Where Probability-Ordered Enumeration Stops Paying: Watchlist Coverage of Rare Contingencies from RTS-GMLC Area 1 to the Full 73-Bus System

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

***Abstract*—Probabilistic contingency analysis on a two-state component model asks which of the 2^n outage states of an n-component power system carry both non-negligible probability and a damaging consequence. Two families of tools answer it: sampling the component availability vector by Monte Carlo (MC), and enumerating outage states in non-increasing probability with a certified bound on what was left out. A predecessor study showed on RTS-GMLC Area 1 (24 buses, 62 components) that the enumeration path — the TERRA-watt engine, TERRA below — reproduces conventional MC numbers exactly and, at one scale and one threshold pair, covers a rare-contingency *watchlist* (states with post-contingency severity $\geq$ 25 MW and probability $\geq 10^{-6}$) that crude MC misses. This paper asks where that coverage advantage stops. We build a four-system ladder — Area 1 replicated $\times 1/\times 2/\times 4$ (62/124/248 components) and the real 73-bus three-area RTS-GMLC (193 components) — and score enumeration against five stronger identification baselines (inflated-Bernoulli importance sampling, cross-entropy, subset simulation, a pruning hybrid, and deterministic line-outage-factor and probability-times-severity screens) at matched unique-evaluation budgets on a 3×3 grid of watchlist definitions (severity $\geq$ 10/25/50 MW $\times$ probability $\geq 10^{-5}/10^{-6}/10^{-7}$). Each cell's reference is a complete best-first sweep (5,028-state continuity with the predecessor to the digit; 328 method logs in total). The result is a cost law rather than a winner: enumeration's evaluations to full coverage equal the number of states above the probability floor, independent of the watchlist, so the deciding parameter is the *watchlist density* ρ — reference states per enumeration evaluation that full coverage costs, a property of the system and the cell, not of any method. Enumeration dominates wherever ρ is large — every cell of the congested replicas, $\rho \approx 0.44$–0.89 ($\times 4$ at $p_{\min} = 10^{-7}$: 0.80 against the best baseline's 0.34, all methods budget-starved) — and inverts on the real system at $p_{\min} = 10^{-7}$, where $\rho \leq 0.008$: a probability-times-severity screen reaches 90% coverage in 4,453 evaluations against enumeration's 422,823 and matches or beats it at budget (0.982 vs 0.957); subset simulation matches it (0.977). Fixed-inflation importance sampling collapses with component count (inflation 2: 0.93 → 0.66 → 0.02). A small AC-physics probe (a *spike*) on 11,336 enumerated states shows the DC watchlist survives AC exactly on the active-power axis at $\times 1$ and that 200 of 1,992 sampled full-system states are DC-benign but AC-violating. All artifacts are committed and an independent implementation in Julia reproduces the full-system reference to the last digit on 20,000 scenarios.**

***Index Terms*—contingency analysis, rare-event simulation, importance sampling, subset simulation, contingency screening, RTS-GMLC, probabilistic reliability, AC power flow**

Companion to the study directory `contingency_risk_at_scale`; every number is read from committed, test-pinned artifacts. TERRA = TERRA-watt. Wall-clock timings are one workstation's; evaluation counts are the portable quantity.

I. Introduction

Setting. A transmission system with n components (branches and generating units) that can each be *available* or *failed* has 2^n outage states. Probabilistic contingency analysis asks which of these states matter — which carry a probability large enough to plan for and a post-contingency consequence severe enough to care about — and how much probability mass they hold in total [1], [2]. Two families of tools answer it. The conventional pipeline draws the availability vector at random (Monte Carlo, MC), evaluates each drawn topology with a power-system tool, and estimates probabilities as frequencies with statistical confidence intervals. The alternative examined here *enumerates* outage states in non-increasing probability, evaluates each once, and carries a deterministic certificate: after m states the total probability of everything not yet visited is known exactly [3]. In this paper the enumeration path is the TERRA-watt engine (TERRA), a graph-resolved physics engine whose contingency workload was validated in a predecessor study; the paper is about the *method* — probability-ordered enumeration with a certificate — not about the engine.

What the predecessor established. On Area 1 of RTS-GMLC — the IEEE Reliability Test System as updated by the Grid Modernization Laboratory Consortium [4], [5] — 24 buses, 38 branches, 62 two-state components with published failure and repair rates — the enumeration path reproduced the conventional MC numbers exactly against an independent implementation of the same conventions written in Julia with the JuMP modelling language (60,000 scenarios matched number-for-number). Its strongest positive finding concerned the *watchlist*: the set of outage states whose post-contingency overload severity is at least 25 MW and whose probability is at least 10^{-6}. Of the 5,028 such states, the 0.99-probability-mass enumeration (11,336 states) captured all, while crude MC at 20,000 draws missed 4,183 ± 20. That finding carried three caveats a referee would raise: it was measured at one

scale, for one (severity, probability) threshold pair, and against the weakest reasonable baseline.

Research questions. *RQ-A:* does watchlist coverage survive scale, and at what scale and watchlist definition does enumeration's advantage over each baseline vanish? *RQ-B:* do stronger identification baselines — importance sampling, cross-entropy, subset simulation, state-space pruning, deterministic $N-k$ screening — close the gap at matched evaluation budgets? *RQ-C (spike):* does the DC watchlist survive AC contingency physics? (Here $N-k$ denotes the loss of k components; DC and AC denote the linearised and the full non-linear power-flow models.)

Contributions. (i) A four-system ladder with complete, certified reference watchlists on a 3×3 (severity × probability floor) grid; (ii) five baselines implemented on one evaluator contract, run at matched unique-evaluation budgets and scored by a union-reference rule that never measures a method against enumeration's own prefix; (iii) the *cost law* — enumeration's evaluations to full coverage equal the number of states above the probability floor, independent of the watchlist — and the crossover it implies, measured; (iv) an AC evaluator on the same contract and a spike quantifying where the DC watchlist survives AC physics; (v) exact continuity with the predecessor and an independent implementation of the full-system reference. We do not claim that enumeration dominates: on the real system's rarest cells it does not, and we say so with the metric, scale and budget attached.

Reading guide. Section II places the work; Section III defines the model, the evaluator and the watchlist metrics; Section IV describes the systems, the reference sweeps, the baselines, the scoring rule and the AC evaluator; Section V the experimental setup and the continuity checks; Section VI the results; Sections VII–IX discuss, list threats to validity, and conclude.

II. Related Work

Methods that decide *which* outage states to evaluate fall into five families. They differ less in machinery than in what they return and what they promise about what they did not return, which is what Table I organises. This paper's method sits in the first family, and its baselines are drawn from the third and fourth.

A. Enumerative methods

Enumerating outage states in non-increasing probability and stopping on a residual-mass criterion is classical in composite-system reliability, and its trade-off against non-sequential Monte Carlo has been studied since at least Rei and Schilling's assessment of the Brazilian system [3]: enumeration wins when the important states can be reached early in the ordering, sampling wins when the phenomena are complex or the state space is not usefully ordered. That statement is qualitative. This paper's contribution to it is quantitative and per-state: instead of comparing aggregate reliability indices, we ask what fraction of an explicitly defined set of individually important states each method has *found* at a matched evaluation budget, with the complete set known.

