Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Flood Response under Crowd Misinformation: Robust Belief Fusion and UAV Verification Prioritization

Hailong Li

11172@zzrvtc.edu.cn

Zhengzhou Railway Vocational and Technical College https://orcid.org/0009-0006-9462-6347

Xiangnan Song

Zhengzhou Railway Vocational and Technical College

Ya Li

Zhengzhou Railway Vocational and Technical College

Research Article

Keywords: flood emergency response, crowdsourced sensing, misinformation, distributionally robust optimization, road passability, unmanned aerial vehicles

Posted Date: July 15th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10104632/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Flood Response under Crowd Misinformation: Robust Belief Fusion and UAV Verification Prioritization

Hailong Li, Xiangnan Song, Ya Li
School of Transportation Management, Zhengzhou Railway Vocational and Technical College, Zhengzhou 451460, China

Corresponding author: 11172@zzrvtc.edu.cn

Acknowledgements. The authors thank Zhengzhou Railway Vocational and Technical College for research support.

Abstract. Flood emergency agencies increasingly combine crowdsourced road reports with scarce unmanned-aerial-vehicle (UAV) inspections, yet panic, stale forwarding, and deliberate rumours can distort road-passability beliefs and redirect relief. This study develops a decision framework for fusing reports when source credibility is uncertain or adversarially exploited and for choosing whether the next trusted observation should inspect a road or recalibrate a reporting source. Credibility is represented by a data-driven ambiguity set and rumours by bounded contamination, yielding a distributionally robust credal posterior and min-max dispatch. We derive the dual value of verification—road-state clarification and calibration spillover—and prove a contamination threshold $\tau^\star$ at which the preferred action switches from road inspection to source recalibration. We further bound robust loss and characterize a strategic relocating attacker. On synthetic benchmarks, non-robust regret rises from 142 to 297 stranded passengers as contamination reaches 0.5, whereas robust regret remains within 105–168; the predicted switch occurs near $\tau^\star = 0.13$–0.16. In a replay of the 20 July 2021 Zhengzhou flood, non-robust fusion reverses a blocked-road assessment and propagates error across about 62 cells, while robust fusion preserves the assessment and one verification recalibrates about 59 reports. In a Zhengzhou road-network decision test, eight baselines incur 233–300 stranded-passenger equivalents under the rumour, whereas the robust policy retains clean-evidence delivery. The framework provides flood-response agencies with a principled rule for allocating scarce verification when crowdsourced road information may be unreliable or manipulated.

Keywords: flood emergency response ; crowdsourced sensing ; misinformation ; distributionally robust optimization ; road passability ; unmanned aerial vehicles

1 Introduction

1.1 Contaminated crowd sensing and the robustness gap

When a city floods, the agency directing rescue and bus-bridging traffic has to decide which road segments are still passable, and it has to decide quickly. Two information streams are

available, and they are asymmetric. One is a free, high-volume stream of crowd reports drawn from social media, mutual-aid documents, and passenger submissions; the other is a scarce and accurate stream of active UAV inspection. The operational problem is therefore not simply to estimate passability: the agency must decide how much to trust each report class, where to spend the next verification, and how the resulting belief should change the relief dispatch.

Yet that abundance is also a vulnerability. During the Zhengzhou rainstorm of 20 July 2021, hourly rainfall reached 201.9 mm, the metro network was suspended, and the emergency response was hampered by roads whose status was unclear; social-media reconstructions of the event show the surface transport network failing in spatially correlated clusters rather than at isolated points (Peng and Zhang, 2024; Qu et al., 2023). Crowd reporting in such events is noisy in structured ways—panic amplifies "blocked" alarms, stale posts are re-forwarded after the water recedes, and fabricated claims circulate—and this misinformation does measurable operational damage to the response (Hilberts et al., 2025; Vosoughi et al., 2018).

Many operational fusion rules make one fragile simplification: they treat the credibility of each source class as a fixed, known quantity. In a real flood that quantity is estimated from limited calibration data, and the estimate can be wrong in two ways at once. It can be wrong by chance, because a class was calibrated on few examples, and it can be wrong by design, because an actor who knows the agency trusts a class can inject a flood of identical claims disguised as that class. A fusion rule that substitutes a point estimate then carries the contamination straight into the dispatch decision, and reports the same confidence whether or not it has been manipulated. This paper therefore asks how much an agency should believe the crowd when credibility is uncertain or adversarial, and which source a scarce verification should be spent calibrating.

We treat misinformation as a problem of allocating sensing and trust. The credibility of a source class is represented by an ambiguity set, the injected reports by a bounded contamination of the evidence, and the belief update is the worst case over both. The resulting posterior is a set of distributions, a credal posterior, and the dispatch decision is the one that is best against the worst member of that set. A UAV verification then plays a double role: it reveals a road state, and by comparing the inspected reports to the truth it corrects the credibility of the class that produced them, which re-weights every other report from that class. Figure 1 summarises the approach.

1.2 Contributions

The paper makes three contributions.

1. *A robust belief update for contaminated credibility.* To keep fusion reliable when source credibilities are estimated with error or injected by an adversary, we represent each credibility by a data-driven ambiguity set and the injections by a bounded contamination, and take the worst-case belief over both. The update returns a conservative credal posterior and its min-max dispatch; it is well posed, conservative, and reduces to the known-credibility point-estimate fusion at zero ambiguity (Sections 3–4, Theorem 1).

2. *A dual value for verification.* To price UAV inspection correctly under contamination,

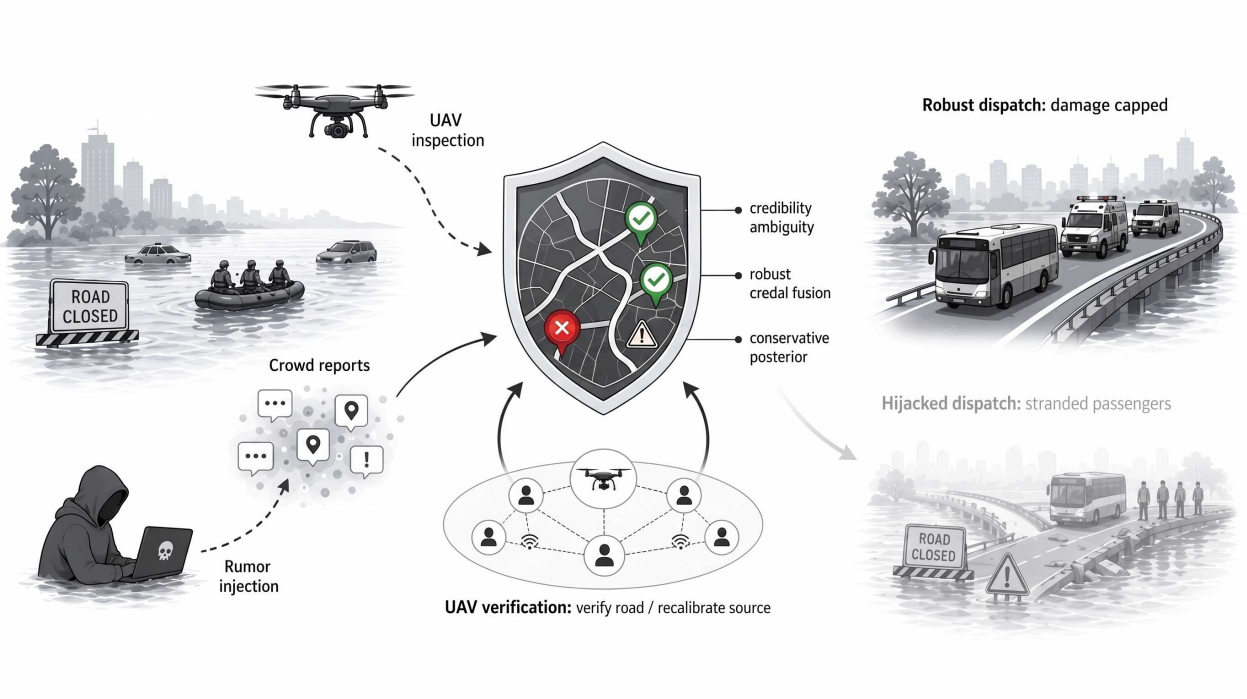


Fig. 1 Conceptual framework. Flood response combines noisy crowd reports and scarce UAV inspection while an adversary may inject coordinated road-status rumours. The proposed layer models credibility ambiguity, applies robust credal fusion, and returns a conservative posterior for dispatch. UAV verification has a dual role: it checks a road state and recalibrates the reporting source, so robust dispatch caps the damage that a hijacked point-estimate dispatch can otherwise turn into stranded passengers

we model each sortie as resolving a road state *and* recalibrating, network-wide, the credibility of the source that reported it. This *calibration spillover* admits a closed form that grows linearly in the number of reports one verification cleans (Section 4.3, Theorem 2); on the Zhengzhou data a single sortie restores a hijacked belief and recalibrates about 59 real reports.

3. *A rule for where to verify.* To tell an operator when to inspect a road and when to calibrate a source, we prove a contamination threshold $\tau^\star$ that switches the next sortie from one to the other, with comparative statics and a closed form, and we bound the loss an adversary can force (Section 4.4, Theorems 3–4). On the benchmark the switch sits at $\tau^\star \approx 0.13$–0.16, and the robust policy caps the adversarial influence that the point estimate lets grow until the pivotal decision flips, including against an adaptive adversary (Section 6); the equilibrium, adversary-optimal form of the threshold against a relocating attacker is Theorem 5.

2 Literature Review

2.1 Misinformation and social-aerial sensing in disaster response

Crowd reports in a flood are fast and free, but a fraction are wrong in structured ways: panic amplifies alarms, stale posts are re-forwarded, and fabricated claims circulate. The rapid online spread of false relative to true news is well documented (Vosoughi et al., 2018), recent reviews trace how misinformation degrades the disaster response itself (Hilberts et al., 2025), and studies of the Zhengzhou "7·20" event reconstruct how official and rumoured signals interleave during the emergency (Qu et al., 2023; Zhang et al., 2024). This work treats misinformation as a diffusion phenomenon to be detected once it spreads; we take it up at the point of fusion, as a contamination of the evidence an agency assimilates, with a bounded adversarial-injection model that follows the data-poisoning attacks and robust-aggregation defences of distributed learning (Khraisat et al., 2025). A complementary line dispatches physical sensors to verify reported events: SocialDrone routes a drone to confirm social-media reports (Rashid et al., 2020b), the damage-aware social-media-driven car-sensing (DASC) framework drives vehicle sensing from social signals of road damage (Rashid et al., 2020a), the social-physical sensing paradigm has been surveyed in detail (Rashid et al., 2023), multimodal variants fuse text and imagery (Yu and Wang, 2024), and UAVs are routed for post-disaster damage assessment under time and battery budgets (Adsanver et al., 2025). These systems verify *events* and optimise coverage or completion; we instead maintain a robust joint posterior over road *passability states* and choose verifications to minimise a downstream rescue and bus-bridging loss. The distinction bites under contamination: signal-driven dispatch chases the loudest reports, which is where an adversary concentrates, while the decision-aware rule spends verification on the source that has been poisoned.

2.2 Crowdsourced credibility, value of information, and evidence fusion

Truth-discovery and information-fusion methods turn unreliable reports into a usable belief, weighting sources by inferred reliability (Alamri et al., 2025) or combining conflicting evidence through Dempster–Shafer and related calculi (Zhang et al., 2025a). The value of a measurement is classically the improvement it brings to a decision (Howard, 1966). In crowd-and-UAV flood response, however, a trusted measurement can do more than resolve the inspected road: it can also correct a contaminated source's credibility and thereby re-weight every report that source produced. We call this externality calibration spillover. Section 6.2 compares the robust update against truth discovery (conflict-resolution-on-heterogeneous-data (CRH)/TruthFinder), Dempster–Shafer fusion, Byzantine-robust aggregation (trimmed mean, multi-Krum, geometric median), and Wasserstein and Kullback–Leibler (KL) distributionally robust updates under the targeted attack. The ordinary point-estimate fusion is defined within our formulation and recovered as the zero-ambiguity, zero-attack limit (Theorem 1(c)); it serves as the non-robust baseline throughout.

2.3 Distributionally robust and credal inference

The fusion layer is a distributionally robust update over a credibility ambiguity set with a Huber-type contamination component (Huber, 1981). Distributionally robust optimisation (DRO) hedges a decision against a set of distributions, classically a Wasserstein ball around the empirical law (Mohajerin Esfahani and Kuhn, 2018), and has become a standard tool in humanitarian and disaster operations, where two-stage formulations with a conditional-value-at-risk objective preposition relief and design networks under demand and routing ambiguity (Zhang et al., 2023a; Wang et al., 2023a; Seif et al., 2026). When the worst case is taken over a set of posteriors rather than a single distribution, the decision is a maximin expected utility over a credal set (Gilboa and Schmeidler, 1989; Walley, 1991), an idea now also used to aggregate ensembles into robust, set-valued predictions (Nguyen et al., 2025). On this machinery we build a decision-loss-calibrated rule: a threshold for when to verify a road and when to recalibrate a source, and a spillover value the same machinery makes computable.

2.4 Flood situational awareness and transport-network resilience

Social media and volunteered geographic information are now routine inputs to flood situational awareness, used to map inundation, infer water depth, and validate hydraulic models, albeit with well-known locational and sampling uncertainties (Peng and Zhang, 2024; Zhang et al., 2025b; Drews et al., 2023; Chow et al., 2026). The downstream concern is the resilience of the transport network the agency must keep running: rainfall and flooding degrade metro and multimodal networks in spatially correlated ways and motivate resilience-oriented recovery planning (Li et al., 2026; Huang et al., 2026). When a rail corridor fails, the response is bus bridging, whose design and dispatch under disruption and demand uncertainty is itself an active optimisation problem (Wang et al., 2025, 2023b; Du et al., 2025; Liu et al., 2025; Zhang et al., 2023b). We take that bridging allocation as the fixed downstream decision whose expected loss scores our information, and ask which verifications most reduce that loss when the crowd evidence is contaminated.

3 Problem Setting and Threat Model

3.1 States, decisions, and loss

Let $E = \{1, \dots, n\}$ index the monitored road segments and let $X \in \{0,1\}^n$ collect their passability states, with $X_e = 1$ meaning segment e is impassable. The states are spatially correlated, because flooding follows drainage basins and elevation bands; we model the prior $\mathbb{P}(X)$ by a Gaussian-copula threshold field (Nelsen, 2006). A downstream bridging decision $a \in \mathcal{A}$ allocates rescue and bus-bridging capacity and incurs a loss $L(a, X)$ measured in stranded passenger-equivalents. For a state distribution μ the Bayes loss is

$$\Phi(\mu) = \min_{a \in \mathcal{A}} \ \mathbb{E}_{X \sim \mu}\big[L(a, X)\big], \tag{1}$$

which is concave in μ as a pointwise minimum of linear functionals. Concretely, with a relief allocation $a = (a_k)$ over cells k of per-unit capacity u and demand d_k, a passable cell serves

$\min(u\,a_k, d_k)$ passengers and an impassable one none, so the loss is the stranded demand $L(a, X) = \sum_k d_k - \sum_{k:\,X_k=0} \min(u\,a_k, d_k)$ (Appendix B gives the full data-generating process). Throughout, *regret* denotes the excess of realized loss over a truth-aware oracle that knows X.

