

Financially inspired methodologies for risk management in electricity markets[☆]

Andres F. Ramirez^a , Alberto J. Lamadrid L.^b

^a Lehigh University, Bethlehem, PA, 18015, USA

^b Lehigh University, 451 Rauch Business Center, Bethlehem, PA, 18015, USA

ARTICLE INFO

JEL classification:

L94
O13
Q4
Q5
Q54

Keywords:

Risk management
Low carbon economy transition
Hedging

ABSTRACT

We present a framework to determine the risk of assets in electricity systems. We train machine learning and estimation models with optimal outputs from simulating Security Constrained Economic Dispatch (SCED) and Security Constraint Unit Commitment (SCUC). The operational score we present provides a quantitative measure of the likelihood that a generating unit will cause imbalances in the grid due to variability and uncertainty. We illustrate our methodology using real world data from the New York Independent System Operator (NYISO).

1. Introduction

The transition to a low carbon economy has accelerated changes in different energy systems, particularly in the electricity system. In the past decades, variable renewable energy sources (VRES) have increased their role in electricity systems, prompted by numerous reasons including the improvement of the economics of the considered technologies. Bloomberg's NEF's New Energy Outlook (NEO) discusses the growing uncertainty amid policy risks and geopolitical tensions that investors and businesses face as the energy transition unfolds. Still, clean energy technologies remain strong, with improving economics and maturity fueling global adoption. At the same time, surging power demand from AI growth presents both challenges and opportunities. [1]. In response, companies are pursuing 24/7 carbon free energy through long term power purchase agreements, on site generation, and innovative procurement strategies to match their electricity loads in real time [2]. Government initiatives such as California's Net Zero Carbon Pollution plan [3], and federal portfolio standards [4] further support integrating renewable energy into the electricity grid. Consequently, there has been a significant upsurge in investments in sustainable energy infrastructure. However, the increased use of VRES also increases the correlation in availability across sources due to events like wind lulls or solar cloud coverage. Therefore, addressing delivery risk for individual assets, and

system-wide portfolios requires new methods, alongside strategies that span the entire life cycle of these technologies, encompassing planning, construction, and operation.

In this work, we present a methodology to counteract two characteristics in existing structures as we rely more on VRES in electricity systems. First, there are frictions in the knowledge of asset operations caused by either imperfect or asymmetric information, which limits the ability to diversify. Second, there are missing mechanisms to manage the associated risks of these informational failures. To address this issue, we develop a methodology for stakeholders in the electricity system based on risk quantification approaches [5], recognizing the limitations on assessing the information uncertainty distributions that affect the agents participating in these systems [6].

This article is organized as follows. In the next section, we put our work in context and review a limited set of works related to our proposal. We also present a brief introduction to the methods used. Section 3 presents the data collected and simulated to estimate the scores. In Section 4 we illustrate the use of the scores based on a realistic model of New York state. We conclude by discussing the implications of adopting our proposed risk management strategy, and future directions for research in Section 5.

[☆] This research was supported by Advanced Research Projects Agency-Energy, ARPA-E of the US Department of Energy through the PERFORM grant #DE-AR0001277. The opinions expressed in this study are the sole responsibility of the authors.

* Corresponding author.

E-mail address: anr422@lehigh.edu (A.F. Ramirez).

2. Review and methodology

In what follows we present a summary of relevant literature and considerations for establishing an electric assets risk bureau.

2.1. Related literature

Previous work stemming from the academic point of view explores the management of uncertainties in the dispatch of electricity. For example, [7] develop theoretical foundations for efficiently coordinating wind power and price responsive demand under uncertainty, while [8] use the MATPOWER Optimal Scheduling Tool to test power system operation methods in uncertain conditions. There are still open questions for the practical establishment of contracts when uncertainty is modeled as part of the market in terms of consensus for the scenarios used and the explicit uncertainty model [9]. The price signals for electricity demand are largely muted, despite competition at the retail level in some geographical areas, due to the way retail markets were implemented. Enabling responsiveness of demand to the conditions at the wholesale level, including VRES availability, would increase market efficiency, and could potentially support VRES integration [10]. Electricity markets in restructured areas have implemented mechanisms to guarantee adequacy from the supply side. These capacity market mechanisms are in part a response to distortions such as caps on prices, potential issues with congestion, and the limited availability of economically viable energy storage. These three aspects make balancing supply and demand in these systems challenging [11]. However, these capacity markets have challenges, including the inability to penalize non-performance, as suppliers may be covered by bankruptcy protection and still leave customers exposed. The price caps imposed on wholesale markets are a mechanism to mitigate supplier market power, but also create incentives distortions for participants.

Our proposed methodology seeks to provide energy market participants with confidential information about their contribution to the risk of the energy grid, shared with the system operator. This information can also be used to compare the performance of individual assets to regional or national averages via the disclosure of information through risk scoring ratings. Our work is related to literature in informational instruments. For example, [12] explores the Hold up problem in the France banking system, showing how access to information through credit scoring ratings improves market efficiency for their case study. The author's main finding is that access to information promotes participation from market participants regardless of their size and characteristics, improves market efficiencies, and reduces market power for bigger institutions that can leverage asymmetrical access information. The electric assets risk bureau we propose provides information about the contributions to the reliability of the system from individual assets with the risk scoring ratings. Similarly to credit scores, the proposed risk scores would be disclosed privately. We propose that market participants and system operators know their respective scores, and have access to aggregated information about competitors. This can serve as a coordination mechanism for transactions between system operators (SOs) and individual market participants. However, the market participant's rating would continue to be private and can only be accessed by SOs or the Electric Assets Risk Bureaus.

[13] demonstrates that investors differ in perceiving uncertainty in stock markets along two key dimensions: the degree to which they perceive future stock movements as either predictable (epistemic) or random (aleatory). Moreover, these perceptions are linked to distinct actions: those seeing uncertainty as predictable seek information, versus those viewing it as random tend to diversify assets. This understanding holds parallels in the wholesale energy market, where participants navigating energy contracts face similar uncertainties. Just as investors seek information or diversify in response to uncertainty, energy market participants managing contracts could benefit from seeking insights or

employing diverse energy portfolios to address similar uncertainties in their domain.

In developing scores for electric assets, the underlying theory is linked to the characteristics of risk measures [14] and the mathematical properties that include (1) Monotonicity. The higher the expected outcome of an asset, the higher its risk measure; (2) Positive Homogeneity. Risk is proportional to the return of the asset and asset position; (3) Translational Invariance. Additive perturbations affect the risk linearly; and (4) Sub Additivity. The combined risk measure of several assets is less than or equal to the risk of the individual assets. However, these theoretical properties do not directly provide guidance on which measures can provide useful information to stakeholders. There is evidence that sometimes measures of risk best aligned with real decisions may not be coherent [15]. Hence, we need to consider implementation aspects in the estimation and trade-offs due to the available information and [16], as well as the codification of the practices and qualitative information control room operators and asset managers.

In the context of sustainable energy system planning and feasibility studies, [17] presents a literature review that explores the realm of risk modeling and uncertainty assessment within this context with particular emphasis on devising optimal energy technology portfolios. The authors review quantitative and semi-quantitative approaches used in modeling risks and uncertainties. Quantitative techniques, for example, measure risks using variance or probability density functions applied to technical and economic parameters. As demonstrated by [18], one can employ an information gap decision theory framework to formulate a risk-constrained, day-ahead scheduling model for gravity energy storage paired with wind turbines. Similarly, [19] explore uncertainty modeling schemes tailored to risk-averse planning and operation of standalone renewable energy microgrids.

Semi-quantitative approaches, such as scenario analysis and multi-criteria decision analysis (MCDA), provide the flexibility to integrate broader socioeconomic factors into risk assessments. For example, [20] develop a Monte Carlo-based scenario framework for renewable power generation, which allows the incorporation of market and policy assumptions alongside technical uncertainties. Likewise, [21] applies a fuzzy MCDA of type 2 interval to investigate the risks of renewable energy in Iran's oil producing context, explicitly accounting for macroeconomic trends and public acceptance in their evaluation. This review discusses various methods employed in modeling and incorporating risk factors in sustainable power generation planning, recognizing the complexity of the risk problem and the need for diverse approaches to capture statistical and non-statistical risk parameters.

For machine learning applications, [22] explores a review of methodological applications in the energy sector that have predominantly focused on regression approaches, particularly in predicting prices and power generation. The reviewed literature highlights that most studies employ regression techniques like support vector machines (SVM), artificial neural networks (ANN), and other similar methods to forecast time series data for crude oil and electricity prices. However, there is a notable gap in the use of classification methods for risk evaluation purposes, which are less frequently applied in evaluating the commitment of units and the flow of electricity following Kirchhoff's laws.

Although these regression models excel at predicting continuous variables, the lack of emphasis on classification approaches leaves an opportunity unexplored. Classification techniques can help identify deviations between actual and planned power generation. This is a critical area for ensuring grid stability and optimizing energy dispatch. This knowledge gap suggests that while ML has advanced the energy forecasting field, there is room for further innovation by incorporating classification techniques to address a broader range of challenges within energy systems.

2.2. Scientific novelty, knowledge gaps, and contributions

Our methodology combines optimization, simulation, and machine learning to generate transparent and interpretable asset-level risk scores for electricity market operations. Inspired by financial risk scoring, it adapts proven techniques to energy system operations. We train statistical models on outputs from Security-Constrained Unit Commitment (SCUC) and Security-Constrained Economic Dispatch (SCED) digital twins that include network congestion. This design provides operationally grounded and auditable risk measures, bridging theoretical modeling with real-world implementation and integrating seamlessly with existing operator practices.

We illustrate the framework with a case study of the New York Independent System Operator (NYISO). Two portfolios are analyzed: the 2019 grid with existing units, and the projected 2030 grid including retirements, changes in generation and transmission capacity, and weather-adjusted VRES inputs. This setup allows us to test the adaptability of the risk scores under changing energy and climate conditions. Risk quantification relies on solving SCUC and SCED problems, with structures validated through expert input.

In summary, the framework advances beyond prior work on market volatility, unit commitment, and policy mechanisms by delivering interpretable, asset-level risk scores directly linked to operator optimization outputs across different locations. The following discussion reviews existing contributions and highlights how this approach expands the literature.

Fogelberg and Lazarczyk study whether wind power volatility raises production failures in the Nordic market and report that short term costs associated with volatility are not significant; this is a market level assessment and does not yield generator-specific delivery risk measures [23], contrasting to our work that provides scores per participant. Cardona-Vasquez and coauthors analyze the Colombian market with a vector error correction model and find that higher penetration of non conventional renewables increases spot price volatility over a one year horizon and under expansion scenarios; the focus is on price volatility rather than calibrated asset level indicators [24]. Wu and coauthors propose a hydro unit commitment model that introduces operation constraint priorities to restore feasibility and show improved convergence relative to penalty methods; this is an optimization formulation that does not provide an interpretable operational risk signal for individual units [25]. Rey-Costa and coauthors examine the capacity oversizing and storage needed to supply one hundred percent renewable power in the Australian market and show that costs can be reduced and that spatiotemporal surplus energy can support new uses; this is a system adequacy and cost study rather than an asset level risk evaluation [26]. Wang and coauthors build an evolutionary game and system dynamics model for green electricity trading in China and show that portfolio standards and perceptions influence strategic behavior and premiums; this addresses market design incentives rather than unit level delivery risk [27].

