Structured Probabilistic Inference for Power-System Contingency Risk: A Comparison with Monte Carlo Simulation

Ricardo Marquez, Ph.D. rmarquez123@gmail.com

***Abstract*—Probabilistic contingency analysis must answer two different questions: *how much risk exists*, and *which contingencies create that risk*. Monte Carlo simulation answers the first well: sampling error shrinks with sample size independent of system dimension. But an individually rare contingency may simply never appear in a finite sample, so the second question—naming the specific outage combinations that carry the risk—is structurally mismatched to sampling. This study evaluates probability-ordered structured inference, implemented in TERRA, as an alternative way of exploring contingency space. On the RTS-GMLC Area 1 system with published two-state reliability parameters, and with both approaches sharing identical reliability data, operating conditions, and an independently cross-validated physical evaluator, we compare the methods on four questions: accuracy, rare-contingency coverage, evidence-conditioned updating, and scaling. TERRA reproduces the conventional estimates wherever both are defined; at the tested scale it deterministically covers a high-consequence watchlist of which matched-budget Monte Carlo misses roughly 83% per run, and it updates risk under new weather evidence by reweighting cached states. The advantage is not universal: as the component space grows, explicit enumeration loses the aggregate-estimation race to sampling, whose precision per draw does not degrade with contingency-space dimension. Structured probability-directed exploration is a strong complement to—not a replacement for—Monte Carlo.**

***Index Terms*—Contingency analysis, power system reliability, Monte Carlo methods, probabilistic methods, rare events, RTS-GMLC.**

I. Introduction

A. Motivation

Operators and planners ask probabilistic contingency analysis for two deliverables. The first is aggregate: the probability of any overload, of load shedding, or the expected severity, under component-availability uncertainty. The second is enumerative: a defensible list of the *specific* outage combinations—single and multiple—that create the risk, so they can be watched, screened, or mitigated. Conventional practice [1], [2] couples a two-state component reliability model with Monte-Carlo (MC) sampling and a power-system solver, and serves the first deliverable excellently. The second is a different kind of problem. A contingency of probability p appears at least once in n independent draws with probability

$$\mathrm{P}(\text{visit}) \;=\; 1-(1-p)^n \;\approx\; np \quad (np \ll 1), \qquad (1)$$

so a $p = 10^{-6}$ contingency has a $\sim$2% chance of appearing in a 20,000-draw study: *a contingency can be known by the model to exist, be consequential, and still have a 98% probability of never appearing in the sample*. For contingency identification, industry practice commonly relies on deterministic $N{-}1/N{-}2$ screening lists [3], [4], which carry no probability accounting at all.

Companion to the audit dossier (`report.html`, `comparison_report.html`); every number in this paper is read from the study's committed, test-pinned result artifacts. TERRA = TERRA-watt. Wall-clock timings were measured on a single workstation and vary by machine; labeled extrapolations assume a unit evaluation cost, while evaluation counts are measured.

B. Research gap

Power-system reliability assessment already includes Monte-Carlo simulation [2], probabilistic state enumeration [5], [6], contingency screening [3], [4], rare-event methods [7], [8], probabilistic $N-k$ identification [9], and Bayesian reliability models [10], [11]. These literatures establish that sampling and structured state-space exploration offer complementary advantages [5]. Rather than proposing probability-directed contingency analysis as a new category, this study investigates how a structured probabilistic inference engine represents and evaluates the same contingency-risk problem within one reusable framework: exploring the contingency space *in probability order*, so that every state above a known probability cutoff is examined exactly once and the total probability of everything not examined is known. This is a structured-inference view of the outage space in the sense of probabilistic graphical models [12]: the joint outage distribution has exploitable structure (independence, or conditional independence given a common cause), and inference can be organized around that structure instead of around random draws.

C. Contributions

(i) An exact-replication validation architecture in which one physical evaluator runs both methodologies and is cross-validated per-scenario against an independent Julia implementation, separating the *validation question* (are both implementations solving the same physics?) from the *research question* (given identical physics, which exploration strategy serves which workload?). (ii) A four-question empirical comparison (accuracy, rare-contingency coverage, evidence-conditioned updating, scaling) on a recognized test system [13], [14] with published reliability data. (iii) An honest crossover characterization, including the regime where Monte Carlo decisively wins.

II. Related Work

A. Probabilistic contingency and composite reliability analysis

Composite-system reliability evaluation (hierarchical level II, HL-II: generation and transmission assessed together) computes load-loss and violation probabilities over joint generator/branch outage spaces; the standard treatments are [1], [2], and the IEEE Reliability Test System family [14], [13] was created largely to benchmark such methods. State enumeration and Monte-Carlo sampling are the two classical exploration strategies whose complementary roles are established practice — Rei and Schilling's assessment of the Brazilian interconnected system explicitly combines them [5]. They are the two strategies treated [2]; this study takes that complementarity as its starting point and evaluates one structured implementation of probability-*ordered* enumeration (with an exact residual bound) head-to-head against sampling on identical physics — claiming empirical characterization, not algorithmic novelty.

B. Probabilistic N–k selection and critical-state identification

Closer to this study's identification workload, Sundar et al. formulate probabilistic $N-k$ failure identification as an optimization seeking outage combinations that maximize disruption weighted by their simultaneous failure probability, demonstrated from IEEE-14 through PEGASE-1354 and a 2,383-bus system [9] — conceptually, contingency probability $\times$ physical consequence, the same quantity our watchlist thresholds. Structured enumeration methods reduce the number of expensive evaluations: impact-increment formulations decouple reliability indices and prune higher-order states [6], and recent lattice-partition methods identify all critical states up to a preset order while deriving converging bounds on loss-of-load probability [15] — work that, like this study, combines named-state coverage with probability bounds. Probability ordering is therefore established prior art; what this study evaluates is its combination, in one engine, with structured dependencies, shared exact physical evaluation, posterior conditioning, and reusable inference.

C. Monte Carlo and rare-event methods

Variance-reduction and rare-event techniques—importance sampling, cross-entropy methods, and subset simulation [7], [8]—attack precisely the tail-visitation weakness of Eq. (1). These methods primarily target efficient estimation of rare-event probabilities rather than deterministic coverage guarantees over individually named contingency states. We use vanilla MC as the baseline here because it is the conventional workflow's estimator; benchmarking against rare-event samplers is identified as necessary future work in Section VIII.