3.2 Sources and credibility

A crowd report is a triple $r = (e, y, c)$: a segment e, a binary claim $y \in \{0, 1\}$ (y=1 reports segment e *blocked*, matching the state X_e=1; y=0 reports it *passable*), and a source class $c \in C = \{\mathsf{eyewitness}, \mathsf{relay}, \mathsf{low_cred}\}$. Class c has a credibility pair $\theta_c = (\alpha_c, \beta_c)$, where the sensitivity $\alpha_c = \mathbb{P}(y{=}1 \mid X_e{=}1, c)$ and the specificity $\beta_c = \mathbb{P}(y{=}0 \mid X_e{=}0, c)$. A report carries a per-report log-likelihood ratio toward X_e=1 that depends on its direction,

$$\ell_c(y) = \begin{cases} \lambda_c = \log \dfrac{\alpha_c}{1-\beta_c} > 0, & y = 1 \text{ (blocked)}, \\ \log \dfrac{1-\alpha_c}{\beta_c} < 0, & y = 0 \text{ (passable)}, \end{cases} \tag{2}$$

so a blocked claim raises the impassability log-odds by λ_c and a passable claim lowers it; we write λ_c for the blocked-direction magnitude used throughout. The credibilities are estimated from a separate labelled *report-level* comparison set, in which each report's claim is checked against a human-assigned correct label for that report; this is distinct from a segment-level passability oracle, which the event does not provide (hence the case study of Section 6.5 reports only oracle-free quantities). On that set the three classes have $\hat{\theta}_{\mathsf{eye}} = (0.90, 0.84)$, $\hat{\theta}_{\mathsf{relay}} = (0.72, 0.68)$, and $\hat{\theta}_{\mathsf{low}} = (0.38, 0.50)$, with calibration sample sizes $n_{\mathsf{eye}} = 2303$, $n_{\mathsf{relay}} = 2066$, and $n_{\mathsf{low}} = 749$.

3.3 Credibility ambiguity and contamination

A point-estimate fusion fixes $\theta_c = \hat{\theta}_c$ and carries any calibration error straight into the dispatch. We replace the point by a rectangular credibility ambiguity set (a box) centred at the estimate,

$$\mathcal{U}_c(\varepsilon_c) = \big\{\theta_c = (\alpha_c, \beta_c) :\ |\alpha_c - \hat{\alpha}_c| \le \varepsilon_c,\ |\beta_c - \hat{\beta}_c| \le \varepsilon_c\big\}, \qquad \varepsilon_c = \eta + \frac{z_{1-\delta}}{\sqrt{n_c}}, \tag{3}$$

a box of half-width ε_c, with n_c the calibration sample size. The worst-case belief over a box is attained at a vertex; because the adversarial tilt reduces a class's discriminating power, the exact min-max dispatch ranges each class between its estimate $\hat{\theta}_c$ and its ambiguity-shrunk credibility, the $2^{|C|} = 8$ vertices used in Section 4.1 and Algorithm 1. The statistical part $z_{1-\delta}/\sqrt{n_c}$ is the half-width of an asymptotic $(1-\delta)$ confidence interval for the calibrated credibility (we take $z_{1-\delta} = 1.6$, $\delta \approx 0.05$), so the ambiguity set is *data-driven*: it is an approximate $(1-\delta)$ confidence region for the true θ_c, contracts at the parametric rate $\Theta(n_c^{-1/2})$, and is widest for the $\mathsf{low_cred}$ class, which is calibrated on the fewest examples. The additive injection budget η widens it to hedge adversarial, not merely statistical, error. Contamination is modelled by a Huber η-neighbourhood: an actor controls a fraction η of

the reports and, on a targeted segment $e^\star$, injects a flood of identical claims disguised as a source class. We write τ for the contamination rate, the fraction of a source's reports on the target that are injected. Operationally, a rate x (the global budget η, or the on-target rate τ) adds $k = \lceil \frac{x}{1-x} N \rceil$ injected reports to a base of N genuine ones, so injected reports form a fraction x of the resulting stream; on a single targeted source η and τ coincide, and Theorem 4 bounds the decision-loss inflation this permits.

3.4 Robust posterior, robust decision, and verification

Given evidence R, the robust update returns the credal posterior

$$\overline{\Pi}(R) = \overline{\mathrm{conv}}\Big\{\pi_\theta(\cdot \mid R') :\ \theta \in \textstyle\prod_c \mathcal{U}_c(\varepsilon_c),\ R' \ \eta\text{-close to } R\Big\}, \tag{4}$$

and the robust decision and value are

$$a^\star_{\mathrm{rob}} \in \arg\min_{a\in\mathcal{A}} \max_{\pi\in\overline{\Pi}(R)} \mathbb{E}_{X\sim\pi}[L(a,X)], \qquad V_{\mathrm{rob}}(R) = \min_{a\in\mathcal{A}} \max_{\pi\in\overline{\Pi}(R)} \mathbb{E}_{\pi}[L(a,X)]. \tag{5}$$

A UAV verification of a segment set S reveals X_S and, by comparing the reports on S to ground truth, shrinks $\mathcal{U}_c$ for every contributing class c. Its value splits into a state-clarification term and a calibration-spillover term,

$$\mathrm{RV}(S \mid R) = \mathrm{RV}_{\mathrm{state}}(S \mid R) + \mathrm{RV}_{\mathrm{calib}}(S \mid R), \tag{6}$$

and the prioritization problem chooses, subject to a budget B,

$$\textbf{(P-B)} \qquad \max_{S:\,|S|\le B}\ \mathrm{RV}_{\mathrm{state}}(S \mid R) + \mathrm{RV}_{\mathrm{calib}}(S \mid R), \tag{7}$$

re-solved as evidence arrives. Here B is a verification-unit budget (the number of segments a sortie can inspect), not a flight trajectory: the paper studies how scarce verification should be *allocated* across sources and segments, and takes UAV routing, endurance, and launch logistics as a separable downstream layer. Appendix C collects the notation.

4 Methodology: Robust Fusion and Verification Prioritization

4.1 Credal posterior on the scenario-particle ensemble

We compute the credal posterior on a scenario-particle ensemble: M particles are sampled from the spatial prior $\mathbb{P}(X)$ and reweighted by the report likelihood (Appendix B). Two ingredients make the update robust. First, the credibility used in the likelihood is tilted to its worst case over the box $\mathcal{U}_c$, which for a box is reached at the ambiguity-shrunk vertex, each class's discriminating power reduced by the half-width ε_c; enumerating the $2^{|C|}$ vertices gives the credal set. Second, aggregation is influence-bounded: reports are trimmed per (segment, source, claim direction) and anomalously concentrated floods are rejected, so no

single source can dominate a segment's posterior. The first ingredient defends against an estimate that is merely uncertain; the second defends against one that is being attacked. This shrink-and-reject posterior is the operational policy used throughout the paper; the exact min-max credal decision over the same set (Appendix E) serves as a tail-risk certificate rather than the default operator, and the two agree closely in practice.

4.2 Well-posedness, conservativeness, and nesting

Theorem 1 (Robust posterior)**.** *Let Φ be the concave Bayes loss* (1) *and let $\overline{\Pi}(R)$ be the credal posterior* (4)*, with each $\mathcal{U}_c(\varepsilon_c)$ compact convex and the likelihood map $\theta \mapsto \pi_\theta(\cdot \mid R)$ continuous.*

(a) The robust value (5) *is attained and finite, and it is conservative: for the unknown true credibility $\theta^\star$, $V_{\mathrm{rob}}(R) \geq \mathbb{E}_{\pi_{\theta^\star}}[L(a^\star_{\mathrm{rob}}, X)] \geq V_{\theta^\star}(R)$. The robust policy never reports an optimistic loss.*

(b) If $\overline{\Pi}'(R) \subseteq \overline{\Pi}(R)$, which results from more calibration samples, a smaller attack budget, or a verification, then $V'_{\mathrm{rob}}(R) \leq V_{\mathrm{rob}}(R)$ and the robust regret bound does not increase.

(c) As $\varepsilon_c \to 0$ for every c and $\eta \to 0$, the set $\overline{\Pi}(R)$ collapses to the single posterior $\{\pi_{\hat{\theta}}(\cdot \mid R)\}$, the robust value tends to $V_{\hat{\theta}}(R)$, and the verification value (6) *reduces to the known-credibility CVoI.*

Part (c) is the reason the non-robust baseline in the experiments coincides with the known-credibility point-estimate fusion: at $\tau = 0$ the robust and non-robust regrets agree up to particle-sampling noise (168 against 142 stranded passengers in Section 6), and they separate only as contamination grows. The proof is in Appendix A.

4.3 Calibration spillover and verification prioritization

Verifying a segment does more than reveal its state. When the inspected segment carried reports from class c, comparing those reports to the truth corrects $\hat{\theta}_c$, and the correction re-weights every other report from c, including reports on segments that were never inspected. The first effect is local and is what CVoI already prices; the second is an externality across the report set.

Theorem 2 (Calibration spillover)**.** *Let S contain m segments reported by class c, let $\Delta\lambda_c$ be the credibility correction the verification induces on c, and let ω_e be the decision weight of segment e. Then the verification value splits as in* (6)*, and the spillover term satisfies*

$$\mathrm{RV}_{\mathrm{calib}}(S \mid R) = \sum_c \Delta\lambda_c \sum_{e \notin S:\, c \to e} \omega_e \;\asymp\; (N_c - m)\,\bar{\kappa}\, p(\tau), \tag{8}$$

where, in the pivotal model formalised in Section 4.4.1, N_c is the number of decision-pivotal segments class c touches, $\bar{\kappa}$ their mean criticality, κ the criticality (decision swing) of the verified segment, and $p(\tau)$ the probability that injection-blind fusion crosses the decision boundary. The spillover grows linearly in the number N_c of reports the recalibration cleans, and the ratio $\mathrm{RV}_{\mathrm{calib}}/\mathrm{RV}_{\mathrm{state}}$ exceeds one once $N_c\bar{\kappa}\, p(\tau) > \kappa$.

The linear dependence on N_c is the lever behind prioritization. An adaptive robust greedy fills the verification budget by marginal robust value per verification unit, and it switches between a road-first branch and a source-first branch according to the threshold of Section 4.4. The proof is in Appendix A.

4.4 Structural results: the contamination threshold and the price of an adversary

4.4.1 A stylised pivotal model

To isolate the verify-versus-calibrate decision we reduce it to one unit of budget spent on a single decision-pivotal corridor served by a contaminated source.

Assumption 1 (Stylised pivotal model)**.** (i) A target segment $e^\star$ is pivotal: moving the posterior belief $\mathbb{P}(X_{e^\star}{=}1)$ across $\frac{1}{2}$ changes the bridging decision, with decision swing $\kappa = \mathrm{CVoI}(\{e^\star\}) > 0$. (ii) Source c contributes reports to N_c pivotal segments, and on each affected segment injection-blind fusion crosses the decision boundary with probability $p(\tau)$, where p is continuous, nondecreasing, and $p(0) = 0$; the affected segments have mean criticality $\bar{\kappa}$. (iii) Verifying one c-segment exposes the flood and drives $\lambda_c \to 0$, neutralizing c on all N_c segments.

4.4.2 The threshold

Theorem 3 (Contamination threshold)**.** *Under Assumption 1, write the marginal robust values of the two actions as*

$$\Delta V_{\mathrm{road}}(\tau) = \kappa\big[\,1 - p(\tau)(1 - c_0)\,\big], \qquad \Delta V_{\mathrm{cal}}(\tau) = N_c\,\bar{\kappa}\,p(\tau),$$

where $c_0 \in [0, 1)$ discounts clarifying one state while the source still poisons the rest. If $N_c\bar{\kappa} > \kappa$ there is a unique threshold $\tau^\star \in (0, 1)$ with $\Delta V_{\mathrm{road}} > \Delta V_{\mathrm{cal}}$ for $\tau < \tau^\star$ (verify the road first) and $\Delta V_{\mathrm{cal}} > \Delta V_{\mathrm{road}}$ for $\tau > \tau^\star$ (calibrate the source first). The threshold is strictly decreasing in N_c and $\bar{\kappa}$ and strictly increasing in κ, and under the logistic crossing law $p(\tau) = \sigma(s(\tau - \tau_0))$, with σ the standard logistic function and $c_0 \to 0$,

$$\tau^\star = \tau_0 + \frac{1}{s}\,\mathrm{logit}\Big(\frac{\kappa}{N_c\bar{\kappa} + \kappa}\Big). \tag{9}$$

4.4.3 The price of an adversary

Theorem 4 (Price of an adversary)**.** *Fix a pivotal target with swing κ and per-report log-likelihood ratio λ_c.*

(a) Injection-blind fusion shifts the target log-odds by $k\lambda_c$ when k reports are injected. With an attack budget η permitting $k = \Theta\big(\frac{\eta}{1-\eta}m\big)$ injected reports, the log-odds manipulation is unbounded in attack magnitude, so the worst-case decision-loss inflation $\Delta L_{\mathrm{nr}}(\eta)$ rises monotonically and saturates at the full decision swing κ (the most a single pivotal flip can cost); no smaller bound holds.

(b) With influence-bounded aggregation (per-class trim cap T_c and flood rejection), the aggregate log-odds shift from any source is at most $\Lambda_T = \sum_c T_c \lambda_c$, independent of η, so the robust decision-loss inflation $\Delta L_{\mathrm{rob}}(\eta)$ is bounded by a constant.

Both proofs are in Appendix A.

4.4.4 From an assumed adversary to an equilibrium one

Making the adversary endogenous turns the policy into a Stackelberg game in which the agency commits and a budgeted attacker best-responds. The fusion half yields a *value of robustness*: the optimal influence cap sits one report below the genuine decision margin M, where no budget can flip the corridor, so the equilibrium loss is a sensitivity tax that is *independent of the attack budget*, whereas point-credibility fusion ($T{=}\infty$) loses the full swing κ (Theorem 6, stated and proved in Appendix A). This is the equilibrium form of the certified-absorption radius (S10): a design rule to set the cap just below the genuine margin, keeping almost all honest evidence while no budget can flip the decision. We state here the verification half.