B. Optimisation and interdiction

A second family poses identification as optimisation. Donde et al. search directly for severe multiple contingencies rather than enumerating them [6]. Sundar et al. formulate probabilistic $N-k$ identification as a bilevel mixed-integer nonlinear program whose upper level chooses the k components to fail and whose lower level redispatches to minimise load shed, maximising the product of the contingency's probability and the resulting shed; convex relaxations, linear approximations and a cutting-plane algorithm make it tractable, and the study reaches systems up to the 1,354-bus PEGASE and 2,383-bus Polish networks [7]. That scale is well beyond ours. The objective, however, is different in kind: it returns the (near-) maximiser of a scalar risk product $P(x)\,L(x)$, whereas a watchlist is the complete set satisfying two *separate* thresholds. The distinction is operational, not cosmetic: a state with $p = 10^{-4}$ and 100 MW of consequence and a state with $p = 10^{-6}$ and 10,000 MW have the same product and demand different responses, and a planner asking "what must my watchlist contain?" needs both listed, not the larger product returned. Sundar et al. also note that their cutting planes rest on a condition that need not hold in general, so those variants are heuristics [7]: no completeness certificate of the kind used here is claimed or implied.

C. Lattice-partition methods

Structurally the closest relatives are lattice-partition methods, which treat the binary outage space as a lattice and recursively split it, identifying critical states (minimal cut sets) and tightening upper and lower bounds on loss-of-load probability while leaving whole sublattices unevaluated [8], [9]. This matters for us specifically, because it attacks the cost law of Section VI: pruning by lattice structure can break the identity between "evaluations to full coverage" and "states above the probability floor" that governs enumeration's cost. Their target quantity is a cut-set/reliability-bound one rather than a two-threshold overload watchlist, so the results are not directly comparable to the coverage numbers here, and we do not implement them as a baseline; we flag the family as the most promising route to relaxing the cost law and return to it in Section VII.

D. Screening, ranking and learned screens

Ranking contingencies by a performance index so that only the top of the list is fully evaluated dates to Ejebe and Wollenberg [10]. Our probability-times-severity (PoS) screen is the plain utility recipe in that tradition — rank candidate states by probability times the sum of their members' single-outage severities — and is deliberately not novel: it is the baseline a practitioner would run, and

TABLE I
FAMILIES OF CONTINGENCY-IDENTIFICATION METHODS, BY WHAT EACH RETURNS AND WHAT IT CERTIFIES ABOUT THE STATES IT DID NOT EVALUATE.

family	question answered	returns	certifies	examples
enumerative	which states satisfy a criterion?	a set	residual probability mass	[3], this paper
optimisation / interdiction	which $N-k$ maximises risk?	an optimiser (or a few)	optimality gap, when the relaxation is exact	[6], [7]
lattice partition	which states are critical (minimal cut sets)?	critical states + reliability bounds	bounds on loss-of-load probability	[8], [9]
screening / ranking	which states first, on a proxy?	an ordered prefix	nothing (or, when learned, a classifier guarantee)	[10], [11]
rare-event sampling	how probable is the rare event?	an estimate + interval	statistical confidence only	[12], [13]
generative inference	generate high-impact states directly	a sample of severe states	coverage under a learned-model assumption	[14]

(Section VI) it is the one baseline that beats enumeration anywhere in our grid. Recent work makes screens *learned* and, in one case, certified: Christianson et al. train an input-convex neural network whose convexity yields a zero-false-negative guarantee for their feasibility classification, with 10–20× screening speedups reported on a 39-bus case [11]. That guarantee and ours are different objects — theirs bounds the classifier's error on states it is shown, ours bounds the probability mass of states nobody has looked at — and they compose: a certified screen inside a probability-ordered enumeration would keep the residual-mass certificate while skipping benign states cheaply.

E. Surrogate evaluators

A separate line accelerates the *physics per state* rather than the choice of states: graph neural networks trained on $N-1$ examples and applied to $N-2/N-3$ screening report inference up to 400× faster than a Newton-Raphson power flow [15]. The two axes multiply — total work $\approx N_{\text{states}} \times C_{\text{physics}}$ — and this paper measures only the first: every number here is in evaluations, so a cheaper evaluator rescales all methods equally and changes no coverage result.

F. Rare-event sampling

Estimating small failure probabilities by simulation rests on importance sampling and its adaptive variants such as the cross-entropy method [13], on subset simulation [12] — still being applied to rare power-system events such as transient instability [16] — on sequential importance sampling for cascading-outage simulation [17], and, in reliability, on enumeration/Monte Carlo hybrids [3], [2]. We implement each in its textbook form on the discrete availability lattice. These methods are built to *estimate a probability*, not to *enumerate a set*, and our results are consistent with that design: they help exactly where the target event is genuinely rare.

G. Generative identification

The newest family skips the combinatorial space altogether and *generates* high-impact states: Mai et al. condition a diffusion model on the operating state, score risk with a topology-aware graph network, and report high-severity yield at near-constant online cost as k grows [14]. Their coverage result is stated under the assumption that the learned distribution is within a bounded Kullback-Leibler divergence of the true high-severity one; ours follows from the stated component probability model by arithmetic. Both are guarantees; only one is unconditional on a fitted model — and the generative approach is aimed squarely at the sparse-watchlist regime where Section VI shows enumeration losing.

H. Dependence in real outage data

All of the above (this paper included) inherit whatever outage model they are given. Utility data show that real multiple line outages are not independent: they concentrate in recurring spatial *motifs*, and contingency lists built from those motifs capture much of the observed probability of multiple initiating outages [18]. Our probability model is independent two-state components (Section III-A), which is the RTS-GMLC convention and the predecessor's; Section VIII treats the consequences and Section VII the route to relaxing it.

I. Power-flow machinery

Post-contingency flows for simultaneous multiple branch outages use exact generalised line-outage distribution factors [19] on the DC model [20]; the load-shedding tier is a DC optimal power flow with per-bus value of lost load solved with the open-source HiGHS linear-programming solver [21]. Region-partitioned inference for the same engine is a separate program [22] and is not used here.

III. Model and Definitions

A. Components, states and probabilities

Each transmission branch and each thermal generating unit with a published forced outage rate is a two-state component: available or failed. A component i with failure rate λ_i (occurrences per year) and repair rate μ_i (the reciprocal of its mean repair time) is failed at a random instant with the steady-state unavailability $q_i = \lambda_i/(\lambda_i + \mu_i)$; the RTS-GMLC data give $q_i < \frac{1}{2}$ for every component.

Failures are independent, so an outage state — a vector $x \in \{0,1\}^n$ with $x_i = 1$ when component i is failed — has probability $p(x) = \prod_i q_i^{x_i}(1-q_i)^{1-x_i}$. Wind, solar, hydro, synchronous condensers and the single storage unit carry no outage probability (they are not components).