D. Contingency screening and $N-k$ selection

Deterministic screening ranks contingencies by severity indices [3]; multiple-element screening extends this to $N-k$ [4]. These methods answer the identification question but carry no probability accounting: a list without a statement of what probability mass it excludes. The enumeration studied here performs screening over the mixed generator/branch $N-k$ lattice ordered by likelihood with excluded-mass accounting — continuing the probability-directed line of Section II-B rather than founding it.

E. Bayesian and graphical models for power-system reliability

Probabilistic graphical modeling of power-system reliability is established prior art: Bayesian networks have been constructed for composite-system reliability assessment and component-importance ranking on the IEEE RTS [10], applied to loss-of-load analysis with explicit attention to rare events on the Roy Billinton Test System (RBTS) and the 24-bus IEEE RTS [11], and extended to weather- and load-conditioned reliability [16]. The engine under test does not introduce graphical inference to this domain. Graphical-model inference [12] exploits distributional structure—here, independence of component failures and conditional independence given a weather common cause [17]—to organize computation. The engine under test (TERRA) applies this at the orchestration level: closed-form product measures, best-first traversal, and posterior reweighting of cached physical evaluations.

III. Problem Formulation

A. Component reliability

Each component i (transmission branch or thermal generator) is a two-state variable $X_i \in \{\text{Available}, \text{Failed}\}$ with failure rate λ_i and repair rate μ_i; its steady-state forced-outage probability is

$$q_i \;=\; \frac{\lambda_i}{\lambda_i+\mu_i} \;=\; \frac{\lambda_i\,\mathrm{MTTR}_i}{8760+\lambda_i\,\mathrm{MTTR}_i}, \tag{2}$$

where λ_i is the failure rate in occurrences per year and MTTR_i is the mean time to repair in hours (so $\mu_i = 8760/\mathrm{MTTR}_i$ per year) [1]. For generating units the same quantity is conventionally tabulated directly as the forced outage rate, $\mathrm{FOR}_i = q_i$. Under independence the outage state S (the set of failed components) has the product measure

$$\mathrm{P}(S) \;=\; \prod_{i\in S} q_i \prod_{i\notin S}(1-q_i). \tag{3}$$

B. Contingency risk

Given an operating point, a physical evaluator maps each outage state to consequences (overloads, islanding, load shed); risk metrics are expectations over Eq. (3): $\mathrm{P}(\text{any overload})$, $\mathrm{P}(\text{shed} > 0)$, $\mathrm{E}[\text{severity}]$, $\mathrm{E}[\text{shed}]$.

C. Aggregate estimation vs. contingency identification

The two deliverables of Section I map to two uses of the same measure. *Monte Carlo* draws $S^{(1)},\dots,S^{(n)} \sim \mathrm{P}(S)$ and estimates any aggregate $\mathrm{P}(A) = \mathrm{E}[\mathbf{1}_A]$ with $n^{-1/2}$ error in the number of draws, independent of contingency-space dimension (although each physical evaluation may itself grow more expensive with network size) — *scale-free sampling precision*, in the sense used throughout this paper—but the same high-probability states are sampled again and again. *Probability-ordered enumeration* generates the distinct

sequence $S_{(1)}, S_{(2)}, \dots$ with $\mathrm{P}(S_{(1)}) \geq \mathrm{P}(S_{(2)}) \geq \cdots$, evaluating each state once. Truncating after the enumerated set $\mathcal{E}$ leaves the residual

$$\varepsilon \;=\; 1 - \sum_{S \in \mathcal{E}} \mathrm{P}(S) \tag{4}$$

known exactly. Let $\hat{P}_A$ be the probability of an event A summed over $\mathcal{E}$ only, and $\hat{C}$ the corresponding partial expectation of a non-negative consequence C bounded by $C_{\max}$; then every event and bounded consequence acquires a deterministic interval

$$\mathrm{P}(A) \in \big[\hat{P}_A,\, \hat{P}_A + \varepsilon\big], \quad \mathrm{E}[C] \in \big[\hat{C},\, \hat{C} + \varepsilon C_{\max}\big]. \tag{5}$$

Figure 1 shows both strategies on one axis, and it frames the central hypothesis of this study: *Monte Carlo should perform well for aggregate probability estimation, while probability-ordered inference may perform better when the objective is systematic coverage of individually rare contingencies.* The rest of the paper tests exactly this.

IV. Methods

A. Test system

We use RTS-GMLC Area 1 [13] (24 buses, 38 branches including five transformers, 52 generators; hydro excluded from dispatch per the established reference convention), shown in Fig. 2. The reliability model has $N = 62$ components: 38 branches with q from Eq. (2) using the published outage-rate and repair-duration columns (transformer repairs $\sim$768 h dominate branch unavailability; transient rates excluded) and 33 thermal units with q_i taken directly from the published forced outage rate. Here $\sum_i q_i = 1.148$ and $\mathrm{P}(\varnothing) = 0.304$. Three committed operating snapshots anchor the analysis: *peak* (hour 4934, 2621 MW), *median*, and *low*. At peak the N-0 economic dispatch (the intact-network, no-contingency operating point) leaves branch A11 at exactly its continuous rating—a deliberately tight operating point, which is why substantial post-contingency overload probability exists there at all; off-peak snapshots carry 110+ MW security margins and $\sim 10^{-5}$ overload mass.

B. Physical contingency evaluator

Each sampled or enumerated contingency is evaluated by an identical three-stage physical model: (1) generation is rebalanced after generator outages by a governor-response proxy; (2) post-contingency flows are computed with DC network physics, exactly, for simultaneous multi-branch outages; (3) states producing overloads, islanding, or generation deficits are passed to a corrective DC optimal power flow (DC-OPF) with load-shedding slack. We refer to stages (1)–(2) as Tier 1 (the security screen, no redispatch) and stage (3) as Tier 2 (corrective redispatch). Fig. 3 summarizes the pipeline; it is identical for both methods and cached by outage state, so evaluation counts compare one-to-one.