Complementary strands in the journal emphasize risk and uncertainty from planning, project, and technology viewpoints. Bakhtiari, Zhong, and Alvarez compare stochastic and deterministic uncertainty models for stand alone microgrids and find that a modified Metropolis coupled Markov chain Monte Carlo approach can reduce energy and power shortage risk while improving performance; the contribution targets microgrid planning and operation and not generator hour delivery within wholesale dispatch [19]. Da Silva Pereira and colleagues use Monte Carlo methods to evaluate economic parameters for photovoltaic projects in Brazil for both grid connected and stand alone systems; the analysis is project oriented and does not translate dispatch outcomes into an operational score [20]. Lin, Chau and coauthors analyze Chinese micro and macro data and report that environmental taxes improve profitability in renewables while oil price volatility and geopolitical risk hinder investment; this addresses financial drivers of expansion rather than operational delivery risk [28]. Shahnazi and

Alimohammadlou identify and prioritize macro and specific risks for renewable energy in Iran with multiple criteria methods based on interval type two fuzzy sets and find that licensing procedures and non renewable energy prices are the top priorities; this is an institutional and policy risk perspective rather than a dispatch linked performance measure [21]. Wu, Li and coauthors develop a day ahead scheduling model for gravity energy storage coordinated with wind using information gap decision theory and show that risk averse and risk seeking objectives can be satisfied with robustness and potential windfall profits; the focus is a technology specific hybrid scheduling model [18].

Across these studies there is no report of a method that converts SCUC or SCED operational outputs into a calibrated and interpretable generator hour score that can be deployed by system operators and market participants. In contrast, our contribution extracts operational labels from dispatch and commitment simulations, applies Weight of Evidence binning with elastic net regression to obtain transparent and auditable scorecards, and evaluates transferability from present day portfolios in 2019 to projected resource mixes in 2030, including congestion and pricing at the nodal level. To our knowledge, no prior work in the literature has combined dispatch optimization outputs with statistically interpretable machine learning to produce calibrated generator-hour risk scores with spatial variation that are transparent, transferable, and suitable for operational deployment. Our framework bridges optimization, simulation, and statistical learning, quantifies delivery risk at the asset level in a manner aligned with operator practice, and supports governance through monitoring, recalibration, and communication.

A further contribution of this paper is to lower information frictions by providing a privacy preserving index that summarizes proprietary operational signals into an asset level risk score without requiring public disclosure of the underlying data. Private market participants and utilities face internal delivery risk; if they do not meet unit commitment, or if actual production deviates from scheduled production, they incur penalties and must manage the associated risk. The score provides a consistent benchmark for self assessment and peer comparison, which supports internal governance and contract design. From the perspective of the system operator, the objective is to maintain grid stability and reliability; system operators are neutral and do not favor individual participants. Rather, they seek to sustain secure operations. Risk scores can inform policy rules, monitoring thresholds, and corrective actions that align individual incentives with system stability while respecting data confidentiality.

2.3. Digital twins for electricity systems

In many jurisdictions, electricity systems managers solve sequences of optimization problems for short term operations. The objective functions of system operators seek to minimize the cost of satisfying a given demand, considering the resources available. These optimization routines approximate in a deterministic setting the numerous uncertainties in the system, including potential failures from elements in the system (sometimes referred to as *contingencies*). Over time, and in response to experiences like blackouts, new organizations have emerged to provide guidelines and in some instances mandates for reliability [see e.g., 29].

Security Constrained Economic Dispatch (SCED) is an example of these optimizations that includes security criteria such as financial contracts to allow for handling realizations of uncertainty. A Security Constrained Unit Commitment (SCUC) includes complexities like the decision to use generating units that have to start up, run for several periods, and potentially shut down, and the associated costs to these starts and shutdowns [see e.g., 30]. The flows in the system can be represented by alternating current (AC) models, creating non linearities, or with different linearizations. We use a Direct Current (DC) linear model of the network flows for computational tractability.

In Eq. (1) we present a generalized formulation of the SCUC/SCED optimization structure:

$$\begin{aligned} & \min_{x,y} F(x) + G(y(x)) \\ & \text{subject to } h_x(x) + h_y(x, y) \leq b \\ & x \in X, \quad y \in Y, \end{aligned} \quad (1)$$

Here, $x \in X \subset \mathbb{R}^{n_1}$ denotes a vector of discrete decision variables, such as unit commitment status and minimum up and down times. The set $y(x) \in Y \subset \mathbb{R}^{n_2}$ denotes a vector of continuous decision variables, such as power dispatch, reserves, and ramping. The objective function is the total welfare of all participants, consisting of three components:

1. Costs associated to discrete variables, $F(x)$, e.g., unit commitment costs u^{ti} for period t , unit i .
2. Costs of continuous variables related to physical deliveries, $G(y(x))$, e.g., energy dispatches p^{ti} .
3. Costs of continuous variables related to contingent services, also represented by $G(y(x))$, e.g., reserves and other ancillary services r^{ti} .

The constraints for the problem can be categorized into four groups:

1. Constraints on discrete variables, $h_x(x)$, e.g., unit commitment events associated with starting or shutting down units, technical constraints that limit the number of periods units need to run, and per period output limits:

$$u^{ti} P_{\min}^{ti} \leq p^{ti} \leq u^{ti} P_{\max}^{ti} \quad (2)$$

2. Optimal power flow constraints, $h_y(x, y)$, e.g., balancing supply and demand, and representing the flow of electricity in the network.
3. Constraints over several time periods, $h_y(x, y)$, e.g., limiting changes in output according to technical ramping capability:

$$p^{ti} - p^{(t-1)i} \leq \delta_+^{(t-1)i} \quad (3)$$

4. Security constraints, guaranteeing the system can operate under deviations, e.g., procurement of reserves:

$$p^{ti} + r^{ti} \leq u^{ti} P_{\max}^{ti} \quad (4)$$

In restructured systems with wholesale markets in place, the objective is to align the goals of a social planner, seeking to minimize the system wide cost, with the profit maximization objectives of market participants [9]. In some markets, there is a valuation of the willingness of consumers to pay for electric reliability, e.g., the Value of Lost Load (VOLL) [31], which supports a total welfare maximization approach.

Participants are primarily involved in managing energy contracts and commitments, directly impacting the SCUC and SCED optimization problems solved by the ISO.

We implement our model using an extension of the MATPOWER Optimal Scheduling Tool [8] and an implementation in python of the simulation for scalability purposes.

2.4. Classification methodologies

We train machine learning and statistical models using the optimal outputs obtained by solving SCED and SCUC optimization problems. For each hour of every day simulated, the outputs for the SCUC optimal dispatch and randomly selected SCEDs out of 1000 randomly generated scenarios are used in the estimation models.

For dimensionality reduction, we use Weight of Evidence (WoE) as a technique to discretize variables. This approach is supported by [32] when defining attributes for scorecard generation in credit scoring. We used the criteria proposed by [33] to ensure a correct binning of the

attributes used in the model. The WoE is defined as the $\ln(\text{Information odds})$, as shown in Eq. (5).

$$\text{WoE} = \ln \frac{\mathbb{P}(\text{Positive})}{\mathbb{P}(\text{Negative})}, \quad (5)$$

where $\mathbb{P}$ denotes the probability operator. WoE accommodates missing values and allows us to maintain a comprehensive dataset to minimize potential biases in our analysis. A divergence measure quantifies how well a binning solution separates [34]. Our next objective is to use a method to optimally maximize a divergence measure e.g., Jeffrey's divergence, or the information value (IV) [35]. The IV is given by the WoE multiplied by the difference in proportion between positively and negatively identified outcomes, as shown in Eq. (6).

$$J(P \parallel N) = IV = \text{WoE} \times (\mathbb{P}(\text{Positive}) - \mathbb{P}(\text{Negative})). \quad (6)$$

We combine WoE, with Information Values (IV) to assess the impact of each variable and identify the main drivers of risk. The scoring developments for behavior and credit applications use logistic regressions among other methods, due to their simplicity and interpretability. This approach aligns with the findings of [36], who offers a comprehensive review of classification techniques used in consumer credit scoring. The score points are calculated on the basis of the coefficients of the fitted model and the transformation of the data from the binning method. The score of an instance x with n features is given by.

$$\text{score}(x) = \sum_{i=1}^n \sum_{j=1}^{b_i} c_i t_{ij} \mathbf{1}_{\{x_i \in S_{ij}\}}, \quad (7)$$

where b_i is the number bins for a feature i and S_{ij} is the set mapping the bin j for feature i .

Alternative estimation methods have been proposed, with associated tradeoffs in terms of computational complexity and sometimes small gains in classification accuracy [37].

2.5. Machine learning estimation methodologies

We applied several methods, both parametric and non parametric to perform the classification. The models include logistic regression with regularization techniques [38], boosting and ensemble methods [39], and neural networks [40]. For the parametric methods, the logistic regression model predicts the probability of the dependent variable being in a particular category based on the values of the predictor variables. For the non parametric methods, XGBoost is a gradient boosting algorithm known for its high accuracy and scalability. XGBoost has been successfully applied in various domains, including finance and risk assessment. Neural networks consist of multiple layers of interconnected nodes that learn to extract features from the input data. Deep learning, a subset of neural networks with multiple layers of nodes to learn hierarchical representations of the input data, has shown impressive results in various applications, including image recognition and natural language processing [41]. Deep learning has been applied in various domains in the energy sector, including energy forecasting and solving SCEDs [42].

The logistic regression model we implemented with elastic net regularization is shown in Eq. (8).

$$p(y = 1|x) = \frac{1}{1 + \exp(-x^T \beta)}, \quad (8)$$

where $p(y = 1|x)$ is the probability of the dependent variable y being in a select category, conditional on the predictor variables x , and β are the coefficients of the logistic regression model. We used the discrete bins derived from WoE as inputs to the logistic regression.

The elastic net regularization combines both L1 and L2 penalties, defined as follows:

$$\lambda_1 \|\beta\|_1 + \lambda_2 \|\beta\|_2^2, \quad (9)$$

where λ_1 and λ_2 are regularization parameters that control the strength of the L1 (lasso) and L2 (ridge) penalties, respectively. Moreover,

$p(y = 1|x)$ is the probability of the dependent variable y being in category 1 given the predictor variables x , and β are the coefficients of the logistic regression model. This regularized logistic regression model helps prevent overfitting and manages multicollinearity in the predictors.

The boosting algorithm we implemented is XGBoost, shown in Eq. (10).

$$\hat{y}_i = \sum_{k=1}^K f_k(x_i), \quad (10)$$

where $\hat{y}_i$ is the predicted value for the i th instance, K is the number of trees in the ensemble, f_k is the k th decision tree, and x_i is the feature vector for the i th instance.

The neural network model we implemented is shown in Eq. (11).

$$\hat{y}_i = f(w_2^T g(w_1^T x_i + b_1) + b_2), \quad (11)$$

where $\hat{y}_i$ is the predicted value for the i th instance, x_i is the feature vector for the i th instance, w_1 and w_2 are the weight matrices for the first and second layers, b_1 and b_2 are the bias vectors for the first and second layers, g is the activation function for the first layer, and f is the activation function for the second layer. The weights and biases of the neural network are learned during the training process using backpropagation.

We measure the performance of the machine learning algorithms using commonly accepted statistical metrics for classification including sensitivity, specificity, Kolmogorov Smirnov (KS) statistics, and Brier Scores. The accuracy of the predictions is measured as the probability that $P(\hat{Y} = Y)$. This is estimated as:

$$\text{acc} = \frac{\text{True Positives} + \text{True Negatives}}{\text{Positive samples} + \text{Negative samples}} \quad (12)$$

To facilitate the adoption of our models, we develop intuitive information interfaces from the estimation and training stages for users and stakeholders. These include receiver operating (ROC) and precision recall curves with diagnostics of the model fitting.

2.6. Selection and considerations

Our selection of the estimation model follows three considerations. First, the accuracy of the models. Second, the interpretability of the model. And third, the computational complexity of the model. In scoring assets for a risk bureau, explainability is very important, as stakeholders need to understand the factors influencing risk to make informed decisions.

