

Article

Resilient Last-Mile Allocation in Humanitarian Supply Chains: A Reproducible Four-Objective Fuzzy Framework

Peyman Rabiei ¹, Daniel Arias-Aranda ¹ and Vladimir Stantchev ^{2,*}

¹ Faculty of Economics and Business, University of Granada, 18071 Granada, Spain; prabiei@correo.ugr.es (P.R.); darias@ugr.es (D.A.-A.)

² Institute of Information Systems, SRH University Heidelberg, 69123 Heidelberg, Germany

* Correspondence: stantchev@computer.org

Abstract

Disaster frequency is rising, and response-phase planners face a binding constraint: relief capacity is always smaller than need, so directing affected people to relief centres is a triage decision under time pressure. Published optimisation models speak operations-research vocabulary (cost, distance, equity); civil-protection planners speak supply-chain-resilience vocabulary (the 4R framework of Bruneau et al.). We close the gap with a four-objective humanitarian last-mile allocation model whose objectives map one-to-one onto the 4R components—Resourcefulness, Robustness, Rapidity, Redundancy—evaluated by Mamdani fuzzy inference systems and solved by NSGA-II, NRGa, and NSGA-III. Across 540 runs over three problem sizes, the four-objective formulation surfaces independent Robustness and Rapidity axes that the baseline three-objective formulation fuses into one, a visibility gain established by rank-correlation analysis; NSGA-III is faster than NSGA-II at every size (27–37%, mean 32%), with the hypervolume ranking depending on the reference-point scheme. Weight- and rule-perturbation sweeps confirm the recommendations are robust to expert-elicitation noise, flagging the three-rule Rapidity classifier as a sensitivity hot-spot. The framework is released as a production-grade, openly licensed decision-support tool—continuously tested, dependency-audited, and reproducible bit-for-bit on stock hardware—treating reproducibility as a first-class resilience criterion and positioning audit-ready operational tooling as the digital-technology innovation supply-chain-resilience practice needs.

Keywords: supply chain resilience; humanitarian logistics; multi-objective optimisation; fuzzy inference systems; NSGA-II; NRGa; NSGA-III; 4R framework; reproducible research



Academic Editor: Mirco Peron

Received: 4 July 2026

Revised: 18 July 2026

Accepted: 24 July 2026

Published: 30 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

The 2023 EM-DAT record lists 399 natural disasters worldwide, claiming 86,473 lives and inflicting US\$202.7 bn in economic damage [1,2]. The arithmetic exceeds the 20-year baseline of 64,148 deaths and is projected to keep rising. Behind the aggregate numbers sits a more compressed timeline: the response-phase mortality literature places the bulk of preventable deaths inside the first 12 h after onset [3], the window in which civil-protection planners must direct affected people to relief centres before the most vulnerable run out of resources or infrastructure damage cuts the routes that remain.

The decision is not whether to direct people; it is whom to direct, where to send them, and what to accept giving up. Disaster relief in the first hours always serves fewer than need help [4,5]. Three structural concerns shape every direction decision: prioritising the

most vulnerable, choosing routes that can actually deliver despite infrastructure damage, and balancing the load across centres so that no single relief site exhausts before the rest. None of the three reduces to the others; the practitioner balances all three under time pressure with incomplete information.

Humanitarian supply-chain research has accumulated tools for this decision class [5–7]. Most published allocation models, including the recent literature in this venue’s community, frame their objectives in operations-research vocabulary: travel distance, total cost, response time, head-count of unmet demand. The vocabulary is well-suited to optimisation, but it is not the vocabulary the policy and planning literature uses. Bruneau et al.’s 4R framework [8]—Robustness, Redundancy, Resourcefulness, Rapidity—has been the dominant resilience taxonomy in disaster risk reduction since 2003 and has spread through commercial supply-chain resilience research [6,9]. Civil-protection planners, FEMA documents, EU Civil Protection Mechanism plans, and academic supply-chain-resilience reviews all use 4R vocabulary. Optimisation models that emit results in cost/distance space force the practitioner to translate; models that emit results in 4R space let the practitioner read the Pareto front directly.

This paper develops a four-objective humanitarian last-mile allocation model in which each objective maps one-to-one onto a 4R component. Three of the four objectives build on the authors’ recent FIS-MOEA line of work [10,11]; the fourth is new and is the result of splitting a baseline transportation-infeasibility index into separate Robustness and Rapidity components. We solve the model with NSGA-II, NRGa, and NSGA-III, evaluate it under rule-base and weight sensitivity analyses, and release an open-source Python reference implementation that any reviewer or practitioner can run on their own laptop in minutes. The combination addresses a gap the related-work survey documents in Section 2: to the best of our knowledge, the recent FIS-MOEA disaster-allocation literature does not include a model that organises its objectives around a named resilience framework and also ships a reproducible artifact.

Problem Statement and Contributions

The contribution of this paper is a reframing and as a tool, not a new optimisation operator. We do not propose a new selection, crossover, or mutation mechanism; we reorganise an existing class of allocation models around a recognised resilience vocabulary and ship an implementation that others can run and check. Concretely, the paper makes five contributions to the literature on humanitarian last-mile allocation and supply chain resilience:

1. A resilience-anchored reframing: an explicit one-to-one (bijective) mapping between four allocation objectives and the four components of Bruneau et al.’s 4R framework, absent—to the best of our search—from the surveyed FIS-MOEA disaster-allocation literature and formalised in Section 3.5. This contribution concerns what the objectives *mean*, independent of how many objectives result.
2. The concrete modelling consequence of that mapping: a four-objective formulation that splits the fused transportation infeasibility level into separate Robustness (TRD) and Rapidity (RPD) components, exposing trade-offs the three-objective formulation cannot represent (Section 6). Where contribution 1 is the reframing, this is its operational realisation in the objective set.
3. An empirical comparison of three established solvers—NSGA-II, NRGa, and NSGA-III—on the four-objective humanitarian allocation problem, used to characterise the model rather than to advance the solvers (Section 5).
4. Rule-base and weight-perturbation sensitivity analyses establishing robustness of the recommendations under expert-elicitation noise (Section 7).

5. A production-grade, openly licensed reference implementation—continuously tested and dependency-audited—that treats bit-for-bit reproducibility on stock hardware as a first-class resilience criterion, positioning audit-ready operational tooling as the digital-technology innovation contributing to supply-chain-resilience practice, and archived with a citable Zenodo DOI (DOI: <https://doi.org/10.5281/zenodo.21083547>).

The implicit hypothesis behind the reframing is that organising the objective set around the 4R framework buys the practitioner something measurable, not just a relabelling. We make it explicit and decompose it into four testable hypotheses, each resolved by the experiments in Sections 5–7:

Hypothesis 1. *Splitting the fused transportation index into separate Robustness (TRD) and Rapidity (RPD) objectives exposes trade-off information that the three-objective formulation cannot represent (Section 6).*

Hypothesis 2. *Four objectives is the multi-/many-objective boundary at which reference-point methods may begin to pay off. We test whether, at four objectives, NSGA-III improves on NSGA-II in Pareto-front quality (H2a) or in computational cost (H2b), without presupposing the direction of either (Section 5.3).*

Hypothesis 3. *The model's recommendations are robust to realistic expert-elicitation noise, both in the fuzzy rule bases (H3a) and in the elicited weights (H3b) (Section 7).*

Hypothesis 4. *The NSGA-II-versus-NRGA cost ranking reported by the earlier unpublished MATLAB prototype of the baseline three-objective formulation (Section 2.3) reproduces in an independent Python implementation (Section 5.3).*

The remainder is organised as follows. Section 2 reviews supply-chain-resilience theory, humanitarian last-mile allocation, FIS-MOEA hybrids, and many-objective evolutionary algorithms. Section 3 formalises the model and introduces the 4R↔objective mapping. Section 4 describes the algorithm design including the NSGA-III reference-point construction. Section 5 reports the algorithm comparison; Section 6 the projection analysis testing whether the 4-obj split exposes new trade-offs; and Section 7 the rule-base and weight sensitivity studies. Section 8 discusses implications for theory and practice; Section 9 concludes.

