Data-Driven Operational Risk Assessment in Renewable-Rich Power System

An Automated Probabilistic Power Flow and SCADA Data Mining Framework

Luong Phu Nhan
Power System Analysis and Planning Department
National Power System and Market Operator Company – Central Regional Power System Operator
Da Nang, Vietnam
luongphunhan93@gmail.com

Le Thi Da Thao
Power System Analysis and Planning Department
National Power System and Market Operator Company – Central Regional Power System Operator
Da Nang, Vietnam
dathaovng@gmail.com

Le Hong Lam
Faculty of Electrical Engineering
University of Science and Technology, The University of Danang
Da Nang, Vietnam
lhlam@dut.udn.vn

Nguyen Luong Minh
Central Electrical Testing Company Limited – Central Power Corporation
Da Nang, Vietnam
minhnl@cpc.vn

Nguyen Huu Hieu *
Faculty of Electrical Engineering
University of Science and Technology, The University of Danang
Da Nang, Vietnam
nhhieu@dut.udn.vn

Abstract*— *The increasing penetration of variable renewable energy (VRE) introduces profound stochastic uncertainties into power systems, rendering traditional deterministic security assessments insufficient. This paper addresses this challenge by developing a comprehensive, AI-ready operational risk assessment framework based on a data-driven Probabilistic Power Flow (PPF) approach. Unlike conventional methods that rely on assumed statistical distributions, the proposed methodology employs a statistical learning pipeline to identify suitable probability density functions directly from historical SCADA telemetry, thereby ensuring high-fidelity input modeling. Furthermore, a cluster-based correlation strategy, rooted in unsupervised pattern recognition, is constructed to rigorously capture the spatiotemporal dependencies among renewable generation sources. The probabilistic framework is executed via an enhanced Monte Carlo simulation (MCS) utilizing Latin Hypercube Sampling (LHS) to improve computational efficiency. Validated through a case study on the Central Vietnam Power System, the results demonstrate that the proposed framework effectively uncovers hidden operational vulnerabilities, specifically identifying a 4.57% probability of exceeding the 90% branch-loading warning threshold on the critical 220 kV Quy Nhon - Tuy Hoa transmission line, while the probability of actual thermal overload above 100% is nearly zero. This near-limit operating risk profile may be masked by conventional deterministic snapshots. The findings emphasize that accurately modeling the correlation structure of renewable sources is indispensable for realistic risk quantification, providing system operators with a robust, algorithm-driven decision-support tool for grid security management.

Keywords: Data-Driven Risk Assessment; Probabilistic Power Flow; SCADA Data Mining; Automated Algorithmic Pipeline; Smart Grid Infrastructure; Statistical Modeling.

I. INTRODUCTION

The global transition toward sustainable energy has fundamentally redefined the operational landscape of modern power grids. In regions experiencing a massive influx of renewable capacities, such as Central Vietnam, the grid operates under unprecedented levels of uncertainty. The integration of VRE sources—primarily solar and wind—introduces complex, weather-dependent stochastic behaviors. For system operators tasked with ensuring continuous grid security, managing these intense fluctuations renders traditional operating states increasingly difficult to forecast and control.

Historically, grid security has been evaluated using Deterministic Power Flow (DPF) methods [1]. However, DPF evaluates the system under a limited set of isolated scenarios, making it structurally inadequate for comprehensively quantifying operational risks in renewable-rich environments [2], [3]. To overcome these limitations, PPF methodologies have been widely adopted. Despite its theoretical advantages, the practical application of standard PPF is frequently bottlenecked by its reliance on rigidly assumed probability distributions - such as assuming normal distributions for load demand or Weibull distributions for wind generation. These a priori assumptions often fail to reflect the true, dynamic statistical behavior of actual grid components. Furthermore, conventional PPF models typically treat uncertainty variables as statistically independent, overlooking the profound spatiotemporal correlations shared by regional renewable sources exposed to identical meteorological conditions.

Ignoring these critical dependencies can dangerously underestimate the probability of severe, cascading systemic risks, such as simultaneous power fluctuations causing critical line overloads.

The increasing availability of massive, high-resolution SCADA telemetry provides a transformative opportunity to bridge the gap between theoretical probability and practical, data-driven grid management. Recent studies have demonstrated this transition from different perspectives: Monte Carlo-based probabilistic power flow methods enable the characterization of system variable distributions and associated risks under uncertainty [4]; data-driven techniques such as clustering and kernel density estimation allow accurate modeling of stochastic renewable energy behaviors [5]; probabilistic approaches address the inherent limitations of traditional DPF in representing real-world system variability [1]; and risk-based probabilistic frameworks provide comprehensive assessment of system stability under high renewable penetration [6]. By leveraging automated algorithmic pipelines and statistical data mining techniques, it is therefore possible to extract highly accurate operational patterns directly from empirical data.

To this end, this paper proposes a comprehensive, data-driven framework for power system operational risk assessment. This framework shifts the paradigm from deterministic and assumption-based models to a robust, automated probabilistic evaluation.

The objective of this paper is not to propose a new sampling algorithm or a new power-flow solver, since Monte Carlo simulation, Latin Hypercube Sampling, Cholesky-based correlation injection, and AC power-flow analysis are well-established techniques. Instead, the main contribution of this work is the development and deployment of an automated, SCADA-driven operational risk assessment framework that integrates these established techniques into a practical end-to-end workflow for renewable-rich power system operation. The specific contributions of this paper are summarized as follows:

1) Automated SCADA-driven probabilistic input modeling: The proposed framework replaces fixed, pre-assumed probability distributions with a data-driven statistical fitting pipeline. Historical SCADA telemetry is automatically cleaned, time-stratified, and used to estimate the marginal probability distribution of each uncertain load and VRE source under specific operating condition.

