHI (W/m^2)	
			Mean	STD.	Mean	STD.	Mean	STD.
PG&E	KSMF	38.70, -121.59	11758	2012	62.91	47.54	217.41	304.31
SDG&E	KSAN	32.73, -117.18	2312	476	63.23	42.76	220.23	305.77
VAE	KLAS	36.07, -115.16	64	20	68.36	51.31	236.07	318.48
SCE	KLAX	33.93, -118.38	11809	2594	65.58	41.57	226.74	310.89
CAISO	N/A	N/A	25943	4949	64.78	44.85	223.38	300.09

A.2 Results of the Benchmark Net Load Forecasting Model for CAISO

Table A.2: MAPE values (in %) of the Vanilla model vs the proposed model on the out-of-sample test data (2019) for CAISO

Zone	Model	0% PV	5% PV	10% PV	15% PV	20% PV	25% PV	30% PV
SCE	Vanilla Model	4.95	5.44	6.15	7.10	8.32	9.98	12.54
	Proposed Model	4.72	4.89	5.10	5.37	5.71	6.17	6.93
VEA	Vanilla Model	9.12	11.97	10.31	11.42	12.68	15.07	20.12
	Proposed Model	8.86	11.24	9.58	10.19	10.78	11.97	14.45
SDGE	Vanilla Model	6.78	7.67	8.84	10.37	12.48	15.81	33.01
	Proposed Model	6.03	6.35	6.76	7.30	8.07	9.37	19.63
PGE	Vanilla Model	4.33	4.79	5.40	6.15	7.07	8.24	9.80
	Proposed Model	3.99	4.13	4.30	4.51	4.78	5.14	5.67
CAISO	Vanilla Model	3.93	4.34	4.89	5.57	6.42	7.51	9.01
	Proposed Model	3.60	3.74	3.91	4.12	4.38	4.72	5.21

A.3 Results for the Benchmark Net Load Forecasting Model for CAISO with Recency Effect

Table A.3: MAPE values (in %) of the Vanilla model vs the proposed model with recency effect modeling on the out-of-sample test data (2019) for CAISO

Model	Vanilla	Vanilla Model	Proposed	Proposed Model
Zone	**Model**	**w/ Recency**	**Model**	**w/ Recency**
SCE	12.54	10.78	6.93	6.18
VEA	29.31	25.06	19.86	17.89
SDGE	19.11	17.27	10.76	10.37
PGE	9.80	8.39	5.67	5.57
CAISO	9.00	7.72	5.21	5.07

APPENDIX B: Electric Reliability Council of Texas (ERCOT)

B.1 Data Description

The Electric Reliability Council of Texas (ERCOT) is an organization responsible for managing Texas's electric grid and wholesale electricity market. Serving more than 26 million customers, ERCOT manages a significant portion of Texas' electricity supply, representing 90% of the state's overall electrical demand. ERCOT divides its service territory into eight weather zones. These zones include Far West Texas (FWEST), West Texas (WEST), North Texas (NORTH), South Texas (SOUTH), Coastal Texas (COAST), North-Central Texas (NCENT), East Texas (EAST), and South-Central Texas (SCENT). Categorizing is based on the varying weather patterns observed in each zone. For instance, the FWEST zone, which includes cities like Midland, is characterized by a hot and dry climate, while the COAST zone, which covers cities like Corpus Christi and Brownsville, is marked by a subtropical climate with hot summers and mild winters.

The data from ERCOT is used as an additional case study to demonstrate the effectiveness of the proposed methodology. We use the total system load and the data from 2019 for all of ERCOT's load zones for out-of-sample testing.

Table B.1: Summary data on load, temperature, and GHI for ERCOT load zones (2016-2019)

Zone	Weather Station	Lat/Lon	Load (MW)		Temperature (°F)		GHI (W/m^2)	
			Mean	STD.	Mean	STD.	Mean	STD.
COAST	KHOU	29.64, -95.28	11413	2685	69.28	46.43	202.84	289.19
EAST	KTYR	32.36, -95.4	1426	341	65.29	49.80	201.63	290.20
FWEST	KMDD	32.04, -102.1	2216	814	64.14	51.04	235.33	320.86
NORTH	KLBB	33.67, -101.82	841	184	60.97	51.52	230.58	316.01
NCENT	KDFW	32.9, -97.02	13089	3569	65.20	50.61	204.45	291.48
SOUTH	KMFE	26.18, -98.24	3280	833	74.80	45.73	220.12	306.25
SCENT	KATT	30.32, -97.77	6490	1758	68.07	49.07	208.13	297.47
WEST	KABI	32.41, -99.68	1142	247	64.46	50.73	219.86	306.14
ERCOT	N/A	N/A	39895	9625	66.52	48.92	215.36	289.88

B.2 Results of the Benchmark Net Load Forecasting Model for ERCOT

Table B.2: MAPE values (in %) of the Vanilla model vs the proposed model on the out-of-sample test data (2019) for ERCOT

Zone	**Model**	**0% PV**	**5% PV**	**10% PV**	**15% PV**	**20% PV**	**25% PV**	**30% PV**
COAST	Vanilla Model	4.60	4.81	5.27	5.92	6.74	7.76	9.01
	Proposed Model	4.56	4.65	4.74	4.85	4.98	5.14	5.33
EAST	Vanilla Model	5.55	5.85	6.40	7.18	8.22	9.54	11.28
	Proposed Model	5.57	5.68	5.80	5.95	6.13	6.36	6.66
NORTH C	Vanilla Model	5.64	5.96	6.53	7.33	8.38	9.72	11.50
	Proposed Model	5.70	5.81	5.95	6.11	6.31	6.55	6.88
NORTH	Vanilla Model	6.94	7.27	7.74	8.40	9.26	10.36	11.80
	Proposed Model	6.92	7.09	7.29	7.53	7.81	8.17	8.64
SOUTH C	Vanilla Model	5.35	5.75	6.42	7.34	8.56	10.19	12.51
	Proposed Model	5.43	5.55	5.70	5.88	6.10	6.39	6.79
SOUTH	Vanilla Model	5.44	5.66	6.06	6.59	7.27	8.14	9.24
	Proposed Model	5.42	5.51	5.63	5.76	5.91	6.10	6.33
WEST	Vanilla Model	4.83	5.04	5.45	6.03	6.77	7.68	8.80
	Proposed Model	4.79	4.89	4.99	5.12	5.27	5.45	5.68
ERCOT	Vanilla Model	4.17	4.37	4.69	5.12	5.66	6.32	7.13
	Proposed Model	4.18	4.27	4.36	4.47	4.60	4.75	4.93

B.3 Results for the Benchmark Net Load Forecasting Model for ERCOT with Recency Effect

Table B.3: MAPE values (in %) of the Vanilla model vs the proposed model with recency effect modeling on the out-of-sample test data (2019) for ERCOT

Model Zone	Vanilla Model	Vanilla Model w/ Recency	Proposed Model	Proposed Model w/ Recency
COAST	9.01	8.33	5.33	4.51
EAST	10.31	9.87	6.49	5.86
NORTH C	11.50	10.14	6.88	5.86
NORTH	11.54	10.75	8.55	7.78
SOUTH C	12.51	11.09	6.79	6.11
SOUTH	9.24	9.00	6.33	5.90
WEST	8.80	7.54	5.68	4.98
ERCOT	7.14	6.46	4.92	4.36