The threshold $\tau^\star$ prices an *assumed* attack on a fixed corridor. A strategic attacker instead observes the verification policy and relocates, turning the verify-or-calibrate choice into a Stackelberg game in which the agency commits and the attacker best-responds. Road verification is a pure strategy the attacker can side-step; source calibration removes the attacker's instrument network-wide and is invariant to relocation. With N_t exchangeable pivotal corridors and a randomising defender, road verification lands on the attacked corridor with catch probability $q = 1/N_t$, which enters its marginal value as $\Delta V_{\mathrm{road}}(\tau) = \kappa(1 - p(\tau)) + q\,\kappa\,c_0\,p(\tau)$ while $\Delta V_{\mathrm{cal}}(\tau) = N_c \bar{\kappa}\, p(\tau)$ is unchanged.

Theorem 5 (Adversary-optimal threshold)**.** *Under the relocating-attacker model with N_t exchangeable corridors and $N_c\bar{\kappa} > \kappa$, the equilibrium verify-to-calibrate threshold under the logistic crossing law is*

$$\tau^\star_{\mathrm{eq}} = \tau_0 + \frac{1}{s}\,\mathrm{logit}\Big(\frac{\kappa}{N_c\bar{\kappa} + \kappa\,(1 - q c_0)}\Big), \qquad q = \tfrac{1}{N_t}. \tag{10}$$

It satisfies $\tau^\star_{\mathrm{eq}} \le \tau^\star$ (with equality iff $N_t = 1$) and is strictly decreasing in N_t: the more corridors a strategic attacker can threaten, the earlier the agency should stop verifying roads and calibrate the source instead. The proof is in Appendix A.

Remark 1 (An equilibrium reading of robustness). Theorem 5 recasts the adaptive-adversary experiment (S11, Section E.3) as an equilibrium rather than a stress test: a best-responding attacker that floods whichever high-demand corridor is left unverified still cannot beat the calibrate-the-source branch, because that branch does not depend on which corridor was hit. A self-contained derivation of Theorems 6 and 5 and a numerical confirmation—the threshold's monotonicity in N_t, the budget-independent value of robustness, and a paired pipeline test in which road verification but not source calibration loses value when the attacker relocates—are in the supplementary theory note.

Table 1: Structural predictions and their computational evidence. S1–S8 are the synthetic experiments of Section 6; R1 is the Zhengzhou case of Section 6.5.

Result	Prediction	Evidence
Thm 1(c)	robust $\to$ point estimate as $\varepsilon, \eta \to 0$	S1 at τ=0: 168 vs 142
Thm 2	one sortie cleans $\Theta(N_c)$ reports	S3: $\sim$68; R1: $\sim$59
Thm 3	a switch at some $\tau^\star$	S2: $\tau^\star \approx 0.13$–0.16
Thm 4	non-robust log-odds unbounded; robust influence capped	S1/S4; R1 shift 1.0 vs 0
Thm 5	relocating attacker lowers the switch, $\tau^\star_{\text{eq}} \leq \tau^\star$	S11; paired pipeline DiD>0
Thm 6	robust loss bounded & budget-independent; non-robust $\to\kappa$	S9/S10; closed-form (A2)

4.4.5 Numerical cross-validation of the structural results

An independent stylised-model script reproduces the four results without the fusion code. It returns a threshold $\tau^\star = 0.162$, matching the empirical switch; a spillover that is linear in N_c; and a non-robust decision that flips under the attack while the robust decision does not. Table 1 lists each prediction against its computational counterpart.

5 Algorithm and Experimental Design

5.1 The robust fusion and verification algorithm

Algorithm 1 assembles the components of Section 4: it returns the robust credal posterior and its min-max dispatch, then spends verification by the contamination threshold, escalating from road to source as τ crosses $\tau^\star$.

The per-replan cost is $O(2^{|C|}M\,|R|)$ for the credal particle ensembles (M scenario particles, Appendix B) and $O(I\,2^{|C|}K)$ for the I-iteration min-max over K corridors; with the three source classes used here the ambiguity box has eight vertices, and a decision-criticality prescreen, which restricts the expansion to segments whose state can change the dispatch, keeps it confined to pivotal segments.

5.2 Testbed, contamination regimes, and the Zhengzhou case

The benchmark is a 10×10 grid of segments with heterogeneous corridor demand, the calibrated three-class source mix of Section 3.2, and the Gaussian-copula prior. We study three contamination regimes: random noise, in which injected claims are placed and signed at random; panic, in which injected "blocked" claims concentrate in report-dense areas; and targeted rumour, in which an actor floods one high-demand corridor with claims opposite to the truth. Regret is the downstream loss relative to a truth-aware oracle, averaged over 12 to 20 random seeds. Appendix B gives the data-generating process.

The regime matters. Averaged over 16 seeds, random contamination leaves non-robust

Input: report stream R; ambiguity sets $\{\mathcal{U}_c(\varepsilon_c)\}$; injection budget η; verification budget B.
Output: robust credal posterior $\overline{\Pi}(R)$; verification set S.

1. Influence-bounded aggregation: trim each (segment, source, claim) group to its cap T_c and reject anomalously concentrated floods.
2. **for** each vertex θ of the ambiguity box $\prod_c \mathcal{U}_c(\varepsilon_c)$ **do** assimilate R to a scenario-particle ensemble and record the marginals π_θ.
3. $\overline{\Pi}(R) \leftarrow \overline{\mathrm{conv}}\{\pi_\theta\}$; form the robust marginal band $[\underline{b}_e, \overline{b}_e]$ per segment.
4. $a^\star_{\mathrm{rob}} \leftarrow \arg\min_a \max_{\pi \in \overline{\Pi}(R)} \mathbb{E}_\pi[L(a, X)]$ by projected supergradient over the vertices.
5. Estimate the contamination rate τ and the threshold $\tau^\star$ (Theorem 3).
6. **if** $\tau < \tau^\star$ **then** (road-first) $S \leftarrow \arg\max_{|S| \le B} \mathrm{RV}_{\mathrm{state}}(S \mid R)$.
7. **else** (source-first) verify one segment of the contaminated source to trigger calibration spillover (Theorem 2), cleaning its reports network-wide.
8. Fly S, observe X_S, fold into R, and re-solve on a rolling horizon.

Algorithm 1: Robust credal fusion and dual-value verification prioritization. Steps 1–4 return the conservative credal posterior and the min-max dispatch; steps 5–8 spend scarce verification by the contamination threshold, escalating from verifying the most decision-critical road to calibrating the contaminated source as τ crosses $\tau^\star$.

regret essentially flat as τ rises from 0 to 0.5 (153 → 121) because random claims cancel, and panic raises it only mildly (153 → 164). The targeted rumour is the regime that breaks non-robust fusion, driving its regret from 153 to 306 while the robust policy stays near 109 (Figure 2). The headline numbers below therefore use the targeted regime as the adversarial worst case.

Zhengzhou case (data and protocol). We replay the de-identified Zhengzhou subset: 52 geocoded reports over 22 occupied cells of an 11×11 grid at roughly 3 km per cell, with source credibility mapped from the native verification labels. The event has no passability oracle, so we report only truth-free quantities. The injected rumour targets the cell that carries the strongest real "blocked" signal, which holds eight eyewitness reports, and disguises itself as a relay claim that the road is passable.

6 Results: Synthetic Benchmark and Zhengzhou Case

We report the targeted-rumour regime throughout, the regime that breaks point-estimate fusion; the testbed, baselines, and protocol are given in Section 5.

6.1 Non-robust collapse versus robust cap (S1)

Figure 4(a) reports the targeted regime. As the contamination rate rises from 0 to 0.5, the regret of non-robust fusion grows from 142 to 297 stranded passengers, while robust fusion stays between 105 and 168. Discarding the crowd is the worst option throughout, which confirms that the crowd should still be used, but used robustly. The robust regret is, if anything, slightly lower at high τ than at low τ: a heavier targeted rumour is also a more concentrated one, so it more often exceeds the flood-rejection threshold and is discarded outright, whereas a light injection sits below the trimming cap and leaks a little bias through;

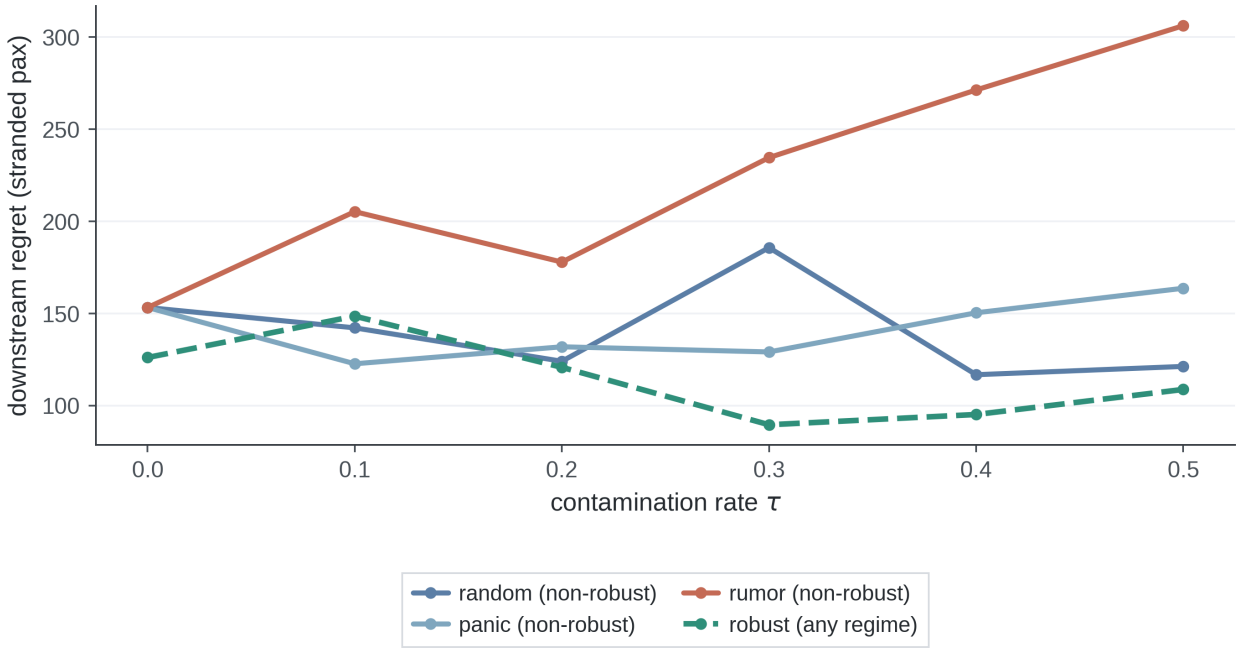


Fig. 2 Downstream regret against contamination rate for the three regimes (16 seeds). Non-robust fusion collapses only under targeted rumour; random and panic contamination are absorbed. The robust policy is capped in every regime

the effect is within the per-seed noise and should be read as "capped and flat," not as a genuine improvement. Per-seed regret is high-variance (per-policy standard errors of about 40–55 over the 12 seeds), so under common random numbers the *paired* gap is the reliable quantity: the non-robust-minus-robust regret gap is 119 at $\tau = 0.3$ and 192 at $\tau = 0.5$, with paired-bootstrap 95% confidence intervals of $[34, 216]$ and $[97, 293]$, both strictly positive. Tabulated runs report mean regret $\pm$ one standard error unless otherwise stated.

6.2 Comparison with competitive baselines (S1b)

The point-estimate and discard-crowd comparators above are deliberately minimal. To test the fusion rule against the methods an analyst would actually reach for, we re-evaluate eight competitors under a single controlled harness: every method shares the same instances, spatial prior, posterior sampler, and dispatch kernel, so only the report-aggregation rule differs. The competitors span robust statistics (coordinate-wise trimmed mean), Byzantine-robust aggregation (multi-Krum and geometric median, i.e. robust federated averaging (RFA) (Khraisat et al., 2025)), evidence fusion (Dempster–Shafer (Zhang et al., 2025a)), truth discovery (CRH/TruthFinder (Alamri et al., 2025)), and two distributionally robust belief updates (a KL/power-likelihood update and a Wasserstein-DRO update (Mohajerin Esfahani and Kuhn, 2018)).

Table 2 reports the result. On clean evidence ($\tau = 0$) all nine methods cluster between 77 and 185 stranded passengers, so none is a straw man, and the robust policy in fact pays a small conservatism tax. Once the rumour appears every competitor collapses toward the non-robust level (247–329 at $\tau = 0.5$), each for a structural reason: trimmed mean and the shard-based

Table 2: S1b: downstream regret of the proposed robust fusion against eight competitive baselines under the targeted rumour (12 seeds, mean±standard error (SE)). All methods share one harness (common instances, spatial prior, posterior sampler, dispatch); only the report-aggregation rule differs. Lower is better.

method	τ=0	τ=0.2	τ=0.3	τ=0.5
non-robust (point estimate)	100±41	249±77	253±83	267±79
coordinate-wise trimmed mean	97±38	241±79	228±73	271±77
multi-Krum	77±40	181±62	270±85	249±80
geometric median (RFA)	113±43	256±63	293±77	247±62
Dempster–Shafer	185±47	252±72	269±74	254±62
CRH / TruthFinder	164±46	310±81	312±81	329±82
KL / power-likelihood DRO	142±54	233±76	255±89	272±75
Wasserstein DRO	97±43	226±83	274±85	296±79
robust (proposed)	**120±54**	**140±59**	**116±53**	**120±54**

aggregators treat the disguised flood as in-distribution per-report values or spread it evenly across shards; Dempster–Shafer saturates under conflict; truth discovery cannot lower a source's *global* weight from a single-segment attack; and the KL and Wasserstein updates damp evidence without capping per-source counts. Among these comparators, the proposed rule—which caps the per-(segment, source, direction) count and rejects floods—is the one that stays capped (≤ 163 across the range—and ≤ 140 for every τ except a single 163 at the lightly contaminated τ=0.1), and its paired-bootstrap advantage over each baseline at $\tau = 0.5$ is strictly positive (for instance +147 $[58, 245]$ over non-robust, +176 $[65, 286]$ over Wasserstein-DRO, and +209 $[73, 349]$ over CRH/TruthFinder stranded passengers). The separation is therefore not an artefact of weak comparators but of the per-source structure the proposed rule exploits and the others ignore. Figure 3 carries the verification of Section 4.3 across all methods: clamping the suspected segment lowers every method's regret, and recalibrating the implicated source lowers it further for the credibility-based methods (Appendix G).

6.3 The verify-versus-calibrate threshold (S2)

Figure 4(b) compares spending one verification on the most decision-critical road against spending it on the most contaminated source. At $\tau = 0$ the two strategies tie (within per-seed noise), verifying the road has lower regret through light contamination, and calibrating the source takes over once contamination is heavier; the threshold is therefore read from the crossing pattern, not from any single grid point. On the coarse grid with 12 seeds the crossing is near $\tau^\star \approx 0.16$, matching the stylised model (0.162); re-estimated with 24 seeds on a finer grid it sits near 0.13. The switch itself is stable; its location lies in the 0.13–0.16 band, with the exact value sensitive to grid and seed resolution.