B. Probability-ordered enumeration and its certificate

Because every $q_i < \frac{1}{2}$, failing an additional component always lowers $p(x)$, and the states can be generated in exactly non-increasing probability by a best-first search over the lattice: a heap of candidate states ordered by p, popping the largest and pushing its two successors (add the next component; or swap the last-added component for the next). After m states have been popped, the unenumerated probability mass is $\varepsilon_m = 1 - \sum_{k \le m} p(x_k)$ exactly — the *certificate*: any event's probability is bracketed by the mass found among enumerated states and that plus ε_m. Enumeration is what TERRA uses; in this paper "TERRA" and "enumeration" are interchangeable and always mean the prefix of this ordering.

C. The evaluator: what a state costs and what it returns

Every method in this paper scores a state through the same two-tier DC evaluator, so that comparisons are about *which states are visited*, never about physics. Tier 1: (a) *governor rebalance* — the generation lost with failed units is redistributed to the available units in proportion to their headroom (rated maximum output $P_{\max}$ minus dispatch), any residual deficit landing at the *slack bus* (the reference bus that balances the network); (b) *exact post-contingency flows* — the DC power flow with all failed branches removed simultaneously, computed from the intact-network flows by the generalised line-outage distribution factors (an exact $k \times k$ correction for k failed branches, not a stack of single-outage factors); (c) *islanding* — connectivity of the remaining graph; (d) *overloads* — branch b with post-contingency flow f_b and continuous rating Rate_b is overloaded when $|f_b| - \text{Rate}_b > 10^{-4}$ MW, and the state's *severity* is $\sum_b (|f_b| - \text{Rate}_b)^+$ in MW; (e) new here, an *island load-loss bound* $\sum_{\text{islands}}(\text{island load} - \sum \text{available } P_{\max})^+$ that needs no linear program (LP), so that a state that islands 300 MW of load is not invisible to a severity-only watchlist. Tier 2 — a DC optimal power flow (DC-OPF) that sheds load at a per-bus value of lost load (VOLL) — is used only for the load-shed risk metrics of the independent-implementation comparison, never for coverage. One *evaluation* is one unique physics call; every method's budget is counted in evaluations, and a cache makes a state cost one evaluation however often it is drawn.

D. Watchlist cells and coverage metrics

A *watchlist cell* is a pair (severity threshold s, probability floor $p_{\min}$) with $s \in \{10, 25, 50\}$ MW and $p_{\min} \in \{10^{-5}, 10^{-6}, 10^{-7}\}$; its *reference* is the set of states with (severity $\ge s$ or island bound $\ge s$) and $p \ge p_{\min}$. For a method with a visit log (its unique evaluated states in order) and a budget B (the number of unique evaluations it may spend): *coverage at B* is the fraction of the reference among its first B unique states, by count (and, secondarily, by probability mass); *false negatives at B* are the reference states not among them; *evaluations to coverage c* is the smallest budget at which the fraction reaches c, reported for $c \in \{0.9, 0.99, 1.0\}$. Figure 1 draws these definitions.

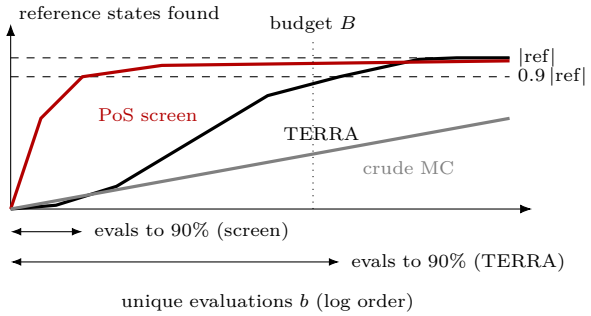


Fig. 1. How a method is scored. Its visit log is a sequence of unique evaluated states; the curve counts how many reference states of a cell have been found after b evaluations. *Coverage at budget* is the curve's height at $b = B$ divided by |ref|; *false negatives at B* the gap to |ref|; *evaluations to 90%* the first b where the curve reaches 0.9 |ref|. Enumeration's own log is the sweep prefix, so its curve reaches |ref| exactly at $|\{x : p(x) \ge p_{\min}\}|$ evaluations. (Schematic.)

E. Budget accounting

Samplers draw until B unique states have been evaluated; the raw number of draws and the cache hits are reported alongside, and a raw-draw cap of $20B$ marks a log as *budget-limited* when the sampler cannot reach B distinct states (crude MC on Area 1 is the case: two million draws yield ≈ 27,000 distinct states).

IV. Methods

A. Systems and operating points

The replication ladder. Area 1 of RTS-GMLC — 24 buses, 38 branches, 46 dispatchable units after the hydro units are dropped, 62 two-state components — is copied ×1, ×2, ×4 and the copies are joined in a chain by slack-to-slack tie lines (500 MW rating, reactance 0.06 per unit, pu) that carry no outage probability. This is a *fixture*: the copies share the same congestion, so ×2/×4 measure how the number of probable states grows with n (combinatorial scaling), not new physics. *The full system.* The three-area RTS-GMLC — 73 buses, 120 branches including the five real inter-area ties, 193 components after the same filters — pulled from the pinned upstream commit with a line-preserving extractor; per-area load profiles are resolved to buses by nameplate share; the single storage unit is dropped like hydro; the concentrating-solar unit is capped at its nameplate from its solar-field inflow series. *One peak snapshot per system:* the hour of maximum load; the base-case dispatch is a DC-OPF that is optimal, sheds nothing and leaves no branch above its rating (full system: hour 5726, 7,961 MW, tightest branch 15.9 MW below rating). Table II lists sizes, budgets and seeds.

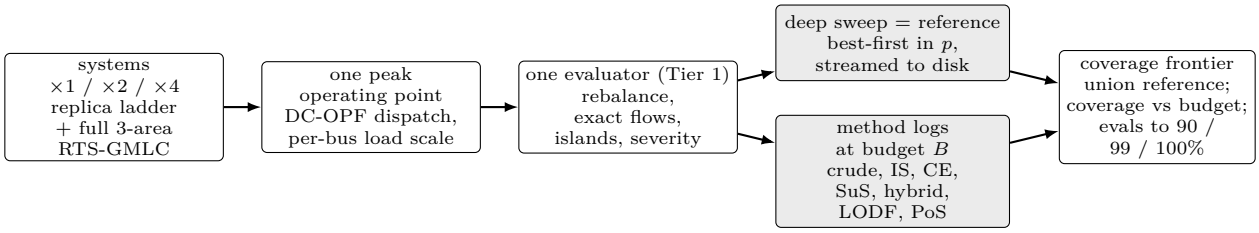


Fig. 2. The measurement design. Every system is reduced to one peak operating point and one evaluator; the reference lattice comes from one streamed best-first sweep per system, every baseline writes a visit log at the same evaluator and budget, and all coverage numbers are pure post-hoc arithmetic over sweep + logs — no method is ever re-run to score a different watchlist definition.