We utilized various evaluation metrics and techniques, and assessed the performance on test and validation sets. These results are summarized in Table 3. The indicators use show how well the models performed on the training and unseen data, ensuring they generalized effectively without becoming overly complex or overfitting the data. We adopt the Logistic regression with L1 and L2 regularization (Elastic Net) as our final model for its interpretability, efficiency, and high accuracy on testing and validation sets. A key component was the use of WoE to discretize attributes, reducing noise and enhancing pattern recognition in the Elastic Net model. These binned variables preserve interpretability while highlighting key risk factors, supporting the methodology's practical relevance. We strike a balance between predictability and explainability. We can explain the model's predictions to stakeholders, regulators, and decision makers with the transformed variables and coefficients in the elastic net model, as shown in Eq. (9) while being able to maintain a high accuracy on the testing and validations sets. The WoE also mitigates the risk of overfitting the model, nonetheless we also used cross validation techniques and assessed the performance of the model in training and validations sets to ensure that we were not overfitting. WoE helped us have a balance between model complexity and predictive power, ensuring that our scoring model generalizes well to new data and remains robust.

When measuring the balance between model complexity and predictive power, we conducted thorough analyses of the fit and accuracy of the models.

XGBoost is particularly accurate for complex and high dimensional datasets due to its ability to handle non linear relationships and its ensemble nature, which enhances predictive accuracy. However, XGBoost also presents some challenges. One of the primary concerns is its increased complexity and the difficulty in interpreting its results. As a black box model, the relationship between input variables and the final prediction is not as transparent or easily explainable as in logistic regression. Although mechanisms like SHAP values exist to improve explainability, the nature of XGBoost can hinder our ability to provide clear insights into the model's decision making process, making it less suitable for regulatory compliance and model validation. Additionally, tuning XGBoost hyperparameters is a challenging task that requires extensive experimentation and significant computational resources. These hyperparameters include the learning rate, maximum tree depth, and subsample ratios, among others. In contrast, the simplicity and well defined coefficients of logistic regression make it easier to fine tune and interpret, saving both time and effort.

Deep learning models effectively capture intricate risk patterns thanks to their multi layer architecture. This allows them to model complex non linear relationships and achieve high predictive accuracy. The computational resources and time required for training deep learning models present practical challenges, especially in scenarios where efficiency and rapid model deployment are crucial.

3. Data and implementation

3.1. Data sources

We use datasets sourced from the Advanced Research Projects Agency Energy (ARPA-E) Performance based Energy Resource Feedback, Optimization, and Risk Management PERFORM project [43]. The reliability and validity of the data underwent assessment by organizations funded by ARPA-E PERFORM program (e.g., National Renewable Energy Laboratory (NREL), Princeton University's PGscen, among others).

The first input is a set of deterministic and probabilistic (quantile) forecasts [44] with time coincident load, wind, and solar generation profiles, encompassing both actual and forecast data. Solar and wind actuals are generated using data from the National Solar Radiation Database (NSRDB), the European Centre for Medium Range Weather Forecasts (ECMWF), and the Weather Research and Forecasting Model (WRF), respectively [45,46]. This data features 5-minute resolution data at the site level, with high temporal and spatial resolution for improved accuracy in risk scoring model development.

We use the PGscen package [47] to create a set of GEMINI trained scenarios on historical grid asset actual and forecasted values input streams and datasets for electricity demand. The networks used to run the SCUCs and SCEDs are reductions of the system corresponding to the NYISO system for the year 2019 [48] and the system corresponding to 2030 [49]. The actual demand is collected from each load zone in NYISO, and in cases where only hourly load data is available, a load downscaling method is employed to generate synthetic 5-minute load data. The temporal aspects of the forecast data include day ahead, intra day, and intra hour forecasts with varying lead times, horizons, and update rates.

We consulted and presented our results to subject matter experts at the NYISO. By involving experts in the development and validation of the optimization framework, our methodology benefits from their understanding of the complexities and nuances inherent in the NYISO grid and renewable energy sources. This expert driven approach ensures that the optimization model accurately captures the interactions between different components of the power system. It also includes adjustments made out of the market module routines.

This study emphasizes transparency in the analysis and interpretation of results. We also discuss and examine the potential implications and consequences of the risk scoring model's application to mitigate any adverse effects on energy producers and consumers in Section 4.

3.2. Data usage in optimization and machine learning

The data introduced in Section 3.1 is used across two main components of the methodology: the power system optimization module and the machine learning based risk scoring model.

In the optimization stage, we used 1000 distinct scenarios representing variations in the loads, wind, and solar generation forecasts. These scenarios serve as input to the Security Constrained Economic Dispatch (SCED) and Security Constrained Unit Commitment (SCUC) models, which simulate grid operations under uncertainty. The resulting outputs include high resolution hourly operational data such as generator commitment decisions, dispatch schedules, ramping activity, marginal production costs, and reserve allocations.

These optimization outputs provide the foundation for building the dataset used in the machine learning model training. For each scenario, generator, and hour of the day, we derive a comprehensive set of features that capture the system state, forecast conditions, and operational responses. The feature set includes temporal descriptors such as calendar date, day of the week, and holiday status; generator specific characteristics including ramp rates, generation limits, setpoints, and fuel type; market related variables such as production costs, nodal prices, and locational marginal prices; and scenario driven metrics including forecast errors and dispatch discrepancies.

To enhance predictive performance while preserving interpretability, features undergo polynomial and lagged transformations, as detailed in Section 3.3. The machine learning model is trained to estimate generator and hour level risk scores for each scenario, enabling the prediction of system risk across a range of unseen forecast conditions.

In this work, model training was performed using 2019 data, which reflects the network topology, generation and demand levels, and renewable penetration of that year. Portability was then assessed using the projected 2030 system, which incorporates planned network changes, retirements, new capacity additions, and weather-adjusted renewable inputs.

While some system characteristics such as physical dispatch constraints, generator attributes, and reserve requirements remain comparable between 2019 and 2030, the passage of time inevitably introduces new uncertainties. These include advances in technology, evolving management and planning practices, and improvements or maintenance in network infrastructure, which cannot be fully captured in the present study. Accordingly, the exercise should be interpreted as providing both an actionable pathway for immediate integration of risk scores into operations and a directional outlook for medium-term portfolio changes, including the accelerated introduction of variable renewable energy sources.

The model accuracy is supported by SCUC and SCED outputs and feature distributions. When larger shifts occur, as anticipated between 2019 and 2030, retraining or recalibration will be required to maintain validity. This ensures that the proposed scorecards remain both interpretable and adaptable as electricity systems evolve.

3.3. Implementation

Fig. 1 presents the complete ABSCoRES methodology pipeline, from input scenario data and system optimization to machine learning driven risk scores and stakeholder interface.

Applicability Across System Operators: Although the methodology was developed and validated using data from the New York Independent System Operator (NYISO), its design is generalizable to other electricity systems and Independent System Operators (ISOs). The core sets of features modeled, such as dispatch decisions, ramping

behavior, and renewable forecast errors, are sets of features commonly tracked in most ISOs. However, customization may be necessary to capture operator-specific market rules, regional generation portfolios, and reserve policies. The pipeline consists of the following main components:

- 1. Renewable energy and load forecasting:** We use detailed datasets containing 1000 scenarios each for load, solar PV, and wind generation forecasts, developed under the ARPA-E PERFORM program as presented in 3.1. These scenarios provide 1-hour resolution data for every generator, each day in the simulation horizon. A random subset of 200 scenarios is selected to serve as input for the optimization process. The selected scenarios include time coincident forecasts across all sources of uncertainty, demand, solar, and wind, capturing the variability and correlation of renewable energy production.
- 2. Optimization and Feature Generation:** The selected scenarios are input into the SCUC and SCED optimization models. The outputs from these optimization runs include generator schedules, marginal costs, ramping activity, reserve allocations, and unit commitment decisions. These optimization results form the basis of the machine learning training dataset. From these outputs, we engineer features to represent the system state and response. These include temporal indicators (day, month, weekday, holiday), generator level attributes (e.g., ramp rates, capacity limits, setpoints), market signals (e.g., cost parameters, nodal prices, LMPs), and scenario driven metrics (e.g., forecast errors and dispatch deltas). To enhance model performance while preserving interpretability, we apply polynomial and lagged transformations. All continuous variables are discretized using the Weight of Evidence (WoE) methodology, as detailed in Section 2.4.
- 3. Machine Learning Framework for Risk Scoring** The core of the ABSCoRES machine learning framework is the transformation of SCUC and SCED outputs into a structured dataset suitable for supervised learning. Each optimization run for each scenario yields hourly generator-level decisions on commitment, dispatch, reserves, and ramping, which are enriched with temporal indicators (day, month, weekday), spatial attributes, and generator characteristics. To reduce redundancy, multicollinearity across features is screened out, and continuous variables are discretized using the Weight of Evidence (WoE) method [50]. The response feature is derived from deviations in generator unit commitment. A generator is labeled as belonging to the positive (risky) class if its deviation exceeds a specified quantile-based threshold; deviations below this level are labeled as the negative (less risky) class. The quantile approach ensures robustness by accounting for distributional properties of deviations across generators, allowing the classification to adapt to heterogeneous operating conditions while preserving comparability of risk levels. As we will explore in the following steps, the resulting dataset is then partitioned into 70/30 train-test sets, with models benchmarked using one fold cross-validation. Hyperparameters are tuned via grid search, and performance is evaluated using AUC, accuracy, F1 score, and RMSE, supplemented by goodness-of-fit checks (AIC, Kolmogorov-Smirnov) and bootstrapping for stability. Hyperparameters for each model were tuned via grid search. Table 1 provides the main tuning ranges:
- 4. Model Evaluation and Accuracy Validation:** Model selection was based on standard performance metrics (i.e., AUC, Accuracy, F1 Score, and RMSE) evaluated on both training and test sets. These metrics were averaged across cross validation folds, and bootstrapping was used to assess stability. Full results are summarized in Table 3.