2. Related Work and Theoretical Background

The literature relevant to this paper sits across four bodies of work: supply-chain-resilience theory; humanitarian last-mile allocation models; fuzzy-inference/multi-objective evolutionary hybrids in disaster operations; and many-objective evolutionary algorithms. We surveyed each across Scopus, Web of Science, and IEEE Xplore using the term sets documented per subsection below, (Representative Boolean queries, run over title/abstract/keywords: supply-chain resilience—"supply chain resilience" OR "4R" OR "Bruneau") AND framework; humanitarian last-mile allocation—"humanitarian" OR "disaster" OR "relief") AND ("allocation" OR "last-mile" OR "shelter") AND ("multi-objective" OR "optimization"); FIS-MOEA hybrids—"fuzzy inference" OR "Mamdani") AND ("NSGA" OR "evolutionary") AND ("disaster" OR "relief" OR "humanitarian"); many-objective algorithms—"many-objective" OR "reference point") AND ("NSGA-III" OR "MOEA/D" OR "RVEA"). Screening was by title/abstract relevance to the four sub-topics; we did not apply a PRISMA-grade protocol, and the negative claims are qualified accordingly, with priority given to peer-reviewed journal articles published between 2014 and 2025 and to landmark foundational works regardless of date. The four sub-surveys converge on a common gap that the present paper addresses, which we state

at the end of the section. We do not claim a systematic-review-grade exhaustive coverage; the relative-work negative claims throughout this paper are qualified accordingly.

2.1. Supply Chain Resilience and the 4R Framework

The 4R framework originates in Bruneau et al.'s seismic-resilience quantification [8]: a community is resilient when it exhibits *Robustness* (capacity to withstand a shock without significant loss of function), *Redundancy* (substitutability of system elements), *Resourcefulness* (capacity to identify and mobilise responses), and *Rapidity* (capacity to act in time). The four components were proposed for community infrastructure but the vocabulary has spread [9]: it is the dominant resilience taxonomy in disaster risk reduction, civil-protection planning documents, and the broader SCM-resilience literature.

Christopher and Peck [6] extended the resilience lens to commercial supply chains, naming flexibility, redundancy, and collaboration as the operational levers. Pettit, Fiksel, and Croxton [12,13] developed an assessment-tool framework that pairs SCR capabilities against vulnerability factors. Tukamuhabwa et al. [7] and Kamalahmadi and Parast [14] review more than a decade of SCR work and consolidate it into taxonomies that map cleanly back onto Bruneau's 4R; Wieland [15] reframes SCR as a continuous adaptive process rather than a static property, but the underlying 4R vocabulary persists. Hosseini et al. [9] provide the most-cited quantitative review, distinguishing measures by what they capture (probabilistic vs. deterministic) and how they aggregate (component-wise vs. system-wide). Behl and Dutta [5] survey humanitarian SCR specifically and identify the absence of computational allocation models that operationalise the 4R framework as an open research direction.

Two observations matter for the present work. First, the 4R framework is operational, not philosophical: each component has tested measurement protocols in the disaster-management literature, and the vocabulary is shared between researchers and practitioners. Second, the operationalisation of 4R *inside computational allocation models*—as opposed to post hoc analysis of a model's outputs—is rare. We address that gap directly in Section 3.5.

2.2. Humanitarian Last-Mile Allocation

The classical humanitarian-allocation literature is dominated by cost, time, and distance objectives [16–19]. Tzeng et al. [20] present a multi-objective relief-delivery model with cost and time; Sheu [21] formulates emergency-logistics priorities and constraints from the field-operations perspective. Abounacer et al. [22] formalise multi-objective location-transportation with cost minimisation and uncovered-demand penalty. Gama et al. [23] introduce a multi-period flood-shelter allocation model with evacuation-order timing and distance objectives. Zhao et al. [24] integrate location and allocation with travel-distance and coverage objectives. Balcik and Beamon [25] introduce facility-location models for humanitarian relief; Pedraza-Martinez et al. [26] study fleet management in humanitarian operations. The Holguín-Veras et al. [27] characterisation of post-disaster logistics' unique features—including the displaced-people allocation problem central to this paper—frames the broader research programme. Sahebjamnia et al. [28] build a hybrid decision support system for humanitarian relief chains; Goldschmidt and Kumar [29] review the field with operational-research emphasis. Most directly related to the decision studied here, Yazdani and Haghani [30] develop a dynamic emergency-planning system that relocates vulnerable people to safe shelters under heat-wave conditions—the same core structure of directing at-risk individuals to protective sites under explicit vulnerability considerations. Our formulation departs from it in three respects the present paper foregrounds: objectives derived as an explicit bijection with the 4R resilience components rather than crisp cost/time terms,

fuzzy-aggregated indices that digest linguistic practitioner judgement, and a bit-for-bit reproducible open artifact.

Equity entered the literature later. Soghrati Ghasbeh et al. [31] develop a robust multi-stage relief-distribution model with explicit fairness terms; Krishna et al. [4] document the practitioner-side experience of inequitable relief in the 2015 South Indian floods, motivating the equity turn. Across this body of work the shared characteristic is that objectives are crisp quantitative measures (seconds, kilometres, dollars, head-counts), not qualitative indices that encode expert reasoning under uncertainty.

Within this surveyed literature, the use of a recognised resilience framework to organise the objective set—particularly the 4R framework [8]—is rare. Recent work increasingly brands disaster-operations models as “resilience” models: Yazdani, Matinrad, and Haghani [32], for instance, present a resilience model for community-led volunteer deployment in flood recovery. On inspection, however, that model operationalises resilience as a mixed-integer cost-minimisation over mobilisation, team-formation, deployment, and outsourcing terms, not as a mapping onto named resilience components—so the distinction we draw survives the comparison and is sharpened by it. This is the pattern throughout: models include resilience as a constraint, an objective label, or a robustness term, but to the best of our knowledge none of the works above structures its objectives as an explicit bijection with the four 4R components. The framing remains Operations Research’s cost/time/equity vocabulary; the SCR community’s vocabulary sits unused on the next shelf over.

2.3. FIS-MOEA Hybrids in Disaster Operations

Mamdani’s original fuzzy-controller paper [33] established that linguistic if–then rules combined with triangular and trapezoidal membership functions could replace tuned crisp models in domains where expert judgement is the canonical input. The pattern fits humanitarian operations management precisely: practitioners reason in linguistic categories (“severely injured,” “partially blocked,” “rapidly depleting”), and translating those to crisp scalars discards practitioner-relevant information.

The FIS-MOEA hybrid pattern in disaster operations was introduced into this line of work by Rabiei and Arias-Aranda [10] for post-disaster vehicle routing and relief supply distribution, and extended to volunteer-to-vacancy assignment in Rabiei, Arias-Aranda, and Stantchev [11]. Both pair Mamdani FIS with NSGA-II and NREGA. The present paper adapts the pattern to the people-to-relief-centre allocation problem, developing the three fuzzy-aggregated objectives—unfairness in prioritisation (ULPP), transportation infeasibility (TIL), and centre allocation imbalance (CAIL)—that constitute the baseline three-objective formulation the four-objective model extends. That baseline was first built as an earlier, unpublished MATLAB prototype in our group; the present paper is its first open, independent Python re-implementation, and hypothesis H4 below asks whether the prototype’s NSGA-II-versus-NREGA cost ranking reproduces here. Adjacent fuzzy or stochastic disaster-allocation work includes Bahmani et al. [34] on fuzzy multi-objective routing under damaged networks and Moreno et al. [35] on two-stage stochastic relief-supply chains with social-equity objectives; neither operationalises a named resilience framework.

Three observations about this line of work motivate the present extension. First, the FIS-MOEA disaster-line papers we surveyed do not reference the 4R framework or any other named resilience taxonomy: the objective functions are introduced on their own terms, with intuitive justification, not as operationalisations of an existing theory. Second, the baseline Transportation Infeasibility Level (TIL) fuses Robustness (road condition, hazard) with Rapidity (travel duration) into a single index—which prevents the practitioner from articulating Robustness/Rapidity trade-offs separately. Third, none of the FIS-MOEA

papers we surveyed releases a reproducible artifact: the prior work in this line uses MATLAB without an accompanying open implementation. This combination is the gap we address in Sections 3–5.

