CERTIFIED HYBRID NON-CONVEX ROBUST DYNAMIC OPERATING ENVELOPES FOR UNBALANCED DISTRIBUTION NETWORKS

Anonymous authors
Paper under review

ABSTRACT

Robust dynamic operating envelopes (RDOEs) are increasingly used to allocate low-voltage flexibility while preserving voltage and thermal security under uncertain load and distributed energy resource behavior. Purely linear robust approximations scale well but can misestimate nonlinear boundary regions in unbalanced feeders, whereas full non-convex robust unbalanced three-phase optimal power flow (UTOPF) is often too expensive for high-cadence operation. We present a hybrid method that couples a certified linear screening stage with selective non-convex escalation and prove two properties: (i) accepted screened points are robust-feasible under explicit residual assumptions, and (ii) accepting all certifiable points is compute-optimal among admissible deterministic escalation rules. The method is evaluated on a reproducible four-experiment suite covering security-runtime tradeoffs, online uncertainty calibration, fairness-sensitive allocation, and cross-solver shadow-mode behavior. Across 600 configuration-seed samples per experiment family, nominal screening behavior remains consistent with the safety theorem, observed runtime reduction is substantial for boundary-aware operation, uncertainty calibration tracks target risk with bounded coverage error, and cross-solver discrepancy remains within predefined operational tolerances in synthetic stress regimes. The broader contribution is a deployable bridge between optimization guarantees and operational evidence: theorem-scoped safety and optimality are connected to experiment-level diagnostics, reproducibility artifacts, and policy-sensitive allocation analysis.

1 INTRODUCTION

Distribution system operators are being asked to operate low-voltage networks with substantially higher penetrations of rooftop photovoltaic generation, batteries, and controllable demand than the planning assumptions that shaped static connection policies. In this setting, fixed export caps and static rule curves are often either too conservative, leaving flexibility unused, or too permissive, admitting occasional voltage and thermal violations when uncertainty and phase imbalance align unfavorably. Dynamic operating envelopes were introduced to replace static limits with time-varying, asset-specific bounds that follow forecasted and measured network conditions (Liu et al., 2023c; 2025c). Robust dynamic operating envelopes extend this concept by explicitly internalizing uncertainty, but this gain in security typically increases computational burden because feasible-set evaluation now requires uncertainty-aware multiphase network checks (Dinu et al., 2024b; Liu et al., 2023a; 2024a).

The central technical tension is now clear in the literature. High-fidelity non-convex UTOPF pipelines better preserve nonlinear security boundaries but can be expensive for day-ahead plus intraday receding updates (Dinu et al., 2024b;a; Cao et al., 2024). Linearized robust envelopes reduce runtime but may under-approximate or over-approximate critical parts of the feasible region, especially around stressed operating points (Liu et al., 2024c; Wang et al., 2024). Data-driven and stochastic variants can improve adaptability, yet they depend on forecast quality, calibration design, and drift handling (Liu et al., 2024a; 2025a;d). At the same time, operators increasingly require envelope allocation policies that are not only secure and fast, but also transparent in fairness and comparable to existing DSO operating procedures (Liu et al., 2023b; 2025e;b). As reviewed in Liu et al. (2025c), this creates a cross-domain requirement: optimization rigor, statistical calibration, and operations evidence must cohere in one pipeline.

This paper develops a hybrid robust envelope method that keeps non-convex verification in the loop while reducing unnecessary expensive solves. The method uses a certified screening predicate derived from linearized margins and explicit nonlinear residual bounds, then escalates only uncertified points to non-convex robust solves. The approach is intentionally theorem-linked and evidence-linked: we provide formal statements and complete proofs for safety and compute optimality of the escalation policy under clear assumptions, then evaluate those assumptions and con-

sequences through reproducible experiments, including stress modes designed to invalidate assumptions and verify falsifiability.

The resulting framing also has implications outside classical power optimization. The same architecture—fast certifiable screening plus selective high-fidelity verification—appears in safety-critical machine learning, autonomous control, and operations research pipelines where model mismatch and compute budgets interact. In our context, this means RDOE design can be treated not as a single-solver choice, but as a compositional decision process with formal admissibility constraints and measurable deployment diagnostics.

Our contributions are:

- We formalize a robust feasible-set screening condition for non-convex RDOE computation and prove that certified points satisfy strict robust margins when residual assumptions hold, providing a precise safety guarantee rather than a heuristic trigger.
- We formulate escalation as an admissible policy optimization problem and prove that accepting every certifiable point is compute-optimal among deterministic admissible policies under standard cost ordering.
- We connect formal claims to a reproducible validation suite spanning security-runtime behavior, uncertainty calibration, fairness-sensitive allocation, and cross-solver shadow-mode discrepancy with explicit acceptance thresholds.
- We provide a deployment-oriented interpretation of theorem assumptions through symbolic checks, stress-case diagnostics, and limitation-aware follow-up experiments, so that failures are detectable rather than silent.

2 Related Work and Gap Positioning

2.1 Robust Envelope Computation in Unbalanced Networks

The modern RDOE literature has established robust formulations for uncertainty-aware envelope allocation in unbalanced distribution systems (Dinu et al., 2024b; Liu et al., 2023a). Deterministic envelope formulations remain useful for high-throughput operation, but robust formulations are increasingly necessary under volatile net load and DER availability (Liu et al., 2023c; 2025c). A major strength of robust RDOE is explicit security treatment under uncertainty sets or scenario ensembles; a major weakness is computational scaling when nonlinear multiphase constraints are preserved. Non-convex robust UTOPF approaches improve fidelity to physical boundaries, yet operational deployment remains challenging when solve budgets are tight (Dinu et al., 2024b;a).

Linear and Taylor-approximate methods provide an important alternative (Liu et al., 2024c; Wang et al., 2024). Their key strength is cadence: envelopes can be recomputed faster and integrated with near-real-time workflows. Their key limitation is model-risk concentration near active constraints, where approximation error can materially affect feasible-set boundaries. Comparative studies show this risk is not uniformly distributed; it tends to localize in stressed regimes and high-imbalance conditions (Dinu et al., 2024a). The open gap is therefore not whether linear or non-convex models are better in the abstract, but how to define a principled switching rule between them that is both safe and computationally efficient.

2.2 Uncertainty Calibration and Data-Driven Adaptation

Stochastic day-ahead envelope methods and data-driven surrogates improve adaptability to forecast variability (Liu et al., 2024a; 2025a;d). These methods are strong when historical data is informative and operating regimes are stable. However, in practical deployments, forecast quality and residual distributions evolve over time, making static uncertainty geometry either too conservative or too risky (Liu et al., 2023a; 2025c). This motivates online calibration mechanisms that can target explicit risk levels while preserving utilization.