Formally, for outage state $X = (B, G)$ with p_g the pre-contingency dispatch of unit g and $\bar{p}_g$ its maximum output: lost generation $L = \sum_{g \in G} p_g$ is redistributed jointly over available thermal units in proportion to headroom $h_g = \bar{p}_g - p_g$,

$$\Delta p_g \;=\; L\, h_g \big/ \textstyle\sum_{g'} h_{g'}, \tag{6}$$

which respects ratings by construction whenever $L \leq \sum h$ (any residual is a deficit and triggers stage 3). Post-contingency flows follow the generalized (multiple) line-outage distribution-factor identity, MLODF [18]. Let Φ be the power transfer distribution factor (PTDF) matrix [19], whose entry $\Phi_{i,b}$ is the change in flow on branch i per unit injection at bus b (withdrawn at the slack bus), let p' be the rebalanced bus-injection vector, and let $\Psi_{ij} = \Phi_{i,\mathrm{from}(j)} - \Phi_{i,\mathrm{to}(j)}$ be the sensitivity of branch i to a unit transfer across branch j; then

$$f^{\mathrm{post}} \;=\; f' + \Psi_{:,B}\left(I - \Psi_{BB}\right)^{-1} f'_B, \qquad f' = \Phi\, p', \tag{7}$$

where f' is the pre-outage flow vector under the rebalanced injections and subscript B selects the outaged branches; the identity is exact within the DC model whenever B is not a cutset. Islanding is classified by graph connectivity, never numerical singularity. The Tier-2 linear program (LP) uses per-bus value-of-lost-load (VOLL) shedding prices (strictly increasing by bus id, making the shed allocation unique) and solves islands as reduced networks; the LP solver is HiGHS [20].

C. Conventional Monte Carlo method

The conventional workload draws each component's availability independently per scenario (committed design: 20,000 draws/snapshot, seed 42; local reference tier: 10^6 draws), evaluates each draw through the pipeline, and reports aggregates with Wilson score intervals [21] for rare-event probabilities and seeded-bootstrap intervals for expectations.

D. TERRA structured inference

When every $q_i < \frac{1}{2}$, adding a component to an outage set strictly decreases its probability, so best-first search over the subset lattice generates states in globally non-increasing probability. TERRA enumerates to a target mass (committed tier: $1 - \varepsilon = 0.99$, i.e. 11,336 states/snapshot; deep tier: 0.999), evaluates each state once through the same pipeline, and reports every metric with the certified intervals of Eq. (5).

E. Validation architecture

Two questions must be separated. The *validation question*: are both implementations solving the same physical problem? The *research question*: once physics is held constant, how do the exploration strategies differ? Fig. 4 shows the resulting two-pillar design. All inputs are committed contract files (reliability data, MC draws, N-0 dispatch, shed prices, renewable caps). For Pillar 1, TERRA's physics surface *runs the conventional MC workload itself* on those draws, and is compared, scenario by scenario, against an independently implemented Julia/JuMP+HiGHS optimization reference (built in the Sienna reference environment and consuming its data conventions; not an off-the-shelf Sienna/PowerSimulations workflow). The result is an exact match on all 60,000 committed scenarios:

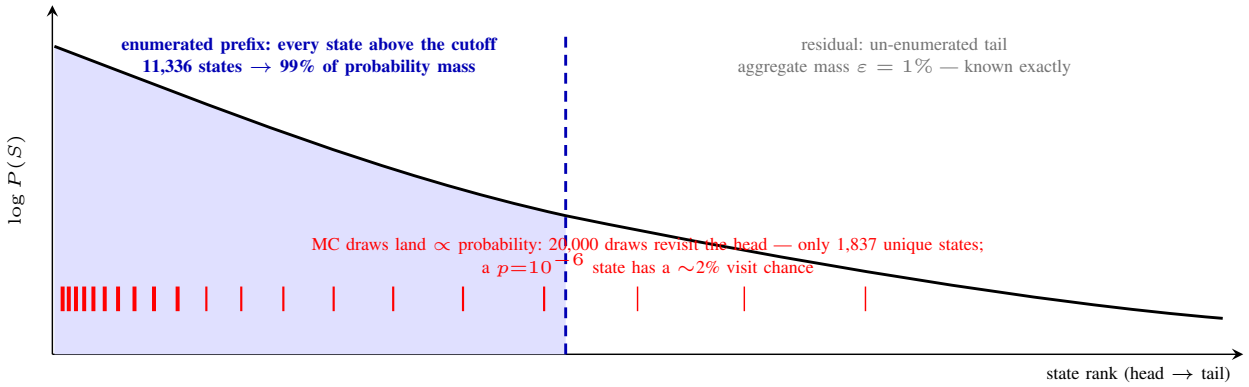


Fig. 1. The central distinction (schematic probability decay; annotations are measured values from the peak snapshot). Sampling effort lands in proportion to probability and revisits the head of the distribution; probability-ordered enumeration covers every state above its cutoff and knows the excluded mass exactly.

TABLE I
THE FOUR RESEARCH QUESTIONS.

RQ	Question	Comparison / metric
RQ1 Accuracy	Does TERRA reproduce conventional risk estimates?	MC (20k, 10^6) vs. TERRA (ε=$10^{-2}, 10^{-3}$); probability agreement
RQ2 Coverage	Which method identifies rare consequential contingencies?	watchlist coverage and unique evaluations at matched budget
RQ3 Updating	How efficiently can risk be updated as evidence changes?	cached TERRA reweighting vs. fresh conditional MC; evaluations + time
RQ4 Scaling	How do the methods behave as the state space grows?	62 vs. 124 components; precision at fixed budget

identical outcome classifications and counts—hence identical probabilities, confidence intervals, and metric scope—with Tier-1 flows within 10^{-9} MW (Appendix A). Only after that proof does Pillar 2 compare workload economics between sampling and the enumeration mode.

V. Experimental Design

Table I is the roadmap: four research questions (RQ1–RQ4), each with its own experiment on the shared model, identical physics throughout.

RQ1 criterion: per metric, $|\,\mathrm{mid}_{\mathrm{TERRA}} - \hat{p}_{\mathrm{MC}}| \le \max(4\,\mathrm{SE} + \varepsilon,\ \mathrm{atol})$, where $\mathrm{mid}_{\mathrm{TERRA}}$ is the midpoint of the certified interval, $\hat{p}_{\mathrm{MC}}$ the MC point estimate, SE its standard error, and atol a metric-family absolute floor (10^{-6} for probabilities, 0.01 MW for MW expectations, \$10 for cost); a stricter 2σ diagnostic and interval-overlap checks are also reported. *RQ2 design:* a watchlist of individually significant, high-consequence states (Tier-1 severity $\ge$ 25 MW *and* $\mathrm{P}(S) \ge 10^{-6}$) built from a deep enumeration at $\varepsilon' = \varepsilon/100$; coverage measured for TERRA at its committed budget and for 20 independent 20,000-draw MC seeds. *RQ3 design:* weather common cause $W \in \{\mathrm{calm}, \mathrm{windstorm}, \mathrm{severe}\}$, prior $(0.94, 0.05, 0.01)$, multiplying exposed 138 kV-line probabilities by $(1, 10, 50)$; measure per-update cost along $W = \mathrm{calm} \to \mathrm{windstorm} \to \mathrm{severe}$, including cold starts, for both methods. *RQ4 design:* replicate

TABLE II
RQ1, PEAK SNAPSHOT: MC POINT ESTIMATES VS. TERRA CERTIFIED INTERVALS. OFF-PEAK SNAPSHOTS AGREE IDENTICALLY (PROBABILITIES AT THE 10^{-5} LEVEL OR EXACT ZEROS).