2.4. Many-Objective Evolutionary Algorithms

NSGA-II [36] remains the most-cited multi-objective evolutionary algorithm for two- and three-objective problems. The combination of non-dominated sorting and crowding-distance tie-breaking is well-suited to the regime where objectives are few enough that the Pareto front retains a clear geometric structure and crowding distance carries usable diversity information.

Reference-point methods entered the literature for many-objective settings (typically ≥ 4 objectives) where crowding distance becomes information-poor [37–39]. Deb and Jain’s NSGA-III [40] replaces crowding distance with niching against a structured reference set, laid down on the normalised objective simplex via the Das–Dennis construction. Decomposition-based methods such as MOEA/D [41] and reference-vector methods such as RVEA [42] pursue the same goal through different mechanisms; we discuss them as candidates for future algorithmic comparison in Section 8.5. The reference-point machinery scales gracefully with objective count and is the modern default for many-objective problems. Recent runtime analysis by Zheng and Doerr [43] quantifies NSGA-II’s degradation as a function of objective count and provides the theoretical underpinning for the empirical observation that reference-point methods take over above about four objectives.

NRGA [44] sits between NSGA-II and NSGA-III conceptually: it inherits NSGA-II’s non-dominated sorting but replaces the crowding-distance tie-breaker with uniform random sampling within the boundary front. The motivation is that crowding distance encodes a particular notion of diversity (gap between consecutive solutions in objective space) that may not match the practitioner’s notion; random sampling avoids that prior at the cost of dropping some non-dominated solutions. NRGA is the second algorithm used in the FIS-MOEA disaster line [10,11] and we retain it for continuity with that line.

This paper’s algorithmic contribution is to add NSGA-III to the comparison, putting it on a four-objective humanitarian allocation problem precisely at the boundary where reference-point methods are expected to start paying off. The empirical findings of Section 5 show that the boundary is crossed for CPU time but not for hypervolume—a finding that maps onto the runtime analysis of [43] and is the substance of the H2 result.

2.5. Gap

The combination of (i) fuzzy-aggregated objectives (quantitative inputs, linguistic outputs) organised around a named resilience framework, (ii) a four-objective formulation that splits Robustness and Rapidity into separate axes following the 4R taxonomy, (iii) NSGA-III alongside NSGA-II and NRGA on a humanitarian last-mile allocation problem, (iv) systematic rule-base and weight sensitivity analyses, and (v) an open-source Python reference implementation released with a citable DOI is, to the best of our search, absent from the published literature. The contributions of this paper map one-to-one onto the five elements of that combination.

3. Model Formulation

3.1. Problem Definition

A disaster has just occurred. Civil-protection planners face a bounded decision: among the m affected people whose locations and conditions are known, choose at most n_{dir} to direct to one of n relief centres. The constraint $n_{\text{dir}} < m$ is operational, not philosophical: relief capacity in the first hours after a major event is always smaller than need [1,4]. The

decision-maker does not get to serve everyone; they have to decide whom to serve, where to send them, and accept that some will wait.

Three concerns shape every direction decision. The first is who: a moderately injured 75-year-old living alone has different priority than a healthy adult with a support network. The second is how: a partially blocked road with moderate hazard delivers a person but at risk; a longer clear road delivers the same person more slowly but more safely. The third is where: a centre already at 80% occupancy will exhaust supplies faster, regardless of how the route looks. Real practitioners weigh all three simultaneously [5].

We model the decision as a multi-objective allocation problem. Each candidate solution is a partial mapping $d : P \rightarrow C \cup \{\perp\}$ that selects exactly n_{dir} people from P and assigns each to a centre in C . The remaining $m - n_{\text{dir}}$ people are not directed in this round. The mapping is feasible by construction; the question is which feasible mappings sit on the Pareto front under the four objectives we develop below.

3.2. Resilience-Anchored Fuzzy Indices

We replace the cost-and-distance objective set common in the allocation literature [20,22–24] with four fuzzy-aggregated indices, each anchored in one component of Bruneau et al.’s [8] 4R supply-chain-resilience framework. Three of the indices form the baseline three-objective formulation; the fourth, and the splitting of the second, are new here.

A point on terminology, since the distinction matters for what follows. The *inputs* to these indices are quantitative measurements: a travel duration in minutes, a distance in kilometres, an occupancy rate, a road-condition category. What is qualitative is the *output*—the linguistic resilience deficit that a fuzzy inference system produces by combining several quantitative inputs through expert-elicited rules. We therefore call the four objectives fuzzy-aggregated rather than qualitative: each digests crisp, measurable signals into a single expert-style judgement on a common $[0, 100]$ scale.

3.2.1. Unfairness Level in People Prioritization (ULPP)—Resourcefulness

ULPP captures whether the right people are being directed first. The 4R framework’s *Resourcefulness* is the system’s capacity to identify problems, mobilise resources, and prioritise correctly under stress [8,9]. In a humanitarian last-mile setting that translates directly into one question: among the m candidates, did we send the most vulnerable? ULPP encodes a need-weighted Rawlsian reading of fairness—prioritisation by greatest individual need rather than equal shares or proportional allocation. We name the index “Unfairness Level in People Prioritization”; readers preferring the SCM literature’s vocabulary may read it as a Rawlsian maximin term [31].

We compute ULPP from three sub-indices: a vulnerability score $VS_j \in [0, 1]$ aggregating age, disability, injury, and living-alone status; an infrastructure damage level $IDL_j \in [0, 100]$ at the person’s location; and a resource-time-remaining estimate RTR_j in hours. Higher VS and IDL with lower RTR mean a person should be directed first; if they are not, ULPP rises. FIS₁ digests the three sub-indices into the unfairness output $ULPP_j \in [0, 100]$ via a 27-rule rule base (Appendix A).

3.2.2. Transportation Robustness (TR)—New

The baseline formulation folds transport robustness and rapidity together into a single Transportation Infeasibility Level (TIL). For supply-chain resilience analysis, that fusion is too coarse: *Robustness* (the network’s capacity to deliver despite damage) and *Rapidity* (the network’s capacity to deliver in time) are distinct 4R components and the practitioner needs to see them separately. We split them.

TR captures network damage along the route. The relevant sub-indices are the road condition score $RCS_{j,i}$ (clear, partially blocked, or blocked) and the possible-hazard score $PHS_{j,i}$, a five-value discrete input ranging from no immediate threat to extreme hazard. For inference both crisp inputs are fuzzified into three linguistic levels (Low, Moderate, High), so FIS_{2a} maps them into a Transport Robustness Deficit $TRD_{j,i} \in [0, 100]$ with a nine-rule base covering the full 3×3 Cartesian over those levels: a clear road with no hazard is very-low TRD; a blocked road with extreme hazard is very-high TRD.

3.2.3. Rapidity (RP)—New

RP captures the speed dimension that TR deliberately excludes. Rapidity in the 4R framework is context-free: how fast the response delivers, independent of who is being delivered to or what they are being delivered through [8]. We model it as a single-input fuzzy classifier on travel duration $TD_{j,i}$ in minutes, producing a Rapidity Deficit $RPD_{j,i} \in [0, 100]$ via three rules: short trip $\rightarrow$ low RPD, moderate $\rightarrow$ moderate, long $\rightarrow$ high.

We considered adding RTR as a second input to FIS_{2b} and rejected it. Resource time remaining already enters FIS_1 to set ULPP; reusing it in FIS_{2b} would double-count Resourcefulness inside Rapidity and break the 4R-objective bijection that motivates the whole formulation. Rapidity is, and remains, a property of the route, not of the person.