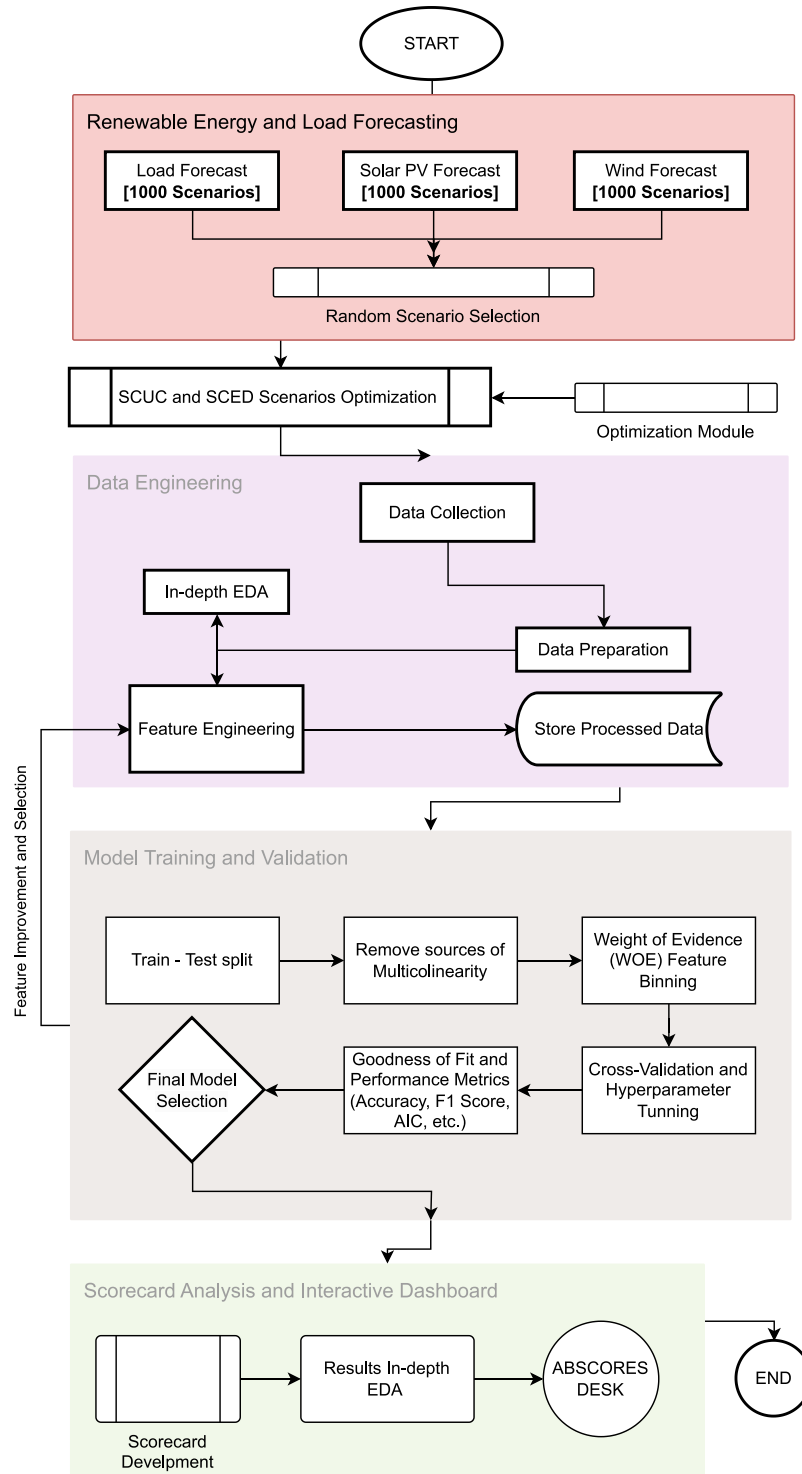


Fig. 1. ABSCoRES methodology pipeline.

The Elastic Net model was ultimately selected; a comprehensive analysis assessing this selection is performed in Section 4.1. While XGBoost and Neural Networks demonstrated marginal gains in predictive accuracy, they lacked transparency, a critical factor for risk scoring in regulated electricity markets. Elastic Net offered a robust compromise between interpretability, computational efficiency, and performance.

5. **Scorecard Deployment:** Model outputs were mapped into a

standardized risk score using calibrated bins aligned with U.S. credit scoring practices [50]. The final scorecard reflects probability bands of imbalance risk and supports comparison across assets, time periods, and forecast scenarios.

The scoring formula applied to each asset is given in Eq. (13). It incorporates the coefficient weighted contribution of each feature (via its Weight of Evidence value), combined with scaling parameters to convert the probability into a score consistent with

Table 1
Representative Hyperparameter grid for model training.

Model	Parameter	Grid range
Elastic Net	α	{0.001, 0.01, 0.1, 1, 10}
	l1_ratio	{0.1, 0.5, 0.9}
Random Forest	n_estimators	{100, 200, 500}
	max_depth	{10, 20}
	min_samples_split	{2, 5, 10}
XGBoost	n_estimators	{100, 200, 500}
	max_depth	{3, 6, 10}
	learning_rate	{0.01, 0.1, 0.3}
Neural Network	hidden_layer_sizes	{{50,}, {100,}, {100, 50}}
	activation	{relu, tanh}
	alpha	{0.0001, 0.001, 0.01}
Logistic/Lasso/Ridge	penalty	{l1, l2}
	C	{0.01, 0.1, 1, 10}

industry standards.

$$\text{Score}_i = \sum_{j=1}^n \left[\left(\beta^j \times \text{WoE}_i + \frac{\alpha}{n} \right) \cdot \text{Factor} + \frac{\text{Offset}}{n} \right] \quad (13)$$

The parameters and definitions used in Eq. (13) are listed in Table 2.

Scorecards in consumer credit translate multi-factor evidence into calibrated default odds with strict transparency. Electricity operations share three properties that make this paradigm attractive: (1) rare but material events such as generator delivery shortfalls; (2) information asymmetries between market participants and system operators; and (3) accountability requirements that favor auditable and interpretable models. These parallels justify adapting a scorecard framework to energy markets, where participants face penalties for deviations and operators must ensure fairness and reliability.

At the same time, electricity systems introduce domain-specific challenges. Dispatch and commitment decisions are constrained by physics and network topology, label prevalence and risk drivers shift with seasonal demand and renewable variability, and operator interventions such as reserves or re-dispatch feed back into realized outcomes. Our design addresses these differences through: (a) SCUC/SCED-derived features that embed physical and operational constraints; (b) Weight-of-Evidence (WoE) bins that enforce monotone and interpretable feature effects; (c) time-aware validation to capture non-stationarity across seasons and planning horizons; and (d) a governance plan with regular recalibration, drift monitoring, and champion-challenger testing.

By combining these elements, the proposed framework preserves the interpretability and comparability of financial scorecards while adapting them to the operational realities of electricity markets. This provides a standardized, transparent, and auditable risk metric that can be used both by private participants to benchmark performance and by system operators to inform policy and maintain system stability.

- User Interface:** The ABCoRES Desk provides stakeholders, operators, aggregators, and planners with real time access to asset risk scores. This enables coordination beyond price signals, helping to anticipate and mitigate delivery risk through informed operational decisions.

4. Results and analysis

We first present information about the fit of the models and then illustrate the characteristics of the operational score and results for the New York system.

Table 2
Scoring methodology parameters.

Parameter	Description
β^j	Coefficient for variable x^j
α	Baseline probability
WoE_i	Weight of Evidence value for variable x^i
n	Number of independent variables x^i in the model
PDO	Points to double the odds (e.g., 20)
Factor	Scaling parameter: $\frac{\text{PDO}}{\ln(2)}$
Offset	Scaling parameter: Target Score – (Factor $\times \ln(\text{Target Odds})$)
Target Score	Maximum score (e.g., 850)
Target Odds	Odds at the maximum score (e.g., 50)

4.1. Risk scoring results

Table 3 summarizes the fit metrics for a subset of the models developed.

We find that the benefits gained from XGBoost's higher accuracy are insufficient to outweigh the advantages of logistic regression's transparency and ease of interpretation for our asset scoring task, considering the tradeoff between predictive accuracy and interpretability.

The main challenge we encountered with deep learning was its inherent complexity and the “blackbox” nature of the model, similar to XGBoost. As the neural network learns representations at different layers, understanding how each input feature contributes to the final prediction becomes increasingly challenging. This lack of interpretability raises concerns, especially for energy dispatch decisions and contracts, where transparency and explainability are paramount for building trust in the model's decisions. Moreover, training deep learning models can be computationally intensive, especially for large datasets or architectures with many layers and nodes.

Therefore, we prioritize the elastic net methodology, as it offers a good balance between predictive performance, ease of interpretation, and the ability to provide understandable and actionable insights, aligning well with the requirements of managing risk in electricity systems.

The choice of the elastic net in conjunction with the WoE binning strategy allows us to strike the optimal balance between predictive accuracy and interpretability. This balance is essential in scoring assets. It ensures that the model performs well in predicting delivery risk, and provides transparent and actionable insights for informed decision making. As a result, our research outcomes are reliable, meaningful, and applicable in real world assessment scenarios of the delivery risk of electric assets. Both boosting and deep learning hold great promise in various applications and will be the subject of future research.

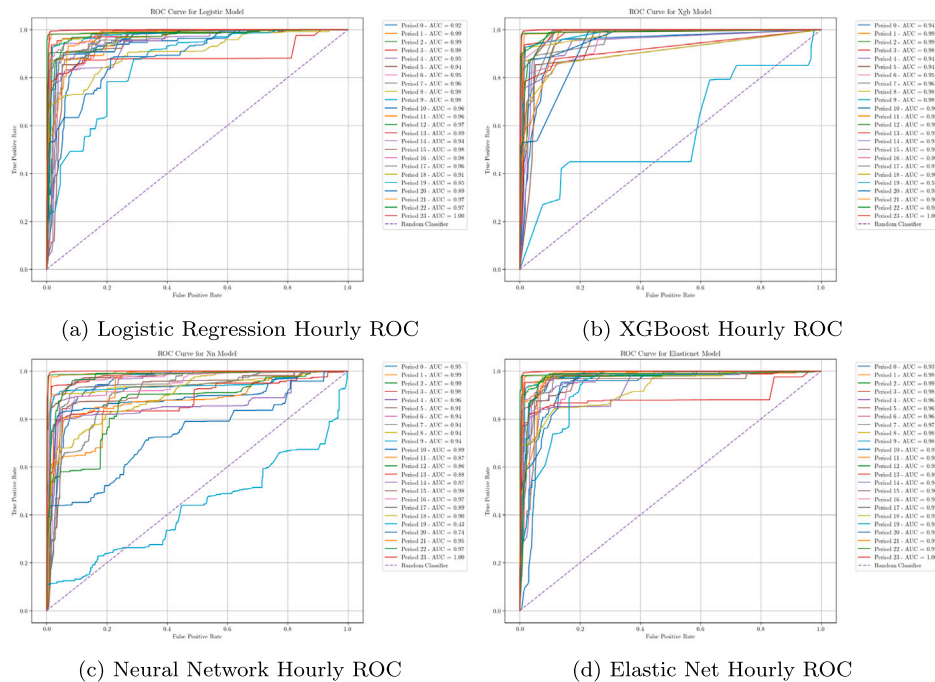
Fig. 2 presents the Receiver Operating Characteristic (ROC) curve for different hours of a random testing day. This visualization provides a measure of our trained model's binary classification performance. This plot depicts the True Positive Rate (TPR, or sensitivity) against the False Positive Rate (FPR, or 1-specificity) at different classification thresholds. Our model Area Under the Curve (AUC), which summarizes the ROC curve into a single number, indicates good model performance. The average hourly AUC for the elastic model used is 0.9625 with a standard deviation of 0.025. Its value is significantly closer to 1 for all the hours of the day, the score for a perfect classifier, signifying that our model's predictions are far more accurate than a random guess (AUC = 0.5). Figs. 2(b) and 2(c) depict similar results overall, but notably, hour 19 shows an AUC ROC below 0.5. We hypothesize that the causes for this issue could be increased variability at hour 19 and these models may be overfitting the training data during this specific time period.

Fig. 3 shows recall for fitted models in the testing set (sensitivity or the true positive rate). The recall represents the model's ability to classify correctly. The recall score in Fig. 3 approaches 1, denoting a high classification performance. Model precision slightly declines during peak demand hours. During times of heightened energy demand,

Table 3

Average hourly metrics for tested methodologies. LR: Logistic Regression, NN: Neural Network.

