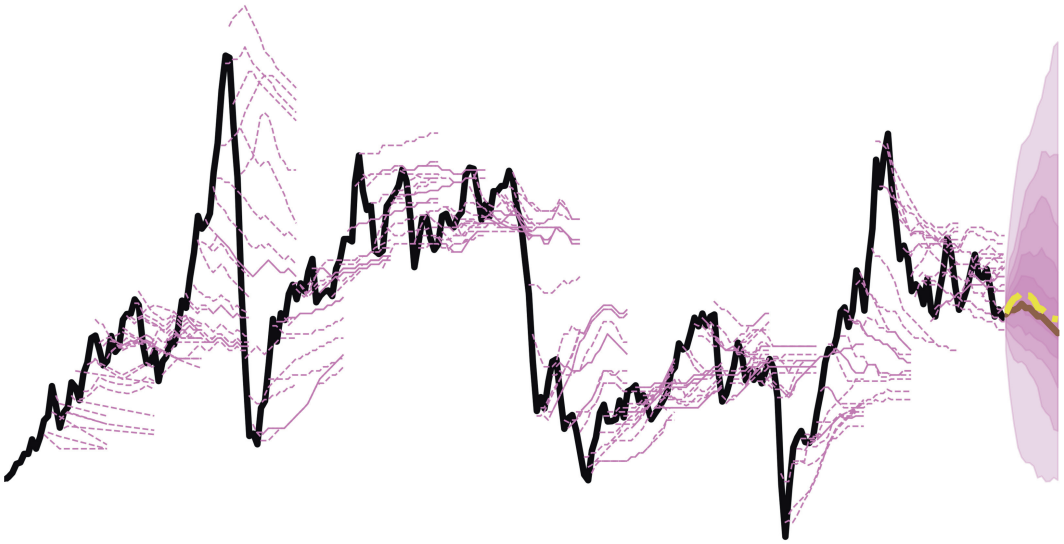
CHALMERS
UNIVERSITY OF TECHNOLOGY



Quantifying Uncertainty in Energy Forecasting

A Study of Forecast Errors, Bias, and the Role of Empirical Prediction Intervals in the U.S. Energy Information Administration's Short-Term Energy Outlook

Master's thesis in Physics

ANESA AGOVIC

DEPARTMENT OF EARTH, SPACE AND ENVIRONMENT

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

Master's thesis 2026

Quantifying Uncertainty in Energy Forecasting

A Study of Forecast Errors, Bias, and the Role of Empirical Prediction Intervals in the U.S. Energy Information Administration's Short-Term Energy Outlook

ANESA AGOVIC

CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Earth, Space and Environment
Division of Physical Resource Theory
Chalmers University of Technology
Gothenburg, Sweden 2026

Quantifying Uncertainty in Energy Forecasting
A Study of Forecast Errors, Bias, and the Role of Empirical Prediction Intervals in the U.S. Energy Information Administration's Short-Term Energy Outlook
ANESA AGOVIC

Supervisor: David Daniels, Swedish National Road and Transport Research Institute
Examiner: Fredrik Hedenus, Physical Resource Theory, Chalmers University of Technology

Master's Thesis 2026
Department of Earth, Space and Environment
Division of Physical Resource Theory
Chalmers University of Technology
SE-412 96 Gothenburg
Telephone +46 31 772 1000

Acknowledgments, dedications, and similar personal statements in this thesis, reflect the author's own views.

Cover: Historical values and forecasts of West Texas Intermediate crude oil prices from the U.S. Energy Information Administration Short-Term Energy Outlook (STEO). Actual outcomes (2001–2024) are shown in black, issued monthly forecasts (2004–2024) in purple, and empirical density forecasts (2024–2025) in purple shaded intervals, with the median indicated in yellow and STEO forecasts (2024-2025) in brown.

Typeset in LaTeX
Printed by Chalmers Reproservice
Gothenburg, Sweden 2026

Quantifying Uncertainty in Energy Forecasting
A Study of Forecast Errors, Bias, and the Role of Empirical Prediction Intervals in the U.S. Energy Information Administration's Short-Term Energy Outlook
ANESA AGOVIC
Department of Space, Earth and Environment
Chalmers University of Technology

Abstract

Recent surges in energy demand have exposed the limitations of widely used forecasts, which often fail to anticipate disruptive shifts in the energy system. Yet, energy projections remain essential for policymaking, investment, and risk assessment, while their accuracy and uncertainty remain insufficiently understood. The U.S. Energy Information Administration's (EIA) Short-Term Energy Outlook (STEO) constitutes one of the most prominent sources for these forecasts, providing monthly data on U.S. energy production, consumption, capacity, and prices. The STEO issues point forecasts over time horizons up to 13-24 months, but does not offer users explicit measures of uncertainty.

This thesis assesses the magnitude, direction, and uncertainty of STEO forecast errors from 2004 to 2024. Key energy quantities were analyzed across five categories: primary production, electricity generation, electricity capacity, fossil fuel prices, and electricity prices. Historical forecast errors were quantified using the mean absolute percentage error (MAPE), exponential mean absolute logarithmic error (EMALE), and mean percentage error (MPE). To represent forecast uncertainty, four empirical prediction interval (EPI) methods, two nonparametric and two parametric (Gaussian), were implemented and compared using the continuous ranked probability score (CRPS).

The analysis shows that forecasts of physical quantities, such as crude oil production and wind capacity, have been comparatively accurate, whereas fossil fuel and electricity price forecasts exhibit much larger errors. Evidence of systematic bias is found in several quantities, including persistent anchoring in electricity price forecasts, inertia in oil price forecasts, and consistent overestimation of solar capacity. To characterize forecast uncertainty, nonparametric EPI methods generally outperform Gaussian approximations, particularly for volatile and heavy-tailed error distributions. Importantly, Gaussian methods systematically underestimate the probability of extreme price spikes, producing a downward bias by smoothing away the asymmetric impact of rare upward shocks.

This work extends the approach by Kaack et al. (2017), who constructed empirical prediction intervals for long-term projections in the EIA's Annual Energy Outlook. By applying a similar framework to the high-frequency STEO data, we provide a complementary perspective on short-term forecasts and their uncertainties. The findings demonstrate how empirical prediction interval methods can enhance point forecasts with explicit uncertainty measures, providing STEO users with a clearer picture of forecast reliability and supporting more informed decision-making.

Keywords: energy forecasting, forecast uncertainty, forecast errors, forecast bias, density forecasts, empirical prediction intervals, continuous ranked probability score, energy prices, Short-Term Energy Outlook, Energy Information Administration

Acknowledgments

I would like to express my sincerest gratitude and thanks to my supervisor, David Daniels, for his unwavering guidance, seemingly endless knowledge, and patience throughout this project. An equal thank you also to my examiner, Fredrik Hedenus, for his engagement and expertise. Finally, to my fellow physics students in Hilbert, countless thanks for all these years.

Anesa Agovic, Gothenburg, September 2025

List of Acronyms

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

AEO	Annual Energy Outlook
Btu	British Thermal Unit
CAISO	California Independent System Operator
CDF	Cumulative Distribution Function
CRPS	Continuous Ranked Probability Score
DOE	Department of Energy
EIA	Energy Information Administration
EMALE	Exponential Mean Absolute Logarithmic Error
EPI	Empirical Prediction Interval
ERCOT	Electric Reliability Council of Texas
GSTOM	Global Short-Term Oil Model
IEA	International Energy Agency
ISO-NE	Independent System Operator New England
LNG	Liquefied Natural Gas
MAE	Mean Absolute Error
MALE	Mean Absolute Logarithmic Error
MAPE	Mean Absolute Percentage Error
MRE	Mean Relative Error
MPE	Mean Percentage Error
NEMS	National Energy Modeling System
NOAA	National Oceanic and Atmospheric Administration
PDF	Probability Density Function
PV	Photovoltaic
STEO	Short-Term Energy Outlook
STIFS	Short-Term Integrated Forecasting System
USSTEM	U.S. Short-Term Energy Model
WEO	World Energy Outlook
WTI	West Texas Intermediate

Table of Contents

List of Figures

List of Tables

1

Introduction

In 2024, energy demand grew faster than anticipated, driven by rising industrial consumption, increased electrification, and the expansion of data centers and artificial intelligence [1]. This surge affected energy sources like oil, natural gas, coal, and renewables. The unforeseen growth in demand revealed that many forecasts had underestimated energy consumption and highlighted a central tension in forecasting, namely that even widely used models fail to predict structural breaks in the energy system. Forecasts, in their simplest form, represent the best estimates or expectation of future outcomes based on current knowledge and historical data [2]. They play a role in guiding policy making, investments planning, and long-term strategic planning [3, 4]. The utility of forecasts lies not only in accurate point predictions, but also in their ability to signal when market conditions are changing in ways that matter for forecast users. In this sense, forecasts serve a role similar to canaries in a coal mine: they are most valuable when they indicate that a structural change may be underway.

In the United States, some of the most widely used energy forecasts are those made by the U.S. Energy Information Administration (EIA), the principal statistical agency for energy within the U.S. government and a part of the U.S. Department of Energy (DOE). Part of EIA's dissemination of information includes the Short-Term Energy Outlook (STEO), a monthly report of short-term energy point forecasts for energy supply, consumption, and prices [5]. The STEO includes forecasts for oil, natural gas, coal, renewables, electricity, and biofuels. Its focus is the U.S. domestic market, but it also covers certain international liquid fuels markets. Each STEO forecast extends through the remainder of the current calendar year and ends the next. Thus, the series of STEO forecasts covers a varying horizon between 1-13 and 1-24 months. These forecasts are widely read by energy-intensive businesses, making them influential in shaping economic expectations [6]. Davis & Brear [7] note that inaccurate short-term forecasts in renewable generation, in particular wind, can lead to scarcity pricing, raise total system costs and average prices, and cause a small rise in greenhouse gas emissions. Forecasting errors can also generate financial imbalances, with their magnitude shaped by the extent to which prevailing pricing mechanisms depend on energy volumes [8]. It is therefore important to understand the accuracy and limitations of these forecasts, particularly with respect to normal volatility in energy markets, as well as their potential to miss structural breaks during extraordinary events, such as the COVID-19 pandemic and the Russian invasion of Ukraine in 2022 [9].

Forecasting in complex socio-technical systems such as energy markets is an inherently

uncertain task [10, 11]. These systems are defined by nonlinearities, feedback loops, and external shocks. Small variations in assumptions about macroeconomic conditions, policy developments, weather, or technological change can result in significant differences between projected and realized outcomes. As Kaminski [10] notes, "Forecasting a specific number is an exercise fraught with danger", and the history of EIA's STEO forecasting is no exception. In a press release from March 2022, the EIA stated that its oil price forecasts have been subjected to heightening uncertainty due to macroeconomic volatility and geopolitical shocks coupled with already low inventories, which together amplified oil price volatility [12, 13]. Strikingly, they stressed "We believe it is more important to understand all the factors generating the uncertainty around our March STEO forecasts than it is to understand the forecasts themselves" [12]. Similarly, industry commentary from the *Oil & Gas Journal* in 2020 noted that the STEO remained susceptible to high levels of uncertainty following mitigation and reopening processes related to COVID-19 [14]. The article highlights that unpredictable macroeconomics and public health developments complicate forecasting and result in more revisions and larger differences between forecasts and realized outcomes.

Seasonality, infrastructure planning, and other long-term trends like cyclical patterns are key determinants of many forecasts [4]. By contrast, volatility is not a model input but an empirical property of the data that reflects the magnitude and frequency of changes in a variable over time. Forecast models attempt to capture this variation in an error term, commonly assumed to be stochastic.

Forecasters tend to avoid predicting the next spike or crash and instead use the current price as the best guess for future prices [15]. This forecasting method is called the naïve assumption and forms the basis of many forecasting techniques, such as the random walk method [4]. The naïve method leads to forecasts that are relatively flat lines, in stark contrast to the oftentimes jagged historical volatility seen in crude oil prices, for example. Figure 1.1 depicts historical values and forecasts for West Texas Intermediate (WTI) crude oil prices, which are notoriously volatile, driven by geopolitics and supply-demand imbalances. This volatility means unforeseen events can rapidly send prices surging or plummeting. In practice, even options markets imply a wide uncertainty in future oil prices, reflecting the market's recognition that forecasting oil prices is difficult [16]. Because of this unpredictability, short-term forecasts of oil prices often appear relatively flat or trendless compared to the jagged historical curve, regardless of horizon. Each red dashed line in Figure 1.1 represents a series of forecasts made in the STEO reports. Together, these overlapping forecasts show a pattern consistent with growing uncertainty over horizon time and raises questions about the structure and behavior of forecast uncertainty in the STEO.

This link between volatility and uncertainty implies that the more erratic the market becomes, the range of plausible outcomes that forecasts must account for should widen. Moreover, repeated over- and underestimation relative to actual values may reveal systematic biases inherent in EIA's forecasting model for the STEO – particularly around periods of market turbulence, such as the 2008-2009 oil price collapse and the COVID-19 pandemic in 2020 – but they could also reflect differences in the timescale at which volatility is characterized. Energy prices are shaped by both short-run fluctuations, often driven by

transitory shocks, and long-run dynamics linked to structural supply and demand factors [17]. If the STEO models are more sensitive to long-term or annual-scale variation, they may under-react to short-term movements, thereby creating the appearance of directional bias even when the forecasts are consistent with longer-term market behavior.

While such visualizations (Figure 1.1) illustrate forecast errors, they do not by themselves reveal whether these deviations are systematic, how large they are on average, or whether the underlying models are adapting over time. A systematic and quantitative analysis is therefore needed to assess whether reliance on the STEO as an unbiased indicator of future energy outcomes is justified. Since the STEO only publishes point forecasts without explicit uncertainty measures, users may place the same degree of confidence in variables that differ widely in accuracy, potentially leading to misinformed decisions. By quantifying the forecast errors, the analysis can uncover potential systematic biases and identify forecasts that are more or less reliable.

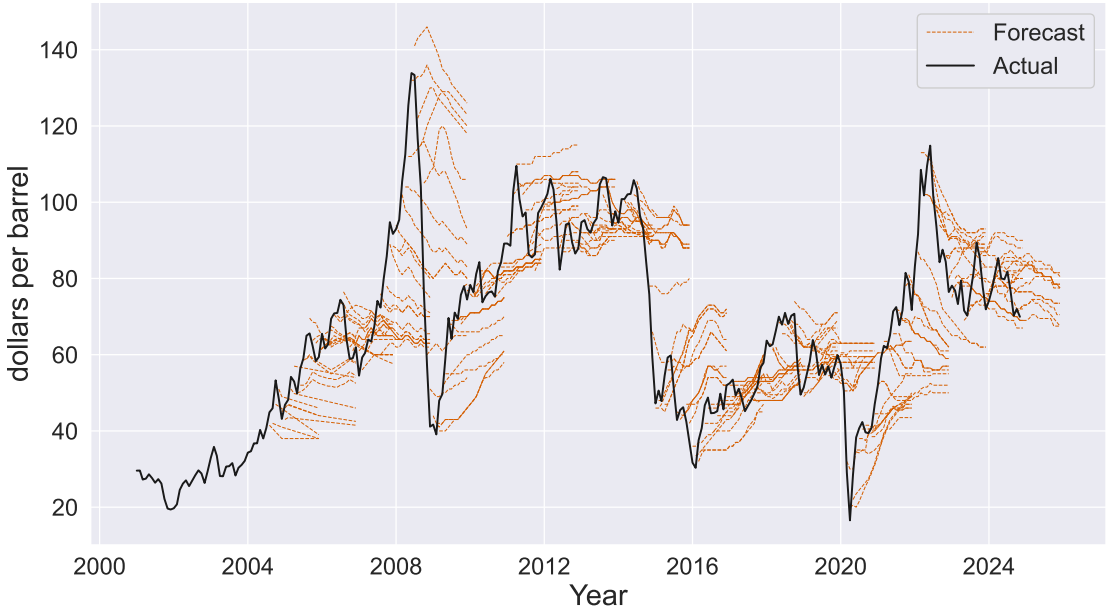


Figure 1.1: Historical West Texas Intermediate crude oil spot (nominal) prices and forecasts published in the EIA's STEO reports from January 2001 - November 2024 (actuals) and August 2004 - December 2024 (forecasts) [18]. The black solid line shows the most recent actuals listed in the STEOs, while the red dashed line denotes monthly forecasts issued over the studied period.

A central motivation for this thesis draws on the work of Kaack et al. [3], who constructed empirical prediction intervals (EPIs) from historical forecast errors and historical deviations in the underlying data to characterize uncertainty in the EIA's Annual Energy Outlook (AEO) projections. Their findings revealed that estimating uncertainty with a forecast using EPIs provided better calibrated and sharper probabilistic uncertainty estimates, compared to the scenario ranges in the AEO. Unlike the AEO, the STEO does not routinely publish side case scenarios, meaning users cannot rely on such ranges to estimate uncertainty. Inspired by this approach, the present work applies a similar empirical methodology to the STEO, offering a high-frequency counterpart to the AEO study. This allows for the investigation of systematic biases in a larger data set that is more

frequently updated by EIA compared to the AEO case, and provides an opportunity to evaluate whether empirical density forecasting methods can improve the interpretability and calibration of uncertainty in the short-term outlook.

This thesis aims to empirically evaluate the uncertainty and error characteristics of forecasts, published in the EIA's STEO. In addition, the performance of different empirical density forecasting methods is compared. The study is guided by the following research questions:

1. What is the magnitude and direction of forecast errors of different energy variables in the U.S. Energy Information Agency's Short-Term Energy Outlook?

2. Are there patterns in the errors that indicate systematic bias or lack of adaption in the Short-Term Energy Outlook modeling system?

3. Can empirical prediction intervals, constructed from past forecasting errors or historical variability in the time series, provide interpretable representations of forecast uncertainty in the Short-Term Energy Outlook?

To address these research questions, the thesis is structured as follows. Chapter 2 provides background on the EIA and the STEO, situating the present work in relation to existing studies. Chapter 3 outlines the methodological framework, including the database construction, error metrics, and EPI methods. The results are presented in Chapter 4, which examines forecast errors, systematic biases, and the performance of the EPI methods across different energy variables in the STEO. These findings are then interpreted in a broader context of energy markets and EIA's forecasting practice in Chapter 5. Finally, Chapter 6 summarizes the main findings of the thesis. This study bridges the gap between retrospective error analysis and forward-looking uncertainty assessment, contributing to a deeper understanding of STEO forecasting performance while also providing guidance to users on how much confidence they should place in its forecasts.

2

Background

This chapter provides the foundation for the analysis by outlining the institutional and methodological setting and challenges of energy forecasting in the U.S. It begins with an overview of the U.S. Energy Information Administration and two of its forecasting reports. Key concepts from forecasting theory are then introduced. Finally, the role of biases in shaping forecasts is addressed, highlighting how human judgment and institutional incentives can interact with econometric models to produce systematic patterns of error.

2.1 The Energy Information Administration and Its Outlooks

EIA's mission is to provide independent and unbiased energy information with the aim of supporting sound policymaking, facilitating efficient energy markets, and public understanding of energy and its relation with the economy and the environment [19]. Parts of its energy information dispersal are the Short-Term Energy Outlook and the Annual Energy Outlook.