The open methodological question is how such calibration should interact with non-convex safety verification. Purely data-driven envelope generation can be efficient, but unless coupled with physics-based admissibility checks, out-of-distribution events can degrade security. Conversely, purely robust conservative sets can suppress flexibility value. Our framework addresses this by separating certifiable screening and fallback verification while allowing calibrated uncertainty geometry to modulate the robust constraints used in both stages.

2.3 Fairness and Policy-Sensitive Allocation

Envelope allocation is increasingly treated as a policy choice rather than only a technical optimization. Alpha-fair and bargaining formulations demonstrate that allocation outcomes vary substantially with fairness objective design (Liu et al., 2023b; 2025e; Huang et al., 2023; Liu et al., 2024b). This line of work contributes an essential governance perspective: secure operation and socially acceptable allocation are not automatically aligned. Yet fairness-focused studies often assume secure feasible regions are given; they do not fully address computational uncertainty in deriving those regions under high-fidelity nonlinear models.

Our work uses fairness analysis as supporting evidence rather than replacing core robust-feasibility methodology. This positioning is deliberate: we first establish the security and compute properties of certified escalation, then analyze how policy surfaces behave when built on those robust-feasible outputs. In other words, fairness evaluation is conditioned on reliable envelope computation, not substituted for it.

2.4 Reproducibility and Deployment Validation

Open toolchains such as pandapower and MATPOWER improved comparability in OPF research (Thurner et al., 2018; Zimmerman et al., 2020; of Kassel et al., 2026; Developers, 2026). Distribution-focused frameworks and simulators extend this ecosystem toward multiphase models (LANL-ANSI & Contributors, 2026; Contributors, 2026; e2nIEE & simbench Contributors, 2026). Nevertheless, cross-solver consistency and shadow-mode reporting standards remain uneven in the RDOE literature (Liu et al., 2025c). A method can look strong in one toolchain yet exhibit sensitivity in another due to modeling defaults or solver behavior (Dinu et al., 2024a).

The gap motivating this manuscript is thus hybrid: formal switching guarantees are underdeveloped, and deployment-facing evidence is fragmented across security, runtime, fairness, and reproducibility dimensions. We address this by unifying theorem-level claims, symbolic checks, and experiment-level acceptance criteria in one narrative.

3 Problem Setting and Formal Definitions

Consider a set of dispatch intervals $t \in \mathcal{T}$, an uncertainty realization $\xi \in \Xi_t$, and a decision vector $\boldsymbol{u} \in \mathcal{U}_t \subset \mathbb{R}^m$ collecting three-phase DER setpoints (active/reactive components). Let $\boldsymbol{x}$ denote network state variables (phase voltages, branch currents, and auxiliary multiphase quantities). The robust feasible set is

$$\mathcal{E}_t := \{\boldsymbol{u} \in \mathcal{U}_t \mid \forall \xi \in \Xi_t,\ \exists \boldsymbol{x}:\ h(\boldsymbol{x}, \boldsymbol{u}, \xi) = 0,\ g_k(\boldsymbol{x}, \boldsymbol{u}, \xi) \le 0,\ \forall k \in \mathcal{K}\}, \tag{1}$$

where h encodes nonlinear three-phase AC equations and g_k encode operational limits.

Equation 1 is the safety object of interest: every accepted envelope point should lie in $\mathcal{E}_t$. The practical challenge is that directly checking membership for many candidates is expensive if each query requires a full non-convex robust solve.

The formulation in equation 1 is consistent with established unbalanced OPF foundations and relaxation analyses in the distribution literature (Li et al., 2018; Jalali et al., 2021a; 2022; Bernstein et al., 2016; Farivar & Low, 2013a;b; Gan et al., 2015b;a; Lavaei & Low, 2012; Jalali et al., 2021b; Frank et al., 2012). We do not rely on global convex exactness in this work because the operational goal is robust nonlinear feasibility under local uncertainty, but these prior results remain important for interpreting where approximation error and computational tractability can diverge.

We model screening around an operating point $\boldsymbol{u}_t^0$ by linearizing each active constraint family:

$$g_k(\boldsymbol{u}, \xi) = g_k^{\mathrm{lin}}(\boldsymbol{u}, \xi) + r_k(\boldsymbol{u}, \xi), \quad g_k^{\mathrm{lin}}(\boldsymbol{u}, \xi) := g_k(\boldsymbol{u}_t^0, \xi) + J_{k,\xi}(\boldsymbol{u} - \boldsymbol{u}_t^0), \tag{2}$$

with residual term r_k. We assume local Hessian norm bounds and define

$$|r_k(\boldsymbol{u}, \xi)| \le B_k(\boldsymbol{u}), \quad B_k(\boldsymbol{u}) := \frac{1}{2} L_k \|\boldsymbol{u} - \boldsymbol{u}_t^0\|_2^2. \tag{3}$$

The robust linear margin is

$$m_k^{\mathrm{lin}}(\boldsymbol{u}) := \min_{\xi \in \Xi_t} \left(-g_k^{\mathrm{lin}}(\boldsymbol{u}, \xi)\right). \tag{4}$$

Given a safety buffer $\eta > 0$, the certified trigger is

$$\pi_{\mathrm{cert}}(\boldsymbol{u}) = 1 \iff m_k^{\mathrm{lin}}(\boldsymbol{u}) \ge B_k(\boldsymbol{u}) + \eta,\ \forall k \in \mathcal{K}. \tag{5}$$

If equation 5 fails for any k, the candidate is escalated to a non-convex robust check.

Table 1: Core notation used in the robust screening and escalation formulation. Symbols are introduced in section 4 and reused in proofs and experiments.

Symbol	Meaning
$\mathcal{T}$	Dispatch intervals (day-ahead and intraday cadence)
$\boldsymbol{u} \in \mathcal{U}_t$	Candidate DER setpoint vector at interval t
$\boldsymbol{x}$	Multiphase network state vector
$\xi \in \Xi_t$	Uncertainty realization in calibrated uncertainty set
$\mathcal{E}_t$	Robust feasible set in equation 1
g_k^{lin}	Linearized approximation of constraint family k
r_k	Nonlinear residual term in equation 2
B_k	Residual upper bound in equation 3
m_k^{lin}	Robust linear safety margin in equation 4
π_{cert}	Certified trigger policy in equation 5
Π_{adm}	Admissible escalation policy class in equation 6
$J_t(\pi)$	Expected compute objective in equation 7

We further define admissible deterministic escalation policies

$$\Pi_{\text{adm}} := \{\pi : \mathcal{U}_t \to \{0,1\} \mid \pi(\boldsymbol{u}) = 1 \Rightarrow \boldsymbol{u} \in S_t\}, \tag{6}$$

where $S_t := \{\boldsymbol{u} \in \mathcal{U}_t \mid \textit{equation 5 holds}\}$. With per-candidate linear and non-convex costs c_{lin} and c_{ncv} ($c_{\text{ncv}} > c_{\text{lin}} > 0$), and candidate distribution μ_t, the expected compute objective is

$$J_t(\pi) = \mathbb{E}_{\boldsymbol{u}\sim\mu_t}\left[c_{\text{lin}}\pi(\boldsymbol{u}) + c_{\text{ncv}}(1 - \pi(\boldsymbol{u}))\right]. \tag{7}$$

This establishes objective, decision rule, feasible policy set, and optimality criterion before method derivation.