Stat	Methodology Metric	Elastic Net	Lasso	LR	NN	Ridge	XGBoost
Average	AUC (Test)	0.979	0.980	0.981	0.991	0.979	0.992
	AUC (Train)	0.980	0.980	0.982	0.992	0.980	0.993
	Accuracy (Test)	0.962	0.962	0.963	0.972	0.962	0.973
	Accuracy (Train)	0.962	0.962	0.964	0.973	0.962	0.974
	F1 Score (Test)	0.801	0.796	0.815	0.868	0.806	0.871
	F1 Score (Train)	0.804	0.798	0.818	0.875	0.808	0.880
	RMSE (Test)	0.186	0.186	0.181	0.157	0.185	0.156
Max	RMSE (Train)	0.185	0.186	0.180	0.154	0.185	0.151
	AUC (Test)	0.996	0.996	0.996	0.999	0.996	0.999
	AUC (Train)	0.996	0.996	0.996	0.999	0.996	0.999
	Accuracy (Test)	0.991	0.991	0.992	0.994	0.991	0.994
	Accuracy (Train)	0.990	0.990	0.991	0.994	0.990	0.994
	F1 Score (Test)	0.975	0.975	0.977	0.985	0.975	0.985
	F1 Score (Train)	0.975	0.975	0.976	0.985	0.975	0.985
Min	RMSE (Test)	0.316	0.316	0.328	0.270	0.316	0.269
	RMSE (Train)	0.316	0.316	0.329	0.270	0.316	0.268
	AUC (Test)	0.940	0.940	0.948	0.972	0.939	0.974
	AUC (Train)	0.939	0.939	0.948	0.973	0.938	0.975
	Accuracy (Test)	0.900	0.900	0.893	0.927	0.900	0.928
	Accuracy (Train)	0.900	0.900	0.892	0.927	0.900	0.928
	F1 Score (Test)	0.367	0.367	0.374	0.705	0.367	0.702
	F1 Score (Train)	0.357	0.357	0.369	0.712	0.357	0.711
	RMSE (Test)	0.096	0.096	0.091	0.075	0.096	0.075
	RMSE (Train)	0.099	0.099	0.094	0.074	0.099	0.074

**Fig. 2.** Receiver Operation Characteristic (ROC) methodology comparison.

such as unexpected spikes or fluctuations in renewable and nonrenewable energy generation, energy producers may deviate more from their committed output, leading to a slight decline in the precision and recall metrics used to evaluate the model's performance. This decline is attributed to the increased complexity and volatility of the energy landscape during these periods, introducing more variability and uncertainty into the prediction process. However, despite these challenges, the model demonstrates robust performance, consistently achieving high AUC ROC scores. This demonstrates the capacity of the model to distinguish between compliance and deviation by energy producers, even under fluctuating demand conditions. Continuous refinement and

adaptation of the model would be crucial to maintaining its effectiveness in predicting deviations across various demand scenarios and changing landscapes.

Fig. 4 shows the Kolmogorov Smirnov (KS) statistic. The KS statistic measures the degree of separation between the distributions to be classified. A high KS indicates a model that separates the distributions more. The KS statistics in our estimations have a consistently high value across all testing hours. Therefore, the models we fit are rated as good classifiers. All these performance metrics, viewed together, offer evidence of our model's robust performance. The high scores in each of the three considered measures, ROC, recall, and KS statistic, show the model's capability of providing accurate, reliable predictions.

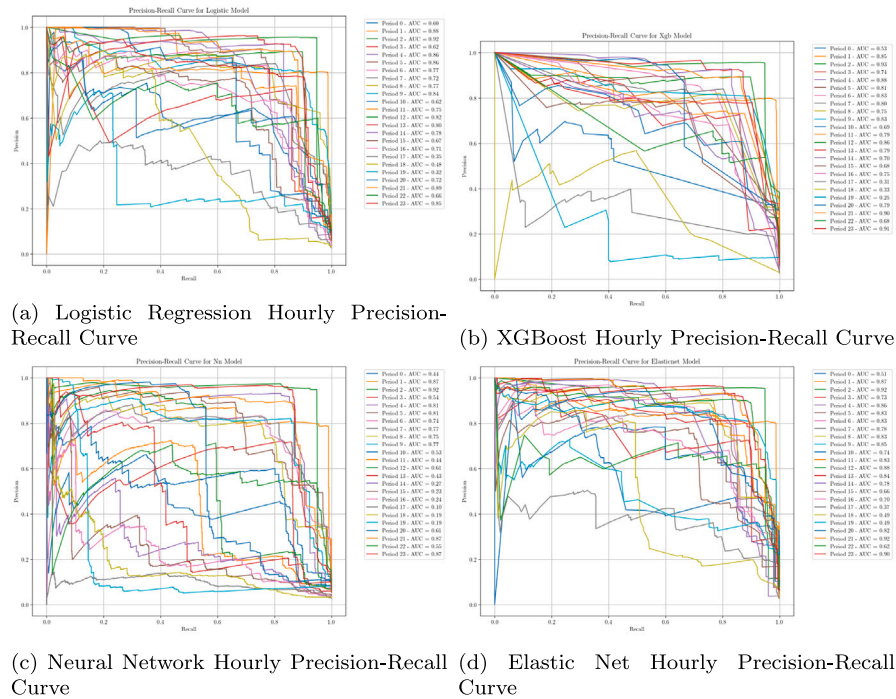


Fig. 3. Precision-Recall Curve methodology comparison.

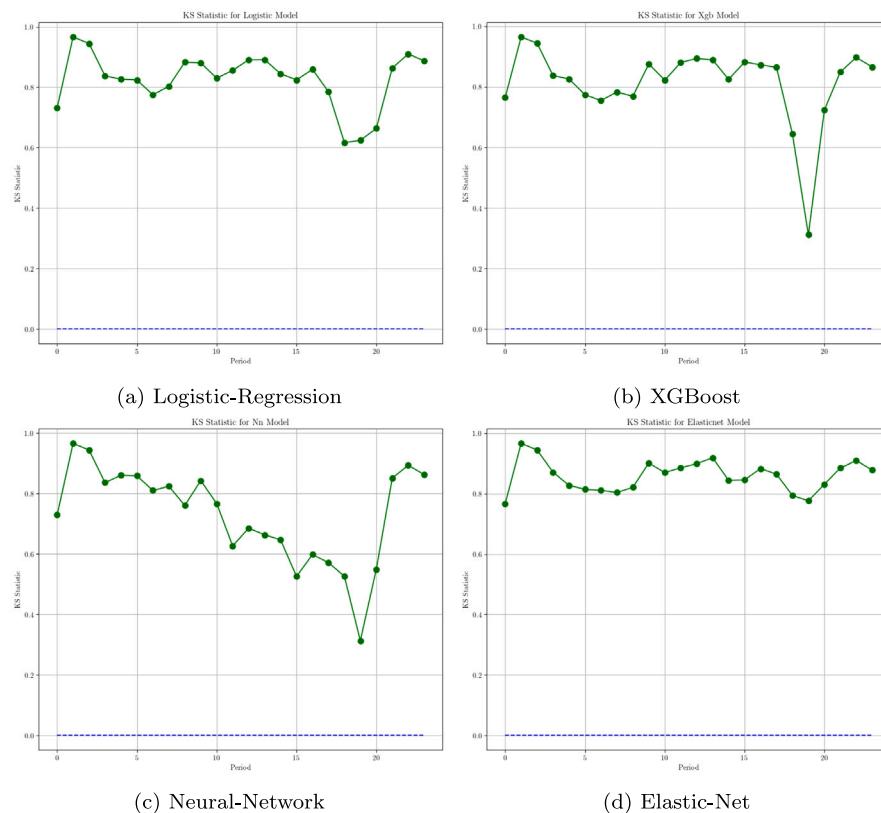


Fig. 4. Kolmogorov-Smirnov (KS) statistic methodology comparison.

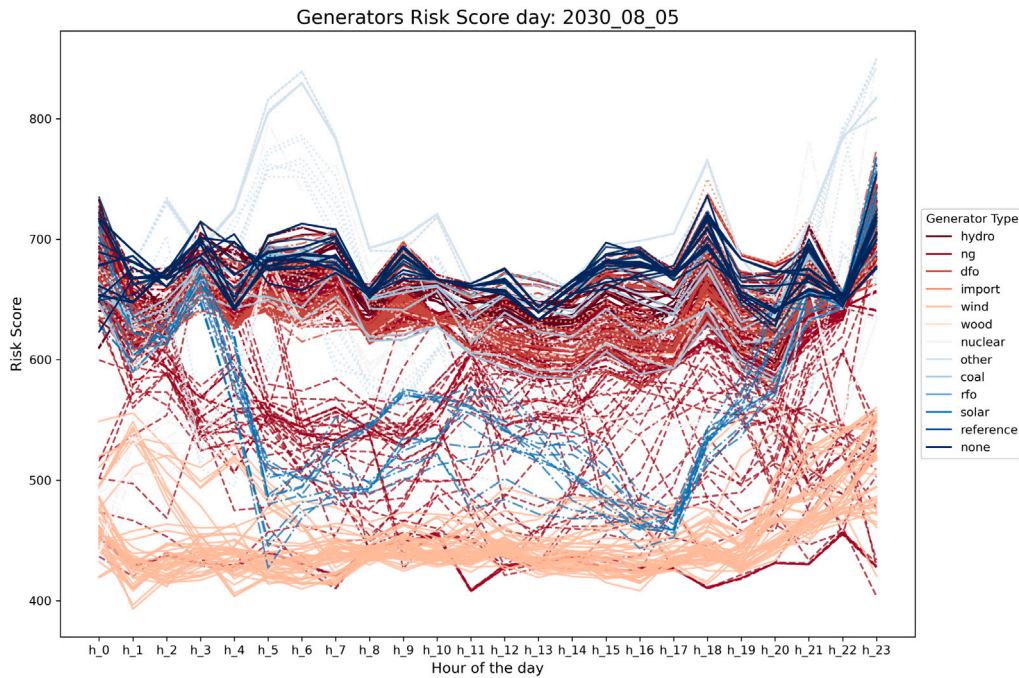
Table 4 compares the distribution of the development (training) and validation (testing) data sets for all available ranges of scores. Column A represents the number of generators located in each bin for all the tested scenarios. Column B is the respective percentage of the total. Columns C and D present the same results for the validation set to

contrast with the development set. Column E presents the difference between Columns D and B. These results show that the distribution of scores is consistent for the development and validation sets. Columns F, G, and H present the Info Odds, WoE, and Information value (IV), respectively for each one of the score ranges. Finally, columns I and J

Table 4

Train and Test score distribution comparison.

Score range	A	B	C	D	E	F	G	H	I	J	K
	Development		Validation		D-B	D/B	ln(F)	E x G	Cummulative		
	Count	Percent	Count	Percent	Difference	Info Odds	WoE	Info Value	Development	Actual	Diff
320–399	47	0.05%	19	0.08%	0.000	1.617	0.480	0.000	0%	0%	0.000
400–479	5735	6.34%	1439	6.36%	0.000	1.003	0.003	0.000	6%	6%	0.001
480–559	4390	4.85%	1139	5.03%	0.002	1.038	0.037	0.000	11%	11%	0.002
560–639	27 446	30.32%	6880	30.40%	0.001	1.002	0.002	0.000	42%	42%	0.003
640–719	44 394	49.05%	11 011	48.65%	−0.004	0.992	−0.008	0.000	91%	91%	−0.001
720–799	7190	7.94%	1838	8.12%	0.002	1.022	0.022	0.000	99%	99%	0.001
800+	1312	1.45%	308	1.36%	−0.001	0.939	−0.063	0.000	100%	100%	0.000

**Fig. 5.** Generators risk scoring profile Testing Date: 2030-08-05.

show the cumulative distribution of the bins, and column K shows the difference between these columns.

Fig. 5 shows the risk profiles of the different energy generators by the hour of the day according to fuel type for the 2030 NYISO forecast system. Fig. 5 has two main takeaways in terms of risk management of VRES. First, there is a relationship between the times of the day when these generators can produce energy and the increase in variability and therefore risk. Solar energy generators, given the installed capacity in the system, have a predictable pattern, with risk scores increasing from sunrise, peaking near midday, and dropping off towards sunset. Although there is variability to this pattern, our results for the testing year 2030 hold for several testing days. Second, wind generators exhibit a lower but more stable risk score pattern during the day. This is a consequence of the inherent variability of wind generation, which operates throughout the day but exhibits pronounced diurnal fluctuations. In contrast, nuclear power generators have a high score (low risk) due to the low variability in their output. Overall, nuclear generators are dispatchable and provide a stable baseline output, irrespective of the time of day.

Natural gas generators provide flexibility and ramping capability to the system. This flexibility allows these generators to cover the high variability periods of renewable energy, creating a buffer during periods of low or high renewable output, and ensuring a consistent power supply.