6.4 Spillover, adversary, and frontier (S3–S5)

One verification with recalibration cleans about 68 same-source-class reports and lowers regret from 325 to 144 (S3, Figure 5(a)). Against a worst-case attack that may target any of

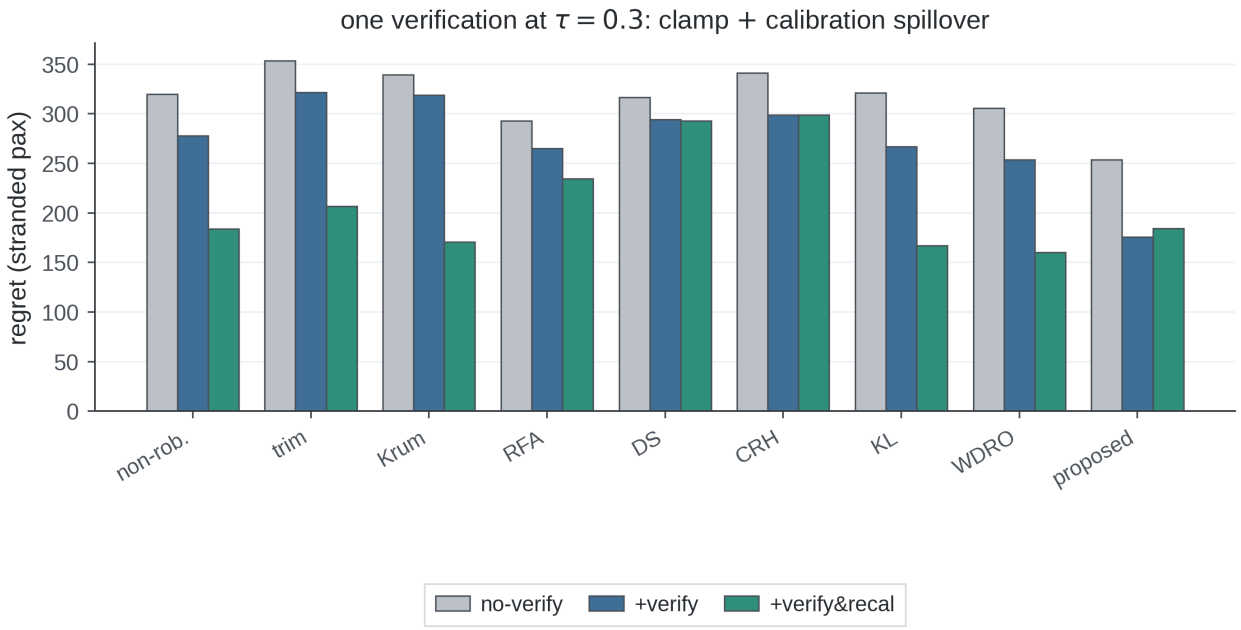


Fig. 3 S1b verification at τ=0.3, across all nine methods. Clamping the suspected segment (+verify) lowers every method's regret, and additionally recalibrating the implicated source (+verify&recal, the calibration-spillover term) lowers it further for the credibility-based methods. The per-method regret-versus-contamination comparison is reported in Table 2 rather than as overlapping curves

the high-demand corridors, non-robust regret rises with the attack budget from 270 to 330, while robust fusion holds between 133 and 208 (S4, Figure 5, centre). The trade-off frontier traces regret as a function of contamination rate and verification budget (S5, Figure 5(c)): with no budget regret worsens as contamination rises, and one to two sorties remove most of the penalty, which reads off the contamination-to-budget substitution an agency can plan against.

Four design-sensitivity ablations – ambiguity radius, source composition, calibration sample size, and spatial correlation length – are reported in Appendix E.5. The headline is that the ambiguity radius trades robustness for sensitivity (the genuine-block miss rate climbs from 0.51 to 0.72 as it grows, with regret minimized near a radius of 0.2), and that a single-cell rumour leaves a network-level footprint of 28–38 cells across correlation lengths.

6.5 Zhengzhou rumour replay

As the injection rate rises, the target belief under injection-blind fusion is dragged from 0.999 toward the wrong state, reaching a shift of 0.91 at an injection rate of 0.4 and a full reversal at 0.5; robust fusion holds the shift at about zero throughout. The contamination does not stay local. Through the spatial prior a single-cell rumour corrupts up to about 62 cells network-wide under injection-blind fusion, against zero under robust fusion (Figure 6). Verification then shows its dual value on real data: even after injection-blind fusion has been hijacked, with the target belief pushed down to 0.11, verifying that cell and recalibrating the implicated source restores the belief to 1.0 and recalibrates about 59 real reports.

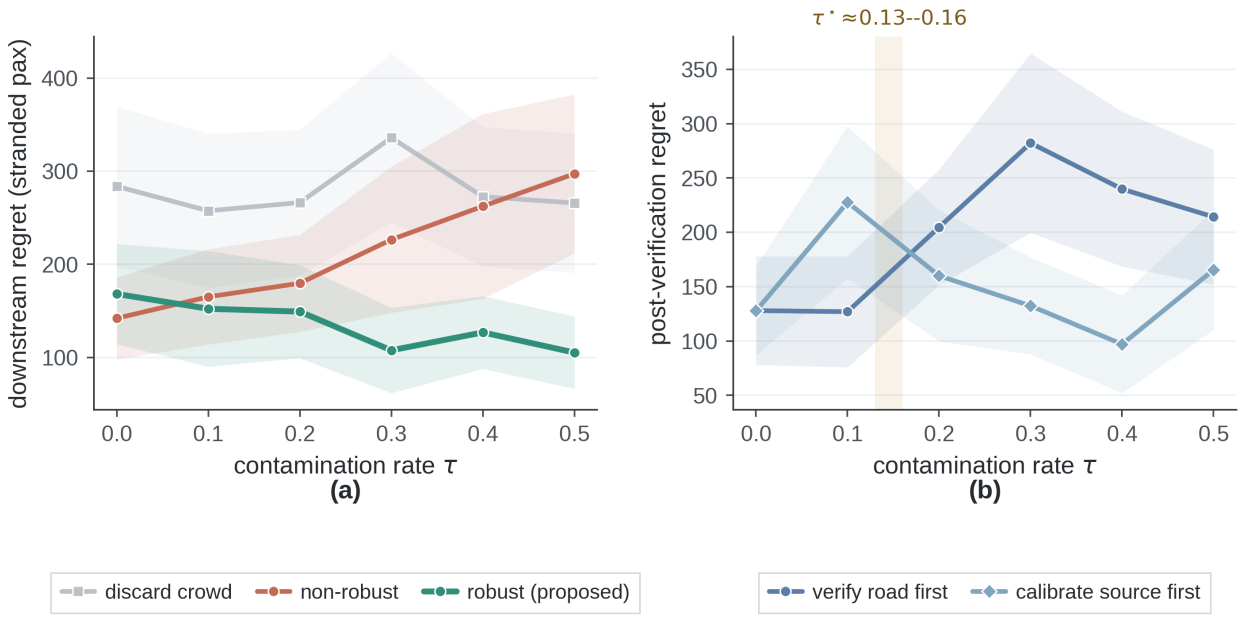


Fig. 4 (a) S1: non-robust regret climbs with contamination while the robust policy is capped. (b) S2: the verify-versus-calibrate switch near $\tau^\star$

6.6 A falsifiable decision metric on the real network

The quantities above are truth-free: they show whether the rumour moves a belief, not whether it strands relief. We therefore add a real-network decision test (R-NET), scored against an independent *trusted-consensus pseudo-oracle* $X^\star$ built once from the trusted verified subset. On the 22 occupied cells, a sparse centroid road graph routes relief from an aid hub to demand cells, and the rumour suppresses the artery carrying the most oracle-usable demand. The score is delivered relief against $X^\star$—real stranded demand rather than belief shift.

Table 3 and Figure 7 report the outcome. Clean evidence lets every method deliver the same 33% of demand. Under the rumour, non-robust fusion and all eight baselines incur 233–300 stranded passenger-equivalents and delivery falls to about 25%, whereas the proposed fusion keeps zero regret and the clean-evidence delivery level across the tested injection range. This zero reflects the concentrated threat geometry: the real-network relief decision turns on one artery, so the suppression attack is exactly the kind of flood the rejection layer drops; a dispersed synthetic attack exposes the complementary failure region (Appendix E.4). A single UAV verification of the targeted artery recovers the lost relief for every method, and a construction sensitivity sweep (Appendix F.4) shows the non-robust gap only where the suppressed artery is decision-critical.

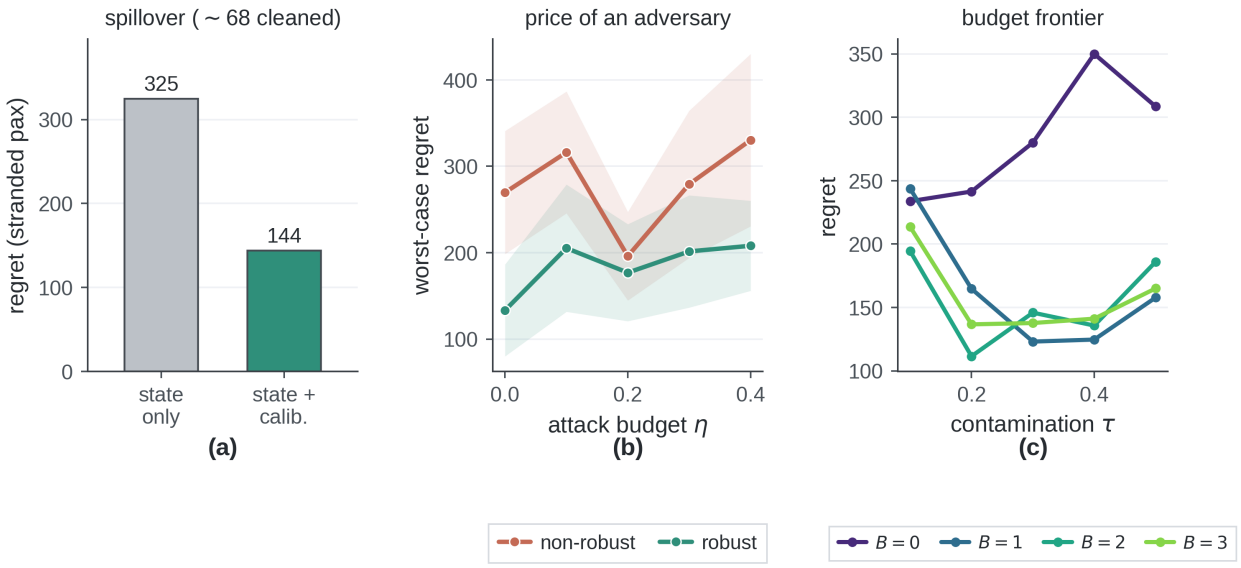


Fig. 5 (a) S3 calibration spillover; (b) S4 price of an adversary; (c) S5 the smoothed contamination × verification-budget frontier

7 Discussion

7.1 Why robustness pays only against targeted manipulation

The three contamination regimes separate cleanly, and only the targeted rumour breaks point-estimate fusion. The mechanism is the log-odds arithmetic of the update. Random claims cancel in expectation and panic adds a bounded, diffuse bias, so a handful of mis-weighted reports moves a segment posterior little; a targeted flood instead stacks identical claims on one pivotal corridor, and because each report adds a fixed log-likelihood increment the posterior crosses the decision boundary once the flood is large enough. The robust layer attacks exactly this arithmetic: trimming caps the per-source log-odds contribution, so an attacker's leverage stops growing with the size of the flood. The same arithmetic explains the magnitudes, not just the direction: the gap between the policies is near zero at $\tau = 0$ and widens monotonically with the injection budget, because robustness buys protection in proportion to how *concentrated* the contamination is rather than how large. That is also why the cap survives an adaptive adversary (S11)—best-responding changes which corridor is flooded, but not the per-source leverage the cap controls.

7.2 Relation to prior work

Three research lines meet in this rule, and setting it against each shows the step it takes. From distributionally robust optimisation it borrows the worst-case-over-a-set machinery that humanitarian models use to hedge demand and routing under a tail objective; the step here is to hedge the road-state *belief* and the source credibility behind it, and to read off the dispatch the credal set implies, so the hedge protects the decision rather than a fixed demand forecast. From truth-discovery and evidence fusion it borrows the down-weighting of

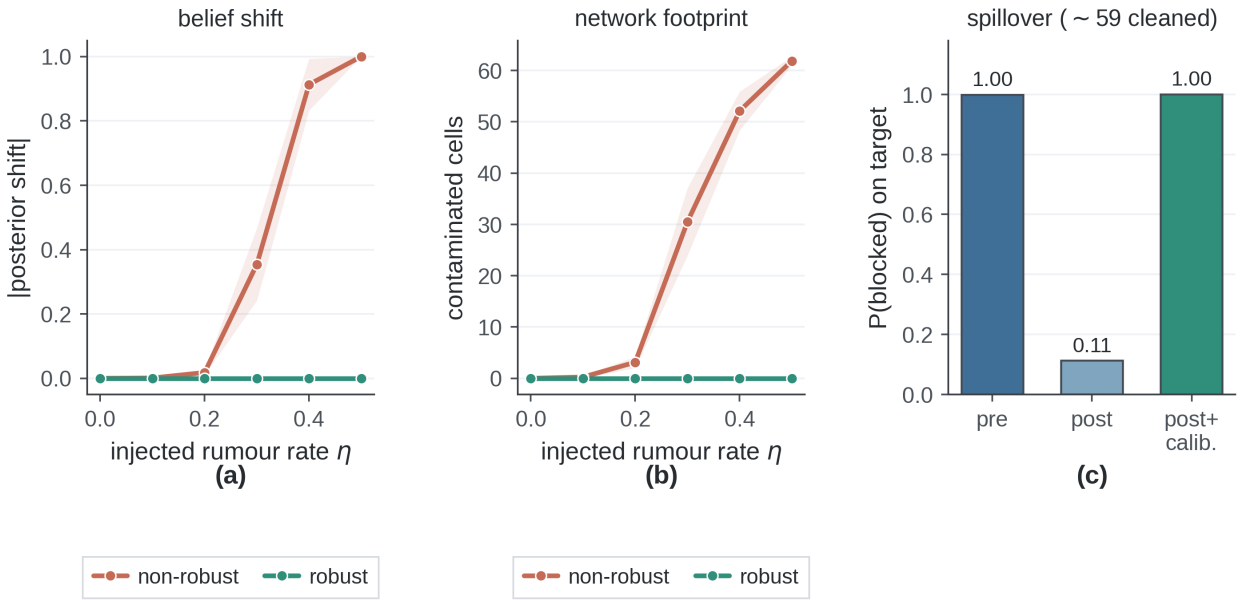


Fig. 6 Zhengzhou case: (a) the rumour-induced target-belief shift; (b) the network contamination footprint; (c) the calibration-spillover reversion

unreliable sources; the step here is to price a verification by the future reports its credibility correction will rescue, a forward-looking value that a static reliability weight cannot carry. From social-aerial sensing it borrows active inspection to confirm crowd signals; the step here is to redirect that inspection, since under a targeted attack the loudest reports are the worst place to look and $\tau^\star$ says when to inspect the source instead. What is new is not any single one of these ingredients but their decision-coupled synthesis: binding the robust credal update, the calibration-spillover value, and the contamination threshold to the downstream dispatch loss is what makes the verify-versus-calibrate switch fall out of the model rather than be imposed on it, and is what yields the closed forms for $\tau^\star$ and the spillover value that a static reliability weight or a generic distributionally robust hedge does not provide.