2) Cluster-based renewable correlation modeling: To capture the spatial and temporal dependence among renewable sources, the framework groups renewable plants according to geographical and meteorological similarity and computes Spearman rank correlation coefficients directly from historical SCADA time series. This enables the generation of correlated renewable scenarios that better reflect actual system operation.

3) Automated correlated PPF co-simulation in PSS/E: The framework integrates correlated Latin Hypercube Sampling with Python-based PSS/E automation to generate, simulate, and evaluate thousands of stochastic operating scenarios [7], [8]. The simulation outputs are then converted into operational risk indices, such as the probability of exceeding branch-loading warning thresholds and voltage-limit violation probabilities.

4) Practical validation on the Central Vietnam power system: The proposed framework is deployed on an actual regional transmission system with high renewable penetration. The case studies demonstrate how SCADA-driven PPF can reveal near-limit operating risks that are masked by deterministic power-flow snapshots.

Ultimately, this automated framework serves as a highly applicable end-to-end decision-support mechanism for secure grid planning and operation. Furthermore, by systematically transforming raw operational telemetry into massive datasets of highly accurate, correlated simulation scenarios, this framework establishes the critical, high-quality data foundation requisite for training future Artificial Intelligence and Deep Learning models dedicated to real-time grid security prediction.

The remainder of this paper is organized as follows:

- **Section II** details the architecture of the data-driven probabilistic framework, encompassing the automated SCADA telemetry processing pipeline, unsupervised statistical input modeling, and cluster - spatiotemporal correlation analysis.
- **Section III** presents the high - performance algorithmic execution, focusing on the enhanced stratified sampling technique and the automated co - simulation orchestration engine between Python and PSS/E software.
- **Section IV** validates the proposed methodology through operational risk case studies on the Central Vietnam power system under extreme scenarios.
- Finally, **Section V** concludes the paper and highlights the main contributions of this research.

II. Architecture of the Data-Driven Probabilistic Framework

The core objective of the proposed framework is to systematically transform raw SCADA telemetry into high-fidelity probabilistic models and correlation structures, thereby eliminating the reliance on theoretical assumptions. The architecture of this data-driven framework is divided into three interconnected modules, fully automated via Python.

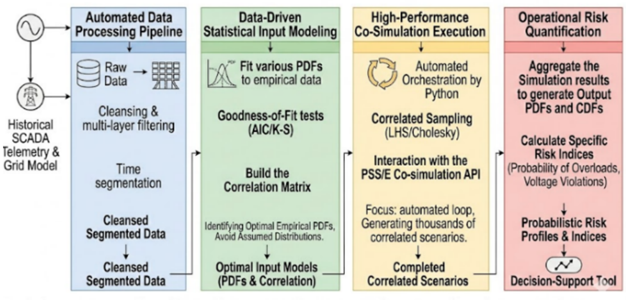

Figure 1. Architecture of Framework

A. Automated SCADA Telemetry Processing Pipeline

The foundation of any intelligent risk assessment model is high-fidelity input data. The raw dataset utilized in this research comprises three years of continuous telemetry collected from the SCADA system of the Central Vietnam power grid. To efficiently manage the millions of recorded data points, the telemetry is stored in the columnar Parquet format, which drastically optimizes query and loading performance for large-scale algorithmic processing.

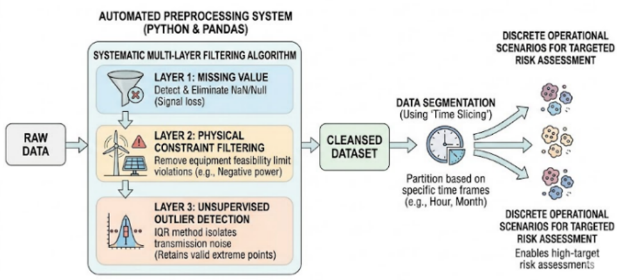

Figure 2. Multi-layer filter

The raw data is ingested into an automated preprocessing pipeline built using the Python programming language and the Pandas library. To guarantee the reliability of the statistical modeling, a systematic multi-layer filtering algorithm is executed:

Layer 1 (Missing Value Detection and Treatment): Detects missing or invalid entries caused by signal losses or communication failures. These entries are flagged and subsequently treated using forward-fill imputation to preserve the continuity of the 5-minute SCADA time series**.**

Layer 2 (Physical Constraint Filtering): Removes instances that violate the physical feasibility limits of the equipment, such as negative active power recordings from solar plants.

Layer 3 (Unsupervised Outlier Detection): To ensure the statistical integrity of the input models, it is imperative to autonomously identify and neutralize data anomalies originating from sensor malfunctions or SCADA transmission noise. We implement an unsupervised outlier detection mechanism utilizing the Interquartile Range (IQR) algorithm. For each targeted dataset, the acceptable operational boundaries are dynamically established as follows:

$$[Limit_{lower}, Limit_{upper}] = [Q_1 - 1.5xIQR, Q_3 + 1.5xIQR] \quad (1)$$

Where:

- $Limit_{lower}$: Lower boundary.
- $Limit_{upper}$: Upper boundary.
- Q_1: First quartile, the value below which 25% of the data fall.
- Q_3: Third quartile, the value below which 75% of the data fall.
- $IQR = Q_3 - Q_1$: Interquartile range.