TABLE II
SYSTEMS, BUDGETS AND SEEDS. B = UNIQUE EVALUATIONS PER METHOD LOG; TIER-1 = MEASURED WALL TIME PER EVALUATION (ONE WORKSTATION).

system	buses	branches	components	B (unique evals)	seeds	Tier-1 ms/eval
×1	24	38	62	100,000	10	0.50
×2	48	77	124	1,000,000	10	0.38
×4	96	155	248	500,000	5	0.57
full	73	120	193	500,000	5	0.44

B. Reference sweeps

The reference of every cell comes from one best-first sweep per system that streams each visited state's probability, severity and island bound to disk (no per-state cache: the frontier heap is the only memory). The sweep stops when the enumerated mass reaches 0.9999 (×1), or when the frontier probability drops below a floor one decade under the grid's smallest $p_{\min}$ (10^{-9}) or 2×10^7 states have been visited (the larger systems). At the stop it *drains the tie group* — replicas produce states of exactly equal probability by symmetry — and records the largest probability of any unvisited state, p'_{stop}. Because the ordering is monotone in p, a state is in the sweep iff $p \geq p_{\text{stop}}$, and a cell is *complete* — its reference exact — iff $p'_{\text{stop}} < p_{\min}$. Every cell of every system is complete here (Table III).

C. Baselines

All five samplers and both screens run on the same evaluator with the same budget B.

- *Crude Monte Carlo* — the availability vector drawn from the nominal law (component i failed with probability q_i), the same draws as the reference implementation's scenario contract.
- *Importance sampling (IS)* — draws from an *inflated* law $q'_i = \min(\kappa q_i, \frac{1}{2})$ with inflation factor $\kappa \geq 1$ ($\kappa = 1$ is crude MC), each drawn state x carrying the exact weight $p(x)/q'(x)$; $\kappa \in 2, 3, 5, 8$ on ×1 and the two best values elsewhere. Inflation makes rare high-order states more likely to be drawn at the price of weight variance.
- *Cross-entropy (CE)* — an adaptive IS: starting from the nominal law, each round draws a batch, keeps the elite fraction $\rho = 0.1$ with the highest *driving variable* $g(x)$ (defined next), and moves each component's draw probability toward the weighted elite frequency with smoothing $\alpha = 0.7$ (probabilities clipped to $[10^{-6}, \frac{1}{2}]$ so none collapses to 0 or 1), until the elite threshold reaches the target severity; a fresh IS run with the adapted law then spends the remaining budget. Its adaptation draws are real evaluations and are logged.
- *Subset simulation (SuS)* — writes the rare event as a nested chain of less rare ones. The driving variable $g(x)$ is the state's largest flow excess $\max_b(|f_b| - \text{Rate}_b)$ when it has no violation (a negative *margin*), and its worst violation $\max(\text{severity}, \text{island bound})$ otherwise, so that the cell criterion is exactly the level set $g \geq s$. Level 0 is crude MC; each further level keeps the top fraction $p_0 = 0.1$ of the previous samples as seeds and runs a Markov chain from each — every component proposes a flip with probability $\frac{1}{2}$ and the flip is accepted with the flipped component's marginal ratio ($q_i/(1-q_i)$ or its inverse), the whole candidate being kept only if it stays above the level — until the level reaches s; the run is repeated with fresh chains until B is spent.
- *Pruning hybrid* — enumeration of the most probable prefix ($\min(B/2$ states, 0.99 mass)) evaluated exactly, plus IS on the residual with a hash of the prefix so that draws landing inside it are spent as zero weight — the classical enumeration/MC hybrid of the reliability literature.
- *Line-outage-factor (LODF) screen* — exhaustive single-component ($N-1$) states, then all pairs ($N-2$) ordered by the sum of the members' $N-1$ severities, until B.
- *Probability × severity (PoS) screen* — the $5B$ most probable states re-ranked by $p(x)$ times the sum of the members' $N-1$ severities (the proxy a screen can know before evaluating), the top B evaluated in

TABLE III
REFERENCE SWEEPS. REFERENCE SIZES ARE THE (25 MW) CELLS AT THE THREE FLOORS; "STOP" IS THE RULE THAT ENDED THE SWEEP.

system	states	mass	p_{stop}	stop	ref 10^{-5}	ref 10^{-6}	ref 10^{-7}	wall
×1	1,045,916	0.9999	1.1e-10	mass	1,421	5,839	24,509	0.1 h
×2	8,018,836	0.9852	1.0e-09	p_floor	4,979	34,508	249,283	1.0 h
×4	20,830,531	0.6344	2.9e-09	max_states	2,070	37,359	492,868	4.1 h
full	20,010,337	0.8883	1.7e-09	max_states	11	222	2,685	2.9 h

that order.

Cross-entropy and subset simulation target one severity s per log (three logs per system and seed); the other methods are threshold-agnostic.

D. Scoring: the coverage frontier

Every method's log is a visit sequence in evaluation order. Coverage is pure post-hoc arithmetic against a *union reference*: the sweep's states meeting the cell criterion, plus any qualifying state a method visited beyond the sweep. The union rule exists so that no method is scored against enumeration's own prefix; because every cell here is complete, the union added nothing. Enumeration's own log is the sweep prefix. Figure 3 states the paper's central observation in this vocabulary.

E. AC evaluator (for the spike)

An `AcTwoTierEvaluator` on the same state contract composes the DC evaluator (governor rebalance, islanding, Tier-2 shed) with an AC power flow of the reduced network (Figure 4): the failed branches removed and the failed units absent; the *slack bus* — the bus that absorbs losses in an AC solution — placed per scenario at the largest available dispatched unit; every bus hosting an available unit a voltage-controlled (PV) bus at 1.0 pu, every other bus a constant-power (PQ) bus with reactive load $Q = (Q/P)\cdot P$ for a stated ratio; transformer taps at 1; per-bus reactive-power limits summed from the units' data. The Newton solve follows a stated retry protocol — warm start from the intact solution, then a flat start, then a damped Newton — and one pass that pins any PV bus outside its reactive limits to the limit (PV→PQ). Islanded states never enter the AC path (they return the DC result), and a state that fails every attempt is a class of its own. Severity is reported both as *active-power excess* $\sum_b(|P_b| - \text{Rate}_b)^+$ (the DC-comparable number) and as *apparent-power excess* (in megavolt-amperes, MVA) $\sum_b(S_b - \text{Rate}_b)^+$ with $S_b = \sqrt{P_b^2 + Q_b^2}$ at the higher-loaded end, the latter both raw and *net of the intact case's own excess* — because, as Section VI shows, the intact case already exceeds an MVA rating.

V. EXPERIMENTAL SETUP

Budgets, seeds and systems are those of Table II (seed base 20260815; cross-entropy and subset simulation one log per severity). 328 method logs were run in total; the sweeps took 8.2 h and the logs 25.9 h on one workstation.

Continuity with the predecessor. The ×1 sweep reproduces the predecessor's depth (prefix lengths 11,336 / 135,516 / 1,045,896 states at enumerated mass 0.99 / 0.999 / 0.9999) and its severity-only (25 MW, 10^{-6}) watchlist to the digit (5,028 states, probability mass 0.232329660242); crude MC at raw draw 20,000 misses 4,172 of them on average over ten seeds (the predecessor reported 4,183 ± 18 with different seeds).