7.3 Generality and operational guidance

Nothing in the construction is specific to floods. It needs three ingredients: a cheap, high-volume human-sensor stream whose per-source reliability is estimated and manipulable; a scarce, trustworthy verifier; and a downstream decision with a measurable loss. Wherever those are present the same robust-credal update, calibration-spillover value, and contamination threshold apply—content moderation that escalates a limited human review, crowdsourced security testing that spends scarce expert audits, supply-chain intelligence that confirms unreliable tips. In each, the operative questions are the same, how much to trust a contaminated stream and where to spend the next trusted measurement, and $\tau^\star$ marks where verifying the object gives way to recalibrating its source. The flood setting supplies the spatial prior that lets one verification clean many reports; in a domain without that correlation the spillover is weaker, which sets the scope discussed next.

Table 3: R-NET: relief stranded on the real Zhengzhou network, scored against the trusted-consensus pseudo-oracle $X^\star$ (regret in stranded passenger-equivalents, 12 seeds); the rumour suppresses the artery carrying the most oracle-usable demand. Lower is better. In this test the proposed policy stays at 0 regret and delivers 33% of demand throughout, against about 25% for the baselines at η=0.5.

method	η=0	η=0.2	η=0.3	η=0.4	η=0.5
non-robust (point estimate)	0	33	175	267	300
coordinate-wise trimmed mean	0	83	225	292	300
multi-Krum	0	58	192	267	300
geometric median (RFA)	0	17	192	292	300
Dempster–Shafer	0	233	233	233	233
CRH / TruthFinder	0	292	300	300	292
KL / power-likelihood DRO	0	0	58	217	275
Wasserstein DRO	0	17	183	292	300
robust (proposed)	**0**	**0**	**0**	**0**	**0**

Operational guidance. The threshold gives an operator a dispatch rule. Below $\tau^\star$, send the UAV to the most decision-critical uncertain road; above it, when a corridor draws a coordinated rumour, spend the next sortie on that corridor to recalibrate the implicated source, which cleans tens of correlated reports at once and can seed debunking material downstream. The contamination-to-budget frontier (Section 6.4) turns an estimate of contamination into a verification budget. The same calibrate-the-source branch holds against an adaptive adversary: an attacker who anticipates a single road verification can evade it by flooding a different high-demand corridor (S11), but recalibrating the source discounts its reports network-wide, so changing the targeted corridor does not help. The practical reading is to escalate from verifying roads to calibrating sources as contamination grows heavy and possibly adaptive.

7.4 Scope and limitations

The boundary of the result is set by its assumptions. The closed-form $\tau^\star$ rests on a stylised two-source pivotal model and is best read as a mechanism and a locator: it explains why the switch exists and places it near $\tau^\star \approx 0.13$–0.16, while the operative value on a given network is the one estimated from data. The rule needs an agency that can attribute reports to a small set of source classes and verify a segment near-perfectly; where attribution is coarse or verification noisy, the spillover shrinks and the switch flattens. The influence cap, moreover, bounds per-source leverage on each segment but not attack *breadth*: an adversary who stays below the flood-rejection threshold on every segment yet disperses injections across many corridors is not bounded by it. Appendix E.4 maps the resulting defensible-versus-collapse boundary—which sits exactly at the rejection threshold—and names the defences that close it (a lower threshold, cross-segment correlation checks, or broader verification). The quantitative study uses a synthetic benchmark with stylised demand and topology, and the Zhengzhou replay has no surveyed road-state ground truth. Its decisions are therefore scored against a trusted-consensus pseudo-oracle built from the verified reports (Section 6.6), the best available reference rather than a survey, while the belief-shift quantities remain truth-free (Appendix F). On this case the shrink-and-reject operator tracks the exact min-max decision

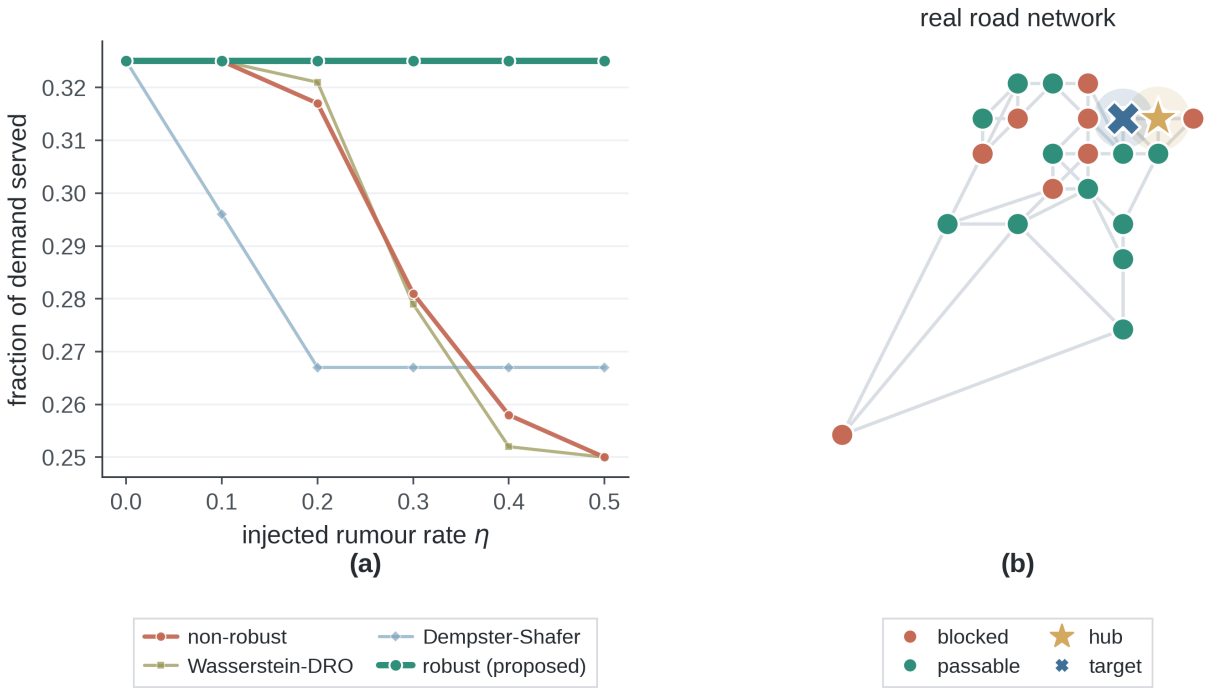


Fig. 7 R-NET on the Zhengzhou network. (a) fraction of true demand actually delivered under the suppression rumour, for non-robust, Wasserstein-DRO and Dempster–Shafer fusion against the proposed policy; (b) the real road network on the 22 occupied cells—blocked (red) and passable (green) under $X^\star$, the relief hub ($\star$) and the rumour-targeted artery ($\times$). Per-method decision regret against the pseudo-oracle is reported in Table 3

within about 27 stranded passengers of mean regret while the exact solve further controls the tail. A field trial in which verifications are flown rather than simulated is the natural next step, as is learning the source attribution online.

8 Conclusions

This paper has shown that crowd misinformation in flood response can be treated as a joint trust-allocation and sensing-allocation problem. The robust belief update converts uncertain source credibility and bounded injection into a conservative credal posterior; a UAV verification carries a calibration-spillover value beyond the road state it reveals; and the threshold $\tau^\star$ gives an operational rule for when to verify a road and when to recalibrate the contaminated source. Across the synthetic benchmark, the Zhengzhou rumour replay, the R-NET pseudo-oracle test, and the real-network sensitivity sweep, targeted rumours damage point-estimate and competing fusion rules, while the robust policy caps the induced dispatch loss in the tested settings.

The practical implication is that emergency managers should not treat crowd reports only as events to be confirmed. Under coordinated contamination, the more valuable action can be to calibrate the reporting source, because that trusted measurement re-weights many related claims and protects the downstream relief decision. The present evidence remains bounded by

stylised theory, a synthetic benchmark, and a Zhengzhou replay without surveyed road-state ground truth. The next step is a field or incident-room trial in which UAV verifications, source attribution, and dispatch choices are logged jointly, so the contamination-to-budget frontier can be turned into an operational sensing-budget valuation.

A Proofs

Lemma (credal posterior is compact convex). If each $\mathcal{U}_c(\varepsilon_c)$ is compact convex, which holds for the rectangular credibility box, and $\theta \mapsto \pi_\theta(\cdot \mid R)$ is continuous, then $\overline{\Pi}(R)$ in (4) is nonempty, compact, and convex. The product $\prod_c \mathcal{U}_c$ is compact convex, and the η-contamination neighbourhood of R is a compact family of finite report multisets up to the injection budget. The image of a compact set under a continuous map is compact, and its closed convex hull is compact convex by Carathéodory's theorem in the finite-dimensional simplex over $\{0,1\}^n$.

Proof of Theorem 1. (a) By the lemma the inner maximum of the continuous concave Φ over the compact convex $\overline{\Pi}(R)$ is attained, and the outer minimum over the compact decision set $\mathcal{A}$ (or its mixed extension) is attained; finiteness follows from boundedness of L. The weak minimax inequality gives $\min_a \max_\pi \mathbb{E}_\pi L \geq \max_\pi \min_a \mathbb{E}_\pi L = \max_\pi \Phi(\pi) \geq \Phi(\pi_{\theta^\star})$, and $a^\star_{\mathrm{rob}}$ is feasible under $\theta^\star$, which is conservativeness. (b) Taking a supremum or infimum over a smaller set cannot increase the worst case or the spread $\sup_{\pi\in\overline{\Pi}} \Phi(\pi) - \inf_{\pi\in\overline{\Pi}} \Phi(\pi)$. (c) The credibility box shrinks to $\{\hat{\theta}\}$ as $\varepsilon \to 0$, and the contamination neighbourhood collapses to $\{R\}$ as $\eta \to 0$; continuity of $\theta \mapsto \pi_\theta$ gives $\overline{\Pi}(R) \to \{\pi_{\hat{\theta}}(\cdot \mid R)\}$, so the verification value (6) reduces to the single-posterior known-credibility CVoI. □

Proof of Theorem 2. Verifying S reveals X_S, so for each contributing class c the empirical confusion of c on S updates $\hat{\theta}_c$ and changes its log-likelihood ratio by $\Delta\lambda_c$, the Beta–Bernoulli posterior-mean shift. In injection-blind fusion the posterior log-odds of every segment e carrying a c-report move by $\Delta\lambda_c$ per such report; weighting by the decision sensitivity $\omega_e = |\partial\Phi/\partial\mathbb{P}(X_e{=}1)|$ and summing over the $N_c - m$ unverified pivotal segments gives the equality in (8). The asymptotic equivalence uses Assumption 1(ii). The ratio statement follows from $\mathrm{RV}_{\mathrm{state}} \approx \kappa$, so the spillover dominates state clarification exactly when $N_c\bar{\kappa}\, p(\tau) > \kappa$. □

Proof of Theorem 3. Let $g(\tau) = \Delta V_{\mathrm{cal}}(\tau) - \Delta V_{\mathrm{road}}(\tau) = N_c\bar{\kappa}\, p(\tau) - \kappa[1 - p(\tau)(1 - c_0)]$. Then $g(0) = -\kappa < 0$ and $g(1^-) = N_c\bar{\kappa} - \kappa c_0 > 0$ because $N_c\bar{\kappa} > \kappa \geq \kappa c_0$. As a nonnegative combination of the continuous nondecreasing p, the function g is continuous and strictly increasing wherever p is, so the intermediate value theorem gives a unique sign change at $\tau^\star$, which is the stated dichotomy. Implicit differentiation of $g(\tau^\star) = 0$ gives $\partial\tau^\star/\partial N_c < 0$, $\partial\tau^\star/\partial\bar{\kappa} < 0$, and $\partial\tau^\star/\partial\kappa > 0$, since g increases in τ and, at fixed τ, increases in N_c and $\bar{\kappa}$ and decreases in κ. Setting $g = 0$ with $p(\tau) = \sigma(s(\tau - \tau_0))$ and $c_0 = 0$ and solving for τ yields (9). □

Proof of Theorem 4. (a) Each injected report adds $\lambda_c > 0$ to the target log-odds, so $k\lambda_c \to \infty$ drives $\mathbb{P}(X_{e^\star}{=}1)$ to a degenerate value against the truth, flipping the pivotal decision and incurring its full swing κ; the worst case over targets selects the maximal-κ pivotal segment, and the achievable k grows with the budget η. (b) Trimming caps the count per (segment, source, claim direction) at T_c, and rejection removes anomalously concentrated floods, so the total contribution of any source to any segment's log-odds is at most $T_c\lambda_c$; summing over the finite class set C bounds the shift by $\Lambda_T = \sum_c T_c\lambda_c$, a constant free of η. Lipschitz continuity of Φ in the target marginal then bounds the robust decision-loss inflation. □

Proof of Theorem 5. Let $g(\tau) = \Delta V_{\mathrm{cal}}(\tau) - \Delta V_{\mathrm{road}}(\tau) = N_c\bar{\kappa}\,p(\tau) - \kappa + \kappa\,p(\tau)(1 - qc_0)$. Then $g(0) = -\kappa < 0$ and $g(1^-) = N_c\bar{\kappa} - \kappa qc_0 > 0$ since $N_c\bar{\kappa} > \kappa \geq \kappa qc_0$, and g is continuous and strictly increasing wherever p is, so the intermediate value theorem gives a unique zero $\tau^\star_{\mathrm{eq}}$ with the stated dichotomy. Solving $g = 0$ for the crossing probability gives $p^\star = \kappa/[N_c\bar{\kappa} + \kappa(1 - qc_0)]$; inverting the logistic law $p = \sigma(s(\tau - \tau_0))$ yields (10). Since $p^\star$ is increasing in q and σ^{-1} is increasing, $\tau^\star_{\mathrm{eq}}$ is increasing in $q = 1/N_t$, hence decreasing in N_t; at $q = 1$ the denominator is that of Theorem 3, so $\tau^\star_{\mathrm{eq}} = \tau^\star$, and for $q < 1$ it is strictly smaller. □