Traditional filtering techniques relying on rigid Gaussian assumptions, such as the Z-score or Modified Z-score, are structurally ill-suited for VRE telemetry. These methods frequently misclassify valid, extreme operational states as errors, or conversely fail to detect subtle noise. A rigorous sensitivity analysis confirmed that the IQR method (employing a 1.5 multiplier threshold) strikes the optimal operational balance. As quantitatively demonstrated in Table I - evaluating reactive power telemetry through bay 231 at the 220kV Tuy Hoa substation—the standard Z-score (threshold 3.0) severely under-filtered by removing merely **0.08%** of samples. Conversely, the Modified Z-score (threshold 3.0) over-penalized the dataset by discarding **2.55%**. The IQR method effectively neutralized noise while preserving the physical truth of the extreme grid states, exhibiting a highly balanced removal rate of **1.52%**.

TABLE I. COMPARISON OF OUTLIER DETECTION METHODS

Method	Threshold	Samples removed	Removal rate (%)
Z-score	3.0	20	0.08%
Modified Z-score	3.0	652	2.55%
IQR	1.5	389	1.52%

Following the multi-layer filtration, the algorithm imputes any excised anomalies by propagating the last known valid observation (forward-fill), preserving time-series continuity across the 5-minute SCADA polling intervals.

A cornerstone of the proposed data mining methodology is transforming continuous raw time-series records into highly contextualized datasets representing characteristic operational scenarios. To accomplish this, an automated "temporal stratification" (time-slicing) technique multidimensionally partitions the cleansed SCADA repository based on diurnal (hours) and seasonal (months) features. By isolating each temporal stratum, the framework enables highly targeted risk profiling. For instance, mining the "midday off-peak in the dry season" cluster is essential for evaluating severe transmission line overload risks, whereas the "evening peak in the rainy season" dataset provides the statistical foundation to assess bus-voltage behavior, particularly high-side voltage deviations.

To improve reproducibility and clarify the scale of the empirical dataset, Table II summarizes the main characteristics of the SCADA telemetry used in this study. The dataset is time-stratified according to season and hour of day so that the probabilistic input models are estimated for specific operating conditions rather than for the entire annual dataset.

TABLE II. DESCRIPTION OF THE SCADA DATASET USED IN THIS STUDY

Item	Value
Historical data period	3 years
SCADA sampling interval	5 minutes
Raw number of records	115,106,400
Number of monitored loads	172
Number of monitored solar power plants	42
Number of monitored wind power plants	52
Number of monitored transmission elements	134

B. *Unsupervised Statistical Learning for Input Modeling*

A common limitation of conventional probabilistic power flow frameworks is the reliance on fixed, pre-assumed

 https://doi.org/10.1051/e3sconf/202672303006

probability distributions, such as assigning a Normal distribution to load demand or a Weibull distribution to wind generation. These assumptions may not accurately represent the empirical behavior of SCADA measurements under specific seasonal and hourly operating conditions. To reduce this modeling bias, the proposed framework deploys an automated statistical learning pipeline to estimate the marginal probability distribution of each uncertainty variable directly from historical SCADA telemetry.

For each temporally stratified dataset, the framework fits a library of candidate parametric probability density functions (PDFs) to the cleansed SCADA samples. The candidate library includes commonly used distributions such as Student-t, Lognormal, Gamma, Beta, Weibull, and Normal distributions. The parameters of each candidate distribution are estimated using maximum likelihood estimation. The fitted models are then evaluated through a two-stage procedure consisting of goodness-of-fit diagnostics and AIC-based model selection.

Step 1: Goodness-of-Fit Diagnostic Testing.

The Kolmogorov–Smirnov (K-S) test [9] is applied as a diagnostic goodness-of-fit indicator to quantify the discrepancy between the empirical cumulative distribution function and the fitted theoretical cumulative distribution function. The K-S test is based on the null hypothesis that the empirical samples are drawn from the considered candidate distribution. In this study, the K-S p-value is not used as a strict rejection filter. This is because the candidate distribution parameters are estimated from the same empirical dataset used for testing, and large SCADA datasets can make even small deviations statistically significant. Therefore, a p-value below the conventional 0.05 significance level does not necessarily imply that the fitted distribution is unusable for probabilistic simulation. Instead, the K-S result is reported as a diagnostic indicator to support the interpretation of the fitted models together with the AIC ranking.

Step 2: AIC-Based Model Selection.

For final model selection, the Akaike Information Criterion (AIC) [10] is adopted as the primary ranking criterion. The AIC evaluates the relative quality of statistical models by balancing likelihood-based goodness-of-fit and model complexity, thereby reducing the risk of selecting an unnecessarily complex model. The AIC is formulated as:

$$\mathrm{AIC} = 2\mathrm{k} - 2\ln(\mathrm{L}) \tag{2}$$

Where:

- k: the number of estimated parameters
- L: the maximum likelihood of the fitted model

The candidate distribution with the lowest AIC value is selected as the best available parametric approximation for the considered uncertainty variable within the predefined distribution library. In this sense, the selected distribution is not claimed to be a perfect representation of the empirical data, but rather the most suitable candidate among the tested parametric models.

To demonstrate the automated statistical fitting pipeline, a representative SCADA operational parameter was evaluated. As shown in Table III, all candidate parametric distributions produced K-S p-values below the conventional 0.05 significance level. This indicates that none of the tested distributions can be regarded as a perfect statistical representation of the empirical SCADA data under a strict K-S acceptance criterion. Therefore, the framework relies on AIC as the primary model selection criterion, while the K-S p-values are retained as diagnostic information. The Normal distribution provides the poorest fit, with the highest AIC value of 1241 and a very small K-S p-value. In contrast, the Student-t distribution achieves the lowest AIC value of 1063 and the highest relative K-S p-value of 0.04706709 among the tested candidates, including Lognormal, Gamma, and Beta distributions. Therefore, the Student-t distribution is selected as the best available parametric approximation for this SCADA variable within the considered distribution library.