Metric	MC_{20k}	MC_{10^6}	TERRA ε=10^{-2}	TERRA ε=10^{-3}
P(any overload)	0.4064	0.4054	[0.3970, 0.4070]	[0.4041, 0.4051]
P(shed > 0)	0.1881	0.1868	[0.1811, 0.1911]	[0.1862, 0.1872]
P(islanding)	0.00055	0.00033	[0.00025, 0.01025]	[0.00033, 0.00133]

Area 1 ×2 (124 components, symmetric tie lines); compare achieved certified width at a fixed 10^5-state budget against the Wilson width at the committed 20k-draw design.

VI. Results

A. RQ1: accuracy

TERRA reproduces the conventional solution at two levels: exactly at the workload level (Pillar 1, Section IV-E), and consistently at the inference level—all 18 metric×snapshot comparisons pass the $4\sigma+\varepsilon$ criterion and the stricter 2σ diagnostic. Table II and Fig. 5 summarize the peak snapshot.

One caution in reading Table II: *the two interval types mean different things*. The MC interval quantifies sampling uncertainty—what the sample happened not to contain. The TERRA interval quantifies the exact probability mass deliberately not yet enumerated, Eq. (4). One is statistical and would change with a different seed; the other is deterministic and would not. Contingency rankings over the mutually resolvable branch set agree (top-5 overlap 5/5, Spearman $\rho = 0.90$); TERRA additionally bounds 9 branch probabilities at the 10^{-7}–10^{-8} level where even the 10^6-draw MC records exact zeros.

B. RQ2: rare-contingency coverage

The methods diverge when the objective shifts from estimating the aggregate to naming the states (Table III, Fig. 6). The watchlist contains 5,028 states carrying 23.2% of the total probability mass. TERRA covers all of them deterministically—every state above its enumeration cutoff is covered by construction, with zero missed enumerated

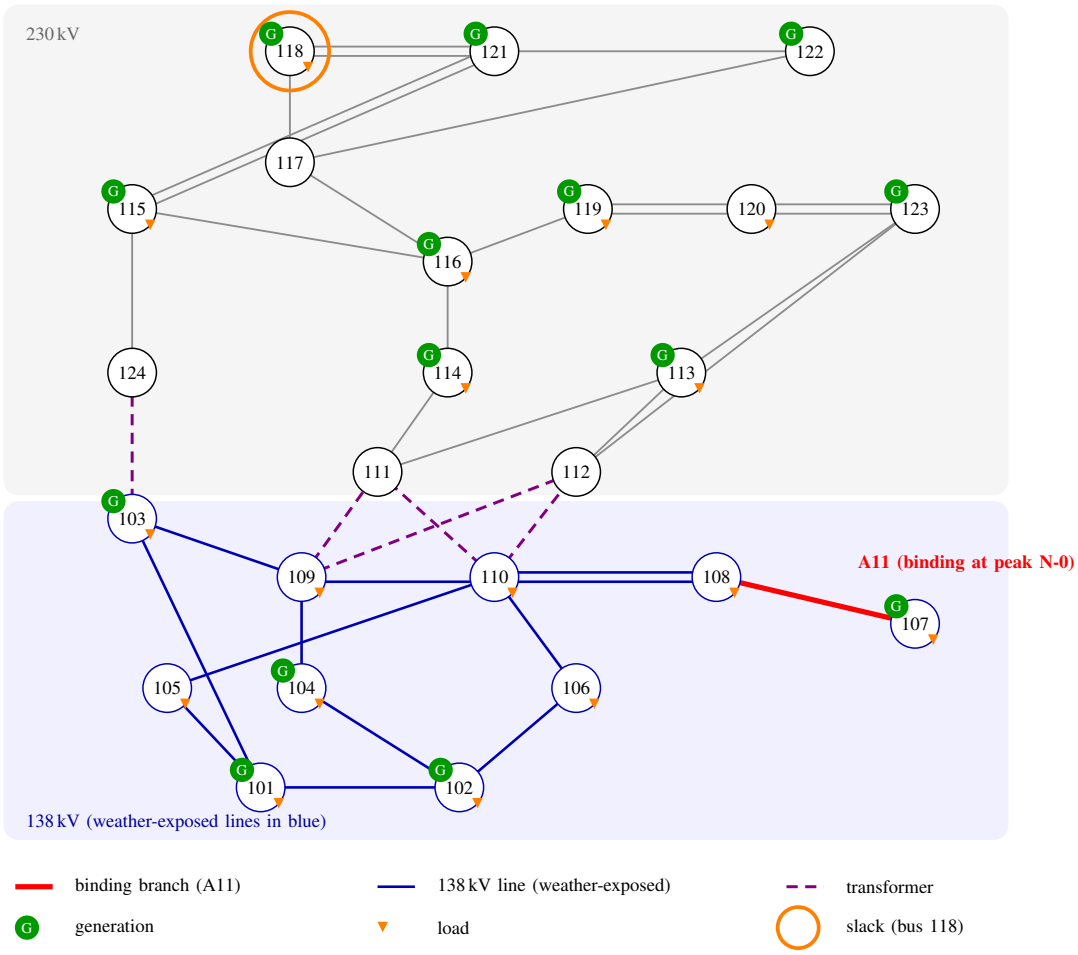


Fig. 2. The study system (RTS-GMLC Area 1; positions follow the classic IEEE RTS-24 one-line layout, edges drawn from the committed network data). A11 (red) is the corridor the peak N-0 economic dispatch leaves at exactly its rating; blue 138 kV lines form the weather-exposed set; dashed violet edges are transformers; bus 118 is the shared slack.

probability and a certified $\leq 10^{-4}$ bound beyond the deep sweep. Twenty independent MC runs at the same committed design miss $\sim$83% each.