Fig. 6 focuses on the hourly risk distribution for a single wind generator. We select a wind generator to illustrate the case of generators scored as risky in our analysis. The highest risk coincides

with the periods with the highest variability in output. Compared to solar energy generation, there is higher variability for wind generation during the day in the NYISO system. Unpredictable wind conditions lead to significant output fluctuations, a challenge for wind power generators. However, note that the variability is stable, maintaining a consistent profile during the day. A key finding of this study is that both VRES and conventional generators, and all other participants, can adjust their policies and strategies to compensate for the underlying stochastic phenomena product of the integration of VRES by leveraging the forecasting capabilities of our scoring method.

Fig. 7 breaks down the previously presented risk profile, unveiling the individual impact of each considered factor. This decomposition provides a multi-dimensional representation of how each factor influences the overall risk score, through a feature engineering and selection process. This insight into variability components enables participants on the NYISO grid to optimize their operations, as it highlights how risk can fluctuate across different times of the day or even on separate days.

Figs. 8 and 9 present the distribution function and cumulative distribution function of ABSCoRES risk scores among all generators for August 2 2030 and 2019.

Fig. 8, shows the hourly distribution function across all generators. The graph reveals a skewness towards the lower end of the risk score spectrum, as indicated by the long tail to the left. Therefore, a set of generators, representing less than the 15% cumulative probability, have relatively low risk scores. The bulk of generators' risk scores are

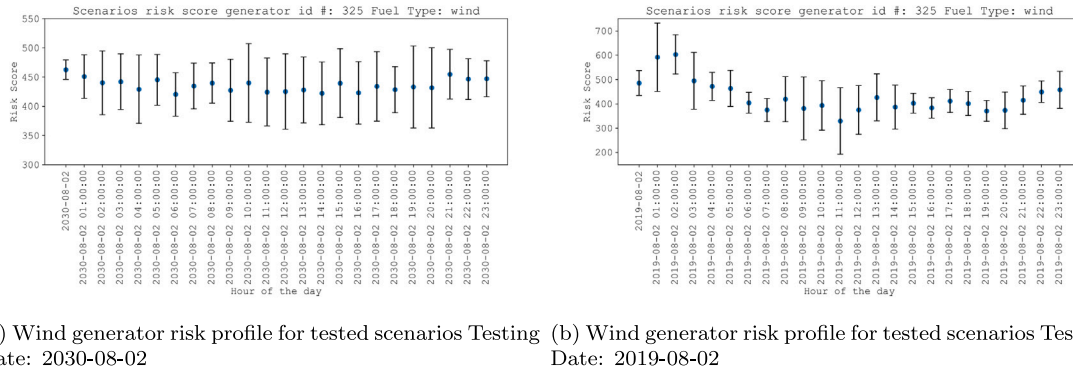


Fig. 6. Comparison of Risk profile for wind generator of risk scores.

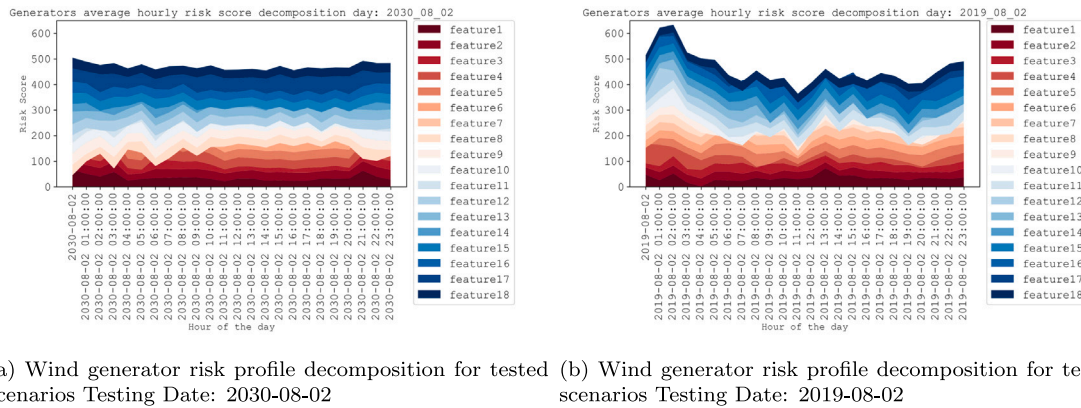


Fig. 7. Comparison of Risk profile decomposition of risk scores.

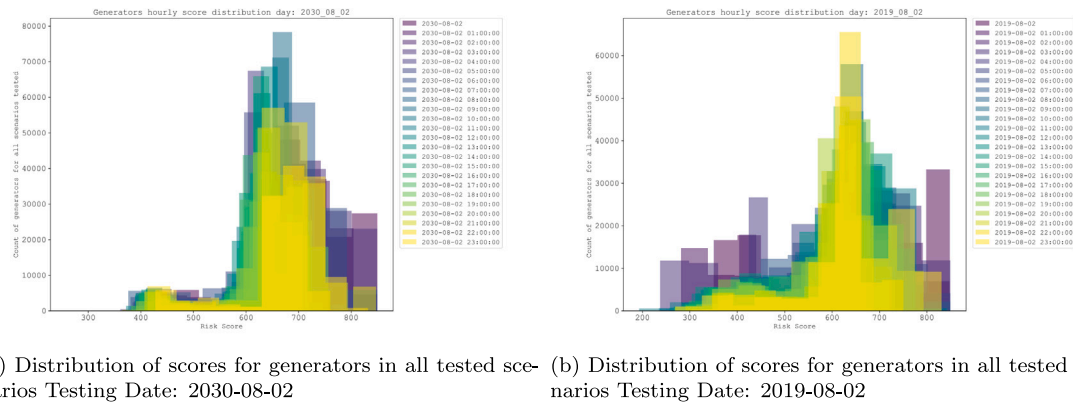


Fig. 8. Comparison of distribution of scores.

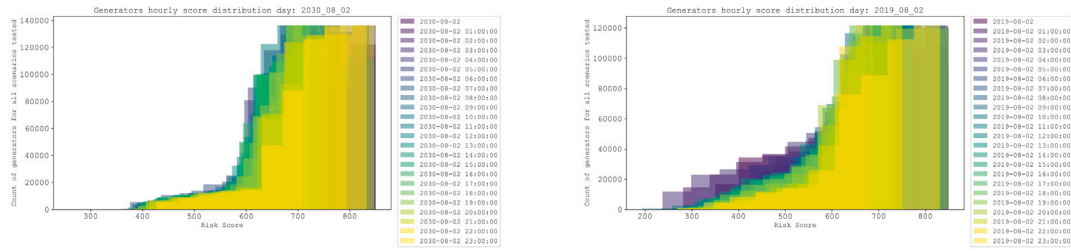
concentrated within the range of 600 to 750, indicating a relatively moderate level of risk for the majority of participants for the testing period.

Fig. 9 presents the cumulative distribution of risk scores. The majority of generators have risk scores above 600, indicating that while there is variability in risk scores, most generators fall within a range of moderate risk of deviation for this illustrative day. It is worth noting that the risk scores align closely with demand patterns within the system. During periods characterized by higher demand and demand variability (16:00–19:00), there is an inherently greater risk of energy generators deviating from their commitments, resulting in lower ABSCoRES. Conversely, in periods of lower demand and reduced variability (00:00–04:00), the risk diminishes, leading to higher ABSCoRES for energy generators.

4.2. Deviation and risk scores

Fig. 10 shows the distribution of deviations from schedules for August 2030. These deviations affect the ABSCoRES operational risk scores. Fig. 10(a), shows the total energy production in megawatts (MW) during the testing period. This graph delineates the energy generation profile for the specific day under analysis, encompassing both renewable and non renewable generators. For this operating period, the peak generation occurs at hour 17, and the lowest generation occurs near midnight.

Figs. 10(b) and 10(c) decompose the deviations from their anticipated energy dispatch between the positive and negative portions for all generators. Maximum positive and negative deviations coincide



(a) Cumulative distribution of scores for generators in all tested scenarios Testing Date: 2030-08-02 (b) Cumulative distribution of scores for generators in all tested scenarios Testing Date: 2019-08-02

Fig. 9. Comparison of Cumulative distribution of scores.

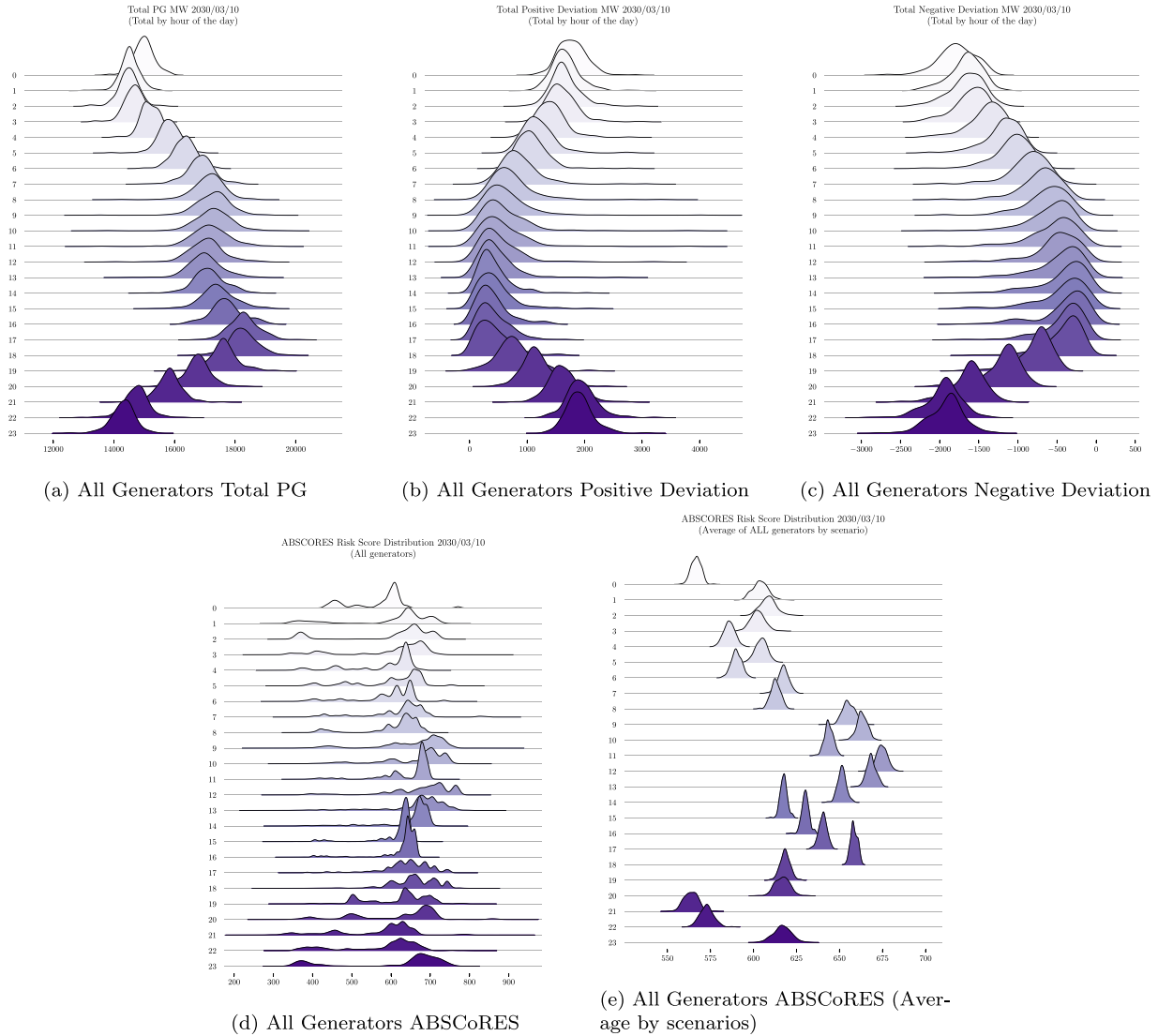


Fig. 10. All energy generators deviation and risk scores.

with periods of lowest energy generation, whereas minimum deviations align with the periods of highest energy output.