3.2.4. Centre Allocation Imbalance Level (CAIL)—Redundancy

CAIL captures whether the load is spread or concentrated. *Redundancy* is the substitutability of system elements under stress: a network where every relief centre carries equal load fails gracefully when one centre saturates; a network where one centre carries everything fails catastrophically [6,8]. CAIL operationalises Redundancy as the mean imbalance of allocations across centres, weighted by each centre's current state.

The sub-indices are the centre occupancy rate $COR_i \in [0, 100]$, the resource depletion rate $RDR_i \in [0, 100]$ at the centre, and travel duration $TD_{j,i}$ to the centre. The intuition encoded in the 27-rule FIS_3 rule base is monotone in the obvious direction: a centre that is already full, depleting fast, and far from the directed person contributes high CAIL; an empty centre with slow depletion and short travel contributes low CAIL.

3.3. Notations and Variables

Table 1 summarises the notation used throughout the model. Every notation carried over from the baseline three-objective formulation is preserved; the additions for the four-objective formulation are flagged in the rightmost column.

The vulnerability score VS_j is derived from age score AS_j , disability DS_j , injury IL_j , and living-status LS_j via the weighted average

$$VS_j = \frac{WAS \cdot AS_j + WDS \cdot DS_j + WIL \cdot IL_j + WLS \cdot LS_j}{WAS + WDS + WIL + WLS}, \quad (1)$$

with the four weights tunable per disaster context (defaults all 1). The roadworthiness score is given by

$$RWS_{j,i} = 1 - \frac{WRC \cdot RCS_{j,i} + WPH \cdot PHS_{j,i}}{WRC + WPH}. \quad (2)$$

Note on Equation (2): orientation of RWS.

We define RWS so that it increases with route quality: the weighted average of RCS and PHS rises with worse conditions (Blocked = 1.0, Extreme = 1.0), so Equation (2) subtracts it from one. This gives $RWS = 1$ for a perfectly safe and clear route and $RWS = 0$ for an impassable one, so that TIL is low exactly when the route is safe and fast.

Table 1. Notation. Symbols carried over from the baseline three-objective formulation are unchanged; new symbols for the four-objective formulation are marked.

Symbol	Description	Origin
P, m	Set of affected people; $m = P $.	baseline
C, n	Set of relief centres; $n = C $.	baseline
n_{dir}	Number of people we direct ($0 < n_{\text{dir}} < m$).	baseline
$d_{j,i}$	Binary indicator: $d_{j,i} = 1$ iff person j directed to centre i .	baseline
$VS_j \in [0, 1]$	Vulnerability score of person j .	baseline
$IDL_j \in [0, 100]$	Infrastructure damage at person j 's location.	baseline
RTR_j (hours)	Resource time remaining for person j .	baseline
$COR_i \in [0, 100]$	Centre i occupancy rate.	baseline
$RDR_i \in [0, 100]$	Centre i resource depletion rate.	baseline
$TD_{j,i}$ (min)	Travel duration $j \rightarrow i$.	baseline
$RCS_{j,i}$	Road condition $\in \{0.1, 0.5, 1.0\}$.	baseline
$PHS_{j,i}$	Possible-hazard score $\in \{0.05, 0.25, 0.5, 0.75, 1.0\}$.	baseline
$RWS_{j,i} \in [0, 1]$	Roadworthiness (high = safe), Equation (2).	baseline
$ULPP_j \in [0, 100]$	Unfairness in prioritising person j .	baseline
$TIL_{j,i} \in [0, 100]$	Transport infeasibility (3-obj only).	baseline
$TRD_{j,i} \in [0, 100]$	Transport Robustness Deficit.	new
$RPD_{j,i} \in [0, 100]$	Rapidity Deficit.	new
$CAIL_{i,td}$	Centre allocation imbalance contribution.	baseline
$\overline{ULPP}, \overline{TRD}, \overline{RPD}, \overline{CAIL}$	Mean of corresponding index over the directed set.	4 of 4 new in 4-obj

3.4. Objective Functions

In words before symbols: each objective averages one resilience deficit over the people we direct, and the optimiser tries to push all four averages down at once. The first, $\overline{ULPP}$, is the mean unfairness of the prioritisation—low when the most vulnerable were directed first. The second and third, $\overline{TRD}$ and $\overline{RPD}$, are the mean robustness deficit of the chosen routes (how exposed they are to damage and hazard) and the mean rapidity deficit (how slow they are). The fourth, $\overline{CAIL}$, is the mean imbalance the decision induces across centres—low when load is spread rather than piled onto already-stressed sites. Each is a number on the same $[0, 100]$ scale where lower is better.

These four objectives conflict, which is why a single best solution does not exist and a Pareto front does. Directing the most vulnerable people (low $ULPP$) often means reaching those who live where infrastructure is most damaged, along the worst routes (high TRD) or the longest ones (high RPD). Choosing the fastest route (low RPD) can mean accepting a more hazardous one (high TRD), and the safest route is rarely the quickest. Sending many vulnerable people to the best-equipped centre lowers $ULPP$ but concentrates load and raises $CAIL$. No allocation minimises all four at once; the model's job is to surface the trade-off surface so the practitioner can choose where on it to sit.

Formally, the model minimises four mean-objective functions over the directed set:

$$\min \overline{ULPP} = \frac{\sum_{j=1}^m \sum_{i=1}^n d_{j,i} ULPP_j}{n_{\text{dir}}}, \quad (3)$$

$$\min \overline{TRD} = \frac{\sum_{j=1}^m \sum_{i=1}^n d_{j,i} TRD_{j,i}}{n_{\text{dir}}}, \quad (4)$$

$$\min \overline{RPD} = \frac{\sum_{j=1}^m \sum_{i=1}^n d_{j,i} RPD_{j,i}}{n_{\text{dir}}}, \quad (5)$$

$$\min \overline{CAIL} = \frac{\sum_{i=1}^n TCAIL_i}{n_{\text{dir}}}, \quad (6)$$

where $\text{TCAIL}_i = \sum_{j=1}^m d_{j,i}$; $\text{CAIL}_{i,\text{TD}_{j,i}}$ collects the per-allocation CAIL contributions at centre i . The model constraints are:

$$\sum_{i=1}^n d_{j,i} \leq 1 \quad \forall j, \quad \sum_{j=1}^m \sum_{i=1}^n d_{j,i} = n_{\text{dir}}, \quad n_{\text{dir}} < m, \quad n_{\text{dir}} > 0, \quad n > 0, \quad (7)$$

i.e., each person is directed to at most one centre, exactly n_{dir} are directed, and the directed set is a strict subset of the affected population.

For the baseline three-objective formulation (used in Section 6 as the H1 comparator), Equations (4) and (5) are replaced by the single $\overline{\text{TIL}}$ objective using the fused FIS_2 pathway of the baseline formulation. Everything else is identical.

3.5. Resilience-Theoretic Mapping (New)

The 4R framework earns its place by doing analytical work, not by labelling. It fixes what each objective is meant to capture and tells the practitioner how to read the Pareto front. Table 2 states the bijection.

Table 2. Bijection between the four objectives and Bruneau et al.’s 4R supply-chain-resilience framework [8]. Each objective operationalises exactly one 4R component; the practitioner reading the Pareto front sees four resilience axes simultaneously.

4R Component	Objective	FIS	Practitioner Reading
Resourcefulness	$\overline{\text{ULPP}}$	FIS_1	“Did we direct the most vulnerable first?”
Robustness	$\overline{\text{TRD}}$	FIS_{2a}	“Will the routes hold under damage?”
Rapidity	$\overline{\text{RPD}}$	FIS_{2b}	“Will the people get there in time?”
Redundancy	$\overline{\text{CAIL}}$	FIS_3	“Is the load spread or concentrated?”

Per-Component Derivation

Table 2 states the bijection in summary form; the mapping is not a label slapped on existing objectives but a derivation from each 4R component’s definition. We walk through the four pairings.

Resourcefulness $\rightarrow$ ULPP via FIS_1 .