The mechanism is Eq. (1), not seed luck: for $\mathrm{P}(C) = 10^{-6}$ and $n = 20{,}000$, $\mathrm{P}(\text{seeing } C) = 1 - (1 - 10^{-6})^{20{,}000} \approx 1.98\%$. MC does not classify the unvisited states as safe—it never observes them at all—but the operational exposure (a consequential contingency absent from every report) is the same; we therefore call this *missed coverage* rather than false negatives. The transportable quantity is the evaluation-count ratio: 11,336 evaluations to certified coverage vs. $\approx$424,661 draws ($37\times$) merely to touch as many unique states, still with random coverage. On this system that is 13 s vs. $\approx$77 s; at a realistic $\sim$1 s AC evaluation cost the same measured counts read $\approx$3 h vs. $\approx$5 days (labeled extrapolation). This result should be read as *validation against the conventional workflow's estimator*, not as competitive benchmarking against the state of the art in contingency identification: vanilla MC is a deliberately weak tool for RQ2's objective, and comparisons against probabilistic $N-k$ selection [9], rare-event samplers [7], [8], and lattice-based critical-state identification [15] remain future work (Section VIII).

TABLE III
RQ2: COVERAGE OF THE 5,028-STATE WATCHLIST (SEVERITY $\geq$ 25 MW, $\mathrm{P}(S) \geq 10^{-6}$).

	TERRA (committed)	MC (20k draws, 20 seeds)
States covered	5,028 / 5,028	mean 845 (810–880)
Missed prob. mass	$0 + \leq 10^{-4}$ certified	mean 0.026
Unique evaluations	11,336	1,837
Draws to match budget	—	$\approx$424,661

C. RQ3: evidence-conditioned updating

Under the weather common cause (Fig. 7) the joint distribution is the finite mixture

$$\mathrm{P}(X) = \sum_{w} \pi_w \prod_{i} q_i(w)^{x_i} \big(1 - q_i(w)\big)^{1-x_i}, \quad \pi'_w \propto \pi_w L(w), \tag{8}$$

where $x_i \in \{0, 1\}$ indicates failure of component i, π_w is the prior weight of weather state w, $q_i(w) = m_w q_i$ with multiplier m_w on exposed lines, and $L(w)$ is the likelihood of observed evidence given $W{=}w$ (hard evidence $W{=}w$ is the degenerate case). Physics is cached by outage state alone, so the three conditional enumerations share 23% of their evaluations, and each posterior update is a pure reweighting.

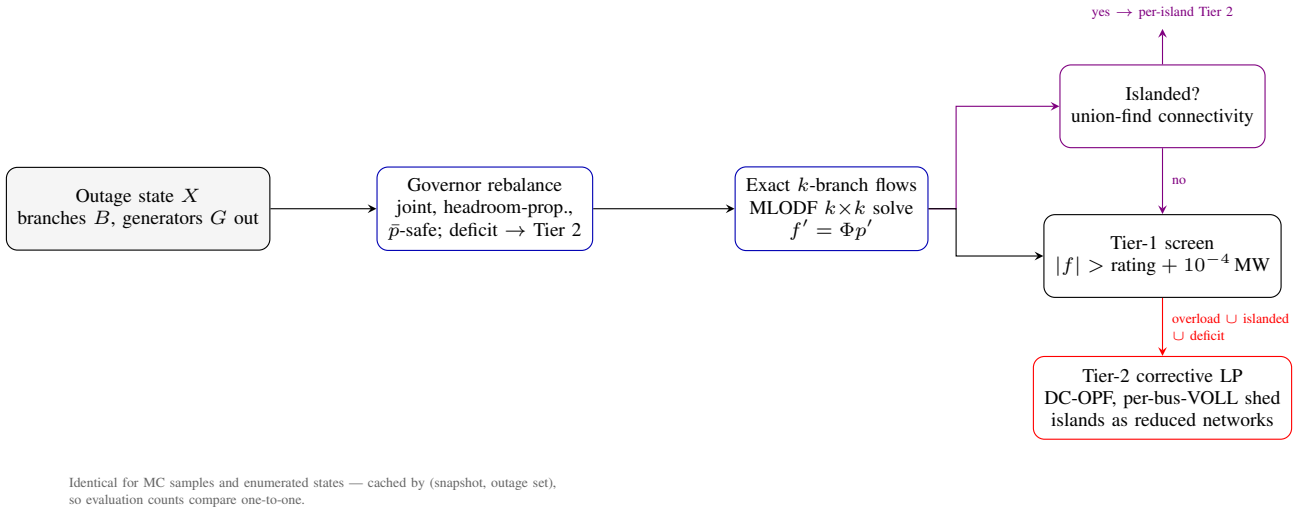


Fig. 3. The per-state evaluation pipeline (identical for sampled and enumerated states; cached by outage set).

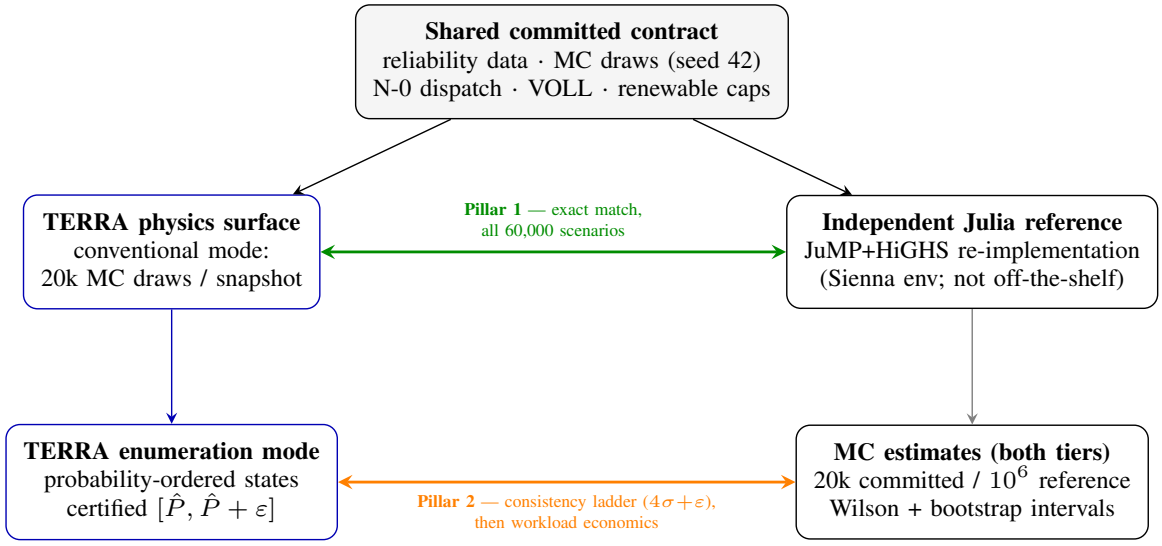


Fig. 4. Validation architecture. Pillar 1 answers the validation question by exact number-for-number replication of the conventional workload against an independent implementation; Pillar 2 answers the research question with physics held constant.