2.1.1 The Short-Term Energy Outlook

The STEO forecasts are based primarily on historical data series from EIA's own statistical data release publications [5]. Other data sources include weather forecasts from the National Oceanic and Atmospheric Administration (NOAA) and macroeconomic forecasts based on the S&P Global model of the U.S. economy [20]. The STEO forecasts are produced using the Short-Term Integrated Forecasting System (STIFS), which combines econometric modeling with expert judgment [5]. STIFS is composed of two main models: the U.S. Short-Term Energy Model (USSTEM), which models electricity, petroleum, natural gas, coal, emissions, and renewables using nearly 600 equations, mostly linear regressions, and the Global Short-Term Oil Model (GSTOM), which projects international crude oil balances and prices. Analysts may modify forecasts through the use of numerical adjustments ("add-factors") based on recent market intelligence or judgment not captured by the models. While this modeling architecture allows for responsive and modular forecasting, it also introduces a layer of opacity that complicates external evaluation.

2.1.2 Long-Term Modeling: The Annual Energy Outlook

Another part of EIA's information supply involves long-term projections of U.S. energy supply, demand, and emissions that are published yearly in the Annual Energy Outlook. Although the present work focuses on the STEO, the AEO is relevant in two respects. First, it is produced by the same institution, implying that any systematic tendencies or institutional biases evident in the STEO may also manifest in the AEO, even though the two outlooks have different modeling frameworks. Second, previous work by Kaack et al. [3] has demonstrated how empirical prediction intervals can be used to evaluate uncertainty in the AEO projections, which provides a methodological framework and point of comparison for the present study.

Released in accordance with the DOE Organization Act of 1977 [21], the AEO is mandated to offer a forward-looking view of national energy trends and serves as a foundational reference for federal, state, and industry-level energy planning [22]. The AEO2025, for instance, utilizes the National Energy Modeling System (NEMS) to generate projections based on demographic, technological, and economic assumptions, as well as current laws and regulations as of December 2024. These include a policy-neutral reference case and a variety of side cases addressing uncertainties in technology cost, oil prices, energy supply, and policy implementation.

The AEO and STEO differ in frequency and time horizon, and they also use mathematical projection frameworks that represent entirely different modeling traditions. The STEO is driven by USSTEM, a collection of econometric regression equations designed to capture statistical relationships in energy supply and demand [5], whereas the AEO relies on structural, scenario-based simulations [23]. Thus, while STEO embodies an econometric forecasting approach focused on short-term dynamics, the AEO emphasizes structural analysis of long-term pathways.

Despite these differences, because the two reports draw on shared data sources and a common understanding of the energy system within the EIA, systematic biases in the STEO may also be present in the AEO. The comparative analysis undertaken by Kaack et al. [3] for the AEO thus serves as a methodological precedent and as a benchmark against which the present study can be positioned.

2.2 Understanding Forecasts, Uncertainty, and Bias

2.2.1 What are Forecasts?

Forecasting is the process of generating quantitative expectations about future outcomes based on historical data and expectations about any future developments that might affect historical trends during the forecast period [2, 4]. Forecasts are related to concepts such as projections and predictions. While these terms are often used interchangeably, the EIA distinguishes forecasts from predictions and projections based on their use of models and attempts at centrality: while projections employ a mathematical model and predictions aim for the most likely outcome, forecasts are both projections and predictions. Forecasts typically rely on statistical or econometric models trained on historical data to produce

short-, medium-, or long-term estimates under two assumptions: that the historical data is available and accurate, and that some aspects of the past patterns will repeat in the future [4].

Forecasts can be classified into two main types, point forecasts and probabilistic forecasts. Point forecasts provide a single value for each future time step. Probabilistic forecasts, by contrast, provide multiple values for each time step and characterize a distribution of possible outcomes with prediction intervals, density estimates, or ensembles to capture uncertainty [2]. When dealing with volatile energy variables, like oil prices or wholesale electricity prices, probabilistic forecasts are useful because the uncertainty is better described, but at the same time, they require more computational resources compared to point forecasts, which often have fewer parameters to learn from in the model. A point forecast may be the mean of the probability distribution produced by a fitted model [4], but there is no way to estimate how accurate the model is by only looking at the point forecasts. In practice, they are attractive because of their simplicity and ease of communication, but they only convey the central tendency and not the range of possibilities, such as the risk of extreme events.

2.2.2 Forecast Horizon and the Growth of Uncertainty

The forecast horizon, denoted H, refers to the number of time steps into the future that the forecast is made. Typically, the absolute magnitude of error grows with larger H. For example, evaluations of AEO projections show that the magnitude of error tends to grow with longer horizons; Fischer et al. [24] demonstrated that the errors for projections 3-5 years ahead were larger than those for 1-year ahead projections; in Kaack et al. [3], the produced EPIs widen with H, reflecting greater uncertainty in long-term projections.

2.2.3 The Nature of Uncertainty in Forecasts

Uncertainty is a natural feature of all forecasting practices. It may arise due to incomplete information or the structure of the forecasting model. Even supposing we have complete information, uncertainty can come from not understanding the relationship between variables, or simplifications and approximations present in the model to make cognitive or computational analysis more manageable. There are different kinds of uncertainty. In time series forecasting, a distinction is commonly made between epistemic and aleatoric uncertainty [25]. Epistemic uncertainty refers to incomplete information about the system, such as insufficient data, and could be reduced, though often impractical or not possible due to lack of time and resources. Aleatoric uncertainty, on the other hand, arises from randomness inherent in the system being modeled and can not be reduced, even with more information or better models. This uncertainty is particularly relevant in energy forecasting, where exogenous shocks can introduce variability that no amount of data or modeling can eliminate.

Some systems exhibit Knightian uncertainty, where even the probability distribution of outcomes is fundamentally unknowable due to ignorance or a limit to knowledge [26]. Typical statistical models rely on the assumption of known unknowns, where randomness can be quantified in probabilistic terms. In the context of energy markets, both measurable

and unmeasurable uncertainties are present. Some fluctuations, like seasonal variations, are less difficult to predict and can be incorporated into the model, while others are much more difficult to predict, like extreme weather events, disruptions, and outages.

To quantify uncertainty, probability is the most used formalism [27]. Sources of uncertainty in empirical quantities include statistical variation, subjective judgment, variability, inherent randomness, and approximation [27]. Variability here refers to the inherent stochasticity of energy systems and their time series. These series are shaped by both deterministic and stochastic trends and random noise, which makes it impossible to exactly predict future values [4]. Given these challenges, forecasters increasingly emphasize probabilistic methods to include uncertainty with the forecasts for effective and more informed decision-making.

Forecasting in real-world settings, like energy systems and markets, benefit from having an expression of the associated uncertainty belonging to point forecasts. Quantifying uncertainty is important for several reasons. First, not all variables exhibit the same degree of forecast error. Without making these differences explicit, users may mistakenly treat all forecasts as equally reliable. Second, without an accompanying estimate of forecast error, critics may dismiss forecasts as wrong when realized outcomes diverge from point predictions, undermining their value. Third, uncertainty quantification helps distinguish between forecast deviations that are consistent with expected variability and those that may signal structural breaks in the energy system.

One widely used approach for quantifying this uncertainty are prediction intervals, which capture a distribution of future outcomes with a prescribed probability, including the best estimate [28]. Empirical prediction intervals can be constructed from historical forecast errors or historical deviations [3]. EPIs can be either non-parametric, meaning that no or very limited assumptions about the underlying error distribution are made, or parametric with a defined distribution.

2.2.4 Bias in Forecasts

Bias in forecasting refers to a persistent tendency for forecasts to overestimate or underestimate actual outcomes. Systematic bias is indicated by a significant deviation of the mean forecast error from zero over many observations. These directional errors arise from the structure of forecasting models, institutional incentives, or judgment errors.

Systematic bias is particularly important in energy forecasting because it can distort expectations across markets, investment decisions, and policy planning [6]. Structural features of forecasting models can reinforce these patterns. Econometric models such as USSTEM, which underpins the STEO, rely heavily on historical statistical relationships represented by coefficients [5]. Since energy markets constantly evolve, these coefficients can lose their forecasting significance, leading to bias in the forecasts. Importantly, systematic biases interact with how forecasts are consumed. Users may assign equal confidence to all point forecasts without accompanying measures of uncertainty, even though some are consistently less reliable than others.

Forecasting is not only shaped by statistical models and structural assumptions but also by analyst judgment, which introduces systematic cognitive biases. These biases can emerge because analysts rely on heuristics or short-cuts to simplify complex judgment tasks [29]. While heuristics can be efficient, they can also introduce predictable biases in the forecasts. There is cognitive bias which arises from human judgment errors in model construction, data interpretation, and assumptions. Forecasters may be influenced by optimism bias, confirmation bias, anchoring, or institutional incentives that push forecasts toward politically or economically favorable outcomes [30].

Optimism bias is when forecasters systematically over-predict favorable outcomes. In energy contexts, this often manifests as forecasts of supply and demand variables that are consistently higher than realized outcomes [24]. For instance, empirical studies of Norwegian oil fields have shown that production forecasts systematically overestimated output and understated uncertainty, consistent with cognitive biases identified in behavioral economics [31]. Another cognitive bias is anchoring, which occurs when forecasters place too much weight on recent historical values when generating new forecasts. This behavior results in forecasts that under-react to new information, thereby producing predictable and systematic forecast errors [32].

In energy forecasting, systematic bias is problematic because it can cause misallocation of resources [33]. Sanders et al. [6] found that EIA's STEO supply forecasts for crude oil, coal, and electricity were largely unbiased, while the forecasts for natural gas consistently showed overestimation. Auffhammer [34] further demonstrated that EIA's forecasts may appear rational only under highly asymmetrical loss functions, where overestimates are penalized more strongly than underestimates, or vice versa. Moreover, cognitive tendencies do not operate in isolation but interact with institutional incentives. Recognizing the role of cognitive bias is thus essential for evaluating forecast errors. As Kucharavy et al. [35] argue, technological forecasts must manage cognitive biases if they are to provide reliable input for strategic decision-making.

3

Method

This chapter outlines the methodology used to evaluate the accuracy of the EIA's STEO forecasts and to represent forecast uncertainty through empirical prediction intervals. A set of error metrics is applied to quantify point forecast performance. Probabilistic forecast accuracy is assessed by constructing parametric and nonparametric empirical prediction intervals. The continuous ranked probability score is used to evaluate the performance of the produced prediction intervals.

3.1 Choosing Energy Variables

To ensure that the analysis captures the most central aspects of the U.S. energy system, 17 of the 864 quantities available in the EIA's December 2024 STEO were selected (Table 3.1). The selected quantities cover the following domains of U.S. energy monitoring: primary production (coal, crude oil, natural gas), electricity generation (coal, gas, hydropower, nuclear, total), electricity generating capacity (solar photovoltaic, wind), fossil fuel prices (coal, natural gas Henry Hub, gasoline, WTI crude oil), and wholesale electricity prices (California independent systems operator, electric reliability council of Texas, independent systems operator New England). Many of these quantities are highlighted in the U.S. Energy Markets Summary in the STEOs, the very first table of each report.

The choice to focus on this subset of variables reflects both practical and conceptual considerations. First, the sheer scale of the full STEO dataset makes a comprehensive evaluation impractical within the scope of this thesis. Second, many variables, such as macroeconomic indicators and weather forecasts, are not produced by the EIA's own modeling system but rather imported from external sources [5]. Including them would not provide direct insight into the performance of the EIA's energy market models. Finally, the selection aimed to ensure balanced coverage across major fuels and technologies, encompassing both fossil fuels and renewables, as well as both physical quantities and prices. However, the methodological framework developed here is general and could be applied to the entire STEO dataset. By focusing on a representative set of variables, the analysis aims to align with the variables that EIA itself prioritizes as indicators of energy market performance.

Figure 1.1 highlights the need for a more systematic analysis of forecast error characteristics, including their magnitude, directionality, and consistency or variability over time. To

address the research questions, we employ error metrics to evaluate the forecast errors: the mean absolute percentage error (MAPE), the mean absolute logarithmic error (MALE), the exponential MALE (EMALE), and the mean percentage error (MPE). These measures provide insight into the size and direction of deviations between forecasts and actual outcomes [3, 4, 36]. To represent uncertainty more explicitly beyond point forecasts, EPIs are constructed using both nonparametric and parametric approaches. The quality of these density forecasting methods is then evaluated using the continuous ranked probability score (CRPS), a strictly proper scoring rule [37].

Table 3.1: EIA's Short-Term Energy Outlook (STEO) naming convention for the 17 quantities analyzed in this study, grouped into five categories: primary production, electricity generation, electricity capacity, fossil fuel prices, and wholesale electricity prices. Series identifiers, descriptions, and units were obtained from EIA's application program interface (API) as of December 2024 [38].

Category	Series Id	Series Description	Unit
● Primary production	CLPRPUS_TON	Coal Production Total U.S.	million short tons
	COPRPUS	U.S. Crude Oil Production	million barrels per day
	NGPRPUS	Natural Gas Total Dry Production	billion cubic feet per day
● Electricity Generation	CLEPGEN_US	Electric power sector net generation from coal, United States	billion kilowatthours
	HVEPGEN_US	Electric power sector net generation from conventional hydropower, United States	billion kilowatthours
	NGEPGEN_US	Electric power sector net generation from natural gas, United States	billion kilowatthours
	NUEPGEN_US	Electric power sector net generation from nuclear, United States	billion kilowatthours
	TOEPGEN_US	Total electric power sector net generation by all energy sources, United States	billion kilowatthours
● Electricity Capacity	SPEPCGWX_US	U.S. electric power sector solar photovoltaic net summer generating capacity	gigawatts
	WNEPCGW_US	U.S. electric power sector wind net summer generating capacity	gigawatts
● Fossil fuel prices	CLEUDUS	Cost of Coal Delivered to Electric Generating Plants	dollars per million Btu
	NGHHMCF	Natural Gas Henry Hub Spot Price	dollars per thousand cubic feet
	MGWHUUS_$	refiner Wholesale Gasoline Price	dollars per gallon
	WTIPUUS	West Texas Intermediate Crude Oil Price	dollars per barrel
● Electricity Prices	ELWHU_CA	Wholesale Electricity Price, CAISO (California ISO) SP15 zone	dollars per megawatt hour
	ELWHU_TX	Wholesale Electricity Price, ERCOT (Texas) ISO North hub	dollars per megawatt hour
	ELWHU_NE	Wholesale Electricity Price, ISO-NE (New England ISO) Internal hub	dollars per megawatt hour

3.2 Data Acquisition

The forecasts and historical values analyzed in this study are drawn from the U.S. EIA's STEO. Each monthly STEO report is published as a multi-sheet Excel workbook, with tables for production, consumption, capacity, prices, and more. In total, 245 reports were collected and systematically reformatted to ease the later data analysis. The complete dataset was reconstructed directly from the original STEO files downloaded from the EIA's public STEO archive [18]. The dataset consists of STEO forecasts from August 2004 - December 2024, with historical values from January 2001 - November 2024.

Unlike the AEO, analyzed by Kaack et al. [3], which is published once per year, the STEO is issued monthly, creating a large set of forecast vintages for each quantity. This enables evaluation of forecasts at horizons H = 1-24 months. Because the STEO is published monthly and extends forecasts to the end of the next calendar year, the longest horizon in any given report depends on the month of publication. Early-year reports span nearly two years ahead, while late-year reports reach only about one year ahead. When all reports are aggregated, the end result is a dataset with dense coverage at short horizons and a steady, linear drop in the number of available forecasts once the horizon exceeds one year ($H > 13$) (Figure 3.1).

PDF versions of STEO tables were excluded from the workflow. Although these provide long historical coverage dating back to 1983, the scanned tables proved difficult to extract reliably and many lacked machine-readable formatting. Automated optical character recognition (OCR) introduced transcription errors that compromised precision in analysis. For this reason, only Excel-based STEO data was used to ensure high-quality data for analysis.

Forecasts and realized values were imported into a SQLite database, with each record structured to include metadata. Specifically, each entry contains: the source STEO file from which the value was extracted; the publication date of the STEO report; the name of the corresponding table in the STEO Excel workbook; the unique series identifier (series Id) used by EIA to label variables; the period for which the forecast is issued, or the period of the realized value; a data type flag indicating whether the entry represents a forecast or a historical value; the numerical value itself; the unit; and the series description summarizing the variable. This allowed cross-vintage comparison of forecasts and ex post validation against realized historical data. The total number of records in the database is 11 939 387 across 1 454 unique quantities. Almost all units and series descriptions were extracted from EIA's open application program interface (API) [38] and imported to the SQLite database. The SQLite-Python workflow has full reproducibility. Records for any of the 1 454 unique STEO quantities can be extracted from the database, analyzed, and visualized.

3.3 Data Analysis

All subsequent data processing and analysis were carried out in Python, using a workflow built on `pandas`, `numpy`, `sqlite3`, and `scipy.stats`. Queries to the SQLite database extracted time series for each of the studied STEO variables. In several cases, the same underlying quantity was reported under different series identifiers at different points in time due to changes made in the variable definition and unit. These overlapping series were therefore combined into a single continuous time series by applying the appropriate scaling factor. For example, dividing by 100 to convert from cents to dollars, or multiplying by days in the month to convert daily rates into monthly totals. In this way, multiple series Ids were merged into a harmonized series to ensure consistency over the entire sample period. Prices reported in the STEO are not adjusted for inflation [20] and no such adjustments were made for the analysis.

Each forecast vintage was compared with actual outcomes as reported in the most recent STEO. Forecast errors were then computed, enabling error metrics to be aggregated by horizon, by variable, and across the full sample. For instance, the WTI crude oil price quantity has a sample of 4 347 forecasts across 245 STEO reports spanning from August 2004 to December 2024, see Figure 3.1.

Figures were designed to highlight cross-horizon trajectories of historical values in juxtaposition to STEO forecasts, error growth patterns across forecast horizons and years, and probability density functions. To evaluate probabilistic accuracy, different EPIs were constructed and compared using the CRPS.

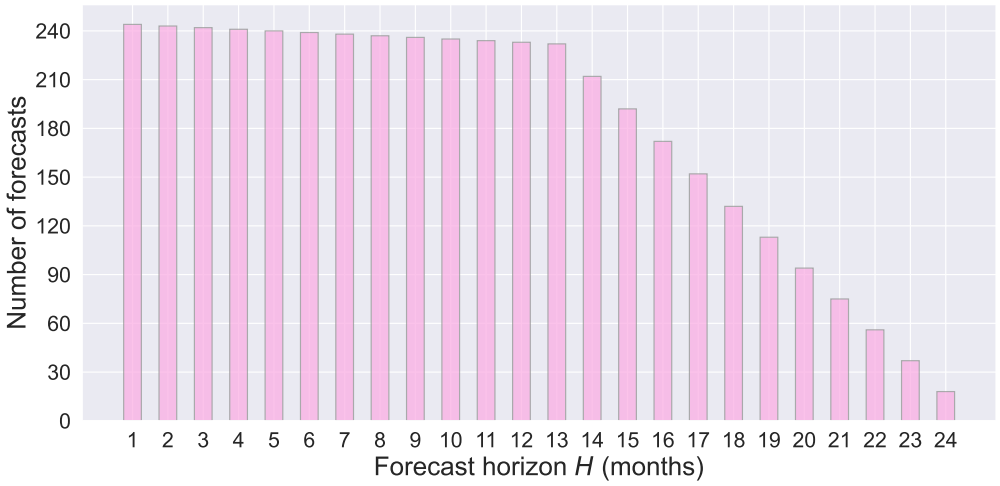


Figure 3.1: The total number of WTI crude oil price forecast points by horizon H from the EIA's Short-Term Energy Outlook from August 2004 to December 2024 used in the analysis.