Bruneau et al. [8] define Resourcefulness as “the capacity to identify problems, establish priorities, and mobilise resources when conditions exist that threaten to disrupt some element [...] of the system.” Three clauses, each present in our model. *Identifying problems* maps to the three sub-indicators feeding FIS_1 (vulnerability score VS, infrastructure damage IDL at the person’s location, resource time remaining RTR); these are precisely the diagnostic signals a triage coordinator would consult. *Establishing priorities* maps to the 27-rule rule base of FIS_1 , which is itself an explicit articulation of expert priority judgement. *Mobilising resources* maps to the directing decision $d_{j,i}$ that the optimisation produces. ULPP is zero exactly when the coordinator’s three diagnostics, the rule base’s priority logic, and the directing decision are mutually consistent—i.e., when the system is being Resourceful in Bruneau’s sense.

Robustness $\rightarrow$ TRD via FIS_{2a} .

Bruneau et al. define Robustness as “the inherent strength or resistance in a system to withstand external demands without degradation or loss of functionality.” For a humanitarian last-mile route, the relevant external demand is infrastructure damage and route hazard, and the relevant function is delivering the directed person to the centre. Two sub-indicators capture this: $\text{RCS}_{j,i}$ encodes the inherent strength of the road (clear/partially blocked/blocked), and $\text{PHS}_{j,i}$ encodes external threat along the route. FIS_{2a} digests both into TRD—the deficit relative to a perfectly robust route. TRD is zero exactly when the

route can deliver the directed person regardless of external conditions, which is the limit case of Bruneau's robustness.

Rapidity $\rightarrow$ RPD via FIS_{2b}.

Bruneau et al. define Rapidity as "the capacity to meet priorities and achieve goals in a timely manner in order to contain losses and avoid future disruption." This is the only 4R component with explicit time content. For our model the canonical time measure is travel duration $TD_{j,i}$. We use TD as the single input to FIS_{2b} deliberately: any multi-input formulation would introduce non-time content into Rapidity and break the framework's component-orthogonality. FIS_{2b} emits RPD, the deficit relative to the limiting case of instantaneous delivery; RPD is zero exactly when delivery is fast enough to contain losses, which is Bruneau's Rapidity in the limit.

Redundancy $\rightarrow$ CAIL via FIS₃.

Bruneau et al. define Redundancy as "the extent to which elements [...] are substitutable, that is, capable of satisfying functional requirements in the event of disruption." For a network of relief centres, substitutability is the property that any one centre saturating does not prevent the others from absorbing displaced load. Three sub-indicators capture the dynamics: COR_i (current occupancy of centre i), RDR_i (rate at which centre i 's resources deplete), and $TD_{j,i}$ (arrival time, which influences when the directed load reaches the centre relative to depletion). FIS₃ digests these into CAIL, the imbalance the directing decision induces. Low CAIL means the directing decision spreads load across centres in a way that preserves substitutability for the next decision round; high CAIL means the decision concentrates load on already-stressed centres, reducing future Redundancy.

Why the derivation matters.

The bijection acts as a design constraint rather than a post hoc gloss: each FIS's input set, rule base, and output range must be explicable in the vocabulary of exactly one 4R component, and each 4R component must be operationalised by exactly one FIS. The derivation above shows that constraint holding across all four pairings. Three theoretical consequences follow, which later sections draw on.

One defensible mapping, not the only one.

The bijection is principled but not uniquely forced. A different modeller could route resource-time-remaining into Rapidity rather than Resourcefulness, or spread Redundancy across two indices; we claim only that the mapping we adopt is defensible under each component's Bruneau-et-al. definition and that it holds the one-FIS-per-component discipline, not that it is the sole admissible operationalisation. The indicator choices—VS, IDL, RTR for Resourcefulness; RCS, PHS for Robustness; TD for Rapidity; COR, RDR, TD for Redundancy—are the diagnostics a triage coordinator already consults, which is the justification we offer for them, and the released tool makes alternative mappings straightforward to specify and test.

The bijection has three theoretical consequences worth stating explicitly.

Translation, not aggregation.

A practitioner who plans in 4R vocabulary can read the Pareto front directly, without translation through a cost/distance/equity intermediary. Asking "what is the cheapest solution that keeps Robustness above 70?" becomes well-formed and answerable. Asking "which solution maximises Rapidity at the expense of Resourcefulness?" becomes well-formed and answerable. The 3-objective formulation, where Robustness and Rapidity are fused into TIL, makes the second question unanswerable by construction. We test this empirically in Section 6.

Theoretical hygiene.

If two of our objectives mapped to the same 4R component, the framework would tell us so by complaining. The bijection constrains the modeller. Adding RTR to FIS_{2b}

would have produced a better-fitting Rapidity index empirically but would have leaked Resourcefulness into Rapidity, which the 4R framework forbids, so we leave RTR out. What the framework does not do.

The 4R framework does not prescribe weights between the four components. Bruneau et al. [8] are explicit on this point: a community planner balancing robustness against rapidity makes a different choice than a triage coordinator on the second day of a flood. The 4R framework provides the *vocabulary* for the trade-off, not the trade-off itself. Our model preserves that property by emitting a four-dimensional Pareto front rather than a single weighted score; the decision-maker selects the trade-off they want, in 4R terms, post hoc.

4. Algorithm Design

4.1. Fuzzy Inference in Brief

Each of our four objectives is produced by a fuzzy inference system (FIS), so we first describe what a FIS is and why we use one, assuming no prior familiarity. A FIS is a way of turning numbers into a graded expert judgement using if-then rules written in words. It suits this problem because the people who run relief operations reason in linguistic categories. A road is “partially blocked,” supplies are “depleting fast,” a person is “highly vulnerable.” Forcing those judgements into a single tuned formula throws away information the practitioner actually has.

An FIS works in three steps. *Fuzzification* reads each crisp input and asks how strongly it belongs to each linguistic level. The levels are defined by *membership functions*: simple shapes over the input range that return a degree of belonging between 0 and 1. We use two standard shapes. A *triangular* membership function, written $\text{tri}[a, b, c]$, rises linearly from 0 at a to 1 at the peak b and falls back to 0 at c ; it models a level with a single most-typical value. A *trapezoidal* function, $\text{trap}[a, b, c, d]$, rises from a to b , stays at 1 across the plateau $[b, c]$, and falls to 0 at d ; it models a level with a range of equally typical values, which is convenient for the ends of a scale. A given input typically belongs partly to two neighbouring levels at once: a vulnerability reading near the boundary between “Medium” and “High” might register, say, 0.27 Medium and 0.73 High rather than snapping to one label.

Rule firing then evaluates a small rule base of the form “if vulnerability is High and infrastructure damage is Severe and resource time remaining is Short, then unfairness is High.” Every rule whose antecedent has non-zero membership contributes, weighted by how strongly it fired, so several rules speak at once. *Defuzzification* collapses those weighted contributions back into one number: we use the centroid (centre of mass) of the aggregated output, which gives a smooth, monotone response on the common $[0, 100]$ output scale. Concretely, a moderately vulnerable person at a location with moderate infrastructure damage and a mid-range resource-time window fuzzifies across the Medium/High levels of each input, fires a handful of rules, and defuzzifies to a modest unfairness score rather than a hard “fair/unfair” verdict. A fully worked numerical example of all three steps is given in the companion documentation accompanying the released code. The remainder of this section specifies the four rule bases and the membership partitions they use.

4.2. Fuzzy Inference Systems

We use Mamdani inference [33] with centroid defuzzification throughout, for the same reason as earlier work in this line [10,11]: linguistic if-then rules elicited from disaster-response practitioners are easier to audit and revise than a tuned crisp model, and the Mamdani conjunction over triangular and trapezoidal membership functions gives the

rule base a clear geometric interpretation. The four-objective formulation uses four control systems (FIS₁ for ULPP, FIS_{2a} for TRD, FIS_{2b} for RPD, FIS₃ for CAIL); the baseline three-objective formulation reuses its FIS₂ for TIL and is invoked only as the H1 comparator in Section 6.