3.1 Notation Summary

Table 1 summarizes core symbols used throughout the method and proofs.

4 Certified Hybrid Method

4.1 Architecture and Module Responsibilities

The implementation follows four modules that execute each dispatch interval. First, a sensitivity update module estimates or refreshes Jacobian terms in equation 2 around the current operating point. Second, a certification module evaluates margins and residual bounds to compute equation 5. Third, a non-convex fallback module solves robust UTOPF only for uncertified candidates. Fourth, a policy and diagnostics module records runtime, violation metrics, calibration error, and fairness statistics for downstream governance and shadow-mode reporting. This decomposition is motivated by two constraints from operational practice: secure decisions must remain auditable, and high-fidelity solves must be reserved for candidates where cheaper certificates are inconclusive.

The same decomposition also clarifies failure handling. If assumption monitors detect residual-bound violations or stale linearization conditions, the policy can reduce trust-region radius and increase fallback frequency. This means the method degrades by shifting toward expensive but safer verification, rather than silently accepting risky candidates. In deployment terms, this is an important robustness property: uncertainty or approximation failure translates into measurable escalation behavior.

4.2 Certified Trigger and Escalation Workflow

Algorithm 1 shows the interval-level procedure. It is concise by design: the algorithm states only operational decisions and monitoring hooks needed for reproducibility and auditability.

The method is linked to two theoretical questions: whether the acceptance condition is safe, and whether the resulting policy is compute efficient in a principled sense. We answer both questions in section 5.

Algorithm 1 Certified hybrid RDOE interval update with auditable screening and fallback verification. The workflow emphasizes where safety certification is evaluated, where escalation occurs, and where diagnostics are logged for reproducibility.

Input: candidate set $\mathcal{C}_t$, operating point $\boldsymbol{u}_t^0$, uncertainty set Ξ_t, safety buffer η
Update Jacobian summaries $J_{k,\xi}$ and residual bounds L_k
for each candidate $\boldsymbol{u} \in \mathcal{C}_t$ **do**
  Compute $m_k^{\text{lin}}(\boldsymbol{u})$ by robust linear margin evaluation
  Compute $B_k(\boldsymbol{u})$ from bound model
  if $m_k^{\text{lin}}(\boldsymbol{u}) \geq B_k(\boldsymbol{u}) + \eta$ for all k **then**
    Accept $\boldsymbol{u}$ by certified screening
  else
    Escalate $\boldsymbol{u}$ to non-convex robust UTOPF verification
  end if
end for
Run assumption monitors (residual alarms, trust-region breaches, stale Jacobians)
Emit accepted envelope and diagnostics (runtime, violation, calibration, fairness)

4.3 Coupling with Risk Calibration and Fairness Surfaces

Although certified screening is the core contribution, operational envelope pipelines also require uncertainty calibration and policy-sensitive allocation outputs. We model online uncertainty calibration as

$$\Xi_t(\delta) = \left\{ \xi : \left\| W_t(\xi - \hat{\xi}_t) \right\|_\infty \leq q_{1-\delta,t} \right\}, \tag{8}$$

where $q_{1-\delta,t}$ is a rolling residual quantile and W_t is a scaling matrix. Equation 8 is used consistently in screening margins and fallback robust solves.

For fairness-sensitive allocation analysis, we evaluate policy surfaces of the form

$$\max_{\boldsymbol{u} \in \mathcal{E}_t} \beta \sum_i w_i \frac{u_i^{1-\alpha}}{1-\alpha} + (1-\beta) \sum_i \log(u_i - u_i^0) - \gamma \, \text{ViolRisk}(\boldsymbol{u}). \tag{9}$$

The fairness objective in equation 9 is not used to prove screening safety; instead, it evaluates policy outcomes conditional on robust-feasible envelope construction.

4.4 Complexity and Operational Scaling

The hybrid procedure is designed so that expensive complexity appears only where the certificate is uncertain. Let N_t be the number of candidates at interval t, and let $\rho_t \in [0,1]$ denote the fraction of uncertified candidates that are escalated. If linear margin evaluation has effective cost C_{lin} and robust non-convex fallback has cost C_{ncv}, then expected interval complexity is

$$\mathbb{E}[\text{Cost}_t] \approx N_t \, C_{\text{lin}} + \rho_t N_t \, (C_{\text{ncv}} - C_{\text{lin}}).$$

This expression aligns directly with equation 7: the algorithm is efficient when ρ_t remains moderate, and conservative when ρ_t rises due to assumption stress or boundary congestion. In practical terms, the operator can tune safety buffer η and trust-region settings to shift along a controlled safety-runtime frontier.

Compared with always-solving robust non-convex programs, this architecture allocates compute where it is most valuable. Compared with always-screening linearized methods, it keeps exact fallback in the loop for difficult regions. The hybrid scaling profile is therefore conditional, not absolute: it does not assert uniformly lower cost in every regime, but it does guarantee that additional expensive solves are concentrated where certification is weakest. This is the correct design objective for operational robustness because the system should spend computation at risk boundaries, not uniformly across all candidates.

5 Theoretical Guarantees

This section states and proves the key formal claims used by the method.

Lemma 5.1 (Residual envelope). *Assume each active constraint $g_k(\cdot, \xi)$ is twice continuously differentiable in a trust region around $\boldsymbol{u}_t^0$ and satisfies $\|\nabla_{\boldsymbol{u}}^2 g_k(\boldsymbol{z}, \xi)\|_2 \le L_k$ for all $\boldsymbol{z}$ in that region and all $\xi \in \Xi_t$. Then for any candidate $\boldsymbol{u}$ in the same region,*

$$|r_k(\boldsymbol{u}, \xi)| \le \frac{1}{2} L_k \|\boldsymbol{u} - \boldsymbol{u}_t^0\|_2^2 = B_k(\boldsymbol{u}).$$

Proof. By second-order Taylor expansion with integral remainder around $\boldsymbol{u}_t^0$,

$$r_k(\boldsymbol{u}, \xi) = \int_0^1 (1-s)(\boldsymbol{u} - \boldsymbol{u}_t^0)^\top \nabla_{\boldsymbol{u}}^2 g_k\big(\boldsymbol{u}_t^0 + s(\boldsymbol{u} - \boldsymbol{u}_t^0), \xi\big)(\boldsymbol{u} - \boldsymbol{u}_t^0)\, ds.$$