Table IV makes the cost structure explicit, and the honest headline is *amortization, not raw speed*: a cold TERRA conditional is *slower* than a cold conventional update; the ∼0.1 ms figure is the marginal cost of every update after the precomputation, with breakeven around 40 updates (or immediately when the mixture analysis is run anyway for the marginal risk numbers). The evidence here is direct observation of the parent variable—the best case for a precomputed mixture; online inference from partial noisy observations remains future work. One physical subtlety (Fig. 8): severe weather *lowers* P(overload) slightly while raising shed and islanding—the dominant overloading branch (A11) is itself increasingly likely to be out, converting overload states into shed or islanded states.

TABLE IV
RQ3: COLD VS. AMORTIZED PER-UPDATE COST (MEASURED). THE TERRA ADVANTAGE IS AMORTIZED REPEATED INFERENCE, NOT A 40,000× SPEEDUP.

	Cold computation	Subsequent update
Conventional (cond. MC)	3.9–6.4 s	3.9–6.4 s
TERRA (cached mixture)	18.6–164.8 s	∼0.1 **ms, 0 new evaluations**

D. RQ4: scaling

Global explicit enumeration does not dominate Monte Carlo. Doubling the component count (62→124 by network replication) sharply degrades explicit enumeration for aggregate estimation (Table V, Fig. 9): at a fixed 10^5-state budget the achieved certified width grows from 10^{-2} to 0.11, while the Wilson width at the committed design is unchanged (scale-free sampling precision). The depth ladder quantifies

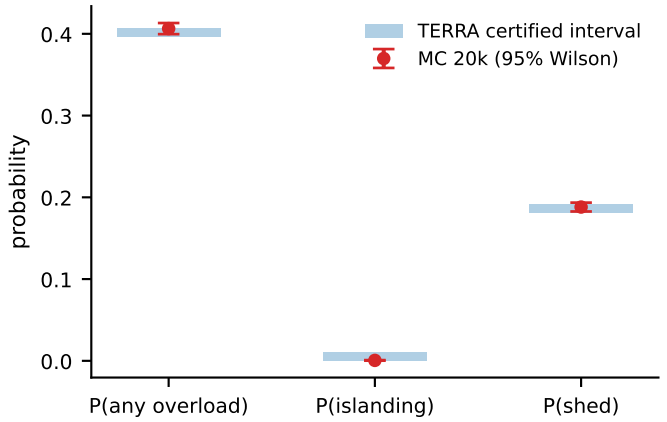


Fig. 5. RQ1 (peak): MC point estimates with 95% Wilson intervals against TERRA certified bands. Every MC interval overlaps its certified band (pinned in the study's tests).

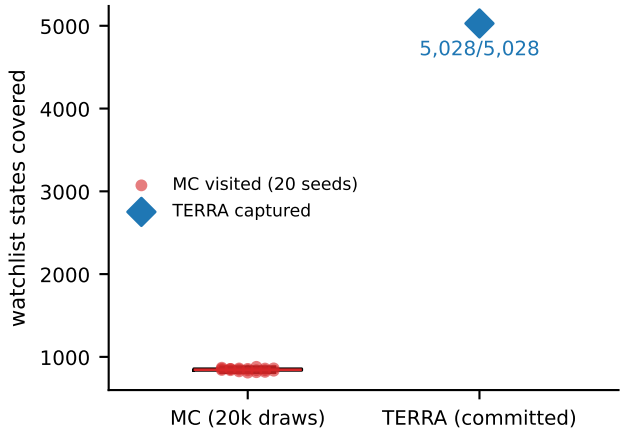


Fig. 6. RQ2: watchlist coverage at matched committed design. TERRA coverage is deterministic; MC coverage is random and budget-starved.

the mechanism: $\varepsilon = 10^{-2}$ requires 11,336 states at $\times 1$ and more than 5×10^6 at $\times 2$ ($\geq 440\times$ per doubling). For aggregate estimation at scale, sampling wins decisively. The identification workload (RQ2) was measured at $\times 1$ only; its behavior at scale is untested here (Section VIII).

The boundary this ladder locates belongs to *global explicit enumeration* — one probability-ordered enumeration over the whole component set — not to structured probabilistic inference as such. Companion work partitions the network into regions, enumerates each region separately, and joins them through physics-derived separator variables (the region's Ward equivalent and its governor statistics), with every joint quantity reported as a certified bracket that includes the un-enumerated mass. On the $\times 2$ replica of this study's Area 1 the two-region decomposition reproduces the exhaustive global result to round-off, and at a per-region truncation of 10^{-2} its certified width is $5\times$ narrower than global enumeration at the same 10^5-state budget with $4.5\times$ fewer physics evaluations. On the RTS-GMLC three-area system, with the three areas as regions and each area enumerated to 10^{-2}, the joint un-enumerated mass is 0.03 where a two-region cut of the same system left 0.13. The decomposition does not remove the combinatorics

TABLE V
RQ4: SCALING WITH COMPONENT COUNT (5M-STATE ENUMERATION CAP; WIDTHS AT FIXED BUDGETS: TERRA 10^5 STATES, MC 20K DRAWS).

	$\times 1$ (62 comp.)	$\times 2$ (124 comp.)
States for $\varepsilon = 10^{-1}$	378	133,031
States for $\varepsilon = 10^{-2}$	11,336	>5,000,000
States for $\varepsilon = 10^{-3}$	135,516	>5,000,000
TERRA certified width at budget	0.010	0.111
MC Wilson width (20k draws)	0.014	0.013
Precision winner at budget	TERRA	MC

— interface pairs replace states and are counted alongside physics evaluations — and it is not faster than enumeration in wall-clock at multi-bus interfaces; what it changes is *coverage*: certified per-branch and per-area numbers on systems whose global state space (10^{12}–10^{24} joint states) no enumeration reaches.