Each input variable is partitioned into three linguistic levels with trapezoidal endpoints and triangular interiors; the output universe is a five-level scale (Very Low, Low, Moderate, High, Very High) on [0, 100] shared across all five FIS for cross-objective comparability. Rule-base sizes are 27 for FIS₁ (the full 3³ Cartesian product over VS, IDL, RTR), 9 for FIS_{2a} (RCS × PHS), 3 for FIS_{2b} (TD alone), and 27 for FIS₃ (the full 3³ over COR, RDR, TD). Appendix A lists every rule and Appendix B every membership-function parameter. FIS₁, FIS₃, and the baseline FIS₂ are inherited unchanged from the three-objective formulation; FIS_{2a} and FIS_{2b} are new to this paper, specified by the authors to the same trapezoidal-endpoint/triangular-interior template as the inherited systems.

A practical detail matters for performance. We pre-compute every FIS evaluation once per problem instance: for the four-objective formulation, m per-person ULPP values and $3 \cdot m \cdot n$ per-pair (TRD, RPD, CAIL) values, all stored as numpy arrays. Chromosome evaluation in the GA loop then reduces to fancy-indexing into those arrays—no scikit-fuzzy calls inside evolution. The cost of pre-computation on the medium instance (8 centres, 225 people) is roughly 10 s; the amortised cost per generation is essentially zero. This is what makes the 540-run experiment matrix in Section 5 tractable.

4.3. Multi-Objective Evolutionary Algorithms

Three MOEAs share the same population-management loop and differ only in their survivor-selection operator. The selection is deliberate rather than exhaustive: NSGA-II and NRGA are retained for direct continuity with the FIS-MOEA disaster line this paper extends [10,11], and NSGA-III is added as the canonical reference-point many-objective solver at the four-objective boundary. Decomposition- and reference-vector-based many-objective algorithms such as MOEA/D [41] and RVEA [42] are the natural next comparators; we scope them to future work (Section 8.5) to keep the present study focused on the resilience reframing rather than on a solver tournament. We summarise each.

4.3.1. NSGA-II [36]

Non-dominated sorting partitions the combined parent–offspring population into Pareto fronts. Survivors are taken front by front; when a front overflows the population quota, the remaining seats are filled by crowding-distance ranking within that front. The crowding-distance machinery is the workhorse of NSGA-II for two- and three-objective problems and the reason it dominates the literature in that regime. Recent runtime analysis [43] shows the mechanism degrades as objective count grows, motivating reference-point methods at higher objective counts.

4.3.2. NRGA [44]

NRGA replaces NSGA-II's crowding-distance tie-breaker with uniform random sampling within the boundary front. The intuition is that crowding distance encodes a particular notion of diversity (gap between consecutive solutions in objective space) which may not match the practitioner's notion; random sampling avoids that prior. The cost is that some Pareto-optimal solutions are dropped at random, lowering NNS in our experiments by a few percent. Whether NRGA helps or hurts depends on the problem geometry and the chromosome encoding—a sensitivity we revisit in Section 5.3.

4.3.3. NSGA-III [40]

NSGA-III replaces crowding distance with reference-point-based niching designed for many-objective problems. A set of structured reference points $\{r_k\}$ is laid down in the normalised objective simplex; survivors are chosen to maintain diversity *relative to those reference points* rather than relative to each other. The mechanism keeps boundary regions populated and is the modern default for objective counts ≥ 4 . We use NSGA-III not because we expect it to dominate NSGA-II at four objectives—the empirical evidence in the literature is mixed at that boundary—but because its CPU cost is independent of population density in front overflow cases, which we test in Section 5.3.

4.4. Chromosome Encoding

Each individual is a flat list of length $2n_{\text{dir}}$, read as two contiguous slices. The first n_{dir} entries are the *person slice*: integer person indices, sampled without replacement from $\{0, \dots, m-1\}$, i.e., a partial permutation that names exactly which n_{dir} of the m candidates are directed in this round. The second n_{dir} entries are the *centre slice*: gene k holds the centre to which the k -th directed person is sent. We store each centre gene as a real value $r_k \in [0, 1)$ and decode it to a centre index by $\lfloor r_k \cdot n \rfloor$ (clamped to $n-1$ at the upper endpoint). The real value is therefore just a mutation-friendly stand-in for an integer in $\{0, \dots, n-1\}$; the centre slice carries exactly a per-person integer centre assignment, no more. Every chromosome the operators can produce is feasible by construction: the person slice is a permutation (no duplicate or out-of-range person), and every centre gene decodes to a valid centre, so no repair against infeasibility is ever needed.

A reviewer of the earlier model in this line asked, reasonably, why not encode an allocation as a single length- n vector with one integer centre per person. That encoding is the natural choice when *every* person is served. Our problem directs a strict subset ($n_{\text{dir}} < m$; relief capacity is smaller than need), so the chromosome must encode two decisions, not one: *which* people are directed and *where* each goes. The person slice carries the first decision and the centre slice the second. Restricted to the directed people, the centre slice *is* exactly the one-integer-per-person encoding that reviewer suggested; the person slice is the additional structure the subset-selection problem requires.

The order of the (person, centre) pairs is not semantically meaningful: the decoder treats them as a set of pairs, not a sequence. Variation operators preserve only the partial-permutation invariant on the person slice. We use uniform crossover with repair (children inherit from each parent at each position with probability 0.5, then duplicate person indices are replaced with random unused indices) and per-gene mutation (replacement with an unused index for the person slice; Gaussian perturbation $\mathcal{N}(0, 0.1)$ clipped to $[0, 1)$ for the centre slice). DEAP's built-in operators do not fit a partial-permutation + real-valued mixed encoding directly, so we implement the two operators ourselves; the implementation is open-source.

4.5. NSGA-III Reference Points (New)

The reference-point set determines NSGA-III's selection geometry. We use Das–Dennis structured points with $p = 4$ divisions, yielding $\binom{p+M-1}{M-1} = \binom{7}{3} = 35$ points uniformly distributed over the four-dimensional objective simplex [40]. The choice is conventional and easy to reproduce; we considered the two-layer scheme that adds points along the simplex boundary for better edge coverage and kept it in reserve, to be adopted only if boundary diversity proved problematic in the algorithm comparison. It did not: NSGA-III retains a full boundary front (NNS = 100/100 at every size, Table 3), so the single-layer set sufficed and the two-layer variant was not needed.

Table 3. Mean $\pm$ s.d. over 30 reps per (size, algorithm) cell on the four-objective formulation. HV (in millions) is the primary quality indicator; CPU time is in seconds. MID is shown last as a diagnostic only (see note). The p -values underlying these means are the Mann–Whitney U (Wilcoxon rank-sum) tests against NSGA-II in Appendix C.

Size	Algorithm	NNS	HV ($\times 10^6$)	SM	CPU (s)	MID [†]
small	NSGA-II	100.0	21.08 $\pm$ 0.47	0.81 $\pm$ 0.06	1.45 $\pm$ 0.07	77.7 $\pm$ 0.5
	NRGA	99.1	20.21 $\pm$ 0.87	0.79 $\pm$ 0.05	1.16 $\pm$ 0.06	77.4 $\pm$ 0.5
	NSGA-III	100.0	20.74 $\pm$ 0.62	0.81 $\pm$ 0.06	0.91 $\pm$ 0.07	77.7 $\pm$ 0.5
medium	NSGA-II	100.0	18.81 $\pm$ 0.40	0.78 $\pm$ 0.05	1.69 $\pm$ 0.06	69.5 $\pm$ 0.4
	NRGA	99.9	17.35 $\pm$ 0.55	0.77 $\pm$ 0.04	1.40 $\pm$ 0.05	69.3 $\pm$ 0.4
	NSGA-III	100.0	18.33 $\pm$ 0.69	0.79 $\pm$ 0.05	1.15 $\pm$ 0.06	69.5 $\pm$ 0.4
large	NSGA-II	100.0	13.98 $\pm$ 0.29	0.81 $\pm$ 0.05	1.99 $\pm$ 0.05	58.6 $\pm$ 0.4
	NRGA	99.8	12.86 $\pm$ 0.29	0.81 $\pm$ 0.04	1.69 $\pm$ 0.05	58.6 $\pm$ 0.4
	NSGA-III	100.0	13.59 $\pm$ 0.33	0.81 $\pm$ 0.05	1.45 $\pm$ 0.05	58.6 $\pm$ 0.4

[†] MID (Mean Ideal Distance) is reported as a diagnostic only and is not used to rank algorithms. MID rewards solutions that sit close to the ideal point, which penalises legitimate Pareto-optimal trade-offs: every solution on a front is equally non-dominated, so a front-quality indicator should not prefer the ones nearest a single corner. We retain MID for backward comparability with the earlier formulation in this line of work, but read front quality from HV, which captures convergence and diversity together.