3.4 Error Metrics

To evaluate the accuracy of the STEO forecasts for the quantities, multiple error metrics were employed. No single metric is sufficient to capture the full range of error characteristics across energy system variables given the differences between physical quantities (production, capacity, generation) and highly volatile price series. Therefore, this study uses multiple point forecast error metrics to characterize the deviations from actual outcomes in historical forecasts. This approach ensures that the magnitude, directionality, and the distribution of forecast errors are assessed. Point error metrics characterize how far historical forecasts deviated from actuals, whereas prediction intervals characterize the probabilistic uncertainty of estimates in the future. EPIs often use information from historical errors to estimate future uncertainties, but not always.

Forecast evaluation is based on the analysis of forecast errors. Let y be the actual value of a quantity and $\hat{y}$ the forecast. The forecast error, or the deviation of the forecast from the actual value, can then be defined as

$$\epsilon = \hat{y} - y. \tag{3.1}$$

To enable comparison across variables with different units, forecast errors are normalized by the actual value. The relative error is defined as

$$\epsilon_{\text{rel}} = \frac{\hat{y} - y}{y} = \frac{\hat{y}}{y} - 1, \tag{3.2}$$

and the relative percent error is

$$\epsilon_{\text{rel}\%} = \frac{\hat{y} - y}{y} \cdot 100. \tag{3.3}$$

From this, the absolute percentage error (APE) can be defined as

$$\text{APE} = \left| \frac{\hat{y} - y}{y} \right| \cdot 100, \tag{3.4}$$

which measures the magnitude of the deviation in percentage terms. The APE is the basis for one of the most widely used forecast evaluation metrics, the mean absolute percentage error (MAPE) [36].

3.4.1 Mean Absolute Percentage Error

The MAPE is defined as the mean of the APE (Equation (3.4)),

$$\text{MAPE}(H) = \frac{1}{n_H} \sum_{t=1}^{n_H} \left| \frac{\hat{y}_{H,t} - y_{H,t}}{y_{H,t}} \right| \cdot 100 \tag{3.5}$$

where n_H denotes the number of errors used to compute the metric at a given horizon H and time t.

MAPE is widely used in forecasting evaluation [36, 39] due to its intuitive expression of average percentage deviation, making it easy to compare errors across quantities. However, MAPE can understate uncertainty for highly volatile series [36]. When actual values approach zero, percentage errors can become extreme, and in volatile series it tends to understate uncertainty by dampening the influence of large deviations. MAPE also places heavier penalties on positive errors than on negative errors [36, 39].

3.4.2 Logarithmic Transformations for Prices

For price variables, which often follow a log-normal distribution owing to the asymmetric nature of their error distributions [3, 40], relative errors are instead transformed into a logarithmic form. This results in a distribution of price forecast errors that is more similar to a normal distribution [3],

$$\epsilon_{\log} = \ln\left(1 + \epsilon_{\text{rel}}\right) = \ln\left(\frac{\hat{y}}{y}\right) = \ln\left(\hat{y}\right) - \ln\left(y\right). \tag{3.6}$$

From this, the absolute logarithmic error (ALE) is defined as

$$\text{ALE} = \left| \ln\left(\frac{\hat{y}}{y}\right) \right| = \left| \ln(\hat{y}) - \ln(y) \right|, \tag{3.7}$$

which measures the absolute difference between forecast and observed values in log space.

To illustrate the contrast between percentage-based and log-based error formulations, Figure 3.2 compares the loss functions for a price. The percentage error loss (APE) increases linearly with relative error, meaning that large deviations dominate the average error. The log error loss (ALE) function grows more gradually and treats proportional over- and under-predictions symmetrically. This is how the loss function changes when the APE is transformed into the ALE. Log-based metrics are therefore more appropriate for volatile and approximately log-normal series like energy prices.

Figure 3.2: Comparison of percentage error loss (APE) (left) and log error loss (ALE) (right) for price forecasts.

We modify the error for price forecasts by taking the mean of Equation (3.7),

$$\text{MALE}(H) = \frac{1}{n_H} \sum_{t=1}^{n_H} \left| \ln\left(\hat{y}_{H,t}\right) - \ln\left(y_{H,t}\right) \right|. \tag{3.8}$$

This transformation makes the measure more robust to volatility [3]. However, the MALE metric is harder to interpret, especially side-by-side with MAPE, since it uses relative logarithmic distances instead of raw distances between forecasts and observed values. For comparability with percentage-based metrics like MAPE, MALE in Equation (3.8) is modified to yield the exponential mean absolute log error (EMALE):

$$\text{EMALE}(H) = \left(\mathrm{e}^{\text{MALE}(H)} - 1\right) \cdot 100. \tag{3.9}$$

3.4.3 Bias

The forecast bias measures the direction of errors, indicating a tendency to produce point forecasts that systematically are too high or too low compared with realized outcomes. Bias is assessed using the mean percentage error (MPE). The MPE makes it possible to look at whether forecasts are systematically overestimated or underestimated, as well as how large and variable the deviations are in relative terms, like the forecast horizon or over time. Bias is crucial in the context of energy systems, because even small but persistent

under- or over-predictions can indicate flaws in the forecasting modeling structure, which can be costly [41].

Bias is typically represented by the mean forecast error [41]. For physical quantities in primary production, electricity generation, and electricity capacity, bias is computed as the mean percentage error (MPE),

$$\mathrm{MPE}(H) = \frac{1}{n_H} \sum_{t=1}^{n_H} \left(\frac{\hat{y}_{H,t} - y_{H,t}}{y_{H,t}} \right) \cdot 100. \tag{3.10}$$

Positive values indicate overestimation, while negative values indicate underestimation.

For price quantities, which are log-normally distributed and subject to sharp spikes, MPEs can become skewed and asymmetric. To address this, bias for price variables is calculated using the mean of logarithmic forecast errors,

$$\mathrm{MPE}_{\log}(H) = \left(\mathrm{e}^{\frac{1}{n_H} \sum_{t=1}^{n_H} \ln\left(\frac{\hat{y}_{H,t}}{y_{H,t}}\right)} - 1 \right) \cdot 100. \tag{3.11}$$

This formulation transforms the mean log error back into percentage terms, making it possible to compare MPEs of price quantities to physical quantities.

3.5 Empirical Prediction Intervals

While the STEO typically publishes only point forecasts for most variables, probability distributions can also be estimated around these forecasts in order to quantify their uncertainty [42]. Such probabilistic forecasts, often referred to as density forecasts [3, 28], specify a full predictive distribution for future outcomes rather than a single expected outcome.

As Gneiting & Raftery [37] emphasize, good probabilistic forecasts should be both calibrated and sharp, meaning that they ought to be consistent with observed outcomes and concentrated around likely values. Furthermore, Kaack et al. [3] remark that "A density prediction is best if it equals the probability density function (PDF) from which this future observation is drawn." Empirical approaches approximate this underlying distribution by drawing on historical forecast errors or volatility in observed outcomes in the historical time series rather than imposing rigorous distributional assumptions.

To evaluate uncertainty representation in the STEO, four empirical prediction interval (EPI) methods are implemented: NP1, NP2, G1, and G2 (Table 3.2).

Table 3.2: Comparison of the four different empirical prediction interval methods used to construct the density forecasts.

Name	EPI method	Based on	Median centered
NP1	Nonparametric	Past forecast errors	No
NP2	Nonparametric	Past forecast errors	Yes
G1	Parametric Gaussian distribution	Past forecast errors	Yes
G2	Parametric Gaussian distribution	Historical variability	Yes

3.5.1 Nonparametric Empirical Prediction Intervals (NP1)

The nonparametric empirical prediction interval (NP1) method relies on the sole assumption that historical forecast errors are representative of future forecast errors. Based on this premise, the distribution of past errors at each forecast horizon H gives a probabilistic estimate of future actual values. A nonparametric approach is used to model the error distributions. It makes next to no assumptions about the underlying distributions of the errors, which makes it more flexible than parametric methods. NP1 directly samples the historical error distribution at each H without centering it. This preserves any median bias observed in past forecasts, making it a suitable approach to reveal biases.

The method collects relative errors ϵ_{rel} for a given H. For each forecast, it samples past ϵ_{rel} with the same H and calculates the empirical percentiles of these historical errors. They are then converted to prediction intervals using the following formula,

$$y_{\text{quantile}} = \frac{\hat{y}}{1 + \epsilon_{\text{rel}}}. \tag{3.12}$$

For logarithmic errors, the density forecast is formulated as:

$$y_{\text{quantile}} = \hat{y}\ \mathrm{e}^{\epsilon_{\log}}. \tag{3.13}$$

3.5.2 Centered Nonparametric Empirical Prediction Intervals (NP2)

The centered nonparametric empirical prediction interval (NP2) method modifies the NP1 method by centering the distribution of historical forecast errors such that the median error ϵ becomes zero.

For each H, the median of the relative errors m_{rel} is computed according to

$$m_{\text{rel}} = \text{median}(\epsilon_{\text{rel}}). \tag{3.14}$$

Each error is then centered,

$$\epsilon'_{\text{rel}} = \epsilon_{\text{rel}} - m_{\text{rel}}, \tag{3.15}$$

which yields a zero-median error distribution, ensuring that the forecast remains the median of the predictive interval. The following formula is used to shift the empirical distribution:

$$y_{\text{quantile}} = \frac{\hat{y}}{1 + \epsilon'_{\text{rel}}}. \tag{3.16}$$

For log-transformed forecast errors, NP2 centers the distribution of log errors,

$$\epsilon'_{\log} = \epsilon_{\log} - \text{median}(\epsilon_{\log}). \tag{3.17}$$

Then, the prediction intervals are constructed as

$$y_{\text{quantile}} = \hat{y}\ \mathrm{e}^{-\epsilon'_{\log}}. \tag{3.18}$$

NP2 symmetrically spreads uncertainty around the point forecast. This ensures that the point forecast is always the median of the predictive distribution, which reduces over-dispersion in contexts where errors are small and symmetric. By centering the distribution around $\epsilon = 0$, the density predictions avoids generating a second point forecast in cases where bias is present in the historical forecast errors, a feature of the NP1 method [3]. However, this formulation treats STEO point forecasts as the best estimates, even though they are not invariably so, and in doing so it also removes evidence of systematic bias from the predictive distribution.

3.5.3 Parametric Gaussian Distribution Based on Forecast Errors (G1)

The parametric Gaussian distribution (G1) method, based on historical forecast errors, estimates a Gaussian probability distribution around each point forecast, using the standard deviation σ of past forecast errors at the same horizon H. The method assumes a mean and a median of $\epsilon = 0$ and that forecast errors are normally distributed. We test the assumption that forecasts are unbiased by measuring whether the density predictions actually center on zero. These assumptions simplify the distribution to one parameter (σ) rather than having to model both bias and variance. Such simplicity can be advantageous in contexts with limited data, since estimating only one parameter reduces estimation error. The risk of this approach is that the Gaussian assumption may fail to capture important features of the true error distribution, such as heavy tails, resulting in underestimation of the probability of extreme events [3]. Thus, while Gaussian EPI methods provide efficient approximations, they do so at the expense of potentially misrepresenting the uncertainty in turbulent energy market conditions.

ϵ_{rel} is estimated and grouped by H. For each H, σ of past errors are computed. The G1 method then assumes that forecasts errors are distributed according to,

$$\epsilon_H \sim \mathcal{N}(0,\ \sigma_H^2), \tag{3.19}$$

where $\mathcal{N}$ denoted the normal distribution with zero mean and variance σ_H^2. Quantiles of this distribution are obtained through the inverse cumulative distribution function (CDF) of the normal distribution. The quantile q for probability level p is given by

$$q_{p,\,H} = z_p\ \sigma_H, \tag{3.20}$$

where z_p denotes the standard normal quantile ($z_{0.16} \approx -1$, $z_{0.84} \approx 1$). These quantiles are then transformed to forecast intervals using the same relation as in Equation (3.12),

$$y_{p,\,H} = \frac{\hat{y}_H}{1 + q_{p,\,H}}. \tag{3.21}$$

This yields prediction intervals at desired coverage levels, meaning the proportion of the probability mass of the distribution is contained within the interval. For a Gaussian distribution centered at zero, the central 68% interval corresponds to one standard deviation $\pm\sigma_H$ around the mean ($p = 0.16$ and $p = 0.84$), while the central 95% interval corresponds to two standard deviations $\pm 2\sigma_H$ ($p = 0.025$ and $p = 0.975$). Figure 3.3 illustrates the

empirical error distributions underlying the G1 method. These distributions are then approximated by Gaussian fits, with σ_H estimated separately for each horizon. While the empirical errors may deviate from perfect normality, the G1 method relies on this parametric assumption to construct horizon-specific prediction intervals. These intervals are then transformed into value space to produce the fan chart representation of forecast uncertainty. The nonlinear mapping from error space to levels introduces asymmetry in the intervals, even though the Gaussian assumption is symmetric in error space.

For price quantities, the same procedure is applied to log-transformed errors instead of the relative errors.

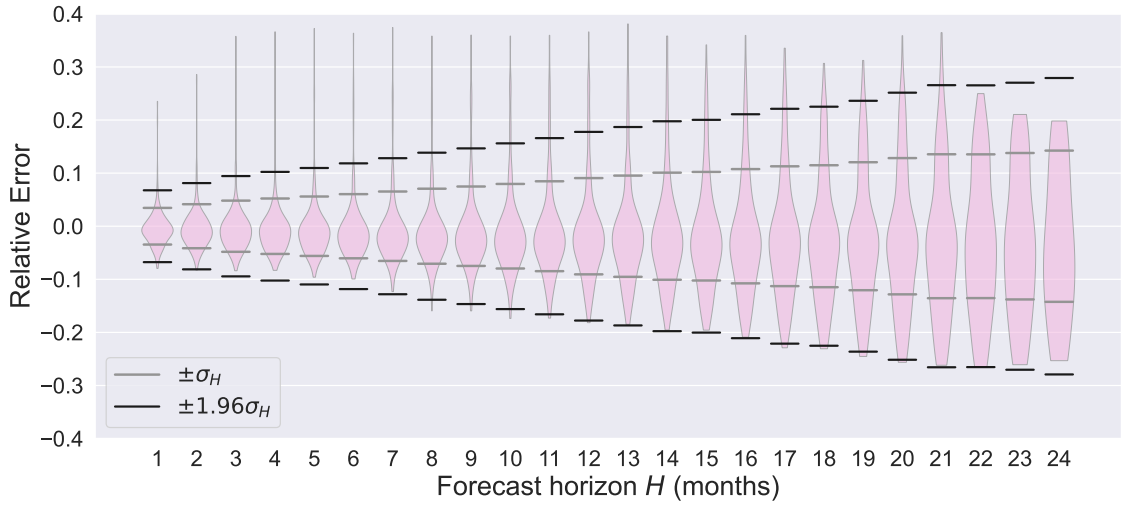


Figure 3.3: Empirical distribution of historical forecast errors by horizon ($H = 1, ..., 24$) for U.S. crude oil production, visualized as violin plots. The violins (pink) represent the observed distributions of relative errors for each forecast horizon. Horizontal lines indicate the fitted Gaussian bands at $\pm\sigma_H$ (gray) and $\pm 1.96\sigma_H$ (black), corresponding to one and two standard deviations of the empirical error distribution, respectively. These Gaussian fits form the basis for the G1 prediction interval method.

3.5.4 Parametric Gaussian Distribution Based on Historical Variations (G2)

The G2 method constructs prediction intervals from the historical volatility in the underlying variable. Uniquely among the four EPIs described here, it does not use forecast errors at all. Thus, it can be seen as the exact opposite of NP1, which uses nothing but historical forecast errors to form an EPI; yet, it retains NP1's parsimony. This makes it a method that assumes that the volatility of the underlying time series itself provides a proxy for forecast uncertainty. The method is based on a standard deviation which takes a pair of historical observations, a horizon H apart, and calculates the relative change in the later value with respect to the earlier value. This is analogous to the relative error. For a historical time series of observed values y_t, the relative change is defined as

$$\Delta_H = \frac{y_t - y_{t-H}}{y_{t-H}}. \tag{3.22}$$

This collection of Δ_H across the sample is then assumed to follow a Gaussian distribution centered at zero, with variance σ_H^2 estimated as the sample variance of the relative changes,

$$\Delta_H = \mathcal{N}(0,\ \sigma_H^2). \tag{3.23}$$

At each H, the standard deviation σ_H is used to parameterize a Gaussian error distribution. Prediction intervals are constructed from the inverse CDF. The quantiles are then transformed into prediction intervals around the STEO point forecasts $\hat{y}_H$ using Equation (3.21), where $q_{p,\,H}$ is estimated from historical relative changes. This ensured that the density forecast expands with increasing empirical variability in the historical series. Figure 3.4 illustrates this construction for an example quantity. The G2 method applies this procedure to all observed pairs in the historical time series for each horizon H, generating a distribution of relative changes Δ_H. The variance of this distribution then provides the parameter σ_H^2, which defines the Gaussian error distribution used to construct the prediction intervals.