VII. Discussion

A. When structured inference helps

Three properties are demonstrated at the tested scale: deterministic residual uncertainty (Eq. (5)); certified coverage of individually rare, consequential contingencies (RQ2); and near-free repeated evidence-conditioned queries over a reusable evaluated state space (RQ3). All three trace to one design decision: physics is keyed by outage state and explored in probability order, so computation is never spent re-visiting a state and never blind to what was excluded.

B. When Monte Carlo is superior

Aggregate estimation on a growing component space (RQ4): sampling precision is scale-free in contingency-space dimension while certified enumeration depth grows combinatorially with total outage intensity. One-shot aggregate estimation at the tested scale is a wash; a single cold evidence-conditioned query favors the conventional pipeline (Table IV).

C. Implications

The results support a division of labor rather than a replacement: sampling for aggregate risk numbers, probability-ordered structured exploration for the watchlist/identification workload and for repeated conditional queries—the workloads where conventional practice currently falls back on deterministic screening without probability accounting. Because both methods here share a milliseconds-cheap DC evaluator, wall-clock ratios are secondary; the measured *evaluation-count* ratios are what transport to expensive AC physics, where fewer evaluations matter most.

VIII. Limitations and Future Work

DC physics only (no AC, voltage, or stability); two-state steady-state components with all $q_i < \frac{1}{2}$ and baseline independence—deliberately favorable to probability-ordered enumeration; a stylized weather-multiplier common cause rather than a fragility model [17]; RQ2 measured at $\times 1$ scale

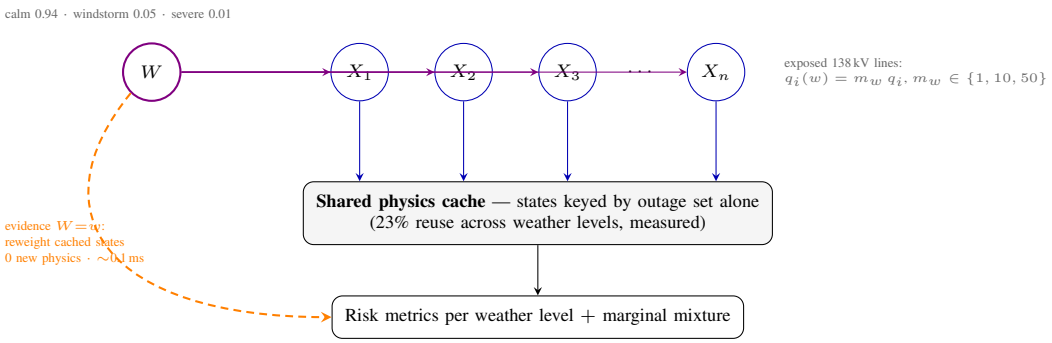


Fig. 7. The common-cause structure: W conditions exposed-line failure probabilities; physics is cached by outage state alone, so observing W reweights cached states rather than triggering new simulation.

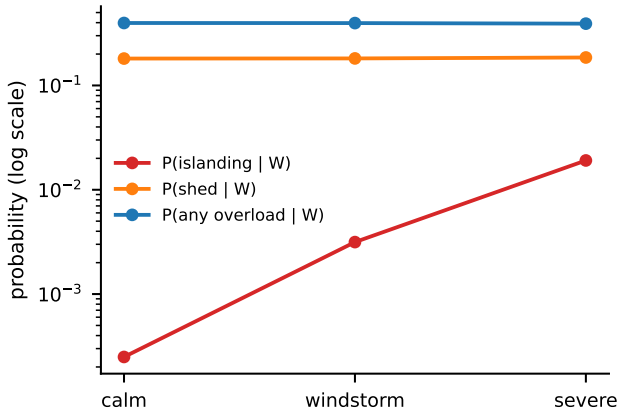


Fig. 8. Conditional risk vs. weather state (certified lower bounds, peak). Islanding rises 76× calm→severe; P(overload) dips via the A11 substitution effect.

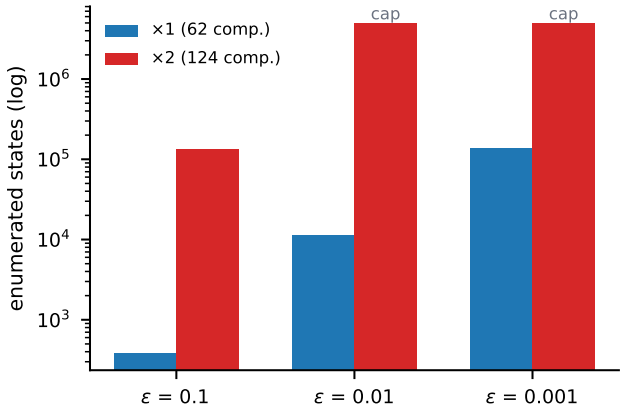


Fig. 9. RQ4: enumeration depth required for certified ε (log scale; bars at the 5×10^6 cap did not reach the target).

only — the priority follow-up experiment, since it is the strongest positive claim; the watchlist definition (severity and probability thresholds) is study-chosen, and RQ2 conclusions should be checked for sensitivity to both; the vanilla-MC baseline understates the strongest conventional identification alternatives (importance sampling, subset simulation [7], [8], deterministic screening [3], [4], probabilistic $N{-}k$ selection [9], and critical-state identification [15]); snapshot-based analysis (the $\sim$40% peak overload probability reflects a deliberately tight N-0 economic dispatch, not real-grid operational risk); conservative severity caps in expectation upper bounds; single-machine wall-clocks with realistic-physics figures labeled as extrapolations. Future work, in priority order: (i) benchmark identification against importance sampling, rare-event simulation, and deterministic $N-k$ screening; (ii) repeat the core experiment with AC contingency physics; (iii) replace the weather multipliers with a defensible fragility model; (iv) apply the local-to-global decomposition — regions enumerated separately, joined through separator messages with certified brackets — where the global state space is out of reach, and close its one open item: an exact joint bracket for any-overload probability at multi-bus interfaces among three or more regions.

The scaling boundary observed here applies to *global explicit state enumeration*. Structured probabilistic inference need not enumerate the global contingency space: aligning the inference partition with the physical network partition — regions enumerated separately, coupled through Ward-equivalent separator variables, as in local-to-global approaches for nonlinear network physics [22] — gives certified per-branch and per-region overload probabilities, expected severity, islanding and deficit for any number of regions, any interface dimension and any interconnection shape (chains, trees, cycles), with the un-enumerated mass and every approximation term stated. Two limits are known and stated: the joint probability of *any* overload is exact for two regions and for single-boundary-bus chains, but only union-bounded when three or more regions meet at multi-bus interfaces; and the certified brackets converge linearly with the interface discretisation, so precision at multi-bus interfaces costs pairs, not states. The configuration studied here does not use the decomposition; the numbers above come from the companion study on the same test system and reliability data.