Figs. 10(d) and 10(e) present the profiles of ABSCoRES Risk scores for all tested scenarios and the average ABSCoRES risk score by scenario for the testing period. The figures illustrate that during periods of high energy output and low deviation, the ABSCoRES reach their maximum average scores. Analogously, in periods of lowest energy output and highest deviation, the ABSCoRES reach their lowest scores. This underscores how ABSCoRES risk scores capture the system's variability

and assess the efficiency of both renewable and non renewable energy generators based on their energy output and deviation. In all cases, the generator characteristics are considered in this evaluation. We refer the reader to Section 3 for further details.

4.3. Seasonal and spatial variation

Fig. 11 shows the expected generation composition by fuel type for 200 randomly generated scenarios simulations of March and August

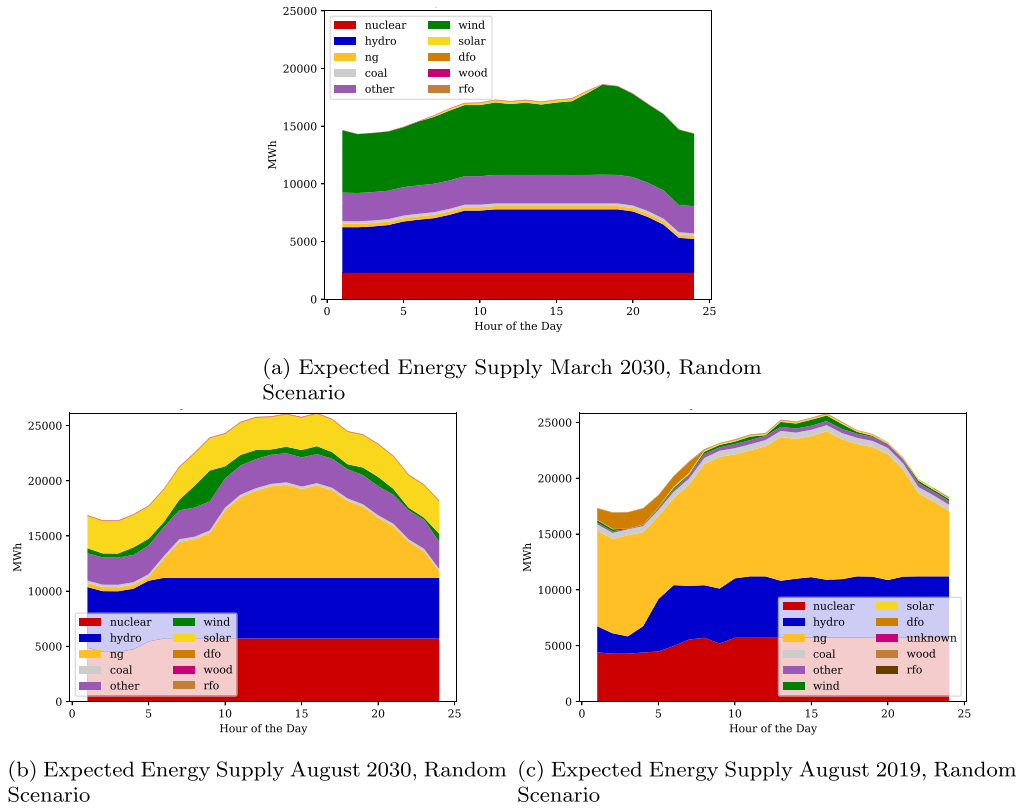


Fig. 11. Comparison of expected daily energy portfolios.

2030. Moreover, we included a random scenario for the same month in 2019. This comparison shows the increased penetration of solar and wind energy. These figures show seasonal variations in the availability of renewable resources. These varying availabilities affect the uncertainty and variability that the system operators (SOs) must balance. The amount of potential VRES generation available alters the system's risk profile, therefore impacting the overall operational conditions.

For instance, in August 2030, ramping requirements increase. New York is a summer peaking system in 2024, driven by air conditioning demand, and VRES generation is lower and more volatile during this period. The demand forecast for March 2030 is lower and has less diurnal variation than the forecast for August 2030, and VRES resources are more consistent than in summer periods. These characteristics lead to decreased ramping needs for conventional generators like the natural gas turbines modeled. These fluctuations in the generation mix directly correlate with shifts in energy generator risk scores, underscoring the intricate interplay between operational dynamics and risk assessment within the energy landscape. Although the ABSCoRES are constructed solely on the efficiency of the generators in meeting their day ahead commitment, the data from the SCED and SCUC optimization models used contains information from the intricate relationships of generation and distribution (e.g., Congestion, ramping).

Figs. 12 and 13 illustrate three selected hours of the day for representative days in the spring and summer of 2030, respectively, in the New York system. We chose these hours based on the results shown in Fig. 10. These periods of the day have the highest and lowest average ABSCoRES risk scores for the periods analyzed. In March, high stress points are concentrated in the northeast region. Although at 5:00 ET, the risk profiles improve, during periods of higher demand and lower energy generation at 2:00 ET and 21:00 ET, energy generators in the capital region near Albany and Watertown experience a dramatic impact on their risk scores. This spatiotemporal analysis underscores a focused stress window during specific hours, with localized challenges. In contrast, August presents a broader and more persistent stress pattern throughout the day and over the system footprint, particularly

affecting generators in Western New York and Long Island. These variations correlate closely with weather phenomena. The temperature gradients are part of the transition from winter to spring in March. The increased cooling demand and reduced VRES availability during August are characteristic of this period.

Load pockets, transmission congestion, and regional energy needs affect the risk at different periods and locations. The ABSCoRES provides the SOs a coordination mechanism and information signal to stakeholders besides the price signal. Overall, the figure hints at actionable insights for fortifying the reliability of the energy system in response to these identified stressors. Besides the operational applications of the risk measures we developed, the ABSCoRES provides long term signals for targeted interventions and infrastructure enhancements, including identifying corridors for transmission line planning areas where generation support for e.g., ramping provision, is needed.

5. Conclusion and policy implications

We present a framework to manage operational risk in electricity systems that integrates the outputs of the Security Constrained Economic Dispatch (SCED) and Security Constraint Unit Commitment (SCUC) models. We calculate a risk score for stakeholders at the asset level for each period. By combining the data output of these two optimization models, we provide a comprehensive understanding of the risk profile in electricity systems. Our analysis involves both simulations and estimation to evaluate the variability and sensitivity of the risk scoring methodology, utilizing different scenarios for solar and wind energy generation, as well as demand.

We use a detailed feature engineering pipeline to integrate SCUC and SCED runs in our framework, converting the outputs of the optimization step for the estimation step. These outputs are then used to construct the dataset for training and validating our machine learning models. We test and compare several estimation techniques, including machine learning methodologies. The evaluation metrics include AUC

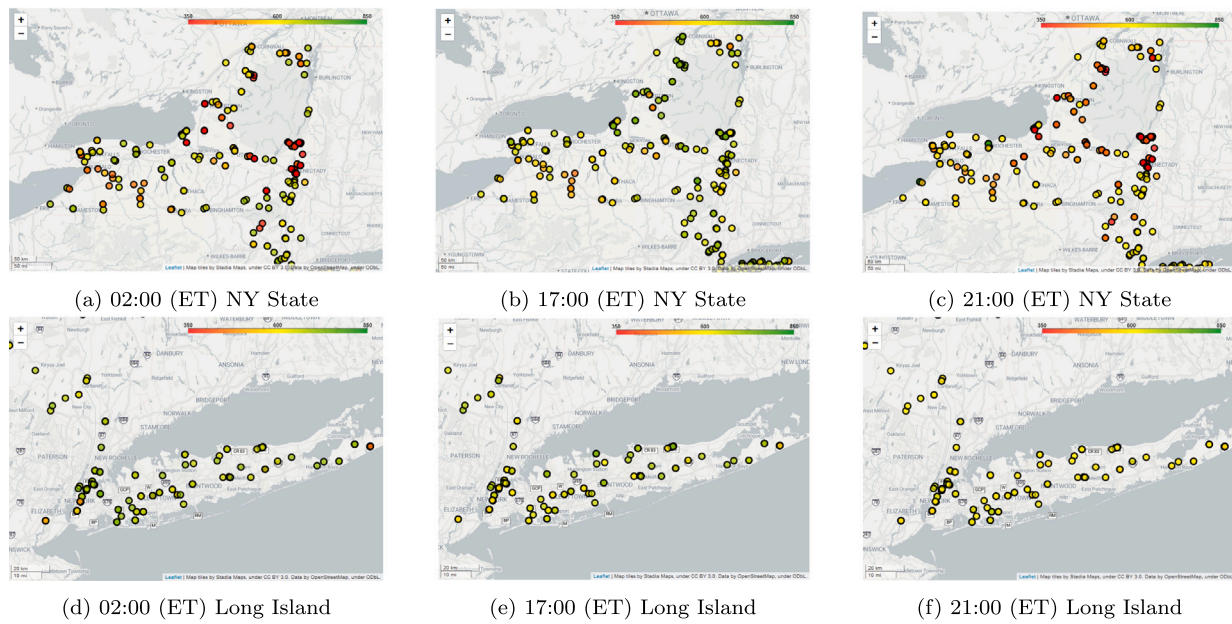


Fig. 12. Spatial risk scores March 2020.

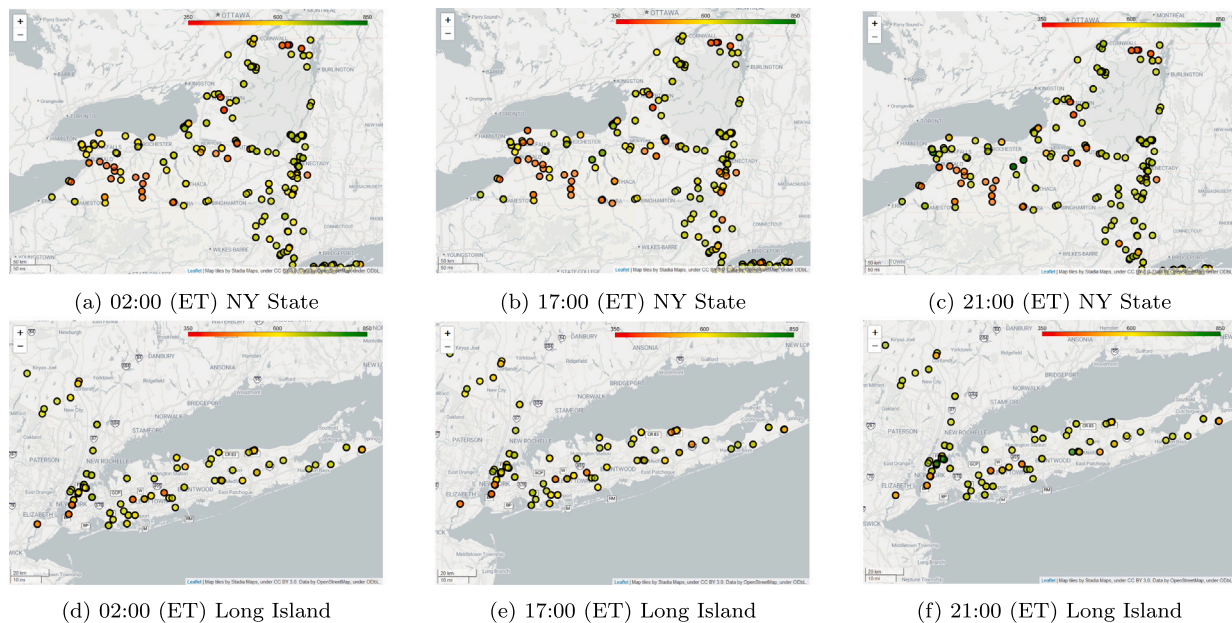


Fig. 13. Spatial risk scores, August 2020.