Independent reference. A Julia (JuMP + HiGHS) implementation of the same conventions, written separately from the Python evaluator, evaluated the committed 20,000-scenario full-system contract: 335 scenarios with an overload, 14 with load shed, 15 islanded, on both sides; the per-scenario maximum absolute difference is 0.0 on severity, shed and cost and on the branch-flow subsample. **Byte-identity.** The predecessor's committed per-scenario result files re-produce byte-for-byte through the evaluator used here.

VI. RESULTS

A. RQ-A: coverage across scale

Enumeration covers every cell of the ×1 and ×2 ladders at B (Fig. 5, Table IV). At ×4 with $B = 5 \times 10^5$ the 10^{-7} watchlists hold 492,868 states and every method is budget-starved — no method's budget can reach the whole reference: enumeration 0.80 (it needs 649,611 evaluations for 100%), the pruning hybrid 0.34, crude 0.18, IS $\kappa = 2$ 0.02. On the real system the 10^{-5} and 10^{-6} cells are covered by every method (they hold 11 and 222 states); at 10^{-7} enumeration reaches 0.961 and 0.957 (25 / 50 MW) within B, the PoS screen 0.915 and 0.982, subset simulation 0.808 and 0.977.

The cost law. Enumeration's evaluations to 100% coverage are the same number in every cell of a system — 34,310 / 316,913 / 649,611 / 581,783 at 10^{-7} for ×1 / ×2 / ×4 / full — because they equal $|\{x : p(x) \geq p_{\min}\}|$: best-first must visit every state above the floor, violation or not (Fig. 3).

B. Watchlist density: the parameter that decides the regime

The cost law says enumeration pays for every probable state and the watchlist says only some of them are worth paying for, so the ratio of the two is the quantity that should predict when enumeration is the right tool. Define the *watchlist density* of a cell as

$$\rho(s, p_{\min}) = \frac{|W(s, p_{\min})|}{E_{100}(s, p_{\min})},$$

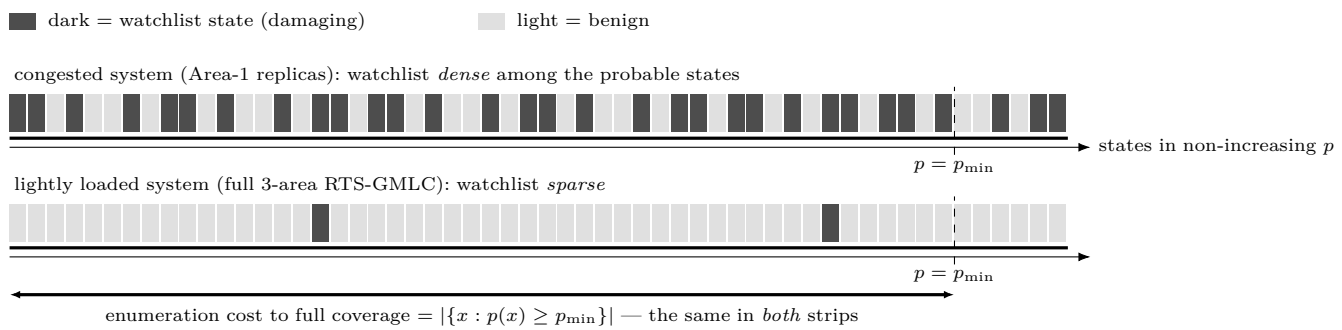


Fig. 3. The cost law. Enumeration visits states in non-increasing probability p; to guarantee every watchlist state with $p \geq p_{\min}$ it must visit *every* state to the left of the dashed line, damaging (dark) or benign (light). That cost is the same in both strips. On a congested system the dark cells are dense, so the visited states are mostly worth visiting; on a lightly loaded system they are sparse, and a screen that ranks states by an $N-1$ severity proxy jumps to the dark cells in a fraction of the evaluations. (Schematic; densities in the two strips are illustrative.)

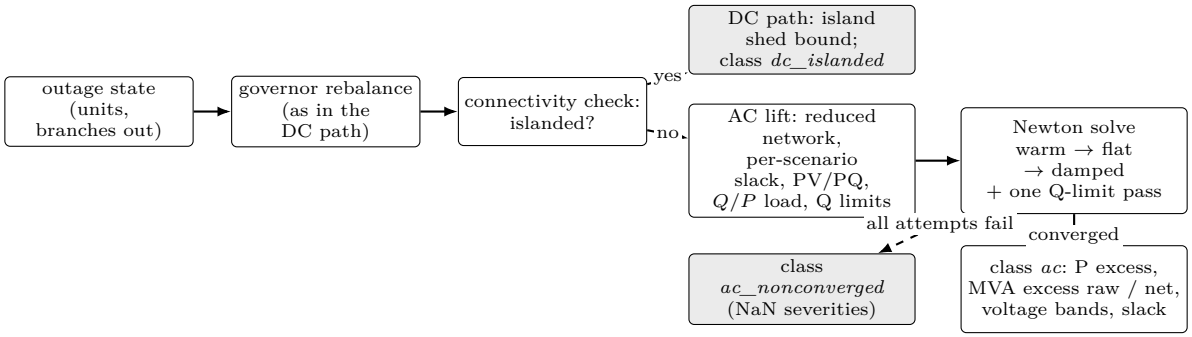


Fig. 4. The AC evaluator on the same state contract. Islanded states never enter the AC path (they return the DC result), so every DC/AC difference is confined to connected states; a state that fails every solve attempt is a class of its own, never scored as benign or violating.

TABLE IV
COVERAGE BY COUNT AT B ON THE 25 MW CELLS (ALL 36 CELLS ARE IN `FRONTIER.JSON`).

system	$p_{\min}$	reference	TERRA	best baseline	its cov.	crude
×1	10^{-5}	1,421	1.000	CE (s=10)	1.000	1.000
×1	10^{-6}	5,839	1.000	pruning hybrid	1.000	0.978
×1	10^{-7}	24,509	1.000	IS $\kappa = 2$	0.927	0.554
×2	10^{-5}	4,979	1.000	CE (s=10)	1.000	1.000
×2	10^{-6}	34,508	1.000	crude MC	1.000	1.000
×2	10^{-7}	249,283	1.000	pruning hybrid	1.000	0.887
×4	10^{-5}	2,070	1.000	pruning hybrid	1.000	1.000
×4	10^{-6}	37,359	1.000	pruning hybrid	1.000	0.696
×4	10^{-7}	492,868	0.804	pruning hybrid	0.341	0.178
full	10^{-5}	11	1.000	CE (s=10)	1.000	1.000
full	10^{-6}	222	1.000	pruning hybrid	1.000	0.862
full	10^{-7}	2,685	0.961	PoS screen	0.915	0.297

where W is the cell's reference watchlist and E_{100} is enumeration's evaluations to full coverage of it. E_{100} is the prefix length at which the last watchlist state appears, so $E_{100} \leq |\{x : p(x) \geq p_{\min}\}|$ with equality when the least probable state above the floor is itself damaging; ρ is thus (an upper bound on) the fraction of enumeration's work that lands on states the watchlist wants. It is a property of the system, the operating point and the cell — not of any method — and it is computable from the reference sweep alone.