TABLE III. AIC-BASED RANKING OF CANDIDATE DISTRIBUTIONS WITH K-S P-VALUES REPORTED AS DIAGNOSTIC INDICATORS

Distribution	AIC	p-value
student	1063	0.04706709
lognorm	1069	3.98E-07
gamma	1077	8.656E-05
beta	1101	5.4639E-06
norm	1241	3.62E-12

The fitted PDF and CDF of the selected distribution are shown in Fig. 3 and Fig. 4, respectively. These plots visually compare the empirical SCADA data with the fitted theoretical model, providing an additional qualitative check of the selected marginal distribution.

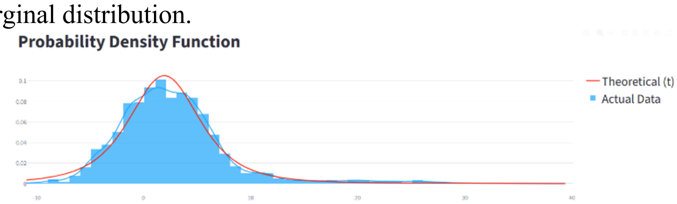

Figure 3. PDF Plot of SCADA actual data and fitted distribution

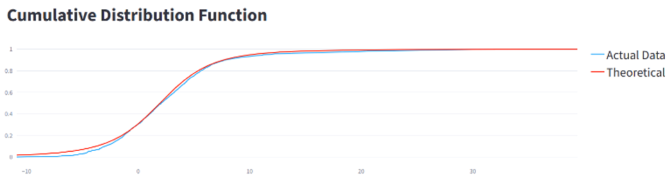

Figure 4. CDF plot of actual SCADA data and estimated

Beyond continuous variables like load and renewable generation, the framework must account for low-probability, high-impact discrete contingencies - specifically, the random forced outages of generator units and transmission branches . In this study, these critical events are mathematically represented using a Bernoulli distribution with two states: in-service or outage. The probability of an outage state is inherently governed by the component's Forced Outage Rate (FOR), a parameter dynamically computed from historical operation and maintenance logs as follows:

$$FOR = \frac{\sum T_{Outage}}{\sum T_{Outage}+\sum T_{In-service}} \tag{3}$$

Where:

- $\sum T_{Outage}$: Total time the device is out of service.
- $\sum T_{In-service}$: Total time the device is in normal operation.

By integrating these empirically derived FORs, the simulation autonomously generates highly realistic N-1 and N-k outage scenarios alongside continuous power fluctuations.

C. Cluster-Based Spatiotemporal Correlation Analysis

Accurately modeling the isolated marginal probability distributions of individual grid components is a necessary but mathematically insufficient condition for rigorous operational risk assessment. In actual grid operation, the outputs of VRE installations - particularly neighboring solar and wind farms - do not fluctuate independently. They exhibit profound spatiotemporal correlations driven by shared large-scale meteorological factors, such as regional cloud cover trajectories, solar irradiance, and wind patterns. Treating these variables as statistically independent can dangerously mask the probability of severe grid emergencies, such as simultaneous power drops across multiple plants leading to critical systemic imbalances.

To algorithmically capture these complex dependencies, the framework integrates a cluster-based correlation layer, functioning as an unsupervised pattern recognition module. In this phase, the algorithm groups loads and renewable generation nodes into distinct clusters based on their geographical coordinates and meteorological profiles. The fundamental heuristic applied is that elements residing within the same spatial cluster exhibit highly correlated behaviors, whereas elements across different clusters are treated as statistically independent. For instance, all wind farms within a specific coastal province can be algorithmically grouped into a unified correlation cluster [11].

To quantify the interdependencies within each cluster, the framework strictly employs the Spearman rank correlation coefficient (r_s) computed directly from the cleansed SCADA time-series data. Traditional power flow approaches frequently rely on the Pearson coefficient; however, Pearson is fundamentally limited to measuring linear correlations and is highly sensitive to non-normal data and outliers. As demonstrated in Section II.B, the empirical telemetry of renewable sources rarely follows a Gaussian distribution and often exhibits complex, non-linear operational relationships. The non-parametric Spearman coefficient overcomes this limitation by operating on the ordinal ranks of the data rather than their absolute numerical values. This ensures exceptional robustness against outliers and enables the algorithm to accurately capture monotonic relationships (i.e., when one variable tends to increase as another does, regardless of linearity). The coefficient between two random variables, X_i and X_j (representing source-source, load-load, or source-load pairs), is mathematically formulated by applying the Pearson equation to their ranked vectors:

$$r_s = \rho_{rg(X_i),rg(X_j)} = \frac{cov(rg(X_i),rg(X_j))}{\sigma_{rg(X_i)}\sigma_{rg(X_j)}} \tag{4}$$

Where:

- X_i and X_j: Time series of random variables
- $rg(X_i)$ and $rg(X_j)$: Ranked data of X_i and X_j
- $cov(rg(X_i), rg(X_j))$: Covariance of ranked values.
- $\sigma_{rg(X_i)}, \sigma_{rg(X_j)}$: Standard deviations of ranked values

Once the Spearman correlation matrices are autonomously computed for every individual cluster, the algorithm aggregates them to construct a comprehensive block-diagonal correlation matrix. This overarching matrix serves as the definitive dependency structure, which is subsequently injected into the high-performance simulation engine to ensure the generated operational scenarios accurately reflect the complex physical realities of the grid.