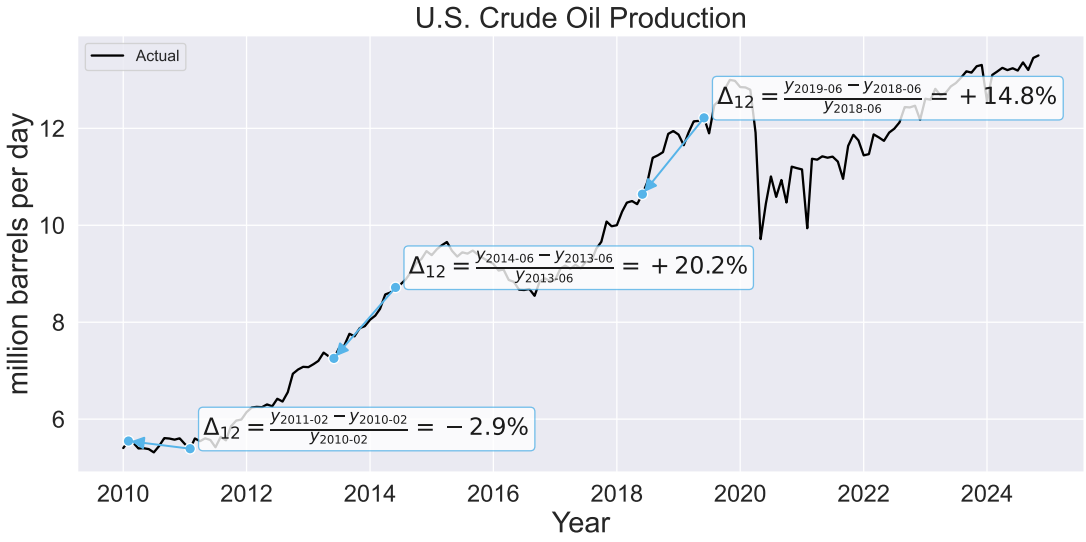
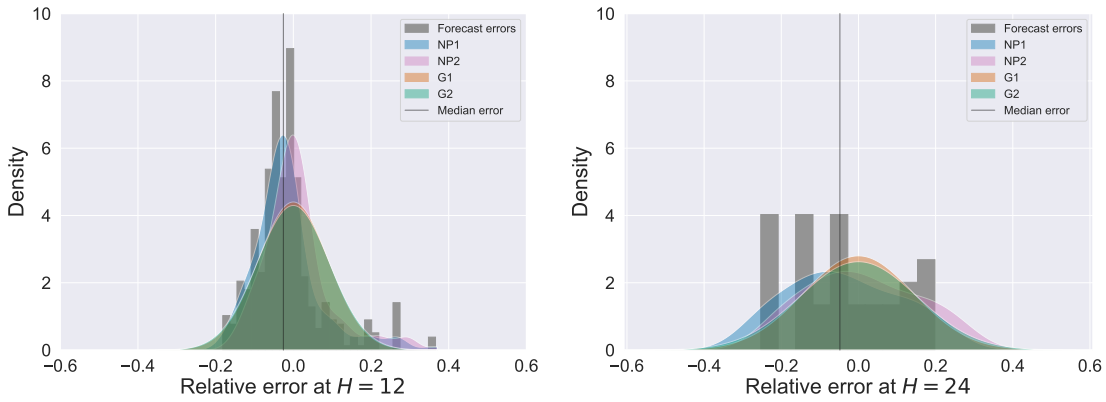


Figure 3.4: Illustration of the calculation of relative changes (Δ_H) in a historical time series for an example variable (U.S. crude oil production). Each Δ_H is defined as the percentage change between two observations separated by a forecast horizon H. In this example, $H = 12$ months. These relative changes provide the empirical variance used in the G2 prediction interval method. In the implementation, this procedure is carried out for all horizons (H = 1-24 months) across the entire historical series.

The PDFs used to construct the EPIs for an example quantity at $H = 12$ are illustrated in Figure 3.5 for NP1, NP2, G1, and G2. At the 12-month forecast horizon, the fits are based on a larger number of forecast-actual pairs, producing sharper distributions. By contrast, at the 24-month horizon ($H = 24$) fewer observations are available, which leads to more dispersed distributions and greater uncertainty in the estimates.



(a) **(b)**

Figure 3.5: Example probability density functions (PDFs) of forecast errors for U.S. crude oil production using the empirical prediction methods NP1, NP2, G1, and G2. The histogram of raw relative errors are in gray. The vertical line indicates the median error. **(a)** The PDFs at forecast horizon $H = 12$ months. **(b)** The PDFs at $H = 24$ months.

3.5.5 Method Ranking: Continuous Ranked Probability Score

A scoring rule assigns a numerical score based on the predicted distribution and the observed outcome [37]. A proper scoring rule is when a forecaster achieves the highest expected score by issuing a predictive distribution F that matches the true distribution of outcomes. It is strictly proper if F uniquely equals the true distribution, meaning that the lowest score is assigned to the forecast who accurately predicts the true distribution. Strictly proper scoring rules incentivize forecasters to provide honest and well-calibrated assessments of their uncertainty. In essence, scoring rules answer the question "How well does the predicted probability distribution align with the observation?"

One widely used scoring rule for density forecasts is the continuous ranked probability score (CRPS) [3, 37]. CRPS is a strictly proper scoring rule. It does not concentrate on any specific point of the probability distribution, but considers the whole forecast distribution. The CRPS generalizes the mean absolute error (MAE) of the forecasts by measuring the squared distance between the predicted CDF F and the empirical step function centered at the observation [43]. Given a forecast distribution F and an observed outcome y, CRPS is defined as [37]:

$$\text{CRPS}_{F,\,y} = \int_{-\infty}^{\infty} (F(z) - \mathbb{1}\{z \geq y\})^2 \, dz \tag{3.24}$$

where $\mathbb{1}\{z \geq y\}$ is the Heaviside step function [44] and z is the integration variable that represents all possible values in F.

To evaluate and compare the performance of the four different EPI methods NP1, NP2, G1, and G2, the CRPS is employed. The CRPS is a loss function that measures the distance between CDF of a forecast and the empirical Heaviside function centered at the observation. It provides a measure of how good a prediction is by comparing predicted probabilities with realized outcomes, resulting in a score in the respective error metric,

with the scale depending on the quantity. A lower CRPS is a better forecast, meaning that the predicted distribution is closer to the actual value [3].

CRPS values were computed across the forecast horizon H by comparing observed forecast errors with the predictive distribution implied by each method. For NP1 and NP2, this distribution was constructed non-parametrically from historical forecast errors, with NP2 centering errors around the median. Equation (3.24) was modified to

$$\text{CRPS}(F, \epsilon_H) = \mathbb{E}_F|\epsilon - \epsilon_H| - \tfrac{1}{2}\mathbb{E}_F|\epsilon - \epsilon'|, \tag{3.25}$$

where $\mathbb{E}$ denotes the empirical error distribution, ϵ and ϵ' are independent draws from $\mathbb{E}$, and ϵ_H is the realized error at horizon H [3, 37].

For G1, a Gaussian distribution was fitted to past forecast errors at each H, with the mean fixed at zero and the standard deviation estimated as the sample standard deviation of historical errors. For G2, Gaussian error distributions were instead estimated from historical relative changes in the underlying time series separated by horizon H. Under a normal distribution, CRPS has the closed-form expression [37]:

$$\text{CRPS}\left(\mathcal{N}(\mu, \sigma^2), \epsilon_H\right) = \sigma\left[\frac{y-\mu}{\sigma}\left(2\Phi\left(\frac{y-\mu}{\sigma}\right) - 1\right) + 2\phi\left(\frac{y-\mu}{\sigma}\right) - \frac{1}{\sqrt{\pi}}\right], \tag{3.26}$$

where Φ and ϕ denote the standard normal CDF and PDF, respectively. The mean μ is set to zero because the errors are assumed to have zero mean when centered.

The final ranking for each variable was obtained by averaging CRPS values across all available forecasts. Using the entire range (H = 1-24 months) captures overall performance across – relative to the STEO – short-, medium-, and long-term forecasts, providing a comprehensive measure of each method. In addition, the CRPS was also computed for shorter horizons (H = 1–6 months) to examine how the ranking changes when forecasts are generally more accurate. This allows a comparison between the performance of methods in the full set of horizons and their behavior in the more reliable short-term range, potentially highlighting differences that would be hidden by a single horizon or by only considering the full range.

4

Results

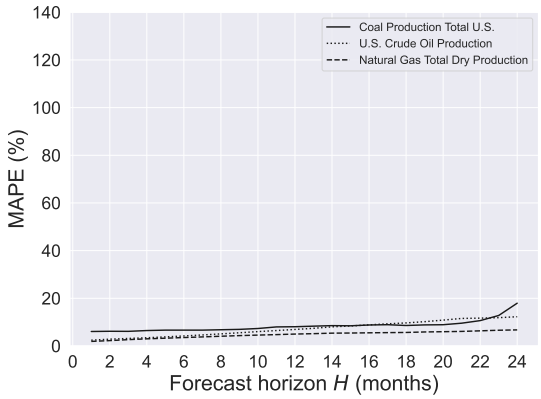
This chapter presents the empirical evaluation of variables from the EIA's STEO using both point forecast metrics and probabilistic methods. The results are structured in three parts. First, point forecast accuracy is assessed with MAPE and EMALE, alongside an analysis of directional tendencies through forecast bias using MPE. These measures highlight systematic over- or under-prediction and the relative magnitude of forecast errors across horizons. Second, the representation of forecast uncertainty is investigated using EPIs, which reveal how well different methods capture the distribution of possible outcomes. Finally, the probabilistic performance of these methods is compared using the CRPS, which provides an aggregate measure of density forecast accuracy across horizons. Together, these analyses provide a comprehensive picture of the accuracy, bias, and uncertainty representation in STEO forecasts.

4.1 Point Forecast Errors

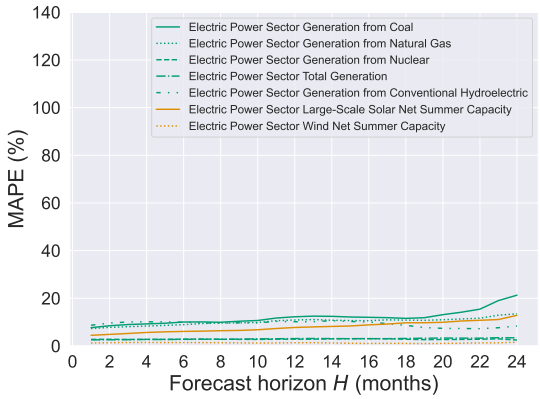
The studied errors reveal a consistent pattern: error magnitudes are substantially higher for price forecasts than for forecasts of physical quantities such as electricity generation or primary production. For primary production variables, see Figure 4.1a, the error is relatively low at shorter horizons and increases gradually with H. The same is observed for the electricity generation and capacity variables in Figure 4.1b, although some variables exhibit larger errors, such as generation from coal, natural gas, and conventional hydroelectric. Fossil fuel prices have higher errors than the errors for the corresponding primary production quantities. This is especially visible in WTI crude oil and natural gas prices, where the errors consistently exceed those measured for oil and natural gas production (Figures 4.1a and 4.1c). Furthermore, the errors in primary production quantities are linear (or constant like coal production) over the forecast horizon H, while the fossil fuel prices (except for the cost of coal) exhibit a logarithmic growth. This is not surprising since the price quantities are logarithmically transformed.

The difficulty in forecasting variables inherently subjected to sudden exogenous changes, like extreme weather conditions, is apparent when contrasting physical quantities (Figures 4.1a and 4.1b) with electricity prices (Figure 4.1d). The physical quantities evolve more predictably, as seen in the linear or nearly constant behavior of the evolution of error over horizon. They are subjected to more predictable changes, like infrastructure developments, which results in lower errors. The forecast errors for electricity prices are more erratic and

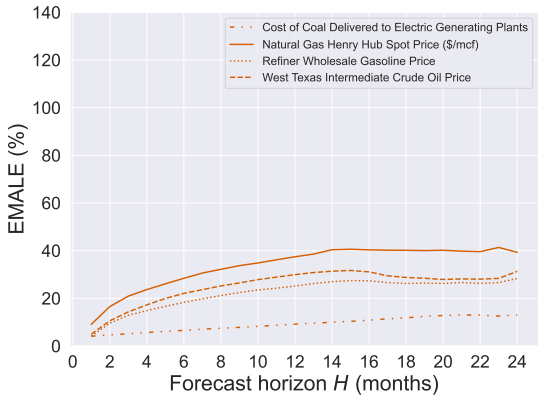
larger, but they follow a somewhat linear growth over increasing H up to around $H = 18$. After $H = 18$ months, the errors are more irregular than those between $H = 1$-18 months. This could be due to the declining numbers of STEO measurements beyond $H = 13$ (see Figure 3.1), which may lead to an uncertainty in the analyzed data itself.



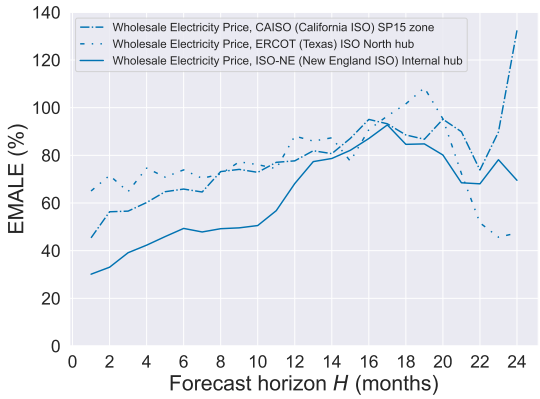
(a) Forecast errors for primary production, August 2004 - November 2024.



(b) Forecast errors for electricity generation and capacity, August 2004 - November 2024 (hydroelectric from October 2007; solar and wind capacity from July 2017).



(c) Forecast errors for fossil fuel prices, August 2004 - November 2024.



(d) Forecast errors for electricity prices, August 2019 - November 2024.

Figure 4.1: Forecast errors across the forecast horizon H for quantities reported in EIA's STEO. Errors for physical quantities are measured using the mean absolute percentage error (MAPE), while errors for price quantities are measured using the exponential mean absolute log error (EMALE).

At the 12-month horizon, a clear difference is seen between forecast errors for physical quantities and prices in the STEO (Figure 4.2). Primary production and electricity generation and capacity forecasts are comparatively accurate, with typical errors below 10%, reflecting the predictability of supply-side developments and seasonal demand patterns over the short term. Price forecasts display larger errors, with fossil fuel prices such as WTI crude oil and natural gas Henry Hub spot price showing errors in the range of 30-40%. Coal supply and price exhibit similar errors (8.1% and 9.1%, respectively), reflecting the

fact that almost all U.S. coal production is used for electricity generation, which naturally leads to correlated errors. The highest errors are observed for electricity prices, where errors exceed 70%, underscoring the difficulty in anticipating short-term price dynamics in electricity markets.

Looking at 12-month ahead forecasts in Figure 4.2 provided a trade-off between a short-term forecast horizon and a horizon where forecasts remain relevant to policy making and investments. While short horizons forecasts are often more accurate, they offer limited strategic value for users as conditions are already largely determined. At the other end, horizons H = 13-24 months are increasingly uncertain, partly due to there being fewer observations at longer horizons (see Figure 3.1). After $H = 13$, the STEO produces fewer forecast-observation pairs, meaning error distributions become noisier and less reliable compared to forecast $H \leq 13$. For instance, this is seen for electricity prices in Figure 4.1d where the errors diverge beyond $H = 18$ from the linear error pattern observed in H = 1-18. Therefore, $H = 12$ capture both the limits of predictability and the time frame where forecasts are actionable for planning and policy.

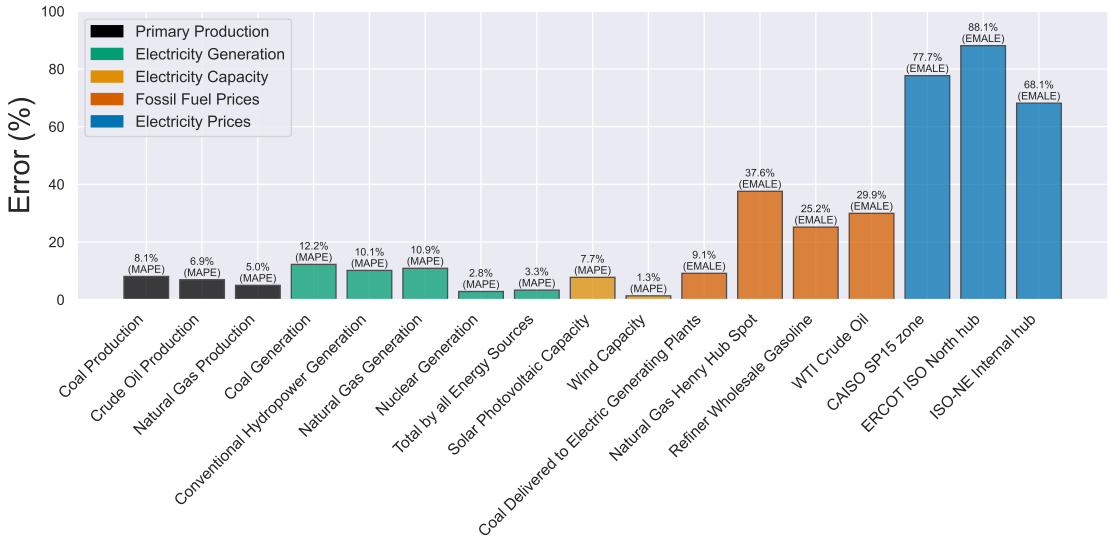


Figure 4.2: Forecast errors at the 12-month horizon ($H = 12$) for quantities reported in EIA's STEO, grouped into primary production (black), electricity generation (green), electricity capacity (orange), fossil fuel prices (red), and electricity prices (blue). For physical quantities, mean absolute percentage error (MAPE) is used, while for price quantities the exponential mean absolute log error (EMALE) is used.

This reflects a general challenge in forecast evaluation: when sample sizes diminish, estimated error distribution becomes unstable, skewing assessments of forecast quality. Kaack et al. [3] note that empirical prediction intervals are sensitive to sample size and that coverage estimates may be distorted when data are sparse. This means that the uncertainty estimation itself becomes more uncertain the further ahead one makes the density forecasts. Thus, part of the irregularity observed beyond $H = 18$ months in Figure 4.1d may not only reflect poor energy price forecasting skill, but also the inherent difficulty of evaluating accuracy with limited data.

To see whether the forecasts have improved or not over time, we look at the evolution of forecast errors for primary production quantities, see Figure 4.3. The errors are normalized to their respective levels in 2004, so each time series illustrates the relative fluctuation across the sample period rather than absolute magnitudes. Coal production exhibits the most pronounced peaks in the normalized error series, with an error ratio exceeding four times the baseline 2004 level in 2015 and more than three times in 2019. Crude oil and natural gas production, by contrast, mostly lie below the 2004 baseline, meaning that the forecast accuracy for these commodities improved relative to 2004. However, this improvement has not been linear – errors have remained on a consistent level until 2020, after which forecast methods and market conditions began to yield steadily better forecasts.

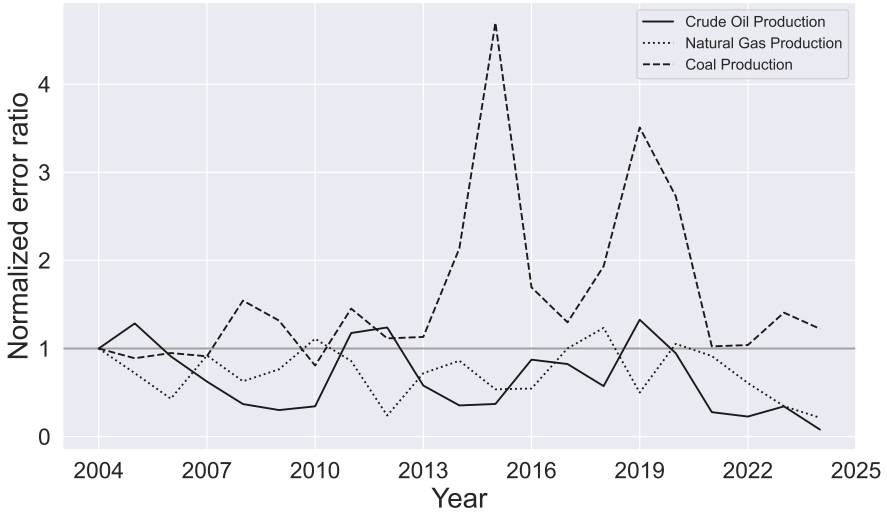


Figure 4.3: Forecast errors for crude oil (solid), natural gas (dotted), and coal (dashed) production in the U.S., normalized to their respective errors in 2004. The errors are yearly averaged and aggregated across all forecast horizons (H = 1-24 months).