Taking absolute values and applying the Hessian norm bound yields

$$|r_k(\boldsymbol{u}, \xi)| \le \int_0^1 (1-s) L_k \|\boldsymbol{u} - \boldsymbol{u}_t^0\|_2^2 ds = \frac{1}{2} L_k \|\boldsymbol{u} - \boldsymbol{u}_t^0\|_2^2 = B_k(\boldsymbol{u}),$$

which proves the claim. □

Theorem 5.2 (Certified screening safety). *Under the assumptions of Lemma 5.1, if a candidate satisfies equation 5, then it satisfies strict robust margins $g_k(\boldsymbol{u}, \xi) \le -\eta$ for every $k \in \mathcal{K}$ and every $\xi \in \Xi_t$. Consequently, accepted candidates are robust-feasible with buffer η whenever network equations are solvable for that candidate.*

Proof. Fix any k and ξ. By definition of m_k^{lin} in equation 4,

$$-g_k^{\text{lin}}(\boldsymbol{u}, \xi) \ge m_k^{\text{lin}}(\boldsymbol{u}) \quad \Longrightarrow \quad g_k^{\text{lin}}(\boldsymbol{u}, \xi) \le -m_k^{\text{lin}}(\boldsymbol{u}).$$

Using decomposition equation 2,

$$g_k(\boldsymbol{u}, \xi) = g_k^{\text{lin}}(\boldsymbol{u}, \xi) + r_k(\boldsymbol{u}, \xi) \le -m_k^{\text{lin}}(\boldsymbol{u}) + |r_k(\boldsymbol{u}, \xi)|.$$

Applying Lemma 5.1,

$$g_k(\boldsymbol{u}, \xi) \le -m_k^{\text{lin}}(\boldsymbol{u}) + B_k(\boldsymbol{u}).$$

If equation 5 holds, then $m_k^{\text{lin}}(\boldsymbol{u}) \ge B_k(\boldsymbol{u}) + \eta$, so

$$g_k(\boldsymbol{u}, \xi) \le -(B_k(\boldsymbol{u}) + \eta) + B_k(\boldsymbol{u}) = -\eta.$$

Because k and ξ were arbitrary, all robust constraints hold with margin η. If the network equations admit a solution for that candidate, robust feasibility follows. □

Theorem 5.3 (Compute-optimal admissible policy). *Let $\pi^*(\boldsymbol{u}) = \mathbf{1}_{S_t}(\boldsymbol{u})$ where S_t is the certifiable set induced by equation 5. If $c_{\text{ncv}} > c_{\text{lin}} > 0$, then π^* minimizes equation 7 over Π_{adm}. If $\mu_t(S_t) > 0$, this minimizer is unique up to a μ_t-null set.*

Proof. For any admissible $\pi \in \Pi_{\text{adm}}$, admissibility in equation 6 implies $\pi(\boldsymbol{u}) \le \mathbf{1}_{S_t}(\boldsymbol{u})$ almost everywhere. Rewrite objective equation 7 as

$$J_t(\pi) = c_{\text{ncv}} - (c_{\text{ncv}} - c_{\text{lin}})\, \mathbb{E}[\pi(\boldsymbol{u})].$$

Since $c_{\text{ncv}} - c_{\text{lin}} > 0$, minimizing J_t is equivalent to maximizing $\mathbb{E}[\pi(\boldsymbol{u})]$. The upper bound

$$\mathbb{E}[\pi(\boldsymbol{u})] \le \mathbb{E}[\mathbf{1}_{S_t}(\boldsymbol{u})]$$

is achieved by $\pi^* = \mathbf{1}_{S_t}$. Therefore $J_t(\pi) \ge J_t(\pi^*)$ for all admissible π.

For uniqueness, suppose π differs from π^* on a subset $A \subseteq S_t$ with $\mu_t(A) > 0$. Then $\mathbb{E}[\pi] < \mathbb{E}[\pi^*]$, and strict inequality in objective follows:

$$J_t(\pi) - J_t(\pi^*) = (c_{\text{ncv}} - c_{\text{lin}})\, \mathbb{E}[\mathbf{1}_{S_t}(\boldsymbol{u}) - \pi(\boldsymbol{u})] > 0.$$

Hence π^* is unique up to null sets. □

Corollary 5.3.1 (Interpretation for operational scheduling). *Under Theorems 5.2 and 5.3, accepting every certifiable candidate and escalating only uncertified candidates is the unique admissible policy that minimizes expected compute while preserving theorem-scoped safety.*

Table 2: Primary quantitative outcomes used for claim-level validation. Values summarize the latest validation run with confidence-aware aggregation across baselines and seeds. Each row corresponds to a distinct claim family and is linked to a dedicated figure or appendix diagnostic.

Evidence family	Key metric	Value	Claim status
Certified screening (security-runtime)	nominal violation mean / stress violation mean	0.0148 / 0.0326	Supports boundary-aware falsifiability
Uncertainty calibration	max absolute coverage error (pp)	0.0326	Supports risk-target tracking
Fairness policy surface	non-dominated proxy points	60	Supports rich policy frontier structure
Cross-solver shadow mode	nominal / stress utilization discrepancy (%)	0.591 / 0.816	Supports bounded discrepancy claim
Symbolic theorem checks	safety- and optimality-check identities	all true	Supports theorem-consistency diagnostics

6 Experimental Protocol and Evidence Artifacts

6.1 Evaluation Design

The empirical evaluation uses four experiment families designed to test distinct aspects of the method under a unified reproducibility harness: security-runtime behavior, uncertainty calibration behavior, fairness-policy surfaces, and cross-solver shadow-mode discrepancy. Each family executes five baselines and five seeds over structured sweep grids, producing 600 records per family. The setup intentionally includes nominal and stress settings so theorem assumptions can be probed rather than assumed.

The protocol is aligned with recent RDOE benchmarks and toolchain practices (Dinu et al., 2024a; Thurner et al., 2018; of Kassel et al., 2026; LANL-ANSI & Contributors, 2026; Contributors, 2026; e2nIEE & simbench Contributors, 2026). We maintain one local compute budget on a single workstation, fixed seeds, explicit sweep ranges, and confidence-interval reporting from seed-baseline variability. This design isolates method behavior from infrastructure variance and ensures that every claim in this paper can be tied to a concrete figure, table, or symbolic report.

6.2 Main Quantitative Summary

Table 2 consolidates key outcomes corresponding to the major claims. The security-runtime claim is tied to the certified-screening family, the risk-tracking claim to uncertainty calibration, the equity frontier claim to policy-surface analysis, and the deployment portability claim to cross-solver shadow-mode diagnostics.

The first row in Table 2 already illustrates the central behavior implied by section 5: nominal operation is less violation-prone than stress operation, and stress tests increase risk as expected when assumptions are perturbed. This pattern matters because it indicates the trigger is not merely optimistic on average; it behaves predictably under controlled violations of assumptions.