III. HIGH-PERFORMANCE ALGORITHMIC EXECUTION

Once the selected probabilistic models and spatiotemporal correlation structures are autonomously extracted from the SCADA data mining phase, they must be rigorously evaluated against the physical constraints of the grid. This section details the automated, high-performance computational engine developed to execute the PPF analysis, seamlessly integrating Python-based algorithmic orchestration with the industry-standard PSS/E simulation platform.

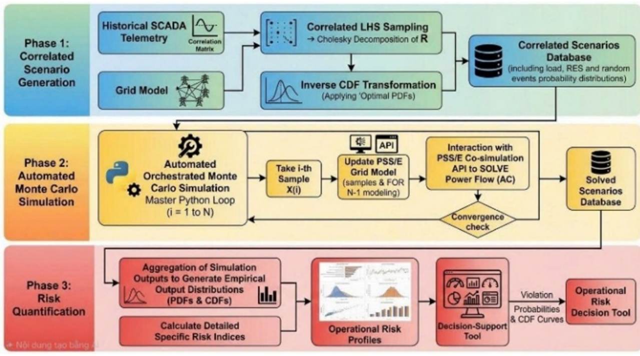

Figure 5. PPF simulation flowchart

A. Enhanced Stratified Sampling and Correlation Injection

The core computational methodology of this framework relies on MCS, selected for its robustness in handling complex, multidimensional uncertainties with non-standard distributions. However, to overcome the heavy computational burden associated with traditional random sampling, this framework implements LHS. By stratifying the cumulative distribution function (CDF) of each random variable into N equal-probability intervals, LHS ensures a highly uniform coverage of the probability space. This stratified approach drastically accelerates the convergence rate of the output statistics compared to standard MCS

To inject the realistic interdependencies discovered in the cluster-based analysis (Section II.C), the algorithm applies Cholesky decomposition [12] to the aggregated Spearman correlation matrix R. The matrix is mathematically factorized into a lower triangular matrix L and its transpose L^T:

$$R = L \times L^T \tag{5}$$

An independent standard normal random vector Z_{ind}, generated via LHS, is then transformed into a correlated normal vector Z_{corr} via matrix multiplication:

$$Z_{corr} = \mathrm{L} \times Z_{ind} \tag{6}$$

Where:

• Z_{ind} : A vector of independent standard normal random variables generated from the LHS procedure.

• L : The lower triangular Cholesky factor of R.

• Z_{corr} : The resulting correlated standard normal vector.

Finally, the actual correlated random samples X_i (representing specific VRE generation and load demands) are synthesized by applying the inverse cumulative distribution function F_i of the AIC-selected marginal models identified in Section II.B:

$$X_i = F_i^{-1}(\Phi(Z_{corr})) \tag{7}$$

Where:

• Φ: CDF of the standard normal distribution

• F_i^{-1}: Inverse CDF of the actual probability distribution (e.g., Student-t, Beta) for element i

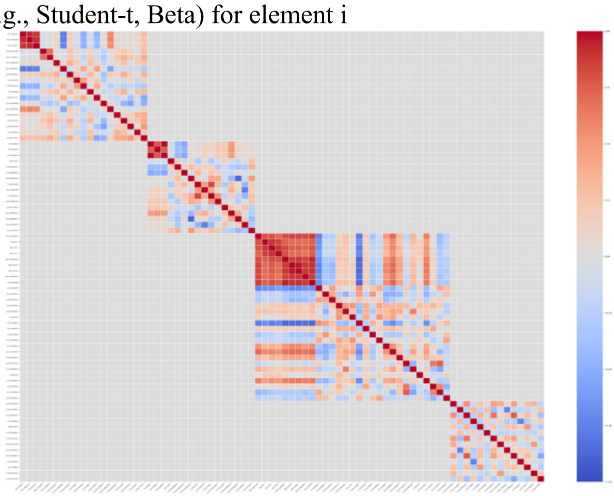

Figure 6. Heatmap of the correlation matrix

B. Python-Driven API Orchestration with PSS/E

Generating and evaluating thousands of complex operating scenarios necessitates highly robust co-simulation architecture. In this framework, a master Python script is deployed as the central orchestrator, executing an automated N-iteration loop.

In each algorithmic iteration, the script dynamically generates a unique, correlated operational scenario encompassing both continuous variables (load, solar, and wind outputs) and discrete stochastic events. For the latter, the operational states of transmission lines and generators are sampled based on their respective Forced Outage Rates (FOR) to simulate random N-1 or N-k contingencies.

Utilizing the psspy Application Programming Interface (API), the Python orchestrator automatically injects these sampled parameters into the PSS/E grid model, overwriting the deterministic baseline. The AC power flow solver is subsequently triggered (e.g., using the psspy.fnsl() function) [13], [14]. Upon successful convergence of the network equations, the Python engine extracts and logs critical system states, including bus voltages and branch active/reactive power flows, into a structured database.