Across the 36 cells ρ separates the two regimes by two orders of magnitude (Table V): 0.44–0.89 on the congested replicas against 0.0009–0.0081 on the real system. Plotting coverage at B against ρ (Fig. 6) makes the regime map explicit: for $\rho \gtrsim 0.4$ enumeration is at or above every baseline, and the only cell in which a baseline beats it — full, (50 MW, 10^{-7}) — is also the cell with the smallest density of all, $\rho = 0.0020$. The mechanism is arithmetic rather than statistical: at $\rho = 0.0020$ enumeration performs about 498 evaluations per watchlist state found, and any ordering with even a weak severity signal beats that ratio.

C. RQ-B: the baselines

Importance sampling. A fixed inflation that helps at 62 components hurts at 124 and collapses at 248: $\kappa = 2$

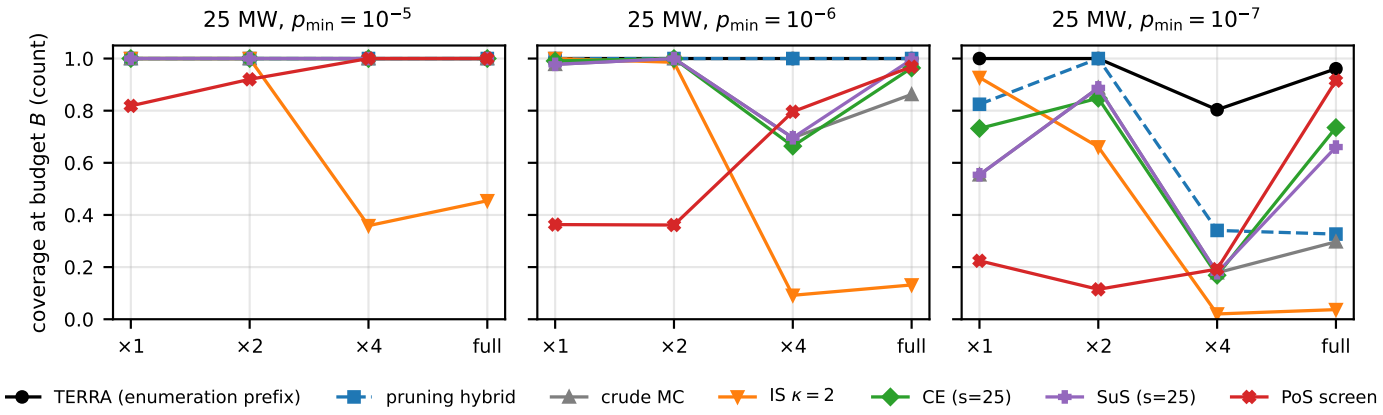


Fig. 5. Coverage of the (25 MW) reference at the matched budget B, by system, for the three probability floors. Enumeration (black) covers every cell on the replication ladder at $p_{\min} \geq 10^{-6}$ and leads at ×4, 10^{-7}; on the real system at 10^{-7} the PoS screen (red) matches or beats it.

TABLE V
WATCHLIST DENSITY $\rho = |W|/E_{100}$ FOR EVERY CELL. THE THREE PROBABILITY FLOORS ARE THE COLUMNS; THE SEVERITY THRESHOLDS THE ROW GROUPS. THE LADDER AND THE REAL SYSTEM DIFFER BY TWO ORDERS OF MAGNITUDE.

system	severity	$p_{\min} = 10^{-5}$	10^{-6}	10^{-7}
×1	10 MW	0.7518	0.8120	0.8448
	25 MW	0.5998	0.6625	0.7143
	50 MW	0.4563	0.5309	0.5428
×2	10 MW	0.7681	0.8305	0.8881
	25 MW	0.7235	0.7138	0.7866
	50 MW	0.5243	0.5181	0.6049
×4	10 MW	0.6726	0.8412	0.8533
	25 MW	0.5085	0.6936	0.7587
	50 MW	0.4387	0.5745	0.6074
full	10 MW	0.0022	0.0050	0.0081
	25 MW	0.0016	0.0036	0.0046
	50 MW	0.0009	0.0016	0.0020

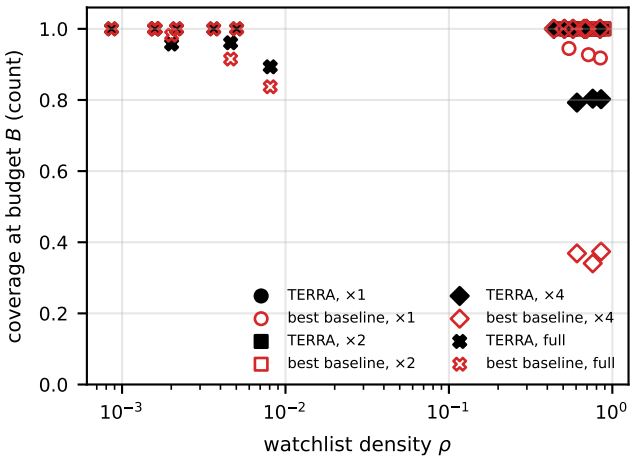


Fig. 6. Coverage at budget B against watchlist density ρ, all 36 cells (filled = enumeration, open red = the best baseline of that cell; marker shape = system). Enumeration's advantage is a function of ρ, not of system size: it is unbeaten for $\rho \gtrsim 0.4$ and loses its only cell at the smallest ρ measured.

gives 0.93 at ×1 (crude 0.55), 0.66 at ×2 (crude 0.89), 0.02 at ×4 (crude 0.18) — doubling every q_i pushes the drawn states toward many simultaneous failures whose weights are negligible (Fig. 8). *Cross-entropy* matches crude wherever the target is not rare (every ladder cell has P(severity $\geq$ s) ≈ 0.25) and helps only on the real system's rare cells, with a large seed spread from proposal collapse. *Subset simulation* never leaves level 0 on the ladder — its logs are draw-for-draw crude MC — and works on the real system's rare cells (0.977 at 50 MW, 10^{-7}). The *hybrid* is enumeration wherever its prefix reaches the watchlist and a weak tail sampler beyond it. The *LODF screen* exhausts its $N-2$ lattice at 1,953–30,876 states — negligible by count, 23–78% by mass. The *PoS screen* is poor on the ladder (Area 1 is $N-1$ secure at peak: every single branch outage is benign, so the proxy leans on generator singles) and decisive on the real system: 90% of the (50 MW, 10^{-7}) watchlist in 4,453 evaluations against enumeration's 422,823, and 0.982 vs 0.957 at B — the one cell in 36 where a baseline beats enumeration at matched budget.