The 35 reference points are recomputed once at solver start and held fixed across generations. With population size $100 \geq 35$ the reference set never overflows and the niching procedure runs without partial-front corner cases. In the four-objective formulation $p = 4$ produces reference spacing of $1/4$ along each axis, which we found empirically sufficient for the Pareto-front densities we observe (typically 100 distinct solutions per front, comfortably above the reference-point count).

5. Performance Analysis

5.1. Experimental Setup

We test three problem sizes drawn at a single seed from a deterministic generator: small (5 centres, 150 affected people, $n_{\text{dir}} = 50$), medium (8/225/75), and large (10/300/100). Small and large match the baseline test instances; medium is new and added to give scaling analysis a third anchor. Each problem instance is materialised as three CSV files (people, centres, travel) and is reproducible bit-exact from the seed via the open-source instance generator at `experiments/generate_instances.py`.

For every cell in the (size $\times$ formulation $\times$ algorithm) matrix we run 30 repetitions with rep-specific seeds drawn at $\text{seed}_r = 1000 + 7919r$. Population size is 100, generations 200, crossover probability 0.7, mutation probability 0.2—baseline hyper-parameters are held fixed across the matrix so the comparison is on algorithm and formulation, not hyper-parameters. NSGA-III uses the Das–Dennis $p = 4$ reference set described in Section 4.5. Implementation is the open-source Python `presidio_vol_assign.allocation` module (The package name `presidio_vol_assign` and the repository `presidio-hardened-vol-assign` retain the “volunteer-assignment” naming of the predecessor tool in this FIS-MOEA line [11], under which the software was first released and archived; we keep it unchanged so the citable Zenodo artifact and its DOI remain stable, even though the model solved here allocates affected people to relief centres.) (scikit-fuzzy 0.5.0 + DEAP 1.4.3 + pymoo 0.6 on Python 3.12, NumPy 2.2, SciPy 1.15; exact versions pinned in the released `uv.lock`). Experiments ran on a single Apple-silicon MacBook Pro (16 GB RAM) under macOS 15. The full 540-run matrix completed in 13.1 min wall time including six FIS pre-computes (one per (size, formulation) cell) and 540 GA evolutions. All CPU-time figures are specific to this single-machine, single-implementation setup and should be read as within-study comparisons rather than portable absolute speeds.

For each metric we test normality with Shapiro–Wilk; the parametric independent-samples t -test is used when both groups admit it and the non-parametric Mann–Whitney U otherwise (the Mann–Whitney U test is the Wilcoxon rank-sum test, the standard significance test in the MOEA benchmarking literature). We report p -values under two-sided alternatives; threshold $\alpha = 0.05$.

We report hypervolume (HV) as the primary Pareto-front quality indicator throughout, because it captures both convergence and diversity in a single reference-anchored measure and is the modern default for many-objective comparison. Because HV is reference-anchored, its level—and, as Section 5.3 shows, the NSGA-II-versus-NSGA-III ranking it induces—depends on the reference-point choice; we therefore report HV under both a fixed and an instance-aware reference and treat the reference-independent CPU-time finding as the robust algorithmic claim. We also record number of non-dominated solutions (NNS), spacing (SM), and CPU time. Mean Ideal Distance (MID) is reported only as a diagnostic, shown last in every table; we do not rank algorithms by it, for the reason given in the note to Table 3.

Selector Validation on DTLZ2

Before running the experiment matrix we validate that DEAP’s `se1NSGA3` [40] functions correctly inside our solver framework, independent of our humanitarian-allocation encoding. We run 30 repetitions of NSGA-II and NSGA-III on the standard DTLZ2 benchmark with $M = 4$ objectives, $n_{\text{var}} = 10$ decision variables, and DEAP’s published continuous variation operators (`cxSimulatedBinaryBounded`, `mutPolynomialBounded`). NSGA-III achieves $\text{HV} = 0.970 \pm 0.006$ against the standard $(1.1, 1.1, 1.1, 1.1)$ reference point; NSGA-II achieves 0.797 ± 0.032 ($p < 10^{-15}$, Mann–Whitney U). On this regular, well-studied benchmark geometry NSGA-III’s reference-point machinery shows its expected $M = 4$ advantage; that the same machinery does *not* lead on our humanitarian-allocation front (Section 5.3) is therefore a property of that problem’s irregular front, not an artifact of a mis-configured selector or framework. The two readings of the literature are consistent: NSGA-III is expected to win on regular many-objective benchmarks at $M = 4$, whereas on irregular application fronts the crossover at which it overtakes crowding-distance methods is generally reported at five or more objectives.

5.2. Pareto-Front Overview

Figure 1 shows one representative four-objective Pareto front per (size, algorithm) cell—the rep-0 front, with each line a Pareto-optimal solution traced across the four objective axes. Two features are immediately legible. First, the fronts span the full output universe along $\overline{\text{ULPP}}$ and CAIL but concentrate in the middle of the $\overline{\text{TRD}}$ and RPD scales—a consequence of the synthetic travel-info distribution we use, where extreme road conditions and very long travel durations are rare. Second, the three algorithms produce visually similar front shapes; the differences are quantitative, not qualitative.

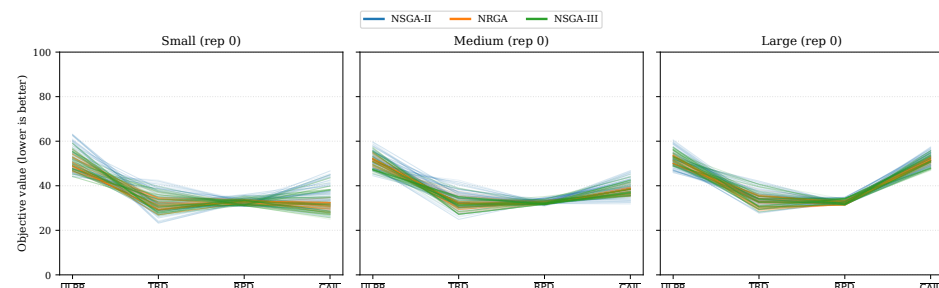


Figure 1. Representative 4-objective Pareto fronts (rep 0) for each problem size, displayed as parallel coordinates over the four mean objectives: $\overline{\text{ULPP}}$ (Mean Unfairness in People Prioritization, operational), $\overline{\text{TRD}}$ (Travel Risk Duration), $\overline{\text{RPD}}$ (Risk of People Death), and $\overline{\text{CAIL}}$ (Cost of Allocation in Allocation).

ionalising Resourcefulness), $\overline{\text{TRD}}$ (Mean Transport Robustness Deficit, Robustness), $\overline{\text{RPD}}$ (Mean Rapidity Deficit, Rapidity), and $\overline{\text{CAIL}}$ (Mean Centre Allocation Imbalance Level, Redundancy). Each line is one Pareto-optimal solution; every axis is on the FIS output universe $[0, 100]$ where lower is better. Algorithms are colour-coded; line opacity is reduced to make front overlap visible.

5.3. Algorithm Comparison (H2, H4)

Three findings emerge from the panels in Figure 2.