Theorem 6 (Value of robustness as a Stackelberg equilibrium)**.** *Consider the fusion subgame on a pivotal corridor with genuine decision margin $M > 0$ and injected per-report log-likelihood ratio $\mu > 0$: the agency commits to an influence cap T (the per-(segment, source, direction) trim cap of Theorem 4; $T = \infty$ is point-credibility fusion) and a budget-η attacker best-responds. Capping discards genuine reports beyond T on a truly blocked corridor at an expected* sensitivity tax *$S(T) \geq 0$, nonincreasing in T with $S(\infty) = 0$.*

- *(a) If $T = \infty$, the attacker flips the corridor for every budget admitting $k \geq M/\mu$ injected reports, so the equilibrium loss is the full swing κ, uncapped in attack magnitude.*
- *(b) If $T < M/\mu$, then $T\mu < M$ and no budget flips the corridor; the equilibrium loss is the sensitivity tax $S(T)$,* independent of η.
- *(c) The agency's optimal cap is the largest safe one, $T^\star = \sup\{T : T\mu < M\}$, and the value of robustness* $\mathrm{VoR} = \kappa - S(T^\star)$ *is positive whenever $S(T^\star) < \kappa$ and constant in the attack budget.*

Proof of Theorem 6. (a) Point-credibility fusion adds μ per injected report, so k reports shift the target log-odds by $k\mu$, which reaches M once $k \geq M/\mu$ and flips the pivotal decision, incurring its full swing κ; further budget cannot raise a single corridor's loss above κ, so the loss saturates at κ. (b) Under the cap the per-source shift is at most $T\mu$; if $T\mu < M$ the margin is never overcome and the decision is unchanged for every k, hence every η, leaving only the discarded genuine evidence $S(T)$. (c) On $\{T : T\mu < M\}$ the loss is $S(T)$, nonincreasing in T, so it is minimised at the supremum $T^\star$; for $T\mu \geq M$ the attacker flips the corridor and the loss jumps to $\kappa \geq S(T^\star)$. Thus $T^\star$ is the leader's best response, and $\mathrm{VoR} = \kappa - S(T^\star)$ follows by subtraction and is constant in η because neither term depends on it. □

Table 4: Per-class calibrated credibility and robustness parameters. The ambiguity half-width is $z_{1-\delta}/\sqrt{n_c}$ with $z_{1-\delta} = 1.6$ ($\delta \approx 0.05$); the radius used in a run is $\varepsilon_c = \eta + z_{1-\delta}/\sqrt{n_c}$ with η the injection budget. T_c is the per-(segment, source, direction) trim cap.

class	α_c	β_c	n_c	$z_{1-\delta}/\sqrt{n_c}$	T_c
eyewitness	0.90	0.84	2303	0.033	8
relay	0.72	0.68	2066	0.035	3
low-credibility	0.38	0.50	749	0.058	2

B Data-generating process and experimental setup

Instances are drawn on an $n_{\text{side}} \times n_{\text{side}}$ grid ($n_{\text{side}} = 10$ for the synthetic study, 11 for the case). The latent passability field is a Gaussian copula with correlation length ρ (default 2) thresholded at a base impassability rate (default 0.35), producing spatially clustered closures. Each segment receives a Poisson number of reports whose source class is sampled from the calibrated mix of Section 3.2 and whose claim is drawn from the class sensitivity and specificity; coverage is denser at the grid centre than at the edge. Each cell k carries a demand d_k (drawn heterogeneously across corridors, with the grid centre weighted higher), and the bridging decision allocates a relief level a_k of per-unit capacity u. A passable cell serves $\min(u\,a_k, d_k)$ passengers and an impassable one serves none, so the loss is the stranded demand $L(a, X) = \sum_k d_k - \sum_{k:\,X_k=0} \min(u\,a_k, d_k)$; this is the $L(a, X)$ of Eq. (1) and the throughput maximised in Appendix F.1. Contamination is injected as in Section 3.3: random places claims uniformly, panic concentrates "blocked" claims in report-dense cells, and targeted rumour floods the highest-demand usable corridor with truth-opposite claims. The robust policy applies the worst-case credibility tilt with ambiguity (shrink) radius $\varepsilon_c = \eta + 1.6/\sqrt{n_c}$ (the $(1-\delta)$ calibration confidence half-width plus the injection budget, $\delta \approx 0.05$) and the influence-bounded aggregation of Section 4.1; the trim caps are 8, 3, and 2 reports per (segment, source, direction) for the eyewitness, relay, and low-credibility classes. Reported regret is the downstream bridging loss relative to a truth-aware oracle, averaged over the seeds noted with each experiment. Beliefs use $M = 1000$ scenario particles, and the exact min-max decision ranges each source class over a credibility box of radius 0.45. Table 4 collects the per-class calibrated values and the robustness parameters used in every run.

Shared settings: grid 10×10 (synthetic) / 11×11 (case); correlation length $\rho = 2$; base impassability rate 0.35; $M = 1000$ particles; exact-min-max credibility box radius 0.45; rectangular credibility ambiguity box; seeds 12 (headline tables), 16 (regime panel), 24 (fine threshold grid). Software: Python 3.10.12, NumPy 2.2.6, pandas 2.3.3, Matplotlib 3.10.9; single-thread Basic Linear Algebra Subprograms (BLAS) and a fixed common-random-number stream make the runs bit-reproducible.

C Notation

Table 5 summarises the notation used in the formulation, theory, and appendix experiments.

Table 5: Main notation.

Symbol	Meaning
E, n	road segments and their number
$X_e \in \{0,1\}$	passability state of segment e (1 = impassable)
$a \in \mathcal{A}$, $L(a, X)$	bridging decision and its loss (stranded passenger-equivalents)
$\Phi(\mu)$	Bayes loss (1), concave in μ
$r = (e, y, c)$	a crowd report: segment, claim, source class
$\theta_c = (\alpha_c, \beta_c)$	sensitivity and specificity of source class c
λ_c	per-report log-likelihood ratio of class c, (2)
$\mathcal{U}_c(\varepsilon_c)$	rectangular credibility ambiguity box, half-width $\varepsilon_c = \eta + z_{1-\delta}/\sqrt{n_c} = \Theta(n_c^{-1/2})$
n_c	calibration sample size of class c
δ, $z_{1-\delta}$	calibration confidence level and its standard-normal quantile
η, τ	attack budget; contamination rate on the target
$\overline{\Pi}(R)$	robust credal posterior (4)
$a^\star_{\mathrm{rob}}$, V_{rob}	robust decision and value (5)
$\kappa = \mathrm{CVoI}(\{e^\star\})$	decision swing (criticality) of pivotal segment $e^\star$
N_c, $\bar{\kappa}$	pivotal segments touched by class c; their mean criticality
$p(\tau)$	boundary-crossing probability of injection-blind fusion
$\mathrm{RV}_{\mathrm{state}}, \mathrm{RV}_{\mathrm{calib}}$	state-clarification and calibration-spillover value (6)
$\tau^\star$	contamination threshold, verify $\leftrightarrow$ calibrate
Λ_T	influence bound under trimming, Theorem 4(b)

D Reproducibility

All code, de-identified data, fixed-seed outputs, and theory checks are supplied as Online Resource 1. The synthetic experiments are produced by `run_b_experiments.py` (S1–S5), the competitive-baseline comparison and its verification extension (S1b, Section 6.2) by `run_b_baselines.py`, and the additive regime and smoothing runs by `run_b_extra.py`; the exact min-max credal decision (S6), the calibration, coverage, and tail-risk metrics (S7), and the ablations (S8) are produced by `run_b_minmax.py`, `run_b_metrics.py`, and `run_b_ablations.py`; the misspecification, certified-absorption, and adaptive-adversary robustness checks (S9–S11, Section E.3) by `run_b_robust_ext.py`, the dispersed-attack failure-region map (S12, Appendix E.4) by `run_b_dispersed.py`; the Zhengzhou case is produced by `run_b_realcase.py`, the real-network falsifiable decision metric (R-NET, Section 6.6) by `run_b_realnet.py` (sensitivity sweep `run_b_realnet_sensitivity.py`), and the truth-free min-max, tail, and ablation extensions (Appendix F) by `run_b_realcase_ext.py`; the structural results are checked numerically by `verify_b_theory.py`, and the Stackelberg-equilibrium results (Theorem 5) by `verify_b_stackelberg.py`. The fusion uses the modules `belief.py`, `downstream.py`, and `cvoi.py` together with the calibrated `credibility_priors.json` and the de-identified `reports_geocoded_real.csv`, which are bundled with the code. The reported runs use Python 3.10.12 with NumPy 2.2.6, pandas 2.3.3, and Matplotlib 3.10.9. To make the runs bit-reproducible, the rumour injection is locked to a fixed common-random-number stream and the numerical libraries are pinned to a single thread; each experiment uses instance seeds base $+ s$ for $s = 0, \ldots, N-1$ (base 100 for S1, 200 for S2, and so on), a belief seed $7 + s$, and a contamination stream offset by $+9000$, with $N = 12$ for the headline tables, 16 for the regime panel, and 24 for the fine threshold grid. Fixed seeds, the data-generating

Table 6: S7: calibration error, credal coverage, and critical-segment miss rate (targeted rumour). Lower ECE and miss rate are better; higher coverage is better.

τ	policy	ECE	credal coverage	critical miss rate
0.0	non-robust	0.099	0.711	0.513
0.0	robust credal	0.051	0.904	0.678
0.3	non-robust	0.135	0.669	0.500
0.3	robust credal	0.058	0.906	0.755

process, and the scripts that regenerate every table and figure are provided in the anonymized supplementary archive and will be deposited on a public repository upon acceptance.

E Extended synthetic results and ablations

These analyses extend the main results of Section 6; they are summarised there and reported in full here.

E.1 Calibration, coverage, and tail risk

Mean regret is not the whole story; Table 6 looks at calibration, coverage, and the tail. The expected calibration error (ECE) of the point posterior against the realized state falls from 0.14 under non-robust fusion to 0.06 under robust fusion at $\tau = 0.3$, and the robust credal band keeps the correct decision admissible for about 91% of segments, against 67% for the non-robust point estimate; this coverage does not degrade as contamination rises. Conservatism has a cost: on genuinely blocked, decision-critical segments the robust posterior calls a road passable more often than the non-robust one (miss rate 0.755 versus 0.500 at $\tau = 0.3$, Table 6), a sensitivity loss the operator can dial through the ambiguity radius (Appendix E.5).

E.2 The exact min-max credal decision

The robust policy above decides against a single shrink-and-reject posterior. Because the bridging loss is separable and decreasing in each corridor's usable probability, the exact min-max credal decision $a^\star_{\mathrm{rob}} \in \arg\min_a \max_{\pi \in \overline{\Pi}(R)} \mathbb{E}_\pi[L(a, X)]$ reduces to a concave maximin over the allocation, which we solve over the vertices of the credibility ambiguity box (Appendix D). Table 7 compares it to the shrink+drop approximation and to non-robust fusion. The exact solve raises mean regret by about 27 stranded passengers relative to shrink+drop, but it lowers the conditional value-at-risk (CVaR) (Rockafellar and Uryasev, 2000) regret tail at $\tau = 0.4$ from 503 to 383, which is what a worst-case decision is built to do; both remain far below the non-robust tail of 594. The shrink+drop operator is therefore a sound mean-optimal surrogate, while the exact min-max is the right choice when the tail dominates the operator's concern. This credal view follows the maximin-expected-utility and imprecise-probability tradition (Gilboa and Schmeidler, 1989; Walley, 1991).

Table 7: S6: downstream regret of the exact min-max credal decision against the shrink+drop approximation and non-robust fusion (targeted rumour, 12 seeds).

τ	0	0.2	0.3	0.4	0.5
non-robust	253±68	318±73	318±67	383±56	390±41
shrink+drop	261±57	264±47	190±55	220±38	221±46
exact min-max	273±41	280±33	258±45	206±37	247±45

E.3 Robustness validation: misspecification, certified absorption, and an adaptive adversary

We stress-test the robustness from three angles (Figure 8). *(S9) Attacker-budget misspecification.* The robust policy hedges against an *assumed* attack budget through its ambiguity radius. Sweeping the assumed budget against the true contamination shows the layer fails safe: under-hedging (an assumed budget near zero) lets regret blow up to 358 at τ=0.4, whereas any positive radius caps it near 90–160, and over-hedging when contamination is absent costs only a small sensitivity tax. The penalty for over-trust is unbounded; the penalty for over-caution is bounded. *(S10) Certified absorption radius.* For each instance we report $\eta^\dagger$, the largest injection rate at which realised regret stays within 1.6× its clean value. Robust fusion certifies a median $\eta^\dagger \approx 0.50$ against 0.30 for non-robust: the realised-decision form of the bounded-versus-unbounded-influence result of Theorem 4. *(S11) Adaptive adversary.* The headline rumour is oblivious; a stronger adversary that knows the verify-the-most-critical response and best-responds, choosing the corridor that maximises post-response regret, drives non-robust regret from 243 to 401. The robust policy still caps it, 273 against 401, a gap of 128 stranded passengers with a paired-bootstrap 95% confidence interval of $[36, 216]$. The cap therefore holds against an adversary that adapts to the verification policy, the equilibrium predicted by Theorem 5: a relocating attacker lowers the verify-to-calibrate threshold (10) yet cannot beat source calibration.

E.4 A dispersed attack the per-segment cap does not bound (S12)

Theorem 4(b) caps the log-odds shift *per segment*; it does not, by itself, bound an adversary who spreads small injections across *many* pivotal segments. We probe this boundary directly, attacking the defence where it is not designed to hold. Instead of flooding one corridor, the adversary injects at most REJECT $- 1 = 2$ identical "blocked" claims per (segment, source, value)—below the flood-rejection threshold and within the trim cap on every segment—but across the segments of the top-K highest-demand corridors. We sweep the attack depth m (claims per triple) and breadth K (corridors hit) and map the robust policy's downstream regret (Figure 9; the clean, no-attack robust regret is 188 for reference).

The result delimits where the defence holds. Below the rejection threshold ($m \leq 2$) the robust regret *grows with breadth*, from 284 at K=1 to 548 at K=10: the per-segment cap lets a stealthy attacker accumulate decision damage across corridors. At or above the threshold ($m \geq 3$) flood-rejection discards the injected groups and the regret falls back toward the clean baseline (to 181–277 at m=3). The phase boundary therefore sits exactly at $m =$ REJECT.