4.2 Directional Forecast Tendencies

Most forecasts for the quantities in Figure 4.4 show underestimation of actual outcomes, with the exception of those in electricity generation and capacity, where many cluster around zero bias. Others exhibit a positive bias. Forecasts for solar photovoltaic (PV) capacity (Figure 4.4b) are positively biased, reflecting historical overestimation of this renewable energy source. In contrast, wind capacity forecasts exhibit almost zero bias across H. Notable outliers are forecasts for coal production (Figure 4.4a) and natural gas Henry Hub spot price (Figure 4.4c), which show positive biases, while forecasts for generation from natural gas show a negative bias, from -3% to -10% across H. The negative bias in gas generation appears to be offset by coal and solar PV generation, reflecting the U.S. trend of declining coal use alongside simultaneous growth in gas and renewable energy. This implies that modeling frameworks used for generation, capacity and production variables are generally better at capturing incremental trends and gradual adjustments in infrastructure compared to price quantities. Although the biases for physical quantities are not as strongly directional as in the case of energy prices, even small but

persistent bias may merit attention, particularly where repeated under-prediction of growth could influence investment decisions or policy analyses [3].

Fossil fuel price forecasts errors, Figure 4.4c, reveal a slight but consistent underestimation for coal and gasoline prices. The largest systematic tendencies are seen in electricity prices, Figure 4.4d, where the forecasted prices for all three independent system operators (ISOs) display prominent negative bias, exceeding -20% to -50% across horizons. This consistent under-prediction suggests systematic tendencies in the STEO modeling system toward underestimating upward price shocks.

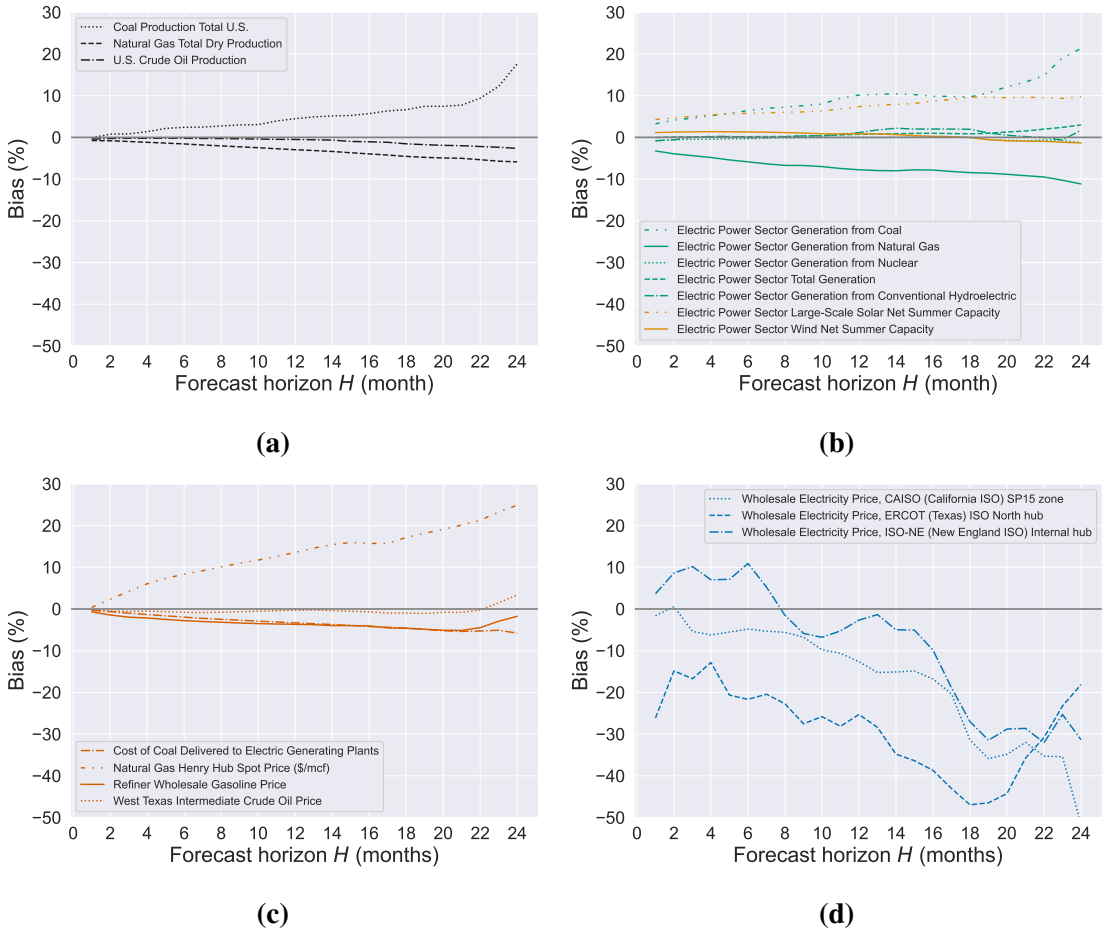


Figure 4.4: Bias in EIA's STEO forecasts across whole the forecast horizon H = 1-24 months, grouped by **(a)** primary production (black), **(b)** electricity generation (green) and capacity (orange), **(c)** fossil fuel prices (red), and **(d)** electricity prices (blue). Bias is measured as the mean percentage error (MPE) for physical quantities, and the log error transformation of the MPE for price quantities. Positive values indicate overestimation and positive values indicate overestimation relative to actual outcomes.

Analyzing U.S. crude oil production forecast bias before and after July 2008, a shift in EIA's STEO forecasting performance is revealed, see Figure 4.5. When examining the entire dataset (Figure 4.4a) crude oil production appears to be among the least biased variables, with relatively small deviations from actuals across the forecast horizon. However, splitting

the sample shows that this overall picture of minor bias masks substantial historical change.

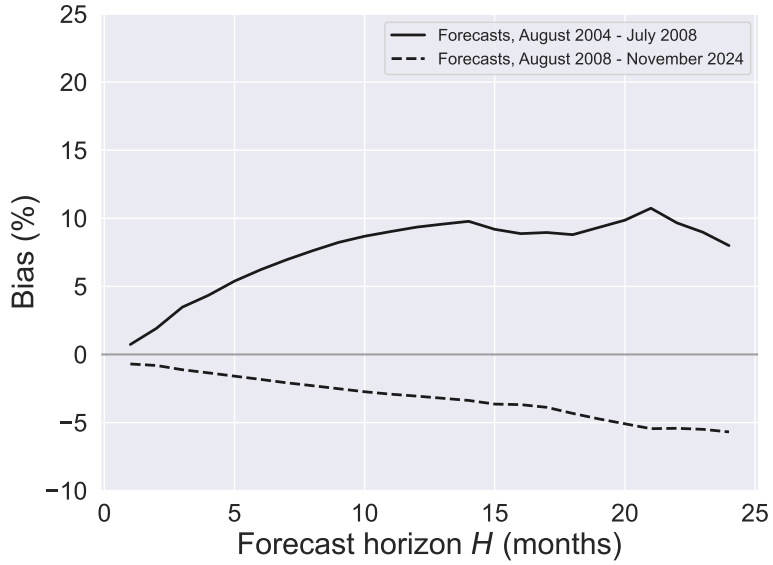


Figure 4.5: Bias in EIA's STEO U.S. oil production forecasts across the whole forecast horizon H = 1-24 months. The data was split into two sample periods: August 2004 - July 2008 (solid), and August 2008 - November 2024 (dashed). Positive values imply overestimation, while negative values imply underestimation.

Before and including July 2008, oil production forecasts are tied to a positive bias, with overestimation increasing with horizon and reaching nearly 10% by two years ahead. After July 2008, the forecasts shifted to to a negative bias, with underestimations going down to about -6% at longer horizons. This shows that the apparent lack of bias in the whole sample period is caused by offsetting effects. Earlier overestimation and later underestimation cancel each other out.

Looking further into forecast bias in STEO quantities pertaining to primary production and fossil fuel prices, a consistent pattern of structural change around July 2008 becomes evident, see Figure 4.6, and many fossil fuel prices quantities display much smaller biases after July 2008 compared to levels before then that were strongly negatively biased.

In primary production, crude oil forecasts shift from a positive bias (8.7%) before 2008 to a negative bias (-3.1%) afterwards. On the other hand, WTI crude oil price forecasts shifted from a strongly negative bias before July 2008 (-11.6%) to a negligible bias after 2008 (-1.2%). Both quantities saw a significant lessening of the bias to very small levels. Since 2008, natural gas and oil production have exhibited similar biases (-3.5% and -3.1%, respectively), reflecting the coupling of their production processes.

Coal production and natural gas Henry Hub spot price both had forecasts that were essentially unbiased at H = 12 before July 2008, resting at 0.2% and 2.5%, respectively. After 2008, both quantities exhibit a large increase in forecast bias: coal production at 5% and natural gas Henry Hub spot price at 16.5%.

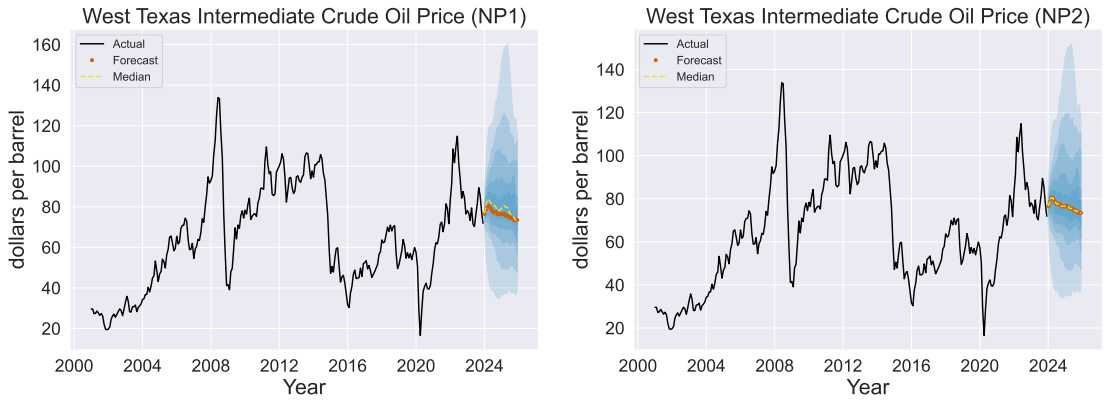
Figure 4.6: Forecast bias at the 12-month horizon ($H = 12$) of EIA's STEO quantities in **(a)** primary production and **(b)** fossil fuel prices, using data that was split into two periods: August 2004 - July 2008 and August 2008 - November 2024. Positive values indicate overestimation and negative values indicate underestimation.

4.3 Empirical Prediction Intervals

4.3.1 Density Forecasts

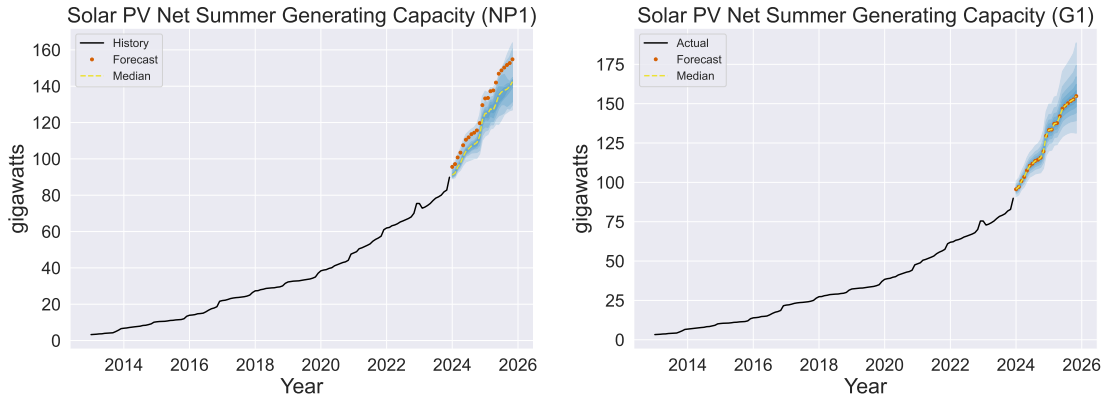
The NP1 and NP2 methods construct their prediction intervals using historical forecast errors without imposing distributional assumptions. A key feature, visible in the empirical density forecasts for WTI crude oil price in Figure 4.7, is the pronounced asymmetry and upper-skewed uncertainty towards higher prices. A heavy upper tail develops as the forecast horizon H increases, especially in the NP1 intervals, consistent with the heavy-tailed distribution of energy prices noted by Kaack et al. [3]. The CRPS in Table 4.1 show that NP1 achieves the best performance for WTI crude oil price forecasts aggregated across the whole forecast horizon H = 1-24 months. At shorter horizons, H = 1-6 months, the NP2 method performs best for this quantitity.

NP1 better preserves the empirical asymmetry in forecast errors, making it more informative for capturing directional risk compared to parametric methods like G1, shown in Figure 4.8 for U.S solar PV net summer generating capacity. EPIs constructed with NP1 and G1 methods reveal the differences in how the two methods capture uncertainty. The G1 method assumes a zero-mean Gaussian error distribution at each horizon, resulting in symmetric intervals around the median. While the spread of the fan chart increases with horizon, the symmetry hides the directional tendencies visible in NP1, especially for wind capacity seen in Figure 4.9. Despite empirical records of forecast errors being more skewed to overestimation in solar PV capacity, the G1 intervals allocate probability mass equally to upper and lower deviations. This risks misrepresenting uncertainty in the density forecasts.



(a) **(b)**

Figure 4.7: Density forecasts for WTI crude oil price constructed using **(a)** the nonparametric and non-median centered empirical prediction interval method NP1, and **(b)** the nonparametric and median centered empirical prediction interval method NP2. The historical data span is between August 2004 and December 2023, with forecasts extending from January 2024 to December 2025. Both approaches rely on historical forecast errors without imposing distributional assumptions. The density forecast is in progressively darker shades of blue, corresponding to central prediction intervals constructed from the percentiles 2, 10, 20, 30, 40, 50, 60, 70, 80, 90, and 98. EIA's STEO forecasts are denoted by red dots. The median line is indicated by a yellow dashed line.



(a) **(b)**

Figure 4.8: Empirical prediction intervals for U.S. solar photovoltaic (PV) net summer generating capacity. The density forecasts were constructed using **(a)** the nonparametric non-median centered NP1 method, and **(b)** the parametric median-centered G1 method based on historical forecast errors.

The best density forecast for U.S. wind net summer generating capacity is constructed with the nonparametric NP1 method, seen in Figure 4.9. The NP1 intervals capture both the growth of wind capacity and the underlying uncertainty about the pace of expansion. The upper bounds are close to the median, suggesting limited upside risk and reflecting a historical tendency of forecasts to slightly overestimate future wind deployments, leaving more probability density mass on the downside.

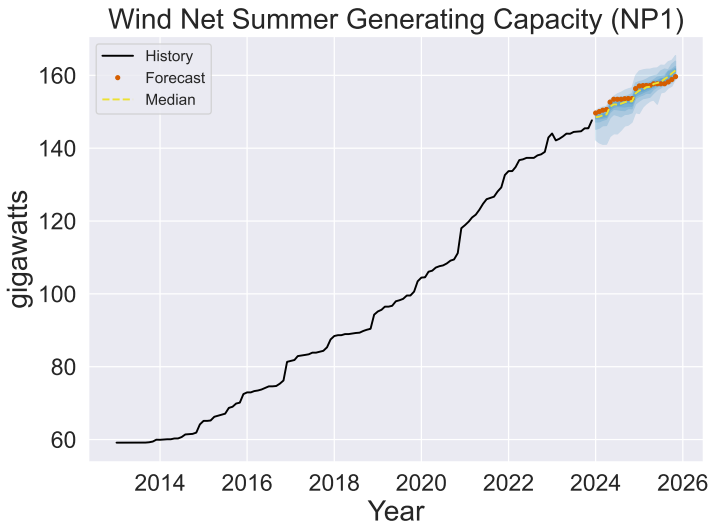
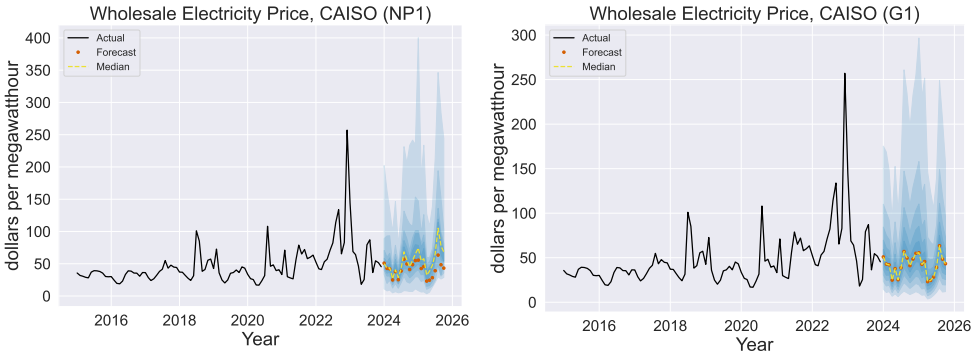


Figure 4.9: Density forecast for U.S. wind net summer generating capacity constructed with the nonparametric and non-median centered NP1 EPI method.

Looking at the density forecasts for highly volatile quantities like wholesale electricity price, the results emphasize that nonparametric methods outperform parametric ones in contexts where error distributions are highly skewed and heavy-tailed. In Figure 4.10, the NP1 method yields prediction intervals for wholesale electricity price in California independent systems operators (CAISO) that widen considerably with larger H and exhibit skewness, particularly toward higher prices. By directly sampling historical forecast errors without a symmetric disposition, NP1 retains the upward skewness in its density forecast, capturing the demand surges, outages, and price shocks in electricity markets [45]. While the G1 method also captures an increase in uncertainty with horizon and the wide upper bounds, it has an imposed symmetry that can lead to an underestimation of the probability of large price spikes.

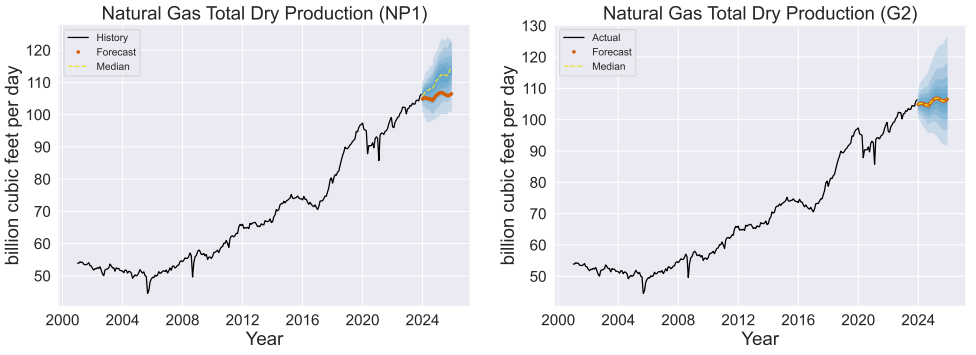
For relatively stable physical quantities, like natural gas production (Figure 4.11), parametric Gaussian methods can perform well, especially at shorter horizons when volatility is lower and distributions more symmetric. Still, nonparametric methods capture nuances more accurately and provide the most reliable density forecasts for natural gas production across the entire forecast horizon as well as for shorter horizons (H = 1-6 months), as seen in Tables 4.1 and 4.2. G2 is the second-best method here. The assumption of zero-centered Gaussian errors with constant variance per horizon is therefore less misleading in this context than for more volatile series.