D. RQ-C: AC physics (spike)

The spike's *precondition* — that the intact operating point be violation-free under the AC lift for some Q/P ratio — fails on both systems: the 106–110 cable A10 (and its copies in Areas 2 and 3) is DC-feasible but 6–26% above its rating in MVA once reactive flow is counted, so MVA severity is reported net of the intact case's own per-branch excess. On Area 1 the DC watchlist survives AC exactly on the active-power axis (0 DC-benign / AC-violating states by P; 3 the other way) and the MVA axis adds 443 DC-benign states net of baseline (raw 1451 — the inflation a raw comparison would produce). On the real system 200 of 1,992 sampled states (10%) are DC-benign but AC-violating on P: the DC-model false-negative class is real there. 18% of ×1's enumerated states fail every solve attempt — concentrated on high-order, heavily overloaded states where the AC equations have no solution (voltage-

TABLE VI
×4, (25 MW, 10^{-7}): reference 492,868 states, $B = 5 \times 10^5$. "—" = never reached within B.

method	cov. (count)	cov. (mass)	FN at B	evals→90%	evals→100%
TERRA	0.804	0.956	96,755	554,413	649,611
pruning hybrid	0.341 ± 0.000	0.807	324,982	—	—
crude MC	0.178 ± 0.000	0.587	405,044	—	—
IS $\kappa = 2$	0.020 ± 0.000	0.183	482,885	—	—
CE (s=25)	0.169 ± 0.001	0.574	409,723	—	—
SuS (s=50)	0.178 ± 0.000	0.587	405,044	—	—
PoS screen	0.192 ± 0.000	0.637	398,122	—	—
LODF N-2 screen	0.009 ± 0.000	0.231	488,678	—	—

TABLE VII
full, (50 MW, 10^{-7}): reference 1,169 states (mass 5.6e-04), $B = 5 \times 10^5$.

method	cov. (count)	cov. (mass)	FN at B	evals→90%	evals→100%
TERRA	0.957	0.990	50	422,823	581,783
pruning hybrid	0.391 ± 0.001	0.769	712	—	—
crude MC	0.315 ± 0.009	0.596	801	—	—
IS $\kappa = 2$	0.032 ± 0.004	0.122	1,132	—	—
CE (s=25)	0.739 ± 0.137	0.863	305	232,943	—
SuS (s=50)	0.977 ± 0.005	0.993	27	289891	—
PoS screen	0.982 ± 0.000	0.994	21	4,453	—
LODF N-2 screen	0.411 ± 0.000	0.775	688	—	—

TABLE VIII
AC spike: precondition, convergence, and DC-vs-AC classes at 25 MW.

	×1 (0.99-mass enum.)	full (2,000 sampled)
states / connected	11,336 / 11,296	2,000 / 2,000
Q/P ratio (best of 0.1 / 0.2 / 0.33)	0.33	0.1
intact case violation-free	no (A10, A11 > MVA rating)	no (A10/B10/C10, voltage)
converged (of connected)	9,222 (81.6%)	1,992 (99.6%)
wall / state (median)	54 ms	206 ms
states with a Q-limit switch	100%	100%
slack above its capacity	94.5%	0.4%
both violating	5970	13
DC-benign / AC-violating (P)	0	200
DC-benign / AC-violating (MVA net; raw)	443; 1451	350; 1979
DC-violating / AC-benign (net / P)	3 / 0	0 / 0

collapse territory) — and are reported as their own class. Losses land on the single slack unit (94.5% of ×1 states push it above its capacity).

VII. Discussion

When enumeration is the right tool. Its cost is set by the probability floor; its value is the certificate (every state above the floor visited, none silently missed) and the coverage that follows wherever damaging states are dense among the probable ones — congested systems, coarse floors. That is the regime of the predecessor's result and of the whole replication ladder, and it generalises there: enumeration is the only method never below 0.79 coverage anywhere. **When it is not.** On a lightly loaded meshed system with a deep floor the probable lattice is mostly benign; a screen that knows the single-outage severities finds the damaging tail in a hundredth of the evaluations, and subset simulation, given a genuinely rare event, is competitive. The honest statement is a regime map, not a ranking.

What the baselines teach. Importance sampling needs its inflation to shrink with component count — a fixed κ is a scale trap; cross-entropy and subset simulation are rare-event tools and add nothing where the event has probability 0.25; the pruning hybrid is enumeration with a tail sampler and inherits enumeration's cost law; the PoS proxy is only as good as the single-outage severities — zero on an $N-1$-secure system, decisive where singles already hurt.

Combinatorial versus physical scaling. The ×4 replicas have zero intact tie flow; their watchlists are product states of the ×1 lattice, so ×2/×4 measure the growth of $|\{p \geq p_{\min}\}|$ with n, not new physics. The full system supplies the physics — and it is precisely there that the regime flips.

Where this sits among the alternatives. Optimisation-based $N-k$ identification returns the maximiser of a scalar risk product and scales to thousands of buses [7]; learned screens classify cheaply, one of them with a zero-false-negative guarantee on the states it is shown [11]; generative models sample severe

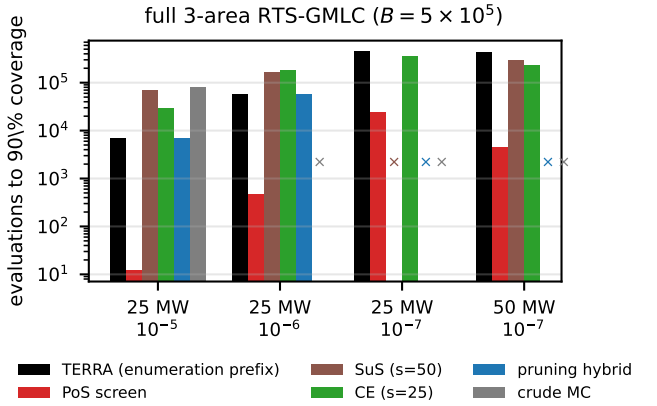


Fig. 7. Evaluations to 90% coverage on the full system (log scale; × = not reached within B). The PoS screen needs 1–5% of enumeration's evaluations on every cell; subset simulation is competitive only where the event is rare (50 MW, 10^{-7}).

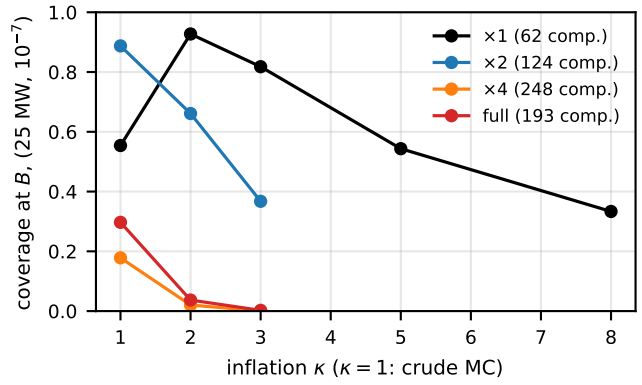


Fig. 8. Coverage at B on (25 MW, 10^{-7}) against the IS inflation κ ($\kappa = 1$ is crude MC): the best κ shrinks with component count and $\kappa = 2$ is catastrophic at 248 components.

states directly at near-constant online cost [14]; surrogate power flows cut the cost of each evaluation by orders of magnitude [15]. None of them returns a set with a bound on the probability of what was never examined, and that — not speed — is what probability-ordered enumeration buys. The measured price of the bound is $1/\rho$ evaluations per watchlist state, which is cheap at $\rho \approx 0.7$ and absurd at $\rho \approx 0.0020$.