IX. Conclusions

On a recognized test system with published reliability parameters, probability-ordered structured inference reproduces conventional Monte-Carlo contingency-risk estimates exactly where exact comparison is possible (the replicated workload) and consistently where it is not (the enumeration mode), while providing two capabilities the sampler lacks: certified

coverage of individually rare, high-consequence contingencies, and near-free repeated evidence-conditioned queries over a precomputed state space. It does not scale as an aggregate estimator in its global explicit-enumeration form, and this study locates that boundary rather than hiding it. The result is a strong proof-of-concept for probabilistic contingency-space reasoning — *estimating risk* and *finding the contingencies that create the risk* are different problems, and each has a different best tool.

APPENDIX

Every number is read from committed JSON artifacts produced by deterministic, seeded compute drivers. Verification layers: (1) the enumeration mathematics (probabilities, ordering, certified bounds) is tested against exhaustive brute force on small systems, including adversarial tail-concentrated events; (2) the multi-branch outage algebra of Eq. (7) matches full PTDF recomputation on reduced networks to 10^{-9} MW, with islanding verified against graph connectivity on all 703 $N-2$ pairs; (3) the full physics evaluator matches the independent Julia reference per-scenario on all 60,000 committed scenarios (identical outcome sets; flows $\leq 10^{-9}$ MW; shed ≤ 0.5 MW); (4) committed artifacts carry invariant tests with floors below observed values, and the HTML dossier regenerates byte-identically. Reproduction: the compute drivers under the study directory run in the order documented in the dossier's §7 (seeds 42/0/1000+/5000+/7000+); the one-time Julia capture is never executed in CI. Data: RTS-GMLC upstream commit `3ece0d37`, CC-BY-4.0 [13].

REFERENCES

[1] R. Billinton and R. N. Allan, *Reliability Evaluation of Power Systems*, 2nd ed. Plenum Press, 1996.

[2] R. Billinton and W. Li, *Reliability Assessment of Electric Power Systems Using Monte Carlo Methods*. Plenum Press, 1994.

[3] G. C. Ejebe and B. F. Wollenberg, "Automatic contingency selection," *IEEE Transactions on Power Apparatus and Systems*, vol. PAS-98, no. 1, pp. 97–109, 1979.

[4] V. Donde, V. López, B. Lesieutre, A. Pinar, C. Tamang, and J. Meza, "Severe multiple contingency screening in electric power systems," *IEEE Transactions on Power Systems*, vol. 23, no. 2, pp. 406–417, 2008.

[5] A. M. Rei and M. T. Schilling, "Reliability assessment of the Brazilian power system using enumeration and Monte Carlo," *IEEE Transactions on Power Systems*, vol. 23, no. 3, pp. 1480–1487, 2008.

[6] K. Hou, H. Jia, X. Li, X. Xu, Y. Mu, T. Jiang, and X. Yu, "Impact-increment based decoupled reliability assessment approach for composite generation and transmission systems," *IET Generation, Transmission & Distribution*, vol. 12, no. 3, 2018, doi:10.1049/iet-gtd.2017.0745.

[7] R. Y. Rubinstein and D. P. Kroese, *Simulation and the Monte Carlo Method*, 3rd ed. Wiley, 2016.

[8] S.-K. Au and J. L. Beck, "Estimation of small failure probabilities in high dimensions by subset simulation," *Probabilistic Engineering Mechanics*, vol. 16, no. 4, pp. 263–277, 2001.

[9] K. Sundar, C. Coffrin, H. Nagarajan, and R. Bent, "Probabilistic N-k failure-identification for power systems," *Networks*, vol. 71, no. 3, pp. 302–321, 2018.

[10] T. Daemi, A. Ebrahimi, and M. Fotuhi-Firuzabad, "Constructing the Bayesian network for components reliability importance ranking in composite power systems," *International Journal of Electrical Power & Energy Systems*, vol. 43, no. 1, pp. 474–480, 2012.

[11] S. R. Khuntia, J. L. Rueda, and M. A. M. M. van der Meijden, "Mutual information based Bayesian analysis of power system reliability," in *Proc. IEEE Eindhoven PowerTech*, 2015.

[12] D. Koller and N. Friedman, *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, 2009.

[13] C. Barrows *et al.*, "The IEEE reliability test system: A proposed 2019 update," *IEEE Transactions on Power Systems*, vol. 35, no. 1, pp. 119–127, 2020.

[14] C. Grigg *et al.*, "The IEEE reliability test system—1996," *IEEE Transactions on Power Systems*, vol. 14, no. 3, pp. 1010–1020, 1999.

[15] F. Chen, H. Hu, W. Wan, H. Zhang, X. Liu, B. Yu, and K. Zhao, "A lattice partition-based approach for critical states identification in a power system," *Journal of Industrial and Management Optimization*, vol. 22, no. 4, pp. 2119–2146, 2026.

[16] T. Daemi and A. Ebrahimi, "Detailed reliability assessment of composite power systems considering load variation and weather conditions using the Bayesian network," *International Transactions on Electrical Energy Systems*, vol. 24, no. 3, pp. 305–317, 2014.

[17] M. Panteli and P. Mancarella, "Influence of extreme weather and climate change on the resilience of power systems: Impacts and possible mitigation strategies," *Electric Power Systems Research*, vol. 127, pp. 259–270, 2015.

[18] T. Guler, G. Gross, and M. Liu, "Generalized line outage distribution factors," *IEEE Transactions on Power Systems*, vol. 22, no. 2, pp. 879–881, 2007.

[19] A. J. Wood, B. F. Wollenberg, and G. B. Sheblé, *Power Generation, Operation, and Control*, 3rd ed. Wiley, 2013.

[20] Q. Huangfu and J. A. J. Hall, "Parallelizing the dual revised simplex method," *Mathematical Programming Computation*, vol. 10, no. 1, pp. 119–142, 2018.

[21] E. B. Wilson, "Probable inference, the law of succession, and statistical inference," *Journal of the American Statistical Association*, vol. 22, no. 158, pp. 209–212, 1927.

[22] S. Srinivasan and K. Sundar, "The solution of potential-driven, steady-state nonlinear network flow equations via graph partitioning," 2025, arXiv:2512.22124.