For natural gas Henry Hub spot prices, NP1 performs best across the whole horizon (Figure 4.12). Aggregating the CRPS over short horizons (H = 1-6 months), however, changes the ranking: G2 outperforms NP1 and becomes the most accurate forecasting density method for this quantity. The Gaussian assumption performs well and does not underestimate the frequency and magnitude of large price spikes, which NP1 appears to do, because G2 samples the historical volatility in the underlying time series data to create the error distribution.



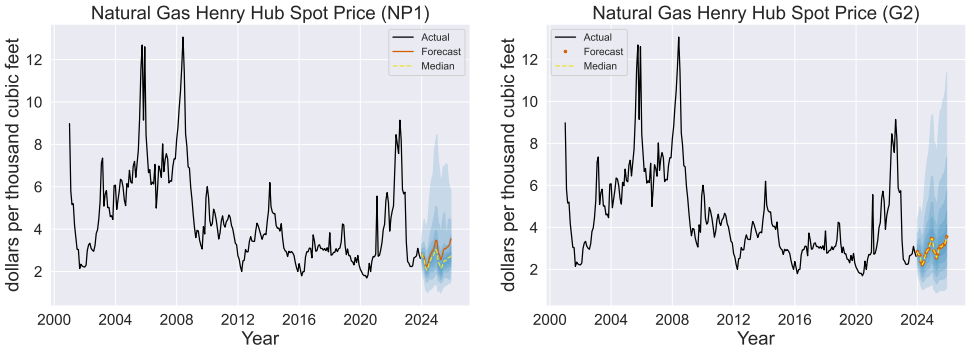
(a) **(b)**

Figure 4.10: Forecast densities for U.S. wholesale electricity prices in the California independent systems operator (CAISO), made using **(a)** the nonparametric non-median centered NP1 method, and **(b)** the parametric median centered G1 method.



(a) **(b)**

Figure 4.11: Prediction intervals for U.S. natural gas total dry production, produced using **(a)** the nonparametric non-centered median NP1 method, and **(b)** the parametric centered median G2 method based on historical deviations.



(a) **(b)**

Figure 4.12: Forecast densities for natural gas Henry Hub spot prices, created using **(a)** the nonparametric NP1 method, and **(b)** the parametric median centered G2 method based on historical variations.

4.3.2 Probabilistic Forecast Accuracy

Table 4.1: The continuous ranked probability score (CRPS) for different density forecasting methods and quantities in EIA's STEO from August 2004 to December 2023, aggregated across all forecast horizons (H = 1-24 months). The lower the score, the better the density [3]. The lowest score in each row is in bold, the second lowest in italics. Rounded to four decimals.

Quantity	NP1	NP2	G1	G2
Coal Production Total U.S.	**0.0553**	*0.0583*	0.0594	0.0595
U.S. Crude Oil Production	**0.0443**	0.0460	0.0464	*0.0459*
Natural Gas Total Dry Production	**0.0256**	0.0313	0.0308	*0.0302*
Electric power sector net generation from coal, United States	**0.0646**	*0.0728*	0.0745	0.0746
Electric power sector net generation from conventional hydropower, United States	**0.0686**	0.0693	*0.0686*	0.0827
Electric power sector net generation from natural gas, United States	**0.0571**	*0.0697*	0.0711	0.1016
Electric power sector net generation from nuclear, United States	**0.0207**	0.0211	*0.0209*	0.0300
Total electric power sector net generation by all energy sources, United States	**0.0260**	0.0273	*0.0271*	0.0396
U.S. electric power sector solar photovoltaic net summer generating capacity	**0.0272**	0.0533	*0.0504*	0.1082
U.S. electric power sector wind net summer generating capacity	**0.0094**	*0.0101*	0.0103	0.0233
Cost of Coal Delivered to Electric Generating Plants	*0.0555*	**0.0564**	0.0584	0.0590
Natural Gas Henry Hub Spot Price	**0.1902**	*0.1968*	0.1996	0.1983
Refiner Wholesale Gasoline Price	**0.1406**	*0.1429*	0.1444	0.1433
WTI Crude Oil Price	**0.1605**	*0.1613*	0.1625	0.1613
Wholesale Electricity Price, CAISO (California ISO) SP15 zone	**0.3246**	0.3663	*0.3659*	0.3630
Wholesale Electricity Price, ERCOT (Texas) ISO North hub	**0.4276**	*0.4604*	0.4686	0.4724
Wholesale Electricity Price, ISO-NE (New England ISO) Internal hub	0.3364	**0.3330**	0.3419	*0.3338*

Table 4.2: The continuous ranked probability score (CRPS) for different density forecasting methods and quantities in EIA's STEO from August 2004 to December 2023, aggregated across forecast horizon H = 1-6 months. The lowest (best) score in each row is in bold, the second lowest in italics. Rounded to four decimals.

Quantity	NP1	NP2	G1	G2
Coal Production Total U.S.	**0.0476**	*0.0477*	0.0483	0.0485
U.S. Crude Oil Production	0.0281	**0.0266**	0.0289	*0.0272*
Natural Gas Total Dry Production	**0.0182**	0.0196	0.0198	*0.0193*
Electric power sector net generation from coal, United States	**0.0559**	*0.0592*	0.0604	0.0641
Electric power sector net generation from conventional hydropower, United States	**0.0659**	0.0686	*0.0680*	0.0810
Electric power sector net generation from natural gas, United States	**0.0565**	0.0617	*0.0605*	0.0966
Electric power sector net generation from nuclear, United States	**0.0204**	0.0210	*0.0206*	0.0298
Total electric power sector net generation by all energy sources, United States	**0.0234**	0.0246	*0.0246*	0.0384
U.S. electric power sector solar photovoltaic net summer generating capacity	**0.0155**	0.0392	*0.0357*	0.0868
U.S. electric power sector wind net summer generating capacity	**0.0088**	*0.0104*	0.0109	0.0204
Cost of Coal Delivered to Electric Generating Plants	0.0451	**0.0371**	*0.0383*	0.0395
Natural Gas Henry Hub Spot Price	0.1516	*0.1392*	0.1408	**0.1390**
Refiner Wholesale Gasoline Price	0.1011	**0.0942**	*0.0999*	0.1002
WTI Crude Oil Price	0.1122	**0.1061**	0.1110	*0.1086*
Wholesale Electricity Price, CAISO (California ISO) SP15 zone	**0.2942**	*0.3093*	0.3110	0.3110
Wholesale Electricity Price, ERCOT (Texas) ISO North hub	**0.3791**	*0.4084*	0.4225	0.4217
Wholesale Electricity Price, ISO-NE (New England ISO) Internal hub	**0.2462**	*0.2502*	0.2506	0.2526

Performance metrics in Tables 4.1 and 4.2 confirm the impressions. Across the full horizon (H = 1-24 months), NP1 achieves the lowest CRPS. At shorter horizons (H = 1-6 months),

NP1 remains the best method, but NP2 also performs well for many quantities, mostly in fossil fuel prices.

The NP1 method is the most reliable forecasting density method overall, particularly in volatile series where capturing skewness and high uncertainty is essential. This method captures the empirical distribution of short-term forecast errors best. Parametric Gaussian methods, G1 and G2, can perform well in primary production variables that are strongly tied to infrastructure, like oil and natural gas production. Table 4.1 provides a full-sample summary of the scores for all 17 quantities, while Table 4.2 focuses on shorter-term forecasts, making it possible to directly compare method performance in near-term horizons and find which method overall provides the most reliable uncertainty estimates.

For physical quantities in primary production and electricity generation and capacity, CRPS values are relatively close across methods. For instance, U.S. crude oil production has values between 0.0443 (NP1) and 0.0464 (G1) at H = 1-24 months, showing that all methods perform comparably in stable series where errors are smaller. Here, even parametric Gaussian-based methods such as G2 perform well. By contrast, fossil fuel and electricity prices show larger spreads across methods, with NP1 still outperforming the Gaussian approaches.

All methods yield similar CRPS values for most physical quantities because uncertainty is low and forecasts are constrained by current infrastructure levels. By contrast, for volatile price series, particularly electricity, the spread of the CRPS values between methods is larger.

5

Discussion

In this chapter, the main findings of the analysis are discussed and placed in a broader context of energy forecasting and uncertainty representation. The results are interpreted in light of the forecasting theory outlined earlier, with respect to systematic bias, error magnitudes, and the treatment of uncertainty in both point and density forecasts. Particular attention is given to structural breaks in forecast performance, differences between physical quantities and price series, and the evidence of anchoring in econometric approaches. The role of empirical prediction intervals and their evaluation through the CRPS is examined, highlighting the trade-offs between parametric and nonparametric approaches. Finally, the implications of these findings for the use of STEO forecasts in energy policy, investment, and risk assessment are considered.

5.1 Forecast Behavior and Predictability

5.1.1 Physical Quantities vs Energy Prices

The results show clear differences between forecasting physical energy quantities and energy prices in terms of predictability. Physical quantities such as primary production, electricity generation, and installed capacity exhibited lower error magnitudes (Figure 4.1) and smaller biases (Figure 4.4) compared to energy prices. These quantities change under infrastructure expansion, technological life cycles, and regulatory environments, all of which follow more gradual and predictable trajectories compared to prices. Energy price forecasts, on the other hand, in particularly short-term electricity prices, are shaped by market-based econometric models that depend on demand, fuel costs for thermal plants, operating costs for renewables and nuclear plants [46, 47]. These models also capture potential policy-driven shifts. Yet, their accuracy remains limited because exogenous shocks and policy-driven changes cannot be anticipated from historical data alone. This may explain why STEO price forecasts display much higher error magnitudes and stronger biases than forecasts for production or capacity. The tendency of STEO forecasts to be disrupted by shocks underlines the need for probabilistic methods such as EPIs, which incorporate the expectation of continued volatility into their uncertainty representation.

Another aspect evident in Figure 4.1 is the systematic relationship between forecast error magnitudes and the forecast horizon H. Consistent with findings by Kaack et al. [3], errors exhibit an upward trajectory as H increases. The further into the future forecasts

extend, the more they are exposed to things like market shocks, unanticipated demand, new energy investments, and changed policies. The effect is especially pronounced in price forecasts, where errors rise steeply for longer horizons. While forecast errors for quantities in primary production, electricity generation, and installed capacity also increase with H, the gradient is typically shallower, suggesting that forecasts of physical quantities retain comparatively higher predictive power over medium- and long-term horizons. This likely reflects the fact that physical quantities evolve along slower-moving paths shaped by infrastructure and regulatory constraints, which impose inertia on the system. Prices, in turn, are far more sensitive to market sentiment and exogenous shocks [48], which drive the steeper error growth over H. For forecast users, this distinction implies that forecasts for physical quantities can provide a relatively reliable basis for planning, whereas price forecasts should be treated with greater caution given their susceptibility to exogenous shocks.

5.1.2 Forecast Patterns: Seasonality and Discrete Expansion

There are quantities that exhibit strong regular seasonal patterns driven by predictable fluctuation in demand. One of these quantities in the STEO is the total electric power sector electricity generation, seen in Figure 5.1a. Electricity consumption, and by extension generation, peaks in the summer and winter months when extreme temperatures lead to high air-conditioning or heating usage, and drops during milder spring and fall periods [49]. These seasonal cycles are clearly visible in the historical data as repeating waves. Importantly, they are also reflected in short-term forecasts, which mirror the ups and downs at the same times of the year. Electricity forecasters can confidently include seasonality because it arises from recurring weather patterns that are well-understood. Unlike the erratic shocks in oil markets, weather-driven demand swings are anticipated each year, so the STEO models clearly incorporate them via seasonal adjustment factors and by explicitly using temperature forecasts and historical profiles [47].

Another forecasting pattern, seen in the wind generating capacity in Figure 5.1b, is a step-like increase over time, as opposed to the flat lines seen for WTI crude oil price in Figure 1.1. Wind power generating capacity, and other renewables like solar, are expanded through discrete developments. For example, the construction of a new wind farm instantly adds a block of capacity when it comes online. Tax incentives play a central role in the timing and scale of wind turbine installations in the United States [50]. They encourage wind installments to begin generating in the earliest calendar year possible, which leads capacity additions to cluster at the end of the calendar year. Consequently, forecasts of installed wind capacity often appear as staircase-like trajectories, with flat stretches punctuated by jumps at the points in time when major projects are expected to come online.

An illustrative example of this staircase pattern is the sharp increase in forecasted wind capacity during 2019 seen in Figure 5.1b. In the January 2019 STEO [51], the EIA forecasted U.S. wind capacity to expand from 96 giga Watts (GW) at the end of 2018 to 107 GW by December 2019, a rapid growth of nearly 20%. The EIA also expected that the annual share of wind electricity generation would, for the first time, exceed the share of hydropower. This expected surge was driven by the large pipeline of projects scheduled

to come online before the phase-out of the federal production tax credit, which provided strong financial incentives for wind power plant developers to complete construction within a certain time window. At the time, there was 232 GW of wind capacity seeking to connect to transmission systems through technical and regulatory processes [52]. The forecasted step-increase was later revised since the forecasted growth did not realize, as seen in later forecasts. This pattern is acknowledged in the EIA's long-term AEO assumptions [53], and known projects and their expected completion dates are compiled and inserted in the STEO models for their short-term forecasts as well [47].

These stair-like and seasonal forecast trajectories contrast sharply with the flat profiles often seen for commodity prices. As seen in Figure 1.1, the forecasts have in various periods been both overestimated and underestimated, but the forecasts are consistently rather flat. During periods of pronounced price volatility, such as the 2008 financial crisis, the forecasts often lagged behind sudden shifts in market conditions. The magnitude of energy price forecast errors in Figure 4.1 reflects the well-documented difficulty of predicting external shocks [9, 12] and potential model bias [6, 34]. Unlike production, generation, and capacity quantities, energy prices have error distributions that are often heavy-tailed, non-linear, and difficult to capture with purely econometric approaches.

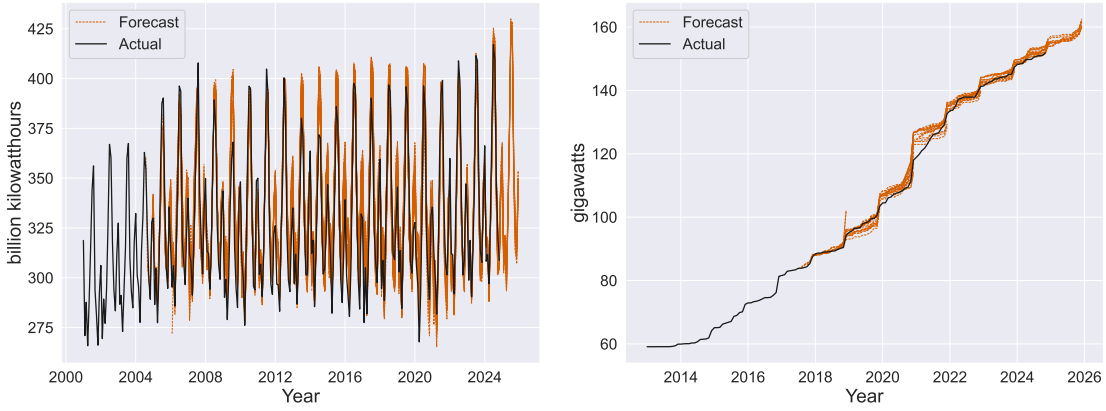


Figure 5.1: Historical values and forecasts of **(a)** U.S. total electric power sector net electricity generation and **(b)** U.S. electric power sector wind net summer generating capacity. The black solid line indicates the latest realized outcomes listed in the EIA's STEO reports, while the red dashed lines show monthly forecasts.

5.1.3 Error Cycles in Coal Production

A further perspective on the robustness of STEO production forecasts is provided in Figure 4.3, which shows forecast errors normalized to their 2004 levels. This normalization shows the relative volatility of forecast performance over time rather than across horizon or in terms of absolute error magnitudes. There is a clear persistence of error cycles in forecasts for coal production, in stark difference to those for natural gas and crude oil production. The sharp increase in error relative to the 2004 base level in 2015 and again in 2019 reflects the extent to which coal production has been shaped by structural changes in the U.S. power sector. The substitution of coal by natural gas and renewables,

the tightening of environmental regulations, and rising mining costs have all lead to a decline in coal demand [54]. Because STEO's econometric models are largely calibrated to historical production levels, they seem to have struggled to forecast the scale of these structural shifts, resulting in large forecast deviations as the coal sector changed.

Crude oil and natural gas production forecasts have been less affected by such structural shifts, most likely because they are more globally traded and subjected to more analyst judgment and predictable demand. Importantly, Figure 4.3 suggests an improvement in forecast performance since 2020, particularly for oil and natural gas. This could reflect both modeling structure refinement and a post-pandemic recovery. Yet, as history has shown, stability may only be temporary.

5.2 Evidence of Bias

5.2.1 Anchoring in Electricity Price Forecasts Leads to Persistent Biases

EIA's STEO electricity price forecasts display characteristics of anchoring to recent historical values rather than adjusting toward structural expectations. The forecasts in Figure 5.2 anchor strongly to recent price spikes, extrapolating similar spikes forward in time, rather than converging toward structural cost expectations. This anchoring behavior shows the reliance on econometric models where forecasts are primarily based on statistical correlations with recent price movements [48], rather than structural models based on supply-demand. Econometric approaches place significant weight on recent data and are not likely to include unprecedented future events unless these are explicitly modeled [55]. While there are benefits to taking such a conservative modeling philosophy, like reducing large forecast errors and reputational risk to the modeling system, it can also entrench systematic bias if market conditions diverge from the historical norm for extended periods.

This anchoring effect can also be linked to the patterns of systematic bias observed in the ISO-NE electricity prices, see Figure 4.4d. At short horizons (H = 1-8 months), the forecasts often overshoot after price spikes, producing positive biases. However, the forecast model consistently fails to predict actual price surges. Because these missed peaks generate very large negative forecast errors, they outweigh the positive errors caused by forecasted spikes. As a result, at longer horizons there is a persistent negative bias. Anchoring, in this case, therefore amplifies transient volatility into systematic forecast bias. Had the STEO modeling structure relied more heavily on structural modeling as opposed to econometric, forecasts might have smoothed over transitory shocks rather than amplifying them forward in time. This suggests that forecast calibration for price variables may benefit from more adaptive mechanisms that can better accommodate rapid changes.