The obvious next algorithm. The two ingredients are separable: the *order* in which states are visited and the *certificate* over what remains. Enumeration ties them together by using probability for both. A severity-directed search (a screen, a learned classifier, a generative proposal) that still tracks the probability mass of the unvisited region would keep the certificate while spending its evaluations where ρ is low — and lattice-partition methods, which already prune whole sublattices while maintaining reliability bounds [8], [9], are the closest existing machinery for exactly that. This paper's contribution to that programme is the measurement it needs: ρ says when the

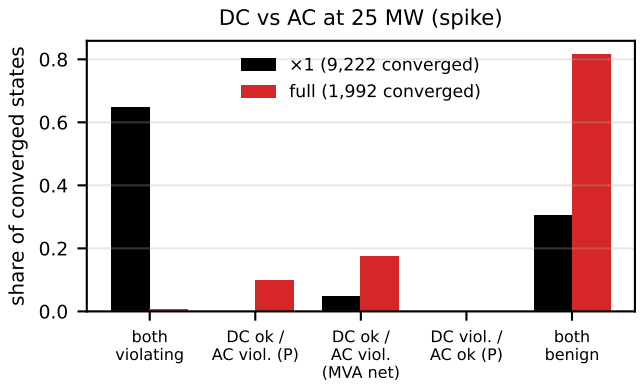


Fig. 9. DC-vs-AC disagreement classes at 25 MW among converged states.

effort is worth spending.

AC. The spike says two things a DC study should carry as caveats: on Area 1 the DC active-power watchlist *is* the AC one; on the real system a tenth of sampled states are DC-benign yet AC-violating on P, and the intact case itself violates an MVA rating — any AC watchlist must be stated net of that baseline. Full RQ-C (top-K per method, an independent AC reference) is deferred.

VIII. Threats to Validity and Limitations

Construct. The watchlist criterion is Tier-1 severity or island bound at one peak snapshot; Tier-2 shed and off-peak snapshots are not scored. Coverage counts states; a mass-weighted view (also reported) narrows the gaps. *Internal.* Budgets are matched on unique evaluations, which favours no method but makes crude MC budget-limited at ×1 (≈ 27,000 unique states per 2×10^6 draws); cross-entropy and subset simulation spend their budget through replication; the PoS candidate window ($5B$ by probability) is a stated choice. Seeds are 10/10/5/5; spreads are reported. *External.* The replication ladder is a fixture; ×2/×4 characterise combinatorial growth under identical congestion; only the full system is a real network, and only at its peak hour. *AC.* The precondition fails, so AC severities are net of the intact baseline; 18% of ×1 states do not converge; the per-scenario single slack absorbs losses; no independent AC reference was run (deferred with the full RQ-C). *Platform.* Wall-clock figures are one workstation's; evaluation counts are the portable quantity.

The probability model. Every number here — the ordering, the certificate, ρ and therefore the regime map — is conditional on the outage model of Section III-A: independent two-state components with published forced-outage rates. Real multiple line outages are not independent; utility data show them concentrating in recurring spatial motifs, and contingency lists drawn from those motifs capture much of the observed probability of multiple initiating outages [18]. Under a dependent model the ranking itself changes: common-mode and weather-correlated failures raise the probability of exactly the

multi-outage states this ordering pushes deep into the tail, which would *increase* ρ (the damaging states move up the ordering) but also invalidate the product form used to compute $p(x)$. The enumeration machinery does not require independence — best-first ordering and the residual-mass certificate need only a joint distribution that can be evaluated and bounded, so a structured joint outage model is the natural extension — but nothing in this paper measures that case, and no result here should be read as evidence about correlated outages.

IX. Conclusion

On a four-system ladder with complete certified references and 328 matched-budget logs, probability-ordered enumeration's watchlist-coverage advantage is a property of the watchlist's density in the probable lattice — $\rho = |W|/E_{100}$, 0.44–0.89 on the replicas against 0.0009–0.0081 on the real system: it holds on every cell of congested replicas up to 248 components (leading 0.80 vs 0.34 where all methods are budget-starved) and inverts on the real 73-bus RTS-GMLC at $p_{\min} = 10^{-7}$, where a probability-times-severity screen reaches 90% coverage in 4,453 evaluations against 422,823 and subset simulation matches enumeration at budget. Enumeration remains the only certified method and the only one never below 0.79. Fixed-inflation importance sampling collapses with component count. The DC watchlist survives AC exactly on active power at $\times 1$ and loses a tenth of sampled states on the real system. Every artifact is committed and test-pinned; the full-system reference is reproduced to the last digit by an independent implementation.

X. Reproducibility

Everything is committed in the study directory and test-pinned:

- artifacts: `sweep_<system>.json`, `logs_<system>.json`, `frontier.json`, `crossover.json`, `ac_spike.json`, `eval_rates.json`, `snapshots_full.json`, `mc_reference_results_rts3area.json`, `crossval_full.json`;
- data: the full 3-area RTS-GMLC pull under `references/sienna/data/input_full` and `timeseries_full` with provenance READMEs and the extractor `references/sienna/scripts/extract_rts_gmlc.py`;
- drivers: `cras_run.py sweep|logs|frontier`, `cras_ac.py -spike`, `cras_paper.py`;
- plugin surface: `plugins.reliability` (`contingency_eval`, `rare_event`, `screening`, `replication`) and `plugins.power_flow` (`islanding`, `overloads`, `ac_lift`, `ac_loading`);
- tests: `tests/plugins/reliability`, `tests/plugins/power_flow`, `tests/case_studies/foundation_power_flow/contingency_risk_at_scale`;
- independent reference: `references/sienna/scripts/run_contingency_risk.jl -system full`.

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] 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.

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

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] H. Hu, W. Wan, F. Chen, X. Liu, B. Yu, and K. Zhao, "Critical states identification in power system via lattice partition and its application in reliability assessment," 2025, arXiv:2510.09420.

[10] 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.

[11] N. Christianson, W. Cui, S. Low, W. Yang, and B. Zhang, "Fast and reliable $n-k$ contingency screening with input-convex neural networks," in *Proceedings of the 7th Annual Learning for Dynamics and Control Conference*, ser. Proceedings of Machine Learning Research, vol. 283. PMLR, 2025, pp. 527–539.

[12] 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.

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

[14] L. Mai, C. Xiao, and Y. Weng, "Scalable and reliable state-aware inference of high-impact $n - k$ contingencies," 2026, arXiv:2602.09461.

[15] A. M. Nakiganda and S. Chatzivasileiadis, "Graph neural networks for fast contingency analysis of power systems," 2023, arXiv:2310.04213.

[16] J. Liu, X. Wang, and X. Wang, "Efficient sampling and sensitivity analysis of rare transient instability events via subset simulation," 2025, arXiv:2503.02858; to appear, IEEE PES General Meeting 2025.

[17] J. Guo, F. Liu, J. Wang, J. Lin, and S. Mei, "Towards high-efficiency cascading outage simulation and analysis in power systems: A sequential importance sampling approach," 2016, arXiv:1611.00582.

[18] K. Zhou, I. Dobson, and Z. Wang, "The most frequent $n-k$ line outages occur in motifs that can improve contingency selection," *IEEE Transactions on Power Systems*, vol. 39, no. 1, pp. 1785–1796, 2024.

[19] 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.

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

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

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