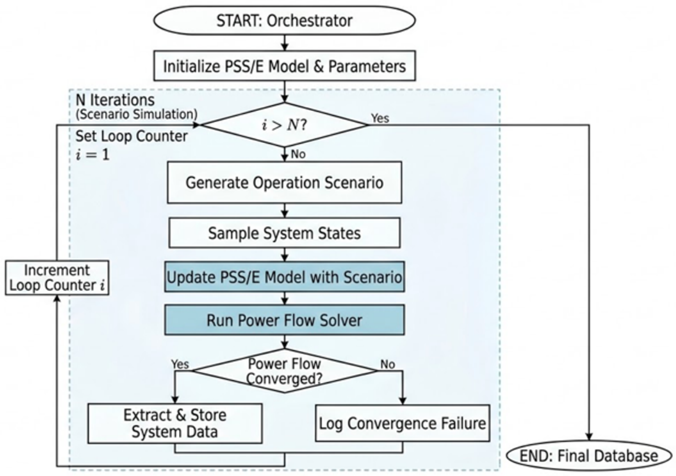

Figure 7. Flowchart of the Automated Co-Simulation Engine

C. Algorithmic Risk Quantification

Upon the completion of the automated co-simulation loop, the framework yields a massive dataset representing *N* statistically probable operational states of the power grid. This raw output is algorithmically aggregated to construct empirical PDFs and cumulative distribution functions (CDFs) for heavily monitored grid variables.

By directly mining these output distributions, system operators can effortlessly quantify precise operational risk indices. For instance, the probability of actual thermal-limit violations, such as P(Loading > 100%), or voltage-limit deviations, such as P(Vbus < Vmin) or P(Vbus > Vmax), can be quantified. Importantly, the extensive dataset generated by this automated PPF framework serves as an ideal, high-quality foundation for training future predictive Machine Learning models dedicated to real-time grid security assessment

IV. CASE STUDY

Before presenting the probabilistic risk results, the main simulation settings are summarized in Table IV to improve reproducibility. The PPF simulation is executed using a Python-based automation script coupled with the PSS/E AC power-flow solver. Each LHS sample corresponds to one stochastic operating scenario, including uncertain load demand, renewable generation outputs, and discrete outage states where applicable.

To empirically validate the robustness and practical utility of the data-driven framework detailed in Sections II and III, the automated pipeline was deployed to evaluate the Central Vietnam power system. Specifically, the algorithmic framework autonomously extracted, modeled, and simulated two extreme operational scenarios directly from the historical SCADA repository: the midday off-peak condition during the dry season and the evening peak condition during the rainy season. By subjecting the grid model to these highly stochastic and vulnerable states, the following subsections demonstrate the framework's capability to accurately quantify hidden operational risks that are typically obscured in traditional deterministic assessments.

TABLE IV. SIMULATION SETTINGS FOR THE PROPOSED PPF FRAMEWORK

Item	Value
Studied system	Central Vietnam power system
Power-flow solver	PSS/E AC power flow
PSS/E version	36.1
Number of buses	88
Number of branches	134
Number of renewable plants	94
Renewable penetration in studied scenario	46.98%
Number of LHS/MCS samples	6000
SCADA sampling interval	5 minutes
Power-flow convergence criterion	Largest mismatch < 0.5MW
Number of non-convergent cases	4
Total computation time	141.9 sec
Hardware configuration	CPU: Intel Core i7-10700 2.90GHz (16 core) RAM: 24 GB OS: Windows 11 Pro 64 bit

A. Midday Off-Peak Scenario – Dry Season (Branch Loading Analysis)

The analyzed scenario corresponds to the midday off-peak operating condition in the dry season. This scenario is characterized by low system demand while solar power generation reaches its maximum output. It represents a typical operating condition of the Central Vietnam power system, where high penetration of photovoltaic (PV) generation significantly influences power flow distribution across the network.

The element under investigation is the 220kV Quy Nhon – Tuy Hoa transmission line, which plays a critical role in both supplying power to the eastern regions of Gia Lai and Dak Lak provinces and evacuating renewable energy, particularly solar power, in the region.

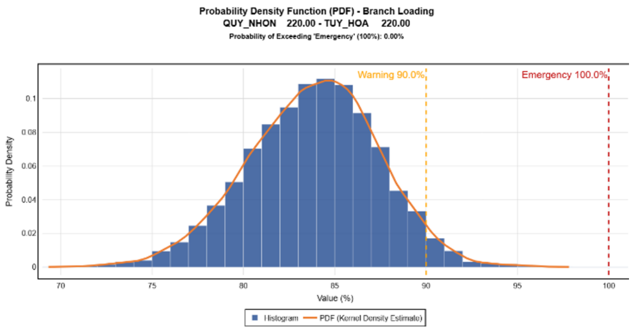

Figure 8. PDF of branch loading for the 220 kV Quy Nhon – Tuy Hoa line

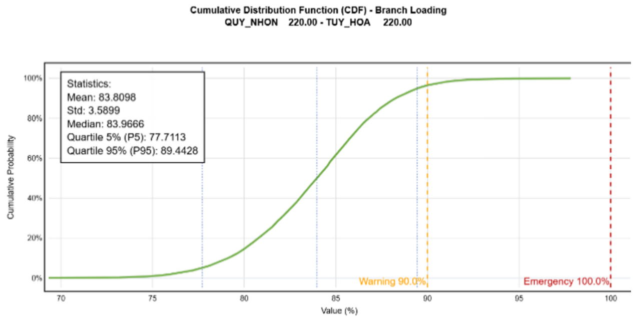

Figure 9. CDF of branch loading for the 220 kV Quy Nhon – Tuy Hoa line

The analysis of the PDF and CDF curves reveals the following:

• Operating trend: The peak of the PDF is concentrated in the range of 80%–85%, indicating the most probable operating condition.

• Variability: The loading distribution spans from approximately 70% to nearly 98%, reflecting significant variability.

• Asymmetry: The distribution exhibits a slight right-skewed behavior.

• Risk assessment: The probability of exceeding the 90% branch-loading warning threshold is approximately 4.57%, whereas the probability of exceeding the 100% thermal loading limit is nearly zero.