ROC, Recall, Accuracy, and KS Statistics. The most parsimonious model included a combination of a binning strategy using Weight of Evidence (WoE) and Information Value (IV), achieving a balance between predictive power and explainability of the method. The risk scores are derived as a numerical transformation of the probability output from the machine learning classification model. To our knowledge, there are no comparable studies or frameworks that have derived interpretable risk measures calculated using SCUCs and SCEDs for stakeholders to use in a distributed manner as a coordination mechanism with SOs.

We employ a quantile strategy to categorize generators based on their likelihood of deviating from their energy contracts. Based on these quantiles, we selected a threshold to characterize generators. Our analysis reveals that the risk scores are sensitive to changes in the portfolio of assets available to the optimization models as they are updated with different types of generators and network characteristics.

The practical applications and implications of this framework for electricity system operators and stakeholders are significant. It provides valuable insights into the risk profiles of energy generators, enabling more informed decision making regarding system operations, resource planning, and market participation, and the electric assets risk bureau can serve as a platform to innovate in risk hedging strategies.

In our future work, we plan to evaluate the impact of alternative changes to the risk scores and the optimization module itself. We are also interested in using scores in planning settings for long term investments.

CRedit authorship contribution statement

Andres F. Ramirez: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Alberto J.**

Lamadrid L.: Writing – review & editing, Writing – original draft, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Andres Felipe Ramirez Arias reports financial support was provided by Advanced Research Projects Agency-Energy. Alberto J. Lamadrid L. reports financial support was provided by Advanced Research Projects Agency-Energy. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

- [1] Bloomberg New Energy Finance, New energy outlook 2025, 2025.
- [2] Google, A Sustainability Moonshot: 24/7 Carbon-Free Energy by 2030, tech. rep, Google, 2024.
- [3] Office of the Governor, State of California, California Releases World's First Plan to Achieve Net Zero Carbon Pollution, tech. rep, Office of the Governor, 2022.
- [4] U.S. Energy Information Administration, Renewable Energy Explained: Portfolio Standards, tech. rep, Energy Information Administration, 2024.
- [5] R. Rockafellar, J. Royset, Measures of residual risk with connections to regression, risk tracking, surrogate models, and ambiguity, *SIAM J. Optim.* 25 (2) (2015) 1179–1208.
- [6] F. Golestaneh, P. Pinson, R. Azizpanah-Abarghoee, H.B. Gooi, Ellipsoidal prediction regions for multivariate uncertainty characterization, *IEEE Trans. Power Syst.* 33 (4) (2018) 4519–4530.
- [7] M. Ilic, L. Xie, J.-Y. Joo, Efficient coordination of wind power and price-responsive demand part I: Theoretical foundations, *Power Syst. IEEE Trans.* 26 (4) (2011) 1875–1884.
- [8] A.J. Lamadrid, D. Muñoz-Álvarez, C.E. Murillo-Sánchez, R.D. Zimmerman, H. Shin, R.J. Thomas, Using the Matpower Optimal Scheduling Tool to test power system operation methodologies under uncertainty, *IEEE Trans. Sustain. Energy* 10 (3) (2019) 1280–1289.
- [9] X. Kuang, A.J. Lamadrid, L.F. Zuluaga, Pricing in non-convex markets with quadratic deliverability costs, *Energy Econ.* 80 (2019) 123–131.
- [10] S. Borenstein, The long-run efficiency of real-time electricity pricing, *Energy J.* 26 (3) (2005) 93–116.
- [11] P. Joskow, J. Tirole, Reliability and competitive electricity markets, *Rand J. Econ.* 38 (1) (2007) pp. 60–84.
- [12] C. Cahn, M. Girotti, F. Salvadé, Credit ratings and the hold-up problem in the loan market, *Manag. Sci.* (2023).
- [13] D.J. Walters, G. Ülkümen, D. Tannenbaum, C. Erner, C.R. Fox, Investor behavior under epistemic vs. Aleatory uncertainty, *Manag. Sci.* (2023).
- [14] P. Artzner, F. Delbaen, J.-M. Eber, D. Heath, Coherent measures of risk, *Math. Finance* 9 (3) (1999) 203–228.
- [15] S. Kou, X. Peng, On the measurement of economic tail risk, *Oper. Res.* 64 (5) (2016) 1056–1072.
- [16] S. Emmer, M. Kratz, D. Tasche, What is the best risk measure in practice? A comparison of standard measures, *J. Risk* (2013).
- [17] A. Ioannou, A. Angus, F. Brennan, Risk-based methods for sustainable energy system planning: A review, *Renew. Sustain. Energy Rev.* 74 (2017) 602–615.
- [18] X. Wu, N. Li, M. He, X. Wang, S. Ma, J. Cao, Risk-constrained day-ahead scheduling for gravity energy storage system and wind turbine based on IGDT, *Renew. Energy* 185 (2022) 904–915.
- [19] H. Bakhtiari, J. Zhong, M. Alvarez, Uncertainty modeling methods for risk-averse planning and operation of stand-alone renewable energy-based microgrids, *Renew. Energy* 199 (2022) 866–880.
- [20] E.J. da Silva Pereira, J.T. Pinho, M.A.B. Galhardo, W.N. Macêdo, Methodology of risk analysis by Monte Carlo method applied to power generation with renewable energy, *Renew. Energy* 69 (2014) 347–355.
- [21] R. Shahnazi, M. Alimohammadlou, Investigating risks in renewable energy in oil-producing countries through multi-criteria decision-making methods based on interval type-2 fuzzy sets: A case study of Iran, *Renew. Energy* 191 (2022) 1009–1027.
- [22] H. Ghoddsi, G.G. Creamer, N. Rafizadeh, Machine learning in energy economics and finance: A review, *Energy Econ.* 81 (2019) 709–727.
- [23] S. Fogelberg, E. Lazarczyk, Wind power volatility and its impact on production failures in the nordic electricity market, *Renew. Energy* 105 (2017) 96–105.
- [24] D. Cardona-Vasquez, J. Garcia-Rendon, A. Arango-Manrique, Effect of the intermittency of non-conventional renewable energy sources on the volatility of the Colombian spot price, *Renew. Energy* 232 (2024) 121073.
- [25] X. Wu, Y. Wu, X. Cheng, C. Cheng, Z. Li, Y. Wu, A mixed-integer linear programming model for hydro unit commitment considering operation constraint priorities, *Renew. Energy* 204 (2023) 507–520.
- [26] E. Rey-Costa, B. Elliston, D. Green, G. Abramowitz, Firming 100% renewable power: Costs and opportunities in Australia's national electricity market, *Renew. Energy* 219 (2023) 119416.
- [27] X. Wang, R. Long, H. Chen, Y. Wang, Y. Shi, S. Yang, M. Wu, How to promote the trading in China's green electricity market? Based on environmental perceptions, renewable portfolio standard and subsidies, *Renew. Energy* 222 (2024) 119784.
- [28] C.-Y. Lin, K.Y. Chau, T.K. Tran, M. Sadiq, L. Van, T.T.H. Phan, Development of renewable energy resources by green finance, volatility and risk: empirical evidence from China, *Renew. Energy* 201 (2022) 821–831.
- [29] NERC, Reliability Standards for the Bulk Electric Systems of North America, North American Electric Reliability Corporation, 3353 Peachtree Road NE Suite 600, North Tower, Atlanta, GA 30326, 2022.
- [30] M.B. Cain, R.P. O'Neill, A. Castillo, History of Optimal Power Flow and Formulations, tech. rep, FERC, 2012.
- [31] D.P. Brown, L. Muehlenbachs, The value of electricity reliability: Evidence from battery adoption, tech. rep, Resources For the Future, 2023.
- [32] D.J. Hand, N.M. Adams, Defining attributes for scorecard construction in credit scoring, *J. Appl. Stat.* 27 (5) (2000) 527–540.
- [33] G. Zeng, A necessary condition for a good binning algorithm in credit scoring, *Appl. Math. Sci.* 8 (65) (2014) 3229–3242.
- [34] G. Navas-Palencia, Optimal binning: mathematical programming formulation, 2020, arXiv preprint arXiv:2001.08025.
- [35] M.J. Bayarri, G. García-Donato, Generalization of Jeffreys Divergence-Based Priors for Bayesian Hypothesis Testing, *J. R. Stat. Soc. Ser. B Stat. Methodol.* 70 (5) (2008) 981–1003.
- [36] D.J. Hand, W.E. Henley, Statistical classification methods in consumer credit scoring: a review, *J. R. Stat. Soc. Ser. A (Statistics Soc.)* 160 (3) (1997) 523–541.
- [37] S. Lessmann, B. Baesens, H.-V. Seow, L.C. Thomas, Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research, *European J. Oper. Res.* 247 (1) (2015) 124–136.
- [38] R.L. Prentice, S. Zhao, Regression models and multivariate life tables, *J. Amer. Statist. Assoc.* 116 (535) (2021) 1330–1345.
- [39] T. Chen, C. Guestrin, Xgboost: A scalable tree boosting system, in: *Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785–794.
- [40] J. Heaton, Ian goodfellow, yoshua bengio, and aaron courville: Deep learning: The MIT press, 2016, 800 pp, ISBN: 0262035618, Genet. Program. Evolvable Mach. 19 (1–2) (2018) 305–307.
- [41] B.R. Gunnarsson, S. Vanden Broucke, B. Baesens, M. Óskarsdóttir, W. Lemahieu, Deep learning for credit scoring: Do or don't? *European J. Oper. Res.* 295 (1) (2021) 292–305.
- [42] X. Pan, T. Zhao, M. Chen, S. Zhang, Deepopf: A deep neural network approach for security-constrained dc optimal power flow, *IEEE Trans. Power Syst.* 36 (3) (2021) 1725–1735.
- [43] PERFORM-Forecasts, Documentation for ARPA-e PERFORM datasets produced by NREL, 2023, <https://github.com/PERFORM-Forecasts/documentation>.
- [44] B. Sergi, C. Feng, F. Zhang, B.-M. Hodge, R.-J. Ross, B. Richard, D. Kate, R. Megan, B. Grant, R. Michael, ARPA-E PERFORM datasets, tech. rep., Advanced Research Projects Agency - Energy, 2022.
- [45] National Renewable Energy Laboratory (NREL), National solar radiation database, 2022, Dataset.
- [46] J.G. Powers, J.B. Klemp, W.C. Skamarock, C.A. Davis, J. Dudhia, D.O. Gill, J.L. Coen, D.J. Gochis, R. Ahmadov, S.E. Peckham, et al., The weather research and forecasting model: Overview, system efforts, and future directions, *Bull. Am. Meteorol. Soc.* 98 (8) (2017) 1717–1737.
- [47] Princeton University, PGscen: A stochastic scenario generation framework for power grid data, 2023, <https://github.com/PrincetonUniversity/PGscen>.
- [48] S. Greene, NYISO Network 2019, Technical Report, University of Wisconsin-Madison, 2022, Available (On-Peak) at <https://drive.google.com/file/d/1-gbaK7vtXIIE7oh-hXn9uhAaDuTW93fz/view?usp=sharing> ; (Off-Peak) at <https://drive.google.com/file/d/1-aluagdHskdkEq08QFFIAT3BWzftMSt-/view?usp=sharing>.
- [49] S. Greene, NYISO Network 2030, Technical Report, University of Wisconsin-Madison, 2023, Available at <https://drive.google.com/file/d/1-nNs17qjkYKkyfQWuiSBiYlr1c02R/view?usp=sharing>.
- [50] N. Siddiqui, Intelligent credit scoring: Building and implementing better credit risk scorecards, John Wiley & Sons, 2017.