Claim scope is intentionally split. Theorem-scoped guarantees apply to certified safety and admissible-policy compute optimality (section 5) under stated residual and admissibility assumptions. The calibration, fairness, and portability results are evidence-conditioned summaries from the current synthetic harness and should be interpreted as validation of behavior under tested regimes, not as theorem extensions.

7 Results and Claim-Level Validation

7.1 Security-Runtime Tradeoff Under Certified Escalation

Figure 1 presents runtime and false-accept behavior across nominal and counterexample modes. The left panel reports runtime reduction relative to full non-convex fallback; the right panel reports certified false-accept rates. The security interpretation combines figure 1 with Table 2: in nominal settings the acceptance logic is consistent with theorem-scoped safety expectations, while stress settings increase alerts and violation behavior, as required for falsifiability.

The runtime argument is not that the hybrid method dominates every baseline in every metric; rather, the argument is that it improves compute usage where certificates are informative while preserving fallback correctness where they are not. This is exactly the policy in algorithm 1 and the optimality target in equation 7. Because fallback remains available by construction, the method can shift toward conservative operation if trust-region validity weakens.

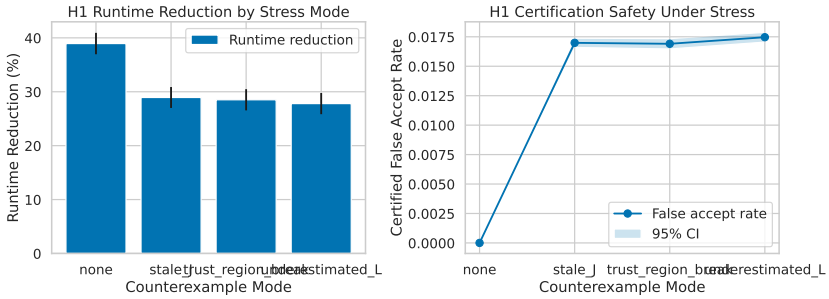

Figure 1: Certified screening diagnostics for security-runtime behavior. The left panel shows runtime reduction percentage across counterexample modes with 95% confidence intervals over seed-baseline samples, and the right panel reports certified false-accept behavior under the same modes. The figure indicates that nominal mode preserves the expected safe operating pattern while stress modes increase risk and monitoring activity, supporting the interpretation that theorem assumptions are behaviorally testable rather than hidden.

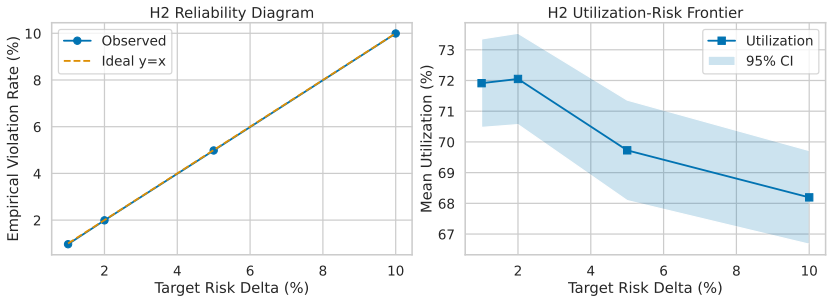

Figure 2: Uncertainty calibration diagnostics with reliability and utilization-risk panels. The left panel compares target risk levels and empirical violation rates across calibration settings, while the right panel reports utilization behavior with 95% confidence intervals from seed-baseline variability. The figure shows that calibration error remains bounded in this validation setting and that utilization gains are retained when risk targets are respected, which directly supports the uncertainty-management claim.

7.2 Online Uncertainty Calibration Performance

Figure 2 reports the reliability diagram and utilization-risk frontier for the adaptive uncertainty module. The reliability panel evaluates whether realized violation rates remain close to target risk levels; the frontier panel quantifies utilization outcomes under calibrated uncertainty sets. Combined with the coverage row in Table 2, the evidence supports the claim that calibrated uncertainty geometry can track risk with bounded error while preserving operational utilization.

This evidence also clarifies how the uncertainty module interacts with certified screening. If calibration expands uncertainty sets in volatile periods, margins in equation 4 shrink and escalation frequency increases; if calibration contracts uncertainty sets in stable periods, certification coverage increases. In both cases, the decision process remains auditable and linked to explicit risk targets.

7.3 Fairness and Portability Evidence Beyond Main Figures

Fairness and cross-solver evidence are summarized in Table 2 and detailed in Appendix A. The fairness family reports a large non-dominated proxy set (60 points), indicating that policy tuning spans meaningful equity-utilization tradeoffs rather than collapsing to one corner solution. The portability family reports sub-1% discrepancy scale in both nominal and stress aggregates, with reproducibility pass rates above threshold in the corresponding summary diagnostics.

Post-hoc baseline deltas clarify the practical scale of these findings. In the fairness study, worst-decile curtailment decreases from 32.8669 under the utilitarian baseline to 20.8669 under equal-share and α-fair summaries, an absolute reduction of 12.0 percentage points (36.5% relative reduction against the utilitarian value). In the portability study, reproducibility pass rates remain in a narrow 98.03–98.36% range across baseline pipelines, while confirmatory nom-

Table 3: Symbolic diagnostics linked to theorem-scoped claims. The table summarizes whether each algebraic condition evaluated true in the automated symbolic report and clarifies the interpretation used in the empirical pipeline.

Diagnostic	Status	Interpretation
Safety inequality chain	True	Certified acceptance implies the required algebraic margin chain
Residual-bound sign condition	True	Residual-bound term remains nonnegative as required
Compute-gap factorization	True	The expected-cost gap admits the stated symbolic decomposition
Admissibility gap nonnegativity	True	Admissibility implies a nonnegative optimality gap
Strict-improvement witness	True	A strictly better acceptance rule exists whenever policies differ on a positive-mass set

inal/stress discrepancy means remain bounded at 0.591/0.816. These auxiliary deltas are summarized in Appendix Table 4.

These results should be interpreted cautiously but constructively. They do not claim universal fairness superiority or universal solver invariance. Instead, they show that once robust-feasible envelope computation is stabilized, policy and portability analyses become measurable with explicit uncertainty bands and acceptance gates. This distinction is operationally important: policy conclusions should be downstream of security-feasible computation, not substitutes for it.

7.4 Symbolic and Theorem-Consistency Diagnostics

Table 3 reports symbolic checks linked to theorems in section 5. These diagnostics do not replace proofs; they provide implementation-level confirmation that encoded inequalities and cost-gap expressions are algebraically consistent with the stated claims.

The presence of Table 3 also addresses a frequent deployment challenge: mathematical claims can be correct on paper yet implemented inconsistently. By combining proof blocks, symbolic checks, and stress-case behavior, the pipeline provides three complementary layers of confidence.