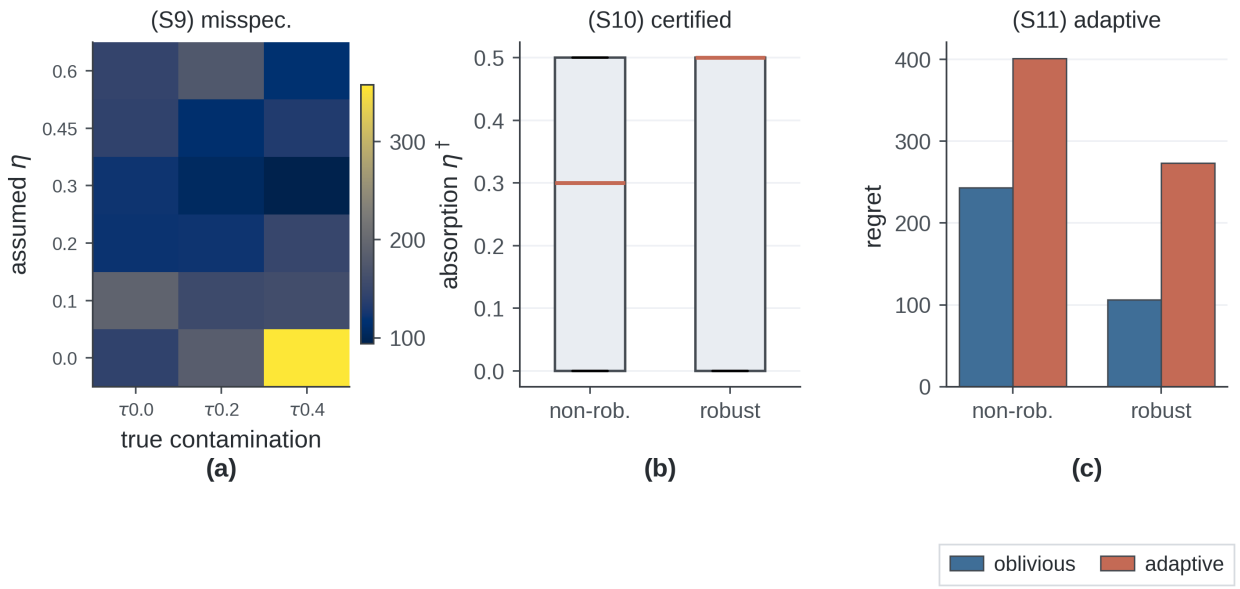


Fig. 8 Robustness validation. (left, S9) regret over assumed attack budget × true contamination: the robust layer degrades gracefully, with an unbounded penalty for over-trust and a bounded one for over-caution. (centre, S10) distribution of the certified absorption radius $\eta^{\dagger}$, robust versus non-robust. (right, S11) oblivious versus adaptive adversary; the robust policy caps the damage even against the best-responding attacker

At a matched total injection budget the contrast is direct: a dispersed stealth attack drives robust regret to 498 at K=10, whereas pouring the same number of reports onto a single corridor—the concentrated flood the layer is built for—stays capped at 207–289.

This is the honest boundary of the rule. The trim-and-reject layer bounds per-source, per-segment leverage, not attack breadth; the defences that would are a lower rejection threshold (at the cost of discarding more genuine bursts), cross-segment correlation checks on injected patterns, or escalating verification breadth as more corridors draw coordinated reports. It also clarifies the real-network result of Section 6.6: that network's relief decision is dominated by a single artery (only a few passable, demand-carrying cells), so the effective suppression threat there is concentrated and the robust policy's near-zero regret reflects the threat geometry, not the pseudo-oracle scoring—on a network with the breadth a dispersed attack needs, the same policy exhibits the non-trivial failure region mapped here.

E.5 Ablation studies

The synthetic design-sensitivity ablations of Section 6, summarised in the main text, are reported in full here.

Figure 10 varies four design quantities at $\tau = 0.3$. The ambiguity radius trades robustness for sensitivity: the genuine-block miss rate climbs from 0.51 to 0.72 as the radius grows, while regret is minimized near a radius of 0.2. Non-robust regret grows with the low-credibility source fraction, from 244 to 321, whereas robust fusion contains it between 100 and 161. Robust regret is mildly U-shaped in the calibration sample size and is lowest near the calibrated value. The footprint of a single-cell rumour stays large, between 28 and 38 cells,

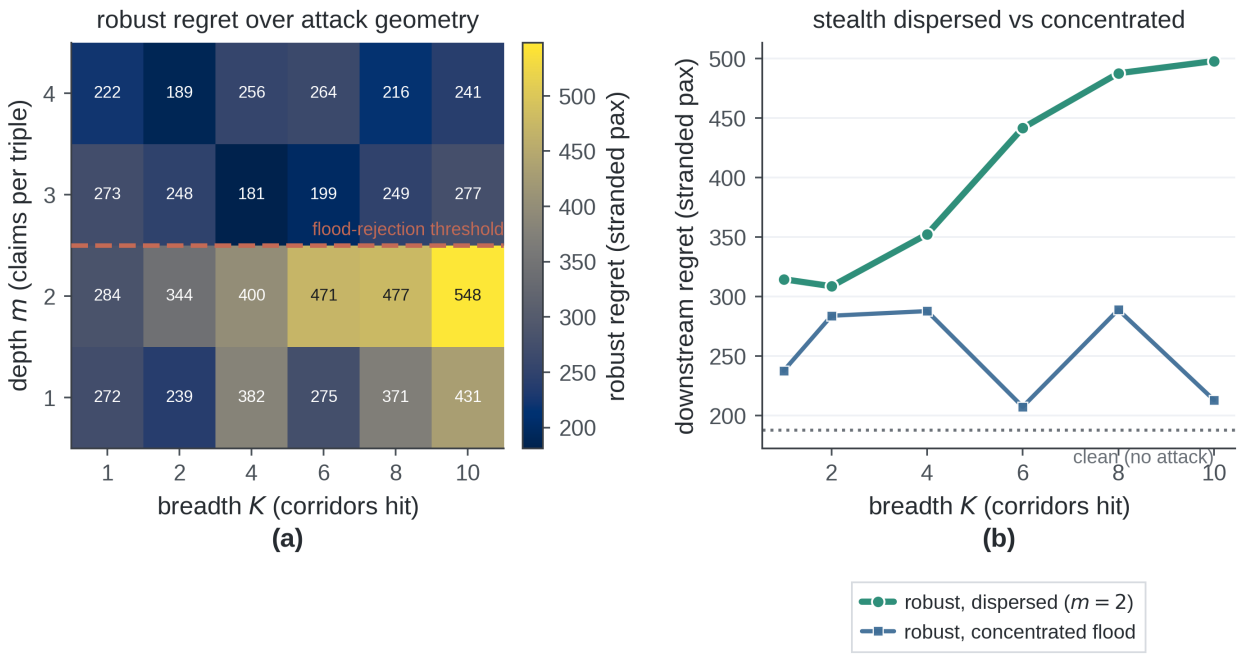


Fig. 9 S12 defensible-versus-collapse boundary. (a) Robust regret over attack depth m (claims per (segment, source, value)) and breadth K (corridors hit): below the flood-rejection threshold ($m \leq 2$) the regret grows with breadth, since the per-segment cap does not bound a dispersed attack; at or above it ($m \geq 3$) the flood is rejected and the regret returns toward the clean baseline. (b) At a matched injection budget, a dispersed stealth attack (m=2) breaks the cap while a concentrated flood of the same size stays bounded

across correlation lengths, which is why a local injection is a network-level threat.

F Real-network robustness extensions

These truth-free analyses carry the synthetic robustness checks of Section 6 onto the Zhengzhou network; they are summarised in Section 6.5 and reported in full here.

F.1 Exact min-max relief allocation on the real network

Because the rumour attacks a single cell, the decision it actually distorts is the allocation of a limited relief fleet across the 22 occupied cells, each serviceable only if passable (full-row corridors are degenerate on the sparse real grid, so the single-cell relief level is used). For each policy we report the served demand *guaranteed* over the credal posterior set, $\min_v \sum_k P_{v,k} \min(u\, a_k, d_k)$—where v indexes the worst-case posterior in $\overline{\Pi}(R)$, $P_{v,k}$ is the probability that cell k is passable under v, and a_k, u, d_k are the relief allocation, per-unit capacity, and demand of Appendix B—which the exact min-max decision maximises by construction. Table 8 shows that once the injection rate reaches 0.4 the non-robust allocation pours the full relief unit into the rumour-hijacked cell (target share $\to 2.0$) and its worst-case guarantee collapses, while the robust and exact min-max decisions refuse to fund it. The exact min-max strictly dominates the shrink+drop approximation on the guarantee (about

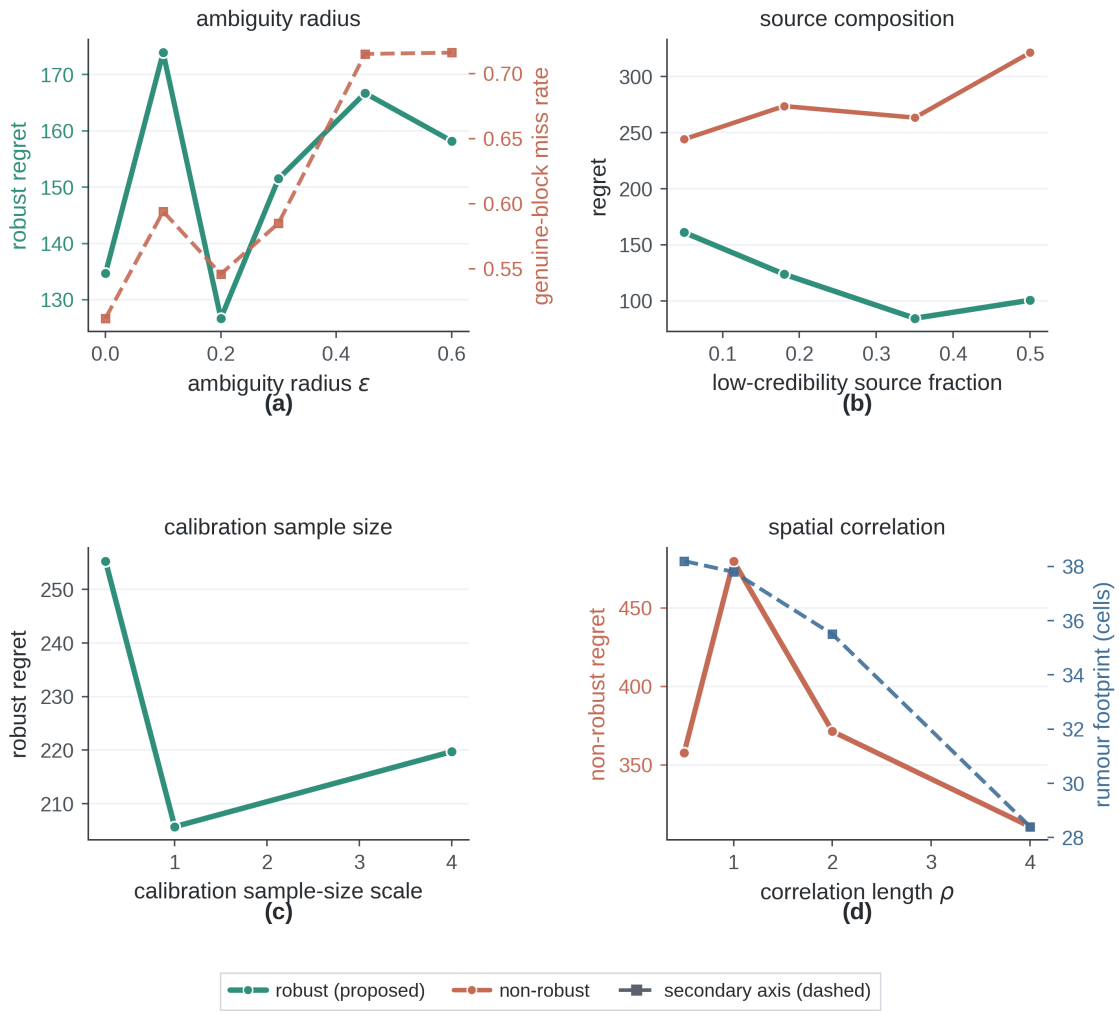


Fig. 10 Ablations at $\tau = 0.3$: (A) ambiguity radius (robustness versus sensitivity), (B) source composition, (C) calibration sample size, (D) spatial correlation length

520 versus 467 of 4400 total demand), and a self-contained dual certificate confirms optimality to within 0.1 stranded passengers: on real data the shrink+drop operator is already near worst-case optimal, and the exact solve both improves and certifies it, mirroring the synthetic S6.

F.2 Tail risk and decision coverage on real data

Figure 11 gives the truth-free analogues of the calibration metrics. The left panel is the tail of the rumour-induced belief shift over injection realisations at $\eta = 0.4$: the non-robust posterior has a mean shift of 0.79 and a CVaR(20%) of 1.00; its worst realisations flip the target outright, while the robust posterior holds both at zero. The right panel is credal coverage of the pre-attack consensus decision: the non-robust point estimate stays admissible only up to $\eta \approx 0.2$ and reaches zero coverage by $\eta = 0.5$, whereas the robust credal band keeps the consensus admissible throughout (coverage 1.0). The robust layer controls exactly

Table 8: R2: worst-case relief throughput *guaranteed* over the credal posterior set on the Zhengzhou network (served passengers out of 4400 total demand; higher is better), and the relief share poured into the rumour-targeted cell.

injection rate η	0	0.1	0.2	0.3	0.4	0.5
non-robust (point estimate)	318	338	425	471	366	360
shrink+drop	467	464	467	467	467	467
exact min-max	520	520	520	520	520	520
relief share on target, non-robust	0	0	0	0	2.0	2.0

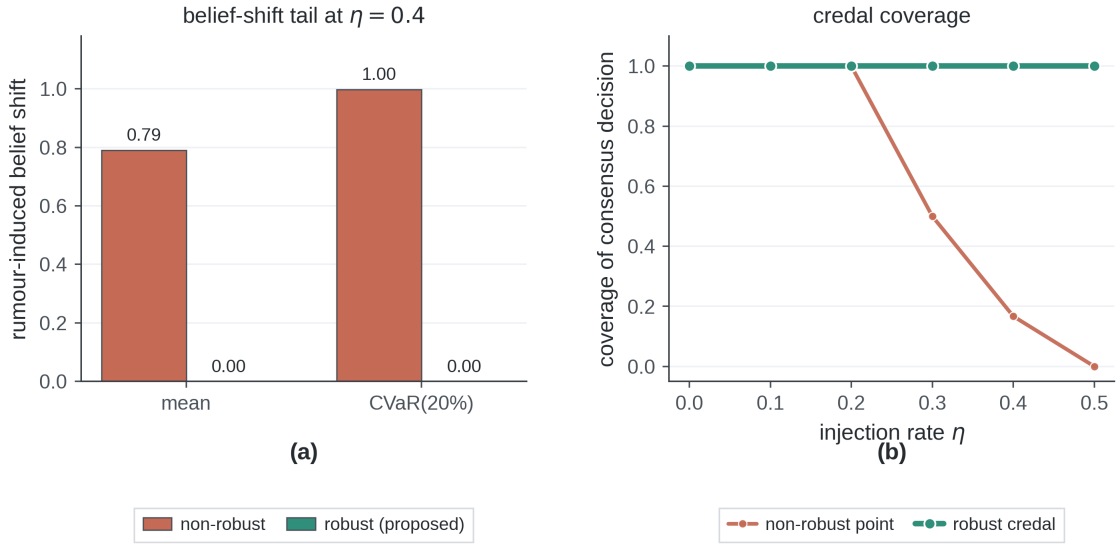


Fig. 11 R3 on Zhengzhou: (left) mean and CVaR(20%) of the rumour-induced target-belief shift at $\eta = 0.4$; (right) credal coverage of the pre-attack consensus decision versus the injection rate. The robust credal band pins both the tail and the coverage; the non-robust point estimate loses both

the decision tail that the min-max objective targets.