A possible explanation for why systematic biases appear more pronounced in ISO-NE is the region's constrained natural gas infrastructure. Unlike most other U.S. electricity markets, ISO-NE sits at the end of several interstate pipelines, and the total capacity to deliver gas to utilities and generators is limited [56]. During times of stress, like cold spells in the winter, the available gas supply is often prioritized for residential heating, leaving

shortages for electricity generation. These supply constraints, combined with competition for liquefied natural gas (LNG) between New England end users and the broader global market [57], make the region's electricity market particularly vulnerable to price spikes.

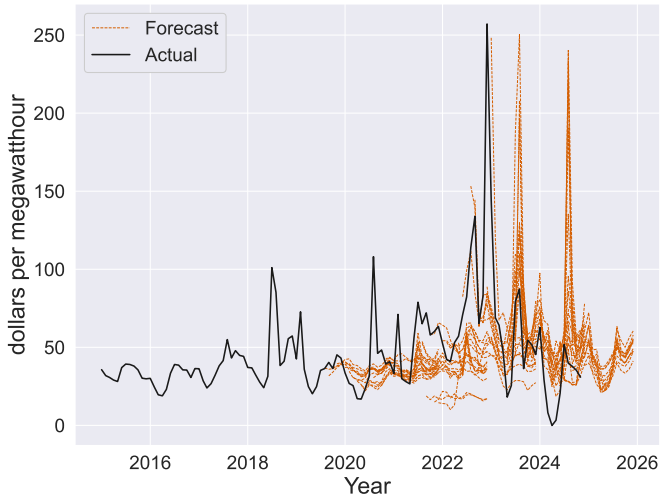


Figure 5.2: EIA's STEO forecasts, denoted by red dashed lines, of wholesale electricity prices in independent system operator New England (ISO-NE) compared with realized outcomes, indicated by the black solid curve. The forecasts extend to December 2025.

5.2.2 Institutional Incentives Drive Inertia in Oil Price Forecasts

Then why do WTI crude oil price forecasts in the STEO appear flat, while wholesale electricity price forecasts exhibit transient surges? The difference lies in the underlying modeling frameworks. Wholesale electricity prices are forecasted using USSTEM, which generates forecasts through econometric relationships. Crude oil price forecasts are not produced within USSTEM but rely on other modeling frameworks like GSTOM and ultimately on EIA analyst judgment [5]. This distinction reflects the fact that oil markets are global and therefore less sensitive to U.S. conditions, whereas domestic gas and coal prices respond more directly to U.S. supply and demand, making econometric modeling frameworks more appropriate for those markets. In politically and economically sensitive markets such as crude oil, there are strong institutional incentives for the EIA to avoid issuing considerable forecast shifts from history. Such conservatism helps reduce the risk of large forecast errors and protects institutional credibility, even if it means that the forecasts respond slowly to new information. The apparent flatness of the forecasts in Figure 1.1 should therefore not be interpreted as a belief that prices will never change, rather, it reflects a rational position given the limitations of forecasting in a volatile market.

5.2.3 Forecast Biases Reversed After 2008

U.S. crude oil production appears relatively unbiased when analyzing the aggregated sample period (Figure 4.4a), but decomposing the data into two periods (August 2004 - July 2008 and August 2008 - November 2024) reveals offsetting tendencies around 2008 (Figure 4.5). Before July 2008, U.S. crude oil production forecasts were consistently

overestimated, reaching 10% at long horizons. After 2008, the bias reverses to persistent underestimation, reaching around -6%.

The choice of 2008 as a dividing line in the time series reflects several converging event that reshaped energy markets. In 2008, U.S. oil production started to grow rapidly [58] and in July the same year, the crude oil price was at an all-time high, which some attribute to an "unprecedented speculative situation" [59] of the crude oil markets at the time. At the same time, the U.S. shale boom was beginning to accelerate [58] and federal policy shifts under the Obama administration expanded subsidies for renewables. These developments justify using July 2008 as a reference point for analysis, even if the precise moment of structural change is not defined. The July 2008 cutoff is used here to show an asymmetry between pre-2008 and post-2008 data, which indicates that at least one structural break occurred, but the precise timing – and whether multiple breaks exist – remains uncertain.

Following the July 2008 oil price surge, the U.S. government proposed a bill, called "Stop Excessive Energy Speculation Act of 2008" [60], intended to restrict speculative trading in oil markets. Inevitably, the size of funds in crude oil markets dwindled and in the meantime, oil prices dropped [59]. This sequence marked a paradigm shift, or structural break, in global oil markets, and seems to have forced a recalibration of the STEO modeling framework affecting many variables. The STEO modeling structure seemingly adapts to structural breaks, and users of the STEO need to be careful in interpreting evaluations of these forecasts. What appears as unbiased in the aggregated data masks substantial historical change, illustrating that averaging data conceals offsetting errors across different breaks.

This asymmetry in the data extended beyond oil production to other fossil fuel and primary production forecasts. Focusing on $H = 12$ (Figure 4.6), before and after July 2008, shows that the recalibration improved the forecast bias for all of the studied fossil fuel prices except for natural gas Henry Hub spot price, which moved from a slight overestimation (2.5%) to a significant overestimation (16.5%). In primary production variables, only U.S. crude oil production displays less biased forecasts, from 8.7% before July 2008 to -3.1% afterwards. U.S. coal production peaked in 2008, and in the following two decades the production has greatly declined due to increased mining costs, more environmental regulations, and overtaking from other electric generation sources [54]. It seems that oil forecasts benefited from recalibration and analyst judgment following the 2008 price crash, whereas coal and gas forecasts remained highly tied to econometric projections that underestimated the depth of structural change, leading to worsening bias. In particular, the emergence of shale represented a structural shift that also affected coal and gas [58], given their substitutability in power generation. STEO price models, in particular natural gas prices, may require re-specification to reflect this new reality.

5.2.4 EIA Overestimates Solar Capacity

The pattern of overestimation in Figure 4.4b, particularly for solar PV capacity, indicates that the STEO modeling framework assumed a continued growth of solar PV deployment that did not materialize. Between 2009 and 2022, installed solar capacity in the U.S. expanded at an average annual rate of 44%, supported by favorable investment tax credits

[61] and cheaper solar PV technology [62]. Despite this rapid expansion, EIA's short-term forecasts for solar PV capacity growth were even more optimistic.

This finding stands in stark contrast to the EIA's long-term AEO and the IEA's World Energy Outlook (WEO), both of which have historically underestimated the rate of renewable adoption, especially for solar PV capacity [63]. While AEO and WEO projections have been overly conservative, the STEO's bias points in the opposite direction: in the short term, the model appears to have overestimated the pace of realized solar capacity growth. The historical rapid growth likely shaped the econometric STEO models, leading to expectations of continued exponential expansion. Cherp et al. [64] show that wind and solar power in the U.S. have already transitioned from an accelerating to a stable growth phase. Wilson et al. [65] similarly find that rapid early growth of new energy technologies gives way to slower expansion once technologies reach "materiality". Taken together, these findings suggest that the persistent overestimation of solar PV capacity in the STEO reflects a lag in recognizing the transition from the high-growth "take-off" phase to a more stable growth trajectory, similar to a "S-curve" [64]. While the long-term forecasting models of the AEO and IEA's WEO historically underestimated the potential of solar energy, the STEO has erred in the opposite direction. Its short-term econometric structure, heavily weighted on recent trends, projects momentum that is no longer justifiable as the market matures.

Unexpectedly, wind capacity does not exhibit a similar error pattern or bias as solar PV. Since both technologies rely on similar project announcement pipelines and construction timelines, one would expect similar biases. The presence of a persistent positive bias in solar forecasts while wind remains relatively unbiased suggests either a modeling deficiency specific to solar or inconsistencies in the underlying data. Alternatively, it is possible that wind power is still in its rapid growing phase, and that forecasts remain aligned for now, but may become overly optimistic once wind power approaches materiality.

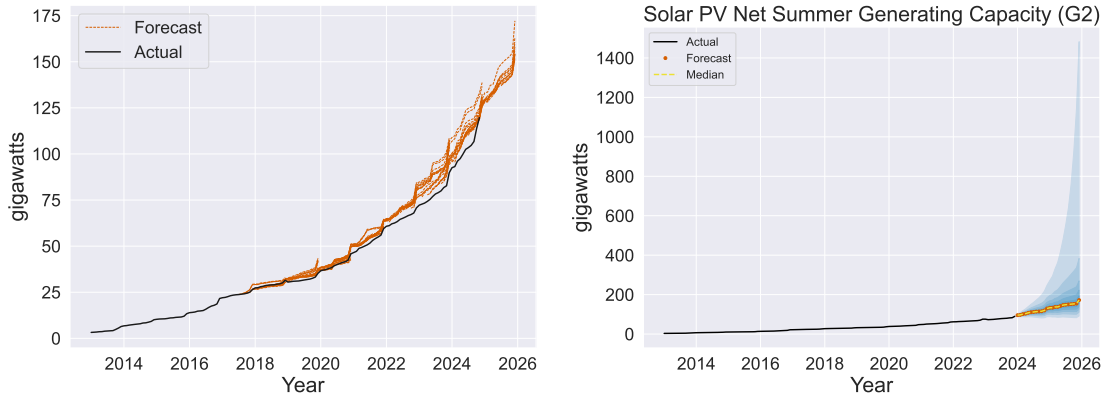
5.3 Uncertainty Representation and Prediction Intervals

5.3.1 Performance of Nonparametric vs Parametric EPIs

The constructed EPIs and resulting density forecasts highlight the trade-off between parametric and nonparametric methods in capturing uncertainty across different energy system variables. Kaack et al. [3] argue that empirical prediction intervals, especially those constructed from historical errors like G2, can provide more realistic and calibrated uncertainty quantification. We have shown that the nonparametric NP1 method generally provides the best performance. By directly sampling historical forecast errors, NP1 preserves empirical skewness and large deviations that Gaussian-based methods tend to underestimate due to imposed symmetry. This is evident in Figure 4.7 for WTI crude oil, where the nonparametric methods NP1 and NP2 capture the heavy-tailed distribution of oil price errors, producing asymmetric intervals critical for energy prices that often exhibit sudden spikes.

By contrast, the parametric G1 and G2 methods impose Gaussian assumptions on the error

distribution at each horizon, yielding symmetric intervals around the median (Figures 4.8 and 4.11). While these methods may be appropriate for infrastructure-related quantities where changes evolve incrementally – which is evident by G1 being the second best method in many of these quantities – it can be misleading for prices and fast-growing renewables. In particular, G2 performs poorly for solar PV capacity forecasts (Figure 5.3) because it builds its intervals on historical deviations that include periods of exponential growth. This produces inflated upper tails, exaggerating upside uncertainty even when historical growth rates are unlikely to persist [64, 65].



(a) **(b)**

Figure 5.3: **(a)** Actuals and forecasts for U.S. electric power sector solar PV net summer generating capacity in EIA's STEOs. Forecasts are denoted by red dashed lines, and outcomes in solid black. **(b)** Forecast densities for U.S. solar PV net summer generating capacity constructed using the parametric median-centered G2 method.

The wind capacity density forecasts in Figure 4.9 constructed using NP1 display a much more symmetric structure across horizons. The G2 method specifically generates smooth bell-shaped densities, which seem fitting for a quantity like installed wind capacity, where structural and policy trends dominate and volatility is low. The G1 method offers similar central coverage, but is somewhat more conservative in the outer bounds. Both methods show, again, that for infrastructure-related quantities, where the dynamics often are gradual, simpler parametric approaches may suffice, though nonparametric densities offer more flexibility should the historical error distribution shift.

The behavior of the CAISO wholesale electricity price density forecasts in Figure 4.10 reflects the volatility and asymmetric dynamics of electricity markets where short-term shocks such as demand spikes or generation outages spread unevenly. The G1 model produces narrower densities compared to NP1 in this case. While this improves interpretability, it risks underestimating the probability of extreme upward movements. NP1 reproduces the real-world probabilities of sharp price spikes due to demand surges or outages. However, for the ISO-NE electricity price, G2 achieves the second-best score (0.3338), only marginally worse than NP2 (0.3330), while NP1 slightly underperforms (0.3364) (Table 4.1). This suggests that for some electricity price series with more symmetric or well-behaved error structures, parametric methods like G2 can still perform

competitively, particularly if the Gaussian approximation is not strongly violated. Nevertheless, NP1 and NP2 tend to offer better robustness in the presence of skewness or heavy tails, which are frequent in energy price forecasts, and produce narrower PDFs compared to parametric Gaussian methods. Electricity prices stand out as the most difficult series to capture probabilistically, evident by the relatively high CRPS values.

5.3.2 Bias from Gaussian Assumptions in Forecasting Volatile Prices

The results suggest that the limitation of Gaussian EPIs in electricity price forecasting is not only the assumed symmetry around zero, but also the thin-tailed shape of the distribution. By doing so, they systematically underestimate the risk of extreme upward price spikes.

Electricity prices are characterized by heavy-tailed distributions, spikes, and asymmetries [48, 66], features that Gaussian assumptions can understate. Even if average volatility may appear symmetric, the empirical evidence suggests that upward deviations are more likely and larger than downward ones. Gaussian EPIs, with their imposed symmetry and thin tails, fail to capture this asymmetry and therefore bias forecasts toward underestimating the risk of price surges.

From a modeling perspective, it may be more intuitive, or institutionally beneficial, to neglect the unpredictable timing of extreme events and instead publish forecasts that smooth over volatility. However, this tendency produces a negative bias on average, as seen for wholesale electricity prices in Figure 4.4d. This effect may be particularly pronounced in high-frequency forecasts such as those from the STEO, where monthly resolution makes price spikes more visible than in annual reports like the AEO, where shocks can be averaged out across the year.

5.3.3 CRPS Evaluation

The CRPS offers an aggregate measure of density forecast accuracy across the forecast horizon, and the results show that NP1 consistently achieves the lowest CRPS values in most variables, while NP2 often ranks second (Table 4.1). This aligns with the notion that nonparametric methods yield better-calibrated distributions when historical error distributions have heavy tails or strong asymmetry, like for the WTI crude oil price (Figure 4.7) or the CAISO electricity price (Figure 4.10), conditions that characterize the STEO data more than the AEO scenarios examined by Kaack et al. [3]. In the AEO, where volatility was comparatively low and error distributions narrower, Kaack et al. [3] found that the G2 method performed best, since the Gaussian assumptions were less problematic. The STEO results suggest that when volatility is higher, symmetric Gaussian assumptions misrepresent risk, making nonparametric methods more flexible.

Yet, some nuances arise. At shorter horizons (H = 1-6 months), NP2 outperforms NP1 in several fossil fuel prices and primary production quantities. This demonstrates the advantage of median-centering in contexts where short-term errors are relatively small and symmetric. By reducing over-dispersion in the intervals, NP2 achieves tighter calibration without sacrificing coverage. Likewise, the G2 method emerged as the best

density forecast for Henry Hub spot prices (Figure 4.12) aggregated over short horizons, in agreement with Kaack et al. [3] finding that Gaussian intervals can perform well when volatility is modest in the time series. While parametric Gaussian distributions may fail to capture heavy tails and thus underestimate extreme event or structural breaks, they can, paradoxically, generate longer tails than nonparametric methods. This property explains why G2 sometimes produces excessively wide upper bounds, as in solar PV capacity forecasts (Figure 5.3), overstating upside risks when historical growth has been exponential.

The CRPS analysis in this study differs from that of Kaack et al. [3]. Their evaluation compared EPI methods against scenario-based projections from the AEO. The STEO provides only point forecasts, meaning the evaluation of probabilistic intervals here rests solely on the realized historical distribution of forecast errors. This limits comparability in two ways. First, it prevents testing the coverage of forecast intervals against any EIA structured scenario ranges. Second, it restricts the assessment of how well the EPI methods anticipate structural breaks, as these are not embedded in a scenario framework but only observed ex post.

5.3.4 EPIs as Bias Indicators

Kaack et al. [3] emphasize the role of forecast bias in shaping interval performance. In NP1, the error distribution is not median-centered, the empirical median of errors is instead preserved, so a non-zero median of past errors manifests as a second point forecast. This feature makes NP1 both a tool for detecting systematic forecast tendencies, and as a bias-corrected alternative to point forecast. When the NP1 median deviates from the STEO reference forecast, it shows that the past forecast errors have been directional.

For renewable capacity variables, especially solar PV, the NP1 median lies below the STEO forecasts, reflecting a pattern of systematic overestimation (Figure 4.8). This aligns with the bias results in Figure 4.4b, where solar capacity forecasts showed persistent positive bias. A similar but weaker pattern is seen for wind capacity, where the NP1 upper bounds are compressed and the median is slightly lower than the STEO forecasts. These patterns show that the STEO models incorporated rapid historical growth rates in renewable installation, which – once expansion stabilized in more recent years [64, 65] – translated into overly optimistic short-term expectations.

By contrast, for price quantities such as WTI crude oil (Figure 4.7) and CAISO electricity prices (Figure 4.10), the NP1 median lies above the STEO forecasts. This exemplifies a historical tendency towards under-prediction, particularly of price spikes. The divergence is most pronounced in CAISO, consistent with the strong negative bias found in Figure 4.4d. NP1, by preserving skewness in the error distribution, exposes this underestimation directly.

For natural gas production (Figure 4.11), the NP1 median also lies above the STEO forecasts, indicating a historically negative systematic bias. This is confirmed by the persistent underestimation of natural gas production displayed in Figure 4.4a.

Kaack et al. [3] caution that while NP1 median shifts can reveal systematic bias, they do not always improve forecast accuracy if the error distribution is non-stationary. For instance, the reversal in crude oil production bias before and after July 2008 (Figure 4.5) shows that systematic errors can change direction over time. Nevertheless, the NP1 method consistently exposes the directional tendencies of forecasts produced by the STEO models.

5.4 Implications for Energy Policy and Risk Assessment

The findings of this study suggest that users of STEO forecasts should be aware of systematic biases, anchoring tendencies, and the limitations of point forecasts in representing uncertainty. Improved uncertainty representation through empirical intervals could support more robust decision-making in policy and investment contexts. As highlighted in the foreword to the AEO 2023, "Thus, our objective must be to identify robust insights rather than precise numbers–think ranges and trends, not predictions and point estimates" [67]. Systematic biases – such as inertia in oil price forecasts, overestimation of solar capacity, and anchoring in electricity price forecasts – indicate that users should not assign the same level of confidence on all STEO point forecasts.

For energy policy, this implies that reliance on deterministic point forecasts risks underestimating both market volatility and the likelihood of structural change in energy markets. For risk assessment, the results show that neglecting the skewed and heavy-tailed nature of price errors, particularly in electricity price markets, can introduce systematic biases. Nonparametric forecasting approaches that preserve error distributions offer a more credible range of possible outcomes than parametric Gaussian methods. This can be used as a complement to traditional modeling and as a way to better align forecast users with the realities of uncertainty, improving resilience in investment, planning, and decision-making.

6

Conclusion

This thesis has systematically quantified forecast errors and uncertainty in the U.S. Energy Information Administration's (EIA) Short-Term Energy Outlook (STEO) of 17 energy quantities between 2004 and 2024 across variables in primary production, electricity generation, electricity capacity, and wholesale electricity prices.