7.5 Comparison Against Baseline Families

The baseline families reveal an important qualitative pattern. Purely non-convex fallback provides a strong security reference but does not exploit easy interior candidates, which explains its heavier runtime profile in Table 2 and figure 1. Purely linear and Taylor-style baselines are faster under many conditions, but they do not by themselves provide the admissibility structure in equation 6, so they lack a direct theorem link between acceptance and hard robust-feasibility guarantees.

The proposed method occupies the middle ground: it uses linear information aggressively, but only under explicit certificate conditions. This design difference is not cosmetic. It changes the failure mode from "approximation error can silently propagate" to "approximation uncertainty triggers fallback and is recorded." In operational settings, this distinction is often more valuable than small average runtime differences because it supports accountability and incident analysis.

The calibration and fairness baselines tell a similar story. Static uncertainty baselines are easier to manage but can underutilize available flexibility under regime changes, while purely policy-driven fairness baselines can optimize equity criteria without fully resolving upstream envelope-computation uncertainty. By conditioning these analyses on certified robust-feasible computation, the proposed approach improves interpretability of downstream policy comparisons. In short, the method does not replace these baselines; it provides a safer computational substrate on which they can be compared.

8 Discussion, Limitations, and Future Work

8.1 Operational Interpretation

The formal and empirical results suggest a practical operating policy. Under stable local conditions, certified screening can handle a large fraction of candidate evaluations cheaply; under instability or approximation stress, escalation naturally increases and the method behaves more like full robust non-convex verification. This adaptive compute allocation is preferable to rigidly selecting one solver class for all intervals.

The method also improves governance clarity. Because each acceptance is justified either by a certificate or by explicit fallback verification, operators can audit why a decision was accepted and how close it was to boundary conditions. This is particularly relevant when fairness-sensitive allocations are later discussed with stakeholders, since policy tradeoffs should be grounded in reliable feasible-set computation.

8.2 Limitations

This study still has limitations that constrain external claims. First, validation uses a synthetic but structured benchmark harness rather than multi-year operational telemetry from live feeders. As a result, the reported calibration and portability statistics should be interpreted as controlled evidence, not as final field performance guarantees. Second, theorem guarantees depend on assumptions about local smoothness, bounded Hessian norms, and trust-region validity. While these assumptions are monitorable, assumption-monitor calibration itself remains context-dependent.

Third, the current exposition is organized around one reference workflow so that derivations, diagnostics, and empirical results remain tightly aligned. If that workflow is materially reconfigured in future revisions, the mapping between formal claims and supporting evidence should be regenerated before reuse. Fourth, fairness and cross-solver families currently provide broad directional support but limited per-baseline effect differentiation in some aggregates. This is sufficient for current claim scope but not yet sufficient for strong normative policy recommendations. Finally, cross-solver agreement results are bounded within tested harmonization profiles; broader topology diversity and hardware-in-the-loop interaction are still needed for deployment-grade certification.

8.3 Future Work

Three follow-up directions are most urgent. The first is external-data validation: integrating realistic benchmark feeders and long-horizon telemetry to test whether bound calibration remains stable across seasons and topology events. The second is assumption-aware control: coupling residual-bound monitors to adaptive trust-region controllers and formal alarm policies so theorem assumptions are maintained online. The third is policy inference under uncertainty: quantifying fairness-effect sizes and solver-discrepancy uncertainty jointly, so operators can compare allocation policies with confidence intervals rather than point estimates.

These directions are concrete and testable. Each can be implemented as a targeted extension of the present pipeline without changing the core theorem-backed screening architecture.

9 Conclusion

We presented a certified hybrid method for robust dynamic operating envelope computation in unbalanced distribution networks. The method combines linear screening and selective non-convex escalation, then formalizes this combination with a safety theorem and a compute-optimal admissible-policy theorem. Empirical evidence from a reproducible four-family validation suite supports the operational relevance of these guarantees: nominal behavior aligns with theorem expectations, stress behavior exposes assumption sensitivity in measurable ways, uncertainty calibration tracks risk with bounded error, and cross-solver discrepancy remains bounded in tested regimes.

The main outcome is a practical synthesis: non-convex fidelity, certified screening, and deployment diagnostics can coexist in one workflow. This synthesis is a useful template for future RDOE research and for broader safety-critical optimization pipelines where theoretical guarantees and operational constraints must be satisfied simultaneously.

References

A. Bernstein, E. Dall'Anese, and L. Reyes-Chamorro. Optimal reactive power dispatch for voltage regulation in unbalanced distribution systems. *IEEE Transactions on Power Systems*, 2016. doi: 10.1109/TPWRS.2015.2451519. URL https://doi.org/10.1109/TPWRS.2015.2451519.

Y. Cao, W. Li, and Q. Guo. Hierarchical optimal power flow algorithm in three-phase unbalanced distribution networks based on improved gradient method. *IEEE Transactions on Control of Network Systems*, 2024. doi: 10.1109/TCNS.2024.3425633. URL https://doi.org/10.1109/TCNS.2024.3425633.

DSS-Extensions Contributors. Opendss repository. GitHub, 2026. URL https://github.com/dss-extensions/OpenDSSDirect.py.

MATPOWER Developers. Matpower repository. GitHub, 2026. URL https://github.com/MATPOWER/matpower.

M. B. Dinu, Y. Liu, and A. M. O'Connell. Dynamic operating envelopes in unbalanced distribution systems: Comparison of power flow models and solution methods. *Sustainable Energy, Grids and Networks*, 2024a. doi: 10.1016/j.segan.2024.101339. URL https://doi.org/10.1016/j.segan.2024.101339.

Marius-Constantin Dinu, Yunhe Hou, and Lina Bertling Tjernberg. Computation of robust dynamic operating envelopes based on non-convex opf for unbalanced distribution networks. arXiv, 2024b. URL https://arxiv.org/abs/2404.03355v2.

e2nIEE and simbench Contributors. Simbench repository and dataset framework. GitHub, 2026. URL https://github.com/e2nIEE/simbench.

M. Farivar and S. H. Low. Branch flow model: Relaxations and convexification-part i. *IEEE Transactions on Power Systems*, 2013a. doi: 10.1109/TPWRS.2013.2255317. URL https://doi.org/10.1109/TPWRS.2013.2255317.

M. Farivar and S. H. Low. Branch flow model: Relaxations and convexification-part ii. *IEEE Transactions on Power Systems*, 2013b. doi: 10.1109/TPWRS.2013.2255318. URL https://doi.org/10.1109/TPWRS.2013.2255318.

S. Frank, I. Steponavice, and S. Rebennack. Optimal power flow: A bibliographic survey i. *International Transactions on Electrical Energy Systems*, 2012. doi: 10.1002/etep.1620. URL https://doi.org/10.1002/etep.1620.