This result indicates that the 220 kV Quy Nhon–Tuy Hoa line does not experience actual thermal overload in the simulated scenarios. However, it operates close to the warning threshold with a non-negligible probability. Such near-limit operation represents a latent operational risk because additional disturbances, renewable forecast errors, or topology changes may push the line closer to its emergency limit. This type of probabilistic warning information cannot be obtained from a single deterministic power-flow snapshot

B. Evening Peak Scenario – Rainy Season (Bus Voltage Analysis)

The evening peak scenario represents the period of highest system demand during the day. At this time, solar generation is negligible, and the system is primarily supplied by hydropower, thermal power plants, and wind generation. During the rainy season, system operation is strongly influenced by weather conditions. Solar generation becomes highly variable due to cloud cover and rainfall, while hydropower and wind generation fluctuate depending on environmental conditions.

The analyzed node is the 220kV bus at the Da Nang 500 kV Substation, which is a key node in the Central Vietnam transmission network. The bus voltage is represented using the PDF and the CDF.

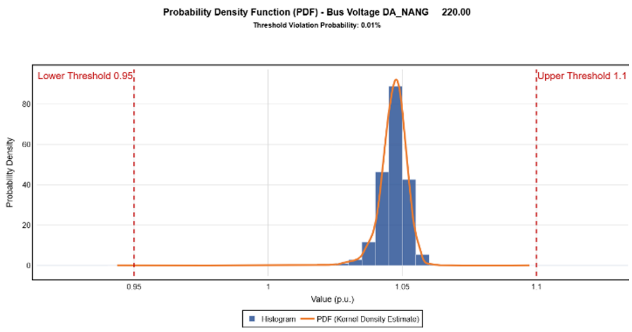

Figure 10. PDF of bus voltage at the 220kV bus of the Da Nang 500 kV Substation

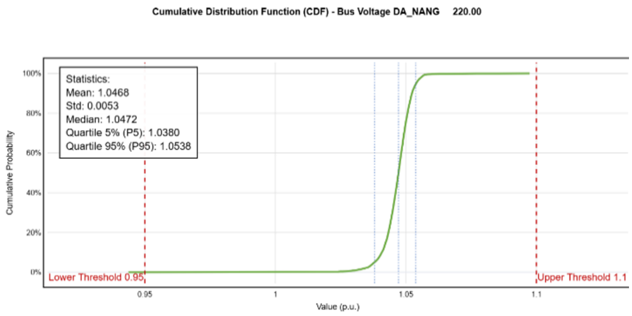

Figure 11. CDF of bus voltage at the 220kV bus of the Da Nang 500 kV Substation

The analysis of the PDF and CDF curves reveals the following:

• Typical operating value: Mean voltage is approximately 1.0468 pu, concentrated in 1.04–1.055 pu.

• Variability: Small standard deviation (~0.0053 pu) indicates stable voltage.

• Characteristic range: Voltage lies within 1.038pu to 1.054 pu (P5–P95).

• Risk assessment: Probability of violating limits (0.95–1.1 pu) is about 0.01%, very low but should still be monitored under adverse conditions.

The voltage distribution shows that the 220 kV bus at the Da Nang 500 kV substation remains within the acceptable operating range in almost all simulated scenarios. The mean voltage is approximately 1.0468 pu, with a small standard deviation of about 0.0053 pu. The P5–P95 interval ranges from approximately 1.038 pu to 1.054 pu, indicating stable voltage behavior during the rainy-season evening peak condition. The probability of violating the voltage limits of 0.95–1.10 pu is approximately 0.01%, which is very low. Therefore, this case does not indicate a significant voltage-security problem under the studied operating condition, although continued monitoring remains necessary under adverse system conditions.

C. Comparative Analysis with Baseline Methods

To further evaluate the benefit of the proposed SCADA-driven correlated PPF framework, two baseline approaches are considered for comparison.

The first baseline is DPF, in which the system is evaluated using a single representative operating snapshot. This approach provides only one operating point and therefore cannot quantify the probability of near-limit or limit-violation events.

The second baseline is an independent PPF model, in which the marginal probability distributions are fitted from SCADA data, but the uncertainty variables are sampled independently without considering the empirical correlation structure among renewable sources. This baseline is used to isolate the impact of the proposed cluster-based Spearman correlation modeling.

The proposed method differs from both baselines by combining SCADA-fitted marginal distributions with correlation-preserving scenario generation. The comparison is performed using the same monitored quantities as in the previous subsections: the loading of the 220 kV Quy Nhon–Tuy Hoa line and the voltage magnitude at the 220 kV bus of the Da Nang 500 kV substation.

TABLE V. COMPARISON BETWEEN THE PROPOSED METHOD AND BASELINE APPROACHES

Method	**Input model**	**Correlation considered**
DPF	Single operating snapshot	No
Independent PPF	SCADA-fitted marginal PDFs	No
Proposed correlated PPF	SCADA-fitted marginal PDFs	Yes

TABLE VI. QUANTITATIVE COMPARISON OF RISK INDICES UNDER DIFFERENT METHODS

Case / monitored variable	**DPF**	**Independent PPF**	**Proposed correlated PPF**
CASE A: Midday Off-Peak Scenario – Dry Season (Branch Loading Analysis)			
Quy Nhon–Tuy Hoa loading, mean (%)	97	83.79	83.81
Quy Nhon–Tuy Hoa loading, std (%)	N/A	3.9984	3.5899
P(Loading > 90%)	N/A	5.7%	4.57%
P(Loading > 100%)	N/A	Nearly 0%	Nearly 0%
CASE B: Evening Peak Scenario – Rainy Season (Bus Voltage Analysis)			
Da Nang 220 kV voltage, mean (pu)	1.0440	1.0469	1.0468
Da Nang 220 kV voltage, std (pu)	N/A	0.0079	0.0053
P(Voltage violation)	N/A	0.03%	0.01%

The DPF value corresponds to the selected deterministic operating snapshot used by system operators for the studied scenario, whereas the PPF results represent statistical indices aggregated over 6000 stochastic scenarios.