The first finding is that forecast accuracy varies strongly across categories. Primary production quantities such as U.S. crude oil and natural gas production exhibit relatively low mean absolute percentage errors (MAPE), between 5-7% at the 12-month horizon ($H = 12$), with error growth remaining linear over horizons of up to two years. Physical quantities like U.S. crude oil production and wind capacity also display low errors that remain relatively constant across the whole horizon. At $H = 12$, the absolute errors are between 1-7%. In contrast, fossil fuel prices (gasoline, crude oil, natural gas) display substantially higher errors, with exponential mean absolute log errors (EMALE) exceeding 25-35% at the 12-month horizon. Wholesale electricity prices also have significantly larger errors than primary production and electricity generation and capacity quantities. At $H = 12$, the errors exceed 70%. This shows that forecasting prices is especially challenging, reflecting their vulnerability to short-term market shocks, rather than the slower dynamics that govern production, generation and capacity.

The second finding concerns systematic biases in the STEO. Forecasts of wholesale electricity prices reveal persistent anchoring to recent values, causing an under-reaction to price surges. Oil price forecasts display inertia by consistently showing less volatility than the underlying data. This reflects institutional incentives at the EIA to avoid projecting oil price swings. A re-estimation of the STEO models appears to have taken place for many variables after 2008. Altered forecast bias in oil production, shifting from consistent overestimation to underestimation, is evident. For capacity, the EIA systematically overestimates growth in solar photovoltaic generating capacity. This suggests that the STEO models are anchored in historical growth rates, causing them to lag in recognizing the more stable expansion trajectory of solar PV.

Third, the evaluation of empirical prediction intervals (EPIs) shows that nonparametric methods based on historical forecast errors (NP1, NP2) generally outperform parametric Gaussian approximations (G1, G2). Non-parametric methods produce sharper and more well-calibrated densities when historical error distributions are heavy-tailed or asymmetric compared to the Gaussian methods, which typifies the STEO data. Gaussian density

probability methods tend to underestimate the probability of extreme price spikes, biasing results downward by averaging out the asymmetric effects of rare upward shocks.

Taken together, these results indicate that STEO forecasts are not unbiased central estimates but carry distinct error structures. Users of the STEO should therefore treat forecasts for physical quantities in production, generation, and capacity as relatively reliable, but interpret price forecasts with greater caution. Importantly, the absence of uncertainty measures in the STEO may mislead users into assigning undue confidence to forecasts with different levels of reliability.

We demonstrate that EPIs provide a feasible way to supply forecasts with measures of uncertainty. By quantifying historical error distributions, these intervals capture both volatility and persistent bias, offering STEO users a more transparent assessment of risks. The forecasts should be interpreted as probabilistic and error-prone estimates rather than deterministic predictions. Making uncertainty explicit is crucial for more robust decision-making in an ever-changing energy landscape.

Bibliography

[1] International Energy Agency. *Global Energy Review 2025*. Report. Paris, Mar. 2025. URL: www.iea.org/reports/global-energy-review-2025 (visited on July 24, 2025).

[2] Stephen Haben, Marcus Voss, and William Holderbaum. "Time Series Forecasting: Core Concepts and Definitions". In: *Core Concepts and Methods in Load Forecasting: With Applications in Distribution Networks*. Cham: Springer International Publishing, 2023, pp. 55–66. DOI: 10.1007/978-3-031-27852-5_5.

[3] Lynn Kaack, Jay Apt, M. Morgan, and Patrick McSharry. "Empirical prediction intervals improve energy forecasting". In: *Proceedings of the National Academy of Sciences* 114 (July 2017), p. 201619938. DOI: 10.1073/pnas.1619938114.

[4] Robin John Hyndman and George Athanasopoulos. *Forecasting: Principles and Practice*. English. 2nd. Australia: OTexts, 2018.

[5] U.S. Energy Information Administration. *Short-Term Energy Forecasting Overview*. Handbook. Washington, DC, 2020. URL: https://www.eia.gov/analysis/handbook/pdf/STEO_Overview.pdf (visited on July 25, 2025).

[6] Dwight R. Sanders, Mark R. Manfredo, and Keith Boris. "Accuracy and efficiency in the U.S. Department of Energy's short-term supply forecasts". In: *Energy Economics* 30.3 (2008), pp. 1192–1207. DOI: 10.1016/j.eneco.2007.01.011.

[7] Dominic Davis and Michael J. Brear. "Impact of short-term wind forecast accuracy on the performance of decarbonising wholesale electricity markets". In: *Energy Economics* 130 (2024), p. 107304. DOI: 10.1016/j.eneco.2024.107304.

[8] Bidan Zhang, Guannan He, Yang Du, Haoran Wen, Xintao Huan, Bowen Xing, and Jingsi Huang. "Assessment of the economic impact of forecasting errors in Peer-to-Peer energy trading". In: *Applied Energy* 374 (2024), p. 123750. DOI: 10.1016/j.apenergy.2024.123750.

[9] Xiufeng Xing, Yingjia Cong, Yu Wang, and Xueqing Wang. "The Impact of COVID-19 and War in Ukraine on Energy Prices of Oil and Natural Gas". In: *Sustainability* 15.19 (2023). DOI: 10.3390/su151914208.

[10] Vincent Kaminski. *Energy Markets*. London: Risk Books, 2013.

[11] Ron Westrum. "The Social Construction of Technological Systems". In: *Social Studies of Science* 19.1 (1989), pp. 189–191.

[12] U.S. Energy Information Agency. *EIA statement on the factors leading to uncertainty in Short-Term Energy Outlooks*. 2022. URL: https://www.eia.gov/pressroom/releases/press497.php (visited on July 24, 2025).

[13] Jeremiah Shelor. "Uncertainty is Theme of Latest STEO as EIA Models Falling Crude Prices in 2H2022". In: *Natural Gas Intelligence* (Mar. 8, 2022). URL: https://www.naturalgasintel.com/news/uncertainty-is-theme-of-latest-steo-as-eia-models-falling-crude-prices-in-2h2022 (visited on July 30, 2025).

[14] OGJ editors. "EIA: Short-term energy outlook remains subject to high uncertainty". In: *Oil & Gas Journal* (Sept. 9, 2020). URL: https://www.ogj.com/general-interest/economics-markets/article/14183070/eia-short-term-energy-outlook-remains-subject-to-high-uncertainty (visited on July 31, 2025).

[15] Knut Are Aastveit, Jamie L. Cross, and Herman K. van Dijk. "Quantifying Time-Varying Forecast Uncertainty and Risk for the Real Price of Oil". In: *Journal of Business & Economic Statistics* 41.2 (2023), pp. 523–537. DOI: 10.1080/07350015.2022.2039159.

[16] James Hamilton. *Q&A on oil prices*. 2012. URL: https://econbrowser.com/archives/2012/08/qa_on_oil_price (visited on Aug. 5, 2025).

[17] Robert S. Pindyck. "The Long-Run Evolution of Energy Prices". In: *The Energy Journal* 20.2 (1999), pp. 1–27.

[18] U.S. Energy Information Administration. *STEO Archives*. 2025. URL: https://www.eia.gov/outlooks/steo/outlook.php (visited on Aug. 25, 2025).

[19] U.S. Energy Information Administration. *Mission and Overview*. URL: https://www.eia.gov/about/mission_overview.php (visited on July 25, 2025).

[20] U.S. Energy Information Administration. *Short-Term Energy Outlook STEO December 2024*. Report. Washington, DC, Dec. 2024.

[21] U.S. Congress. *S.826 — 95th Congress: Department of Energy Organization Act*. 1977. URL: https://www.congress.gov/bill/95th-congress/senate-bill/826 (visited on July 31, 2025).

[22] U.S. Energy Information Administration. *Annual Energy Outlook 2025*. Report. Washington, DC, Apr. 2025.

[23] U.S. Energy Information Administration. *The National Energy Modeling System: An Overview*. Report. Washington, DC, May 2023. URL: https://www.eia.gov/outlooks/aeo/nems/overview/pdf/0581(2023).pdf (visited on Aug. 26, 2025).

[24] Carolyn Fischer, Evan Herrnstadt, and Richard Morgenstern. "Understanding errors in EIA projections of energy demand". In: *Resource and Energy Economics* 31.3 (2009), pp. 198–209. DOI: 10.1016/j.reseneeco.2009.04.003.

[25] Hermann G. Matthies. "Quantifying Uncertainty: Modern Computational Representation of Probability and Applications". In: *Extreme Man-Made and Natural Hazards in Dynamics of Structures*. Ed. by Adnan Ibrahimbegovic and Ivica Kozar. Dordrecht: Springer Netherlands, 2007, pp. 105–135.

[26] Frank H. Knight. *Risk, Uncertainty and Profit*. Boston, MA: Houghton Mifflin Co, 1921.

[27] Millett Granger Morgan and Max Henrion. *Uncertainty: A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis*. Cambridge: Cambridge University Press, 1990.

[28] Anthony S. Tay and Kenneth F. Wallis. "Density forecasting: A survey". In: *A Companion to Economic Forecasting*. Wiley, Nov. 2007, pp. 45–68. DOI: 10.1002/9780470996430.ch3.

[29] J. Ehrlinger, W.O. Readinger, and B. Kim. "Decision-Making and Cognitive Biases". In: *Encyclopedia of Mental Health (Second Edition)*. Ed. by Howard S. Friedman. Second Edition. Oxford: Academic Press, 2016, pp. 5–12. DOI: 10.1016/B978-0-12-397045-9.00206-8.

[30] Amos Tversky and Daniel Kahneman. "Judgment under Uncertainty: Heuristics and Biases". In: *Science* 185.4157 (1974), pp. 1124–1131.

[31] Erik Nesvold and Reidar B. Bratvold. "Debiasing probabilistic oil production forecasts". In: *Energy* 258 (2022), p. 124744. DOI: 10.1016/j.energy.2022.124744.

[32] Sean Campbell and Steven Sharpe. "Anchoring Bias in Consensus Forecasts and Its Effect on Market Prices". In: *Journal of Financial and Quantitative Analysis* 44 (Apr. 2009), pp. 369–390. DOI: 10.1017/S0022109009090127.

[33] Jerome L Stein. “Speculative Price: Economic Welfare and the Idiot of Chance”. In: *The Review of Economics and Statistics* 63.2 (May 1981), pp. 223–232.

[34] Maximilian Auffhammer. “The rationality of EIA forecasts under symmetric and asymmetric loss”. In: *Resource and Energy Economics* 29.2 (2007), pp. 102–121. DOI: 10.1016/j.reseneeco.2006.05.001.

[35] Dmitry Kucharavy, David Damand, and Marc Barth. “Technological forecasting using mixed methods approach”. In: *International Journal of Production Research* 61.16 (2023), pp. 5411–5435. DOI: 10.1080/00207543.2022.2102447.

[36] Rob J. Hyndman and Anne B. Koehler. “Another look at measures of forecast accuracy”. In: *International Journal of Forecasting* 22.4 (2006), pp. 679–688. DOI: 10.1016/j.ijforecast.2006.03.001.

[37] Tilmann Gneiting and Adrian Raftery. “Strictly Proper Scoring Rules, Prediction, and Estimation”. In: *Journal of the American Statistical Association* 102 (Mar. 2007), pp. 359–378. DOI: 10.1198/016214506000001437.

[38] U.S. Energy Information Administration. *API Dashboard*. URL: https://www.eia.gov/opendata/browser/ (visited on Aug. 25, 2025).

[39] Sungil Kim and Heeyoung Kim. “A new metric of absolute percentage error for intermittent demand forecasts”. In: *International Journal of Forecasting* 32.3 (2016), pp. 669–679. DOI: 10.1016/j.ijforecast.2015.12.003.

[40] S.T. Rachev, M. Hoechstoetter, F.J. Fabozzi, and S.M. Focardi. *Probability and Statistics for Finance*. Frank J. Fabozzi Series. Hoboken, New Jersey: Wiley, 2010.

[41] Andrey Davydenko and Goodwin Paul. “Assessing Point Forecast Bias Across Multiple Time Series: Measures and Visual Tools”. In: *International Journal of Statistics and Probability* 10 (Dec. 2024), pp. 46–46. DOI: 10.5539/ijsp.v10n5p46.

[42] Yun Shin Lee and Stefan Scholtes. “Empirical prediction intervals revisited”. In: *International Journal of Forecasting* 30.2 (2014), pp. 217–234. DOI: 10.1016/j.ijforecast.2013.07.018.

[43] Hans Hersbach. “Decomposition of the Continuous Ranked Probability Score for Ensemble Prediction Systems”. In: *Weather and Forecasting* 15.5 (2000), pp. 559–570. DOI: 10.1175/1520-0434(2000)015<0559:DOTCRP>2.0.CO;2.

[44] Eric W. Weisstein. *Heaviside Step Function*. From Mathworld – A Wolfram Resource. URL: https://mathworld.wolfram.com/HeavisideStepFunction.html (visited on Aug. 7, 2025).

[45] U.S. Energy Information Administration. *Electricity explained: Factors affecting electricity prices*. 2023. URL: https://www.eia.gov/energyexplained/electricity/prices-and-factors-affecting-prices.php (visited on July 16, 2025).

[46] Paolo Gabrielli, Moritz Wüthrich, Steffen Blume, and Giovanni Sansavini. "Data-driven modeling for long-term electricity price forecasting". In: *Energy* 244 (2022), p. 123107. DOI: 10.1016/j.energy.2022.123107.

[47] U.S. Energy Information Administration. *Short-Term Energy Outlook Model Documentation: Electricity Supply module*. Handbook. Washington, DC, 2021. URL: https://www.eia.gov/analysis/handbook/pdf/STEO_Electricity_Supply.pdf (visited on Aug. 23, 2025).

[48] Rafał Weron. "Electricity price forecasting: A review of the state-of-the-art with a look into the future". In: *International Journal of Forecasting* 30.4 (2014), pp. 1030–1081. DOI: 10.1016/j.ijforecast.2014.08.008.

[49] U.S. Energy Information Administration. *What is the shoulder season in electricity markets?* 2024. URL: https://www.eia.gov/todayinenergy/detail.php?id=64044 (visited on Aug. 5, 2025).

[50] Richard Bowers. "U.S. wind energy production tax credit extended through 2021". In: *U.S. Energy Information Administration, Today in Energy* (Jan. 28, 2021). URL: https://www.eia.gov/todayinenergy/detail.php?id=46576 (visited on Sept. 17, 2025).

[51] U.S. Energy Information Administration. *Short-Term Energy Outlook STEO January 2019*. Report. Washington, DC, Jan. 2019.

[52] Ryan Wiser and Mark Bollinger. *2018 Wind Technologies Market Report*. Tech. rep. Lawrence Berkeley National Laboratory (LBNL), Berkeley, CA (United States), Aug. 2019. DOI: 10.2172/1559881.

[53] U.S. Energy Information Administration. *Assumptions to the Annual Energy Outlook 2025: Renewable Fuels Module*. Report. Washington, DC, Apr. 2025.

[54] Jonathan Church. "U.S. production of all types of coal has declined over the past two decades". In: *U.S. Energy Information Administration, Today in Energy* (Apr. 8, 2025). URL: https://www.eia.gov/todayinenergy/detail.php?id=64924 (visited on Aug. 24, 2025).

[55] Jennifer Castle, Michael Clements, and David Hendry. "An Overview of Forecasting Facing Breaks". In: *Journal of Business Cycle Research* 12 (Sept. 2016), pp. 3–23. DOI: 10.1007/s41549-016-0005-2.

[56] David B. Patton, Pallas LeeVanSchaick, Jie Chen, and Joseph Coscia. *2024 Assessment of the ISO New England Electricity Markets*. Report. Potomac Economics, June 2025. URL: https://www.iso-ne.com/static-assets/documents/100025/iso-ne-2024-emm-report-final.pdf (visited on Sept. 20, 2025).

[57] Chris Peterson, Mark Morey, and Victoria Zaretskaya. "New England natural gas and electricity prices increase on supply constraints, high demand". In: *U.S. Energy Information Administration, Today in Energy* (Feb. 3, 2022). URL: https://www.eia.gov/todayinenergy/detail.php?id=51158 (visited on Sept. 20, 2025).

[58] Robert Rapier. "How The Shale Boom Turned The World Upside Down". In: *Forbes* (Apr. 21, 2017). URL: https://www.forbes.com/sites/rrapier/2017/04/21/how-the-shale-boom-turned-the-world-upside-down/ (visited on Aug. 24, 2025).

[59] Yue-Jun Zhang. "Speculative trading and WTI crude oil futures price movement: An empirical analysis". In: *Applied Energy* 107 (2013), pp. 394–402. DOI: 10.1016/j.apenergy.2013.02.060.

[60] U.S. Congress. *S.3268 — 110th Congress: Stop Excessive Energy Speculation Act of 2008*. 2008. URL: https://www.congress.gov/bill/110th-congress/senate-bill/3268 (visited on Aug. 24, 2025).

[61] Katherine Antonio Jonathan DeVilbiss. "EIA expects U.S. annual solar electricity generation to surpass hydropower in 2024". In: *U.S. Energy Information Administration, Today in Energy* (Nov. 7, 2023). URL: https://www.eia.gov/todayinenergy/detail.php?id=60922 (visited on Aug. 24, 2025).

[62] Noah Kaufman. "Annual Energy Outlook Projections and the Future of Solar Photovoltaic Electricity". Working Paper No. 2014/4. 2014.

[63] Gabriel Lopez, Yousef Pourjamal, and Christian Breyer. "Paving the way towards a sustainable future or lagging behind? An ex-post analysis of the International Energy Agency's World Energy Outlook". In: *Renewable and Sustainable Energy Reviews* 212 (2025), p. 115371. DOI: 10.1016/j.rser.2025.115371.

[64] Aleh Cherp, Vadim Vinichenko, Jale Tosun, Joel A. Gordon, and Jessica Jewell. "National growth dynamics of wind and solar power compared to the growth required for global climate targets". In: *Nature Energy* 6.7 (July 2021), pp. 742–754. DOI: 10.1038/s41560-021-00863-0.

[65] C. Wilson, A. Grubler, N. Bauer, V. Krey, and K. Riahi. "Future capacity growth of energy technologies: are scenarios consistent with historical evidence?" In: *Climatic Change* 118.2 (May 2013), pp. 381–395. DOI: 10.1007/s10584-012-0618-y.

[66] Rafał Weron and Adam Misiorek. "Heavy tails and electricity prices: Do time series models with non-Gaussian noise forecast better than their Gaussian counterparts?" In: *Research Papers of Wroclaw University of Economics and Business* 1076 (Apr. 2007), pp. 472–480.

[67] U.S. Energy Information Agency. *Annual Energy Outlook 2023 – Narrative: Administrator's Foreword*. 2023. URL: https://www.eia.gov/outlooks/aeo/narrative/index.php#AdministratorsForeword (visited on Aug. 13, 2025).

DEPARTMENT OF SPACE, EARTH AND ENVIRONMENT
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden
www.chalmers.se

CHALMERS
UNIVERSITY OF TECHNOLOGY