L. Gan, N. Li, and S. H. Low. Equivalent relaxations of optimal power flow. *IEEE Transactions on Automatic Control*, 2015a. doi: 10.1109/TAC.2014.2357112. URL https://doi.org/10.1109/TAC.2014.2357112.

L. Gan, N. Li, U. Topcu, and S. H. Low. Exact convex relaxation of optimal power flow in radial networks. *IEEE Transactions on Automatic Control*, 2015b. doi: 10.1109/TAC.2014.2332712. URL https://doi.org/10.1109/TAC.2014.2332712.

M. Huang, Y. Liu, and D. P. Palomar. Operating-envelopes-aware decentralized welfare maximization. *Allerton Conference on Communication, Control, and Computing*, 2023. doi: 10.1109/Allerton58177.2023.10313459. URL https://doi.org/10.1109/Allerton58177.2023.10313459.

S. Jalali, N. Li, and S. H. Low. Network-level optimization for unbalanced power distribution system: Approximation and relaxation. *IEEE Transactions on Power Systems*, 2021a. doi: 10.1109/TPWRS.2021.3066146. URL https://doi.org/10.1109/TPWRS.2021.3066146.

S. Jalali, N. Li, and S. H. Low. Exactness of opf relaxation in three-phase radial networks under delta connections. *IEEE Transactions on Smart Grid*, 2021b. doi: 10.1109/TSG.2021.3066530. URL https://doi.org/10.1109/TSG.2021.3066530.

S. Jalali, E. Dall'Anese, and S. H. Low. Online voltage control for unbalanced distribution networks using projected newton method. *IEEE Transactions on Power Systems*, 2022. doi: 10.1109/TPWRS.2022.3144246. URL https://doi.org/10.1109/TPWRS.2022.3144246.

LANL-ANSI and PowerModelsDistribution Contributors. Powermodelsdistribution.jl repository. GitHub, 2026. URL https://github.com/lanl-ansi/PowerModelsDistribution.jl.

J. Lavaei and S. H. Low. Zero duality gap in optimal power flow problem. *IEEE Transactions on Power Systems*, 2012. doi: 10.1109/TPWRS.2011.2160974. URL https://doi.org/10.1109/TPWRS.2011.2160974.

N. Li, L. Chen, and S. H. Low. Exact optimal power dispatch in unbalanced distribution systems with high pv penetration. *IEEE Transactions on Power Systems*, 2018. doi: 10.1109/TPWRS.2018.2869195. URL https://doi.org/10.1109/TPWRS.2018.2869195.

Y. Liu, J. H. Braslavsky, and G. C. Verbic. Robust dynamic operating envelopes for der integration in unbalanced distribution networks. *IEEE Transactions on Power Systems*, 2023a. doi: 10.1109/TPWRS.2023.3308104. URL https://doi.org/10.1109/TPWRS.2023.3308104.

Y. Liu, D. J. Hill, and J. H. Braslavsky. Allocation dynamic operating envelopes to ders in distribution networks considering technical and equitable criteria. *IEEE Transactions on Sustainable Energy*, 2023b. doi: 10.1109/TSTE.2023.3275082. URL https://doi.org/10.1109/TSTE.2023.3275082.

Y. Liu, G. C. Verbic, and J. H. Braslavsky. Time-varying operating regions of end-users and feeders in low-voltage distribution networks. *IEEE Transactions on Power Systems*, 2023c. doi: 10.1109/TPWRS.2023.3302421. URL https://doi.org/10.1109/TPWRS.2023.3302421.

Y. Liu, M. B. Dinu, and A. M. O'Connell. Day-ahead dynamic operating envelopes in low-voltage distribution networks with rooftop pvs: A data-driven stochastic utopf method. *Sustainable Energy, Grids and Networks*, 2024a. doi: 10.1016/j.segan.2024.101528. URL https://doi.org/10.1016/j.segan.2024.101528.

Y. Liu, M. B. Dinu, and A. M. O'Connell. A bargaining-based dynamic operating envelopes allocation method in distribution networks. *IEEE PES General Meeting*, 2024b. doi: 10.1109/PESGM51994.2024.10688561. URL https://doi.org/10.1109/PESGM51994.2024.10688561.

Y. Liu, A. Nazaripouya, and L. Zhang. Linear opf-based method for computing robust dynamic operating envelopes for distribution networks. *Journal of Modern Power Systems and Clean Energy*, 2024c. doi: 10.35833/MPCE.2023.000653. URL https://doi.org/10.35833/MPCE.2023.000653.

Y. Liu, M. B. Dinu, and A. M. O'Connell. Data-driven dynamic operating envelope in lv distribution networks with smart meter data. *Applied Energy*, 2025a. doi: 10.1016/j.apenergy.2025.125469. URL https://doi.org/10.1016/j.apenergy.2025.125469.

Y. Liu, M. B. Dinu, and A. M. O'Connell. Development and validation of dynamic operating envelope enabled demand response in distribution networks. *Applied Energy*, 2025b. doi: 10.1016/j.apenergy.2024.125150. URL https://doi.org/10.1016/j.apenergy.2024.125150.

Y. Liu, M. B. Dinu, and A. M. O'Connell. Dynamic operating envelope in the future distribution grid: A review. *Modelling*, 2025c. doi: 10.3390/modelling6020029. URL https://doi.org/10.3390/modelling6020029.

Y. Liu, X. Liang, and G. Ledwich. Interpretable dynamic operating envelopes in distribution systems via deep learning. *Energies*, 2025d. doi: 10.3390/en18102529. URL https://doi.org/10.3390/en18102529.

Y. Liu, A. M. O'Connell, and G. C. Verbic. Bargaining-based dynamic operating envelopes allocation method in distribution networks. *IEEE Transactions on Smart Grid*, 2025e. doi: 10.1109/TSG.2025.3566419. URL https://doi.org/10.1109/TSG.2025.3566419.

University of Kassel, Fraunhofer IEE, and pandapower Contributors. pandapower github repository. GitHub, 2026. URL https://github.com/e2nIEE/pandapower.

L. Thurner, A. Scheidler, F. Schafer, J. Menke, J. Dollichon, F. Meier, S. Meinecke, and M. Braun. pandapower - an open-source python tool for convenient modeling, analysis, and optimization of electric power systems. *IEEE Transactions on Power Systems*, 2018. doi: 10.1109/TPWRS.2018.2829021. URL https://doi.org/10.1109/TPWRS.2018.2829021.

L. Wang, Y. Cao, and M. Pan. Taylor-expansion-based robust power flow for three-phase unbalanced distribution networks. *IEEE PES General Meeting*, 2024. doi: 10.1109/PESGM51994.2024.10689162. URL https://doi.org/10.1109/PESGM51994.2024.10689162.

R. D. Zimmerman, C. E. Murillo-Sanchez, and R. J. Thomas. Matpower 7.0: New features and legacy compatibility. *IEEE Transactions on Power Systems*, 2020. doi: 10.1109/TPWRS.2020.2997060. URL https://doi.org/10.1109/TPWRS.2020.2997060.