F.3 Ablations on the real network

Figure 12 ablates the two robustness levers on the real data. The left panel decomposes the robust pipeline: injection-rejection trimming is the load-bearing defence, holding the rumour shift near zero across all ambiguity radii, whereas credibility shrink alone (trimming disabled) cannot stop the numerical flood (shift 0.6–0.7) because it also dilutes the genuine evidence. The ambiguity radius instead governs sensitivity: the real "blocked" signal is retained until the radius exceeds about 0.6, beyond which even eyewitness evidence is washed toward 0.5, which bounds the usable hedging range. The right panel varies the copula correlation length: a single-cell rumour leaves a footprint of 25 to 82 cells, largest near $\rho \approx 1$, confirming on real data that a local injection is a network-level threat, in the same direction as the synthetic S8(D).

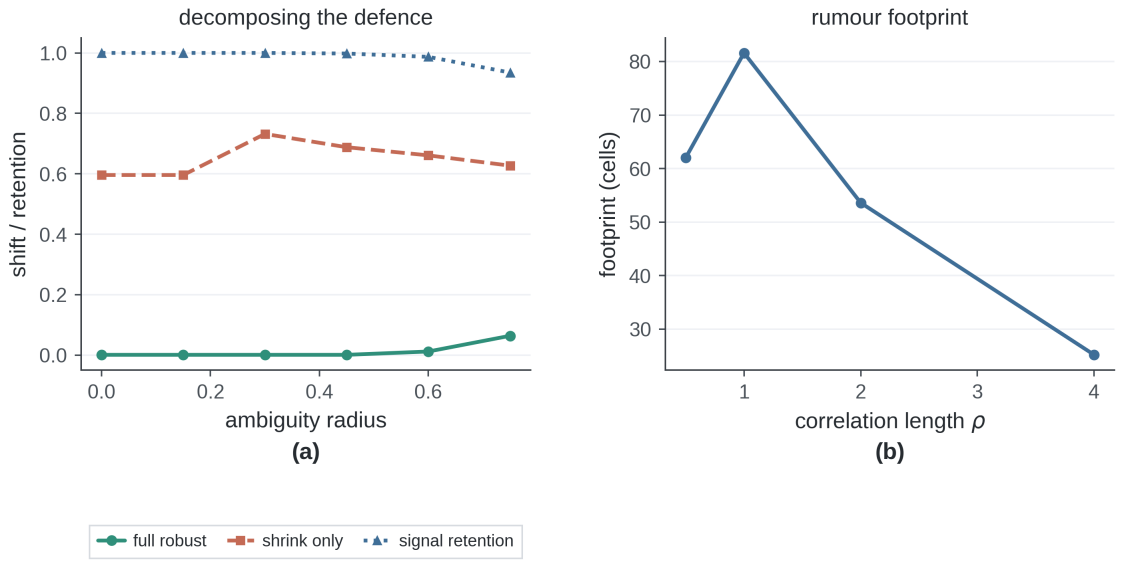


Fig. 12 R4 on Zhengzhou: (left) decomposing the robust pipeline, where injection-rejection trimming carries the rumour resistance, while the credibility-shrink radius governs sensitivity (real-signal retention); (right) a single-cell rumour's network footprint against the spatial correlation length

F.4 Sensitivity of the real-network result to its construction

Because the real-network testbed fixes a block fraction (0.35), a graph connectivity (k=3), and a fleet scarcity (0.6), we vary each in turn over fourteen configurations and re-run the suppression rumour at $\eta = 0.4$. The proposed policy has zero regret against the pseudo-oracle in every configuration; the non-robust policy strands 90–270 passenger-equivalents wherever the suppressed artery is decision-critical, and the gap closes to zero only when a denser graph (k=4) gives the artery an alternate route or a scarcer fleet (≤ 0.5) never funds the contested corridors—in both cases the attack cannot change the dispatch and the two policies coincide. The headline is therefore not an artefact of a single construction choice; the full sweep is provided in the supplementary material.

G Competitive baselines: mapping and verification extension

Each baseline of Section 6.2 maps the report multiset to a per-segment evidence log-odds that then enters the shared spatial prior and dispatch, so the comparison isolates the aggregation rule from the prior and the decision: trimmed mean trims the extreme 20% of each segment's per-report log-likelihood ratios; multi-Krum and geometric-median (RFA) shard the reports and aggregate by Krum selection or the Weiszfeld median; Dempster–Shafer discounts by class reliability and combines by Dempster's rule; CRH/TruthFinder iterates reliability weights against the consensus; and the KL and Wasserstein-DRO updates temper or soft-threshold each report's log-likelihood ratio. The verification extension at $\tau = 0.3$—one sortie on the suspected-flood segment, clamping its state and then recalibrating the implicated source—is

reported across all methods in Figure 3: clamping helps every method, while the calibration-spillover recalibration (Theorem 2) adds a further large reduction for the credibility-based methods (a further 148 stranded passengers for multi-Krum and 100 for the KL update) and is naturally small for the proposed rule, which has already neutralised the flood, and for Dempster–Shafer and truth discovery, whose aggregation does not re-weight reports by the corrected credibility.

References

Adsanver, B., Coban, E., Balcik, B., 2025. A predictive multistage postdisaster damage assessment framework for drone routing. International Transactions in Operational Research 32, 626–668. https://doi.org/10.1111/itor.13429.

Alamri, B.H.S., Monowar, M.M., Alshehri, S., Zafar, M.H., 2025. A fog-assisted group-based truth discovery framework over mobile crowdsensing data streams. PLOS One 20, e0330656. https://doi.org/10.1371/journal.pone.0330656.

Chow, T.E., Yang, T., Yan, Y., 2026. The uncertainties and potential of locational-based social media for floodscape mapping. Big Earth Data https://doi.org/10.1080/20964471.2026.2646046. published online.

Drews, M., Steinhausen, M., Larsen, M.A.D., Dømgaard, M.L., Huszti, L., Rácz, T., Wortmann, M., Hattermann, F.F., Schröter, K., 2023. The utility of using volunteered geographic information (VGI) for evaluating pluvial flood models. Science of The Total Environment 894, 164962. https://doi.org/10.1016/j.scitotenv.2023.164962.

Du, X., Cheng, J., Wang, S., Li, H., Zhen, L., 2025. Bus bridging service optimisation for urban rail transit disruption under uncertainty. Transportmetrica A: Transport Science https://doi.org/10.1080/23249935.2025.2521638. published online.

Gilboa, I., Schmeidler, D., 1989. Maxmin expected utility with non-unique prior. Journal of Mathematical Economics 18, 141–153. https://doi.org/10.1016/0304-4068(89)90018-9.

Hilberts, S., Govers, M., Petelos, E., Evers, S., 2025. The impact of misinformation on social media in the context of natural disasters: Narrative review. JMIR Infodemiology 5, e70413. https://doi.org/10.2196/70413.

Howard, R.A., 1966. Information value theory. IEEE Transactions on Systems Science and Cybernetics 2, 22–26. https://doi.org/10.1109/TSSC.1966.300074.

Huang, L., Liu, Z., Yu, C., Yan, B., 2026. Resilience-oriented recovery optimization of metro systems under extreme rainfall-induced urban flooding disruptions. Sustainability 18, 2597. https://doi.org/10.3390/su18052597.

Huber, P.J., 1981. Robust Statistics. Wiley, New York.

Khraisat, A., Alazab, A., Alazab, M., Jan, T., Singh, S., Uddin, M.A., 2025. Securing federated learning: a defense strategy against targeted data poisoning attack. Discover Internet of Things 5, 18. https://doi.org/10.1007/s43926-025-00108-6.

Li, J., Ou, J., Luo, Y., Zhou, S., Lu, C., Zhou, Y., 2026. Resilience assessment of urban multimodal transportation networks under the impact of rainfall. Transportation https://doi.org/10.1007/s11116-025-10715-5. published online.

Liu, Y., Yang, T., Su, J., 2025. Optimal bus bridging service for urban rail transit disruptions with stochastic passenger demand. PLOS One 20, e0333686. https://doi.org/10.1371/journal.pone.0333686.

Mohajerin Esfahani, P., Kuhn, D., 2018. Data-driven distributionally robust optimization using the Wasserstein metric: performance guarantees and tractable reformulations. Mathematical Programming 171, 115–166. https://doi.org/10.1007/s10107-017-1172-1.

Nelsen, R.B., 2006. An Introduction to Copulas. 2nd ed., Springer, New York.

Nguyen, V.L., Zhang, H., Destercke, S., 2025. Credal ensembling in multi-class classification. Machine Learning 114, 19. https://doi.org/10.1007/s10994-024-06703-y.

Peng, J., Zhang, J., 2024. Spatiotemporal assessment of urban flooding hazard using social media: A case study of Zhengzhou "7·20". Environmental Modelling & Software 176, 106021. https://doi.org/10.1016/j.envsoft.2024.106021.

Qu, Z., Wang, J., Zhang, M., 2023. Mining and analysis of public sentiment during disaster events: The extreme rainstorm disaster in megacities of China in 2021. Heliyon 9, e18272. https://doi.org/10.1016/j.heliyon.2023.e18272.

Rashid, M.T., Wei, N., Wang, D., 2023. A survey on social-physical sensing: An emerging sensing paradigm that explores the collective intelligence of humans and machines. Collective Intelligence 2. https://doi.org/10.1177/26339137231170825.

Rashid, M.T., Zhang, D.Y., Wang, D., 2020a. DASC: Towards a road damage-aware social-media-driven car sensing framework for disaster response applications. Pervasive and Mobile Computing 67, 101207. https://doi.org/10.1016/j.pmcj.2020.101207.

Rashid, M.T., Zhang, D.Y., Wang, D., 2020b. SocialDrone: An integrated social media and drone sensing system for reliable disaster response, in: IEEE INFOCOM 2020 – IEEE Conference on Computer Communications, pp. 218–227. https://doi.org/10.1109/INFOCOM41043.2020.9155522.

Rockafellar, R.T., Uryasev, S., 2000. Optimization of conditional value-at-risk. Journal of Risk 2, 21–41. https://doi.org/10.21314/JOR.2000.038.

Seif, M., Tosarkani, B.M., Zolfagharinia, H., 2026. Enhancing humanitarian logistics under uncertainty: A data-driven distributionally robust optimization approach with worst-case mean-CVaR. Transportation Research Part E: Logistics and Transportation Review 205, 104516. https://doi.org/10.1016/j.tre.2025.104516.

Vosoughi, S., Roy, D., Aral, S., 2018. The spread of true and false news online. Science 359, 1146–1151. https://doi.org/10.1126/science.aap9559.

Walley, P., 1991. Statistical Reasoning with Imprecise Probabilities. Chapman and Hall, London.

Wang, D., Yang, K., Yang, L., Dong, J., 2023a. Two-stage distributionally robust optimization for disaster relief logistics under option contract and demand ambiguity. Transportation Research Part E: Logistics and Transportation Review 170, 103025. https://doi.org/10.1016/j.tre.2023.103025.

Wang, Y., Zhou, Y., Yang, H., Yan, X., 2023b. Integrated optimization of bus bridging service design and passenger assignment in response to urban rail transit disruptions. Transportation Research Part C: Emerging Technologies 150, 104098. https://doi.org/10.1016/j.trc.2023.104098.

Wang, Y., Zhou, Y., Yang, H., Yu, B., Liu, X., 2025. Optimizing bus bridging services with mode choice in response to urban rail transit emergencies. European Journal of Operational Research 323, 108–124. https://doi.org/10.1016/j.ejor.2024.11.042.

Yu, C., Wang, Z., 2024. Multimodal social sensing for the spatio-temporal evolution and assessment of natural disasters. Sensors 24, 5889. https://doi.org/10.3390/s24185889.

Zhang, G., Jia, N., Zhu, N., He, L., Adulyasak, Y., 2023a. Humanitarian transportation network design via two-stage distributionally robust optimization. Transportation Research Part B: Methodological 176, 102805. https://doi.org/10.1016/j.trb.2023.102805.

Zhang, J., Ren, G., Song, J., 2023b. Resilience-based optimization model for emergency bus bridging and dispatching in response to metro operational disruptions. PLOS One 18, e0277577. https://doi.org/10.1371/journal.pone.0277577.

Zhang, P., Zhang, H., Kong, F., 2024. Study on the evolution of online public opinion and government response strategies for the "7–20" extraordinary rainstorm and flooding disaster in Zhengzhou, China. Natural Hazards 121, 2849–2872. https://doi.org/10.1007/s11069-024-06904-7.

Zhang, Q., Zhang, P., Li, T., 2025a. Information fusion for large-scale multi-source data based on the Dempster–Shafer evidence theory. Information Fusion 115, 102754. https://doi.org/10.1016/j.inffus.2024.102754.

Zhang, Y., Fang, J., Shen, D., Yang, W., Wang, X., Lyu, L., 2025b. Urban flood risk evaluation using social media data and Bayesian network approach: A spatial-temporal dynamic analysis in Wuhan city, China. Sustainable Cities and Society 126, 106388. https://doi.org/10.1016/j.scs.2025.106388.

Statements and Declarations

Funding. This work was supported by the General Project of Humanities and Social Sciences Research in Higher Education of Henan Province, China (Grant No. 2026-ZZJH-066), and by the Soft Science Research Program of Henan Province (Grant No. 262400410679). The funders had no role in study design, data generation, analysis, interpretation, manuscript preparation, or the decision to submit the work for publication.

Competing interests. The authors have no relevant financial or non-financial competing interests to disclose.

Ethics approval. Not applicable. This study did not involve human participants or animals. It uses publicly available, de-identified, social-media-derived reports of the 20 July 2021 Zhengzhou flood event.

Consent to participate and for publication. Not applicable.

Data availability. The de-identified report data and all derived quantities needed to reproduce the tables and figures are provided as Online Resource 1 and will be deposited in a public repository upon acceptance.

Code availability. The code that regenerates every table and figure—the fusion modules, the experiment scripts, and the fixed random seeds—is provided as Online Resource 1 and will be released in a public repository upon acceptance.

Author contributions. Hailong Li, Xiangnan Song, and Ya Li contributed to the conception and design of the study. All authors reviewed the manuscript, read and approved the final version, and agree to be accountable for the work.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- ESM1README.txt
- ESM2coredataresults.json
- ESM3corereproducibilitycheck.py