V. CONCLUSIONS

This paper has successfully developed and demonstrated a comprehensive, fully automated data-driven framework for operational risk assessment in renewable-rich power systems. By systematically mining high-resolution SCADA telemetry, the proposed methodology substantially reduces the reliance on rigid, pre-assumed probability distributions that constrain traditional PPF studies. The integration of unsupervised statistical learning enables the stochastic behavior of monitored grid elements to be modeled directly from SCADA data, while the cluster-based Spearman correlation layer captures the non-

https://doi.org/10.1051/e3sconf/202672303006

linear spatiotemporal dependencies inherent in regional renewable generation.

The deployment of a Python-orchestrated co-simulation engine, utilizing Correlated LHS seamlessly integrated with PSS/E, proved highly efficient for evaluating thousands of stochastic operational states. As evidenced by the simulated case studies, the framework successfully quantified hidden operational vulnerabilities that conventional deterministic boundary assessments structurally fail to capture - specifically, the non-negligible probability of operating near thermal loading warning thresholds during the dry season midday off-peak, and the low-probability voltage deviations during the rainy season evening peak. By accurately reflecting the extreme volatility and concurrent fluctuations of renewable sources, the model provides a much more realistic operational risk profile.

Ultimately, this data-driven approach translates abstract multidimensional uncertainties into precise, quantifiable risk indices, empowering system operators with a robust decision-support tool for preemptive grid management. Furthermore, the massive datasets of correlated, physically validated operational scenarios generated by this automated framework establish an ideal, high-fidelity data foundation for training future Artificial Intelligence and Deep Learning models dedicated to real-time grid security prediction and automated dispatch optimization.

References

[1] D. D. Le, V. D. Ngo, T. A. N. Nguyen, and V. K. Huynh, “Study on uncertainty of operating parameters to calculate and analyze working condition of power systems", *Journal of Science and Technology - The University of Danang*, vol. 5, no. 90, pp. 34-38, May 2015.

[2] P. Chen, Z. Chen, and B. Bak-Jensen, "Probabilistic load flow: A review," 2008 Third International Conference on Electric Utility Deregulation and Restructuring and Power Technologies, Nanjing, China, 2008, pp. 1586-1591.

[3] D. D. Le, N. T. A. Nguyen, V. D. Ngo, and A. Berizzi, "Advanced probabilistic power flow methodology for power systems with renewable resources," Turkish Journal of Electrical Engineering and Computer Sciences, vol. 25, no. 2, pp. 1154-1162, 2017.

[4] A. İskenderoğlu and Ö. Gül, “Probabilistic Power Flow Analysis of Power Systems with PV and Wind Using Monte Carlo Method”, in *Proc. 10th Int. Conf. Electr. Electron. Eng. (ICEEE)*, 2023.

[5] Z. Ren, K. Wang, W. Li, L. Jin, and Y. Dai, "Probabilistic Power Flow Analysis of Power Systems Incorporating Tidal Current Generation", *IEEE Trans. Sustain. Energy*, vol. 8, no. 3, pp. 1195-1203, Jul. 2017.

[6] D. S. Kumar, H. Quan, K. Y. Wen, and D. Srinivasan, "Probabilistic risk and severity analysis of power systems with high penetration of photovoltaics," *Solar Energy*, vol. 230, pp. 1156-1164, Dec. 2021.

[7] R. Y. Rubinstein and D. P. Kroese, *Simulation and the Monte Carlo Method*, 2nd Ed. Wiley, Hoboken, NJ, USA, 2008.

[8] J. C. Helton and F. J. Davis, “Latin hypercube sampling and the propagation of uncertainty in analyses of complex systems”, *Reliability Engineering & System Safety*, vol. 81, no. 1, pp. 23-69, Jul. 2003.

[9] V. W. Berger and Y. Zhou, “Kolmogorov–Smirnov Test: Overview”, in *Encyclopedia of Statistics in Behavioral Science*. Wiley, 2014.

[10] J. E. Cavanaugh, “The Akaike information criterion: Background, derivation, properties, application, interpretation, and refinements”, *Wiley Interdiscip. Rev. Comput. Stat.*, vol. 11, no. 3, Art. no. e1460, May/Jun. 2019.

[11] J. Zang, B. M. Hodge, and A. Florita, “Investigating the Correlation Between Wind and Solar Power Forecast Errors in the Western Interconnection”, in *Proc. ASME 7th Int. Conf. Energy Sustain. and 11th Fuel Cell Sci., Eng., Technol. Conf.*, USA, Jul. 2013.

[12] L. Li, Y. Fan, X. Su, and G. Qiu, “Probabilistic Load Flow Calculation of Power System Integrated with Wind Farm Based on Kriging Model”, Energy Eng., vol. 118, no. 2, pp. 565-580, 2021.

[13] Siemens Power Technologies International, *Program Operation Manual PSS®E 36.1.0*. USA: Siemens PTI, Jul. 2024.

[14] Siemens Power Technologies International, Application Program Interface PSS®E 36.1.0. USA: Siemens PTI, Jul. 2024.