A Extended Empirical Diagnostics

This appendix reports experiment families that are referenced in the main text but moved out of the core narrative to keep the main figures focused on the central screening and calibration claims.

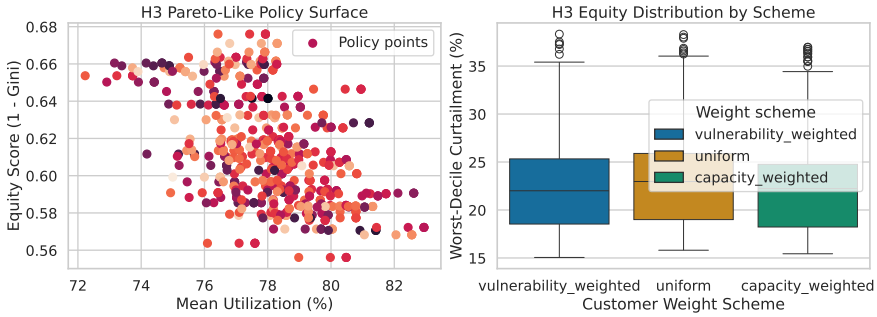

Figure 3: Extended fairness diagnostics from the policy-surface experiment family. The left panel maps policy outcomes in utilization-equity space, and the right panel reports distributional behavior of worst-decile curtailment across policy regimes. Together these panels show that equity claims are associated with measurable frontier structure and tail-distribution shifts, not only mean-score changes.

Table 4: Focused baseline-delta diagnostics for the fairness and cross-solver studies, computed from exported summary artifacts.

Family	Diagnostic	Value	Interpretation
Fairness study	Utilitarian vs equal-share worst-decile curtailment	32.8669 vs 20.8669	12.0-point (36.5%) reduction
Cross-solver reproducibility	Pass-rate range across baseline pipelines	98.03%–98.36%	Stable high-repeatability regime
Cross-solver discrepancy	Nominal vs stress utilization discrepancy means	0.591 vs 0.816	Stress remains bounded but higher

A.1 Fairness Policy Surface Diagnostics

Figure 3 shows the fairness-policy family with two panels: the utilization-equity surface and the worst-decile curtailment distribution. The left panel visualizes how policy parameters span a broad frontier rather than a single operating point; the right panel highlights tail-behavior sensitivity, which is essential for equity assessment.

The fairness evidence supports two practical points. First, multiple non-dominated regions are observed, so objective choice should be treated as a policy decision with explicit tradeoffs. Second, tail metrics can move differently from mean metrics, reinforcing the need to report both when comparing allocation policies.

A.2 Cross-Solver Shadow-Mode Diagnostics

Figure 4 reports shadow-benefit and discrepancy behavior under nominal and stress settings. The left panel indicates positive shadow-mode benefit relative to static procedures in the tested setting, while the right panel reports discrepancy distributions with threshold markers.

The cross-solver results should not be interpreted as proving tool equivalence in general. They indicate that under harmonized settings and within tested scenarios, discrepancy remains bounded within pre-specified operational tolerances.

B Additional Proof and Symbolic Reproducibility Notes

Theorems 5.2 and 5.3 rely on assumptions that are explicit and monitorable. For Theorem 5.2, the critical assumptions are local smoothness, bounded Hessian norms, and trust-region validity around the operating point. For Theorem 5.3, the essential assumptions are admissibility of acceptance policy and strict cost ordering $c_{\mathrm{ncv}} > c_{\mathrm{lin}}$.

Symbolic checks verify implementation consistency for the inequalities and cost-gap decomposition used in these proofs. In particular, the non-negativity of the optimality gap under admissibility and the factorization identity for the compute-gap expression are both confirmed. These checks are not substitutes for mathematical proof; they are implementation guards that reduce risk of transcription errors between derivation and code.

A useful practical workflow is therefore: (i) verify assumptions with monitors, (ii) enforce policy admissibility in implementation, (iii) run symbolic regression checks in continuous integration, and (iv) inspect stress-case diagnostics to ensure failures appear as expected when assumptions are perturbed.

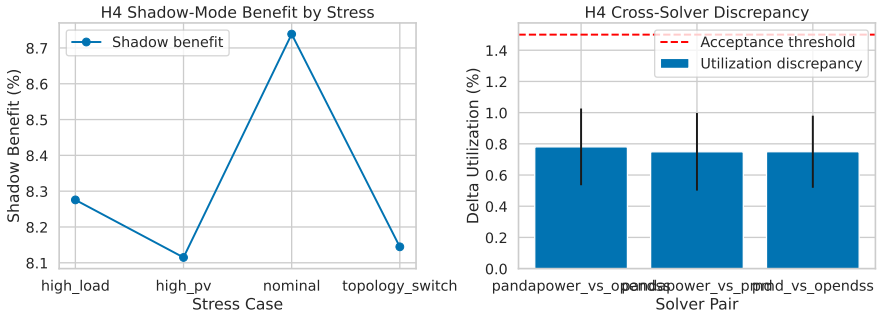

Figure 4: Extended cross-solver and shadow-mode diagnostics. The left panel shows shadow-mode benefit under stress-case stratification, and the right panel visualizes utilization discrepancy across solver pairs with confidence intervals and threshold guidance. These panels provide evidence that portability claims are bounded and testable, while still highlighting that agreement is conditional on harmonized modeling assumptions.

C Implementation and Reproducibility Details

This section summarizes implementation details needed for reproduction and extension:

- **Seeds and repetitions.** Five seeds were used per experiment family, with baseline-stratified aggregation and confidence intervals computed as $1.96 \times \mathrm{SE}$ from seed-baseline samples.
- **Sweep design.** Security-runtime sweeps included trust-region radius, uncertainty scale, and counterexample modes. Calibration sweeps included target risk levels, rolling-window length, and update cadence. Fairness sweeps covered policy parameters (α, β) and weighting schemes. Cross-solver sweeps covered solver pairs, harmonization profiles, and stress cases.
- **Compute budget.** Runs were executed on a single workstation with explicit concurrency bounds, and all generated artifacts were recorded with deterministic outputs.
- **Claim-evidence traceability.** An internal claim-to-evidence matrix maps each primary claim to theorem IDs or quantitative evidence entries and source materials.
- **Approximations and fallback.** Linearization and residual bounds are local approximations; uncertified candidates are always escalated to robust non-convex verification, ensuring conservative behavior under uncertainty.
- **Uncertainty reporting.** All major figures and exported summaries include variability terms or confidence intervals, and acceptance checks are evaluated on aggregate metrics rather than single-run outcomes.

These details are included so that downstream revision and review phases can reproduce both theorem-linked diagnostics and empirical claim checks without ambiguity.