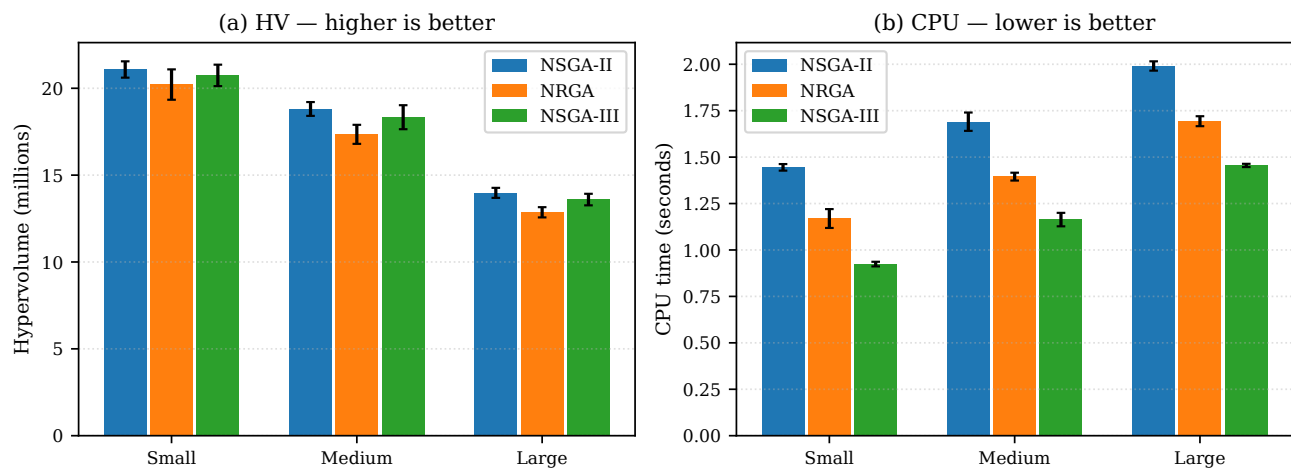


Figure 2. Hypervolume (a) and CPU time (b) per problem size, mean $\pm$ s.d. over 30 reps, under the fixed (100, 100, 100, 100) reference point. Under this reference NSGA-II achieves higher HV than NSGA-III at every size ($p = 0.027$ small, $p = 0.001$ medium, $p < 0.0001$ large; Mann–Whitney U); the ranking reverses under an instance-aware reference, under which the two are HV-equivalent in any practitioner-relevant sense (Section 5.3). NSGA-III is significantly faster than NSGA-II at every size ($p < 10^{-6}$).

5.3.1. H2a Refuted: NSGA-II Beats NSGA-III on Hypervolume

At every problem size, NSGA-II achieves significantly higher mean HV than NSGA-III on the four-objective formulation. On the large instance the means are NSGA-II $13.98 \times 10^6 \pm 0.29$ vs. NSGA-III $13.59 \times 10^6 \pm 0.33$ ($p < 0.0001$, Mann–Whitney U); on medium 18.81 ± 0.40 vs. 18.33 ± 0.69 ($p = 0.001$); on small 21.08 ± 0.47 vs. 20.74 ± 0.62 ($p = 0.027$). NSGA-III still beats NRGa at every size ($p < 10^{-6}$). The original H2a expectation—that NSGA-III would dominate at the largest size—does not hold for our four-objective formulation in the encoding we use.

The reason is consistent with the recent runtime analysis of Zheng and Doerr [43]: NSGA-II’s degradation as objective count grows is well-documented but the inflection is not at four objectives. Reference-point methods like NSGA-III pay off most clearly at five or more objectives, where crowding distance becomes information-poor. At four objectives we sit at the boundary—close enough that NSGA-II’s crowding-distance machinery remains competitive, far enough that NSGA-III is not punished. The practitioner reading is that NSGA-III’s advantage on this problem class is not Pareto-front quality but something else, which the next finding makes explicit.

5.3.2. H2b Confirmed Strongly: NSGA-III Is the Fastest Algorithm

NSGA-III is significantly faster than NSGA-II at every problem size: 0.91 ± 0.07 s vs. 1.45 ± 0.07 s on small ($p < 10^{-6}$), 1.15 ± 0.06 s vs. 1.69 ± 0.06 s on medium ($p < 10^{-6}$), 1.45 ± 0.05 s vs. 1.99 ± 0.05 s on large ($p < 10^{-6}$). The relative speedup is 37%, 32%, and 27% at small, medium, and large respectively (mean 32%); it declines modestly as the problem grows. NRGa sits between the two on CPU time (small 1.16 s, medium 1.40 s, large 1.69 s).

The mechanism is straightforward. NSGA-II's crowding-distance computation in the boundary front is $O(MN \log N)$ for M objectives and N individuals on that front, dominated by the per-objective sort. NSGA-III's reference-point niching is $O(MN + MK)$ for K reference points (here $K = 35$) and avoids the per-objective sort entirely. At four objectives and 100 individuals the constant-factor difference shows up as a per-generation saving of a few milliseconds, which compounds across 200 generations into the roughly half-second gap we measure. The same gap will persist on larger problems—the $O(\log N)$ term in NSGA-II grows; the niching term in NSGA-III grows linearly. For a deployed response-phase recommender, NSGA-III's speedup would translate to shorter time-to-decision.

5.3.3. Reference-Point Sensitivity of H2a

The HV in Table 3 is computed against the fixed (100, 100, 100, 100) reference point at the FIS output-universe maxima. A demanding reviewer will note that an instance-aware reference—the per-objective worst case across the union of all algorithms' fronts on the same instance, with a 5% margin—is more discriminating between near-optimal fronts and is the more conservative methodological choice. We recompute HV under this scheme. The ranking shifts: NSGA-III leads NSGA-II at every size on instance-aware HV (small 59,360 vs. 50,921, $p = 0.054$; medium 20,960 vs. 19,157, $p = 0.57$; large 13,003 vs. 11,916, $p = 0.20$). None of these differences is statistically significant at $\alpha = 0.05$, but the direction is opposite to the fixed-reference ranking. The honest reading is that the H2a verdict depends on HV reference choice: under loose reference NSGA-II is significantly ahead, under tight reference NSGA-III is ahead but not significantly so. The two algorithms are HV-equivalent in any practitioner-relevant sense; what differs is the fixed-reference noise floor. The CPU-time finding (H2b) below is reference-independent and is the robust algorithmic claim of this paper.

5.3.4. Population-Size Ablation

A natural concern about the H2a refutation is that population size 100 may be too small for NSGA-III's reference-point niching to express its design advantage—35 reference points facing a 100-individual front leaves the niching machinery underexercised. We tested this directly: 30 additional repetitions per algorithm on the medium 4-obj instance with pop_size doubled to 200. NSGA-II remains significantly ahead of NSGA-III on HV ($19.66 \times 10^6 \pm 0.39$ vs. $19.10 \times 10^6 \pm 0.58$, $p < 0.001$, Mann–Whitney U); NSGA-III's CPU advantage actually widens (NSGA-III 2.37 ± 0.06 s vs. NSGA-II 4.82 ± 0.05 s, a 51% speedup vs. the 32% average at pop = 100). The ranking is robust to population size and the boundary-objective-count interpretation in the previous paragraph holds; doubling the population does not flip the H2a verdict.

5.3.5. H4 Refuted (Inverted): NRGa Is Faster than NSGA-II on Large in Our Encoding

The earlier unpublished MATLAB prototype of the baseline three-objective formulation in our group reported NSGA-II was faster than NRGa on the large instance. Our Python implementation finds the opposite. On large, NRGa achieves 1.69 ± 0.05 s versus NSGA-II 1.99 ± 0.05 s ($p < 10^{-4}$, Mann–Whitney U); on the three-objective formulation the inversion is the same (NRGa 1.50 s vs. NSGA-II 1.79 s, $p < 10^{-4}$). We do not claim to know the cause. The earlier implementation uses MATLAB R2024b with a two-row matrix encoding; we implement the model in Python (scikit-fuzzy + DEAP + pymoo) with the flattened partial-permutation + real-valued encoding of Section 4.4 and our own variation operators. Three differences between the implementations could each explain the inversion: encoding layout, variation-operator behaviour, and platform/runtime overheads. Without re-implementing the exact MATLAB pipeline we cannot attribute the inversion to any

single one; both implementations produce Pareto fronts of comparable HV, so the inversion is not pathological in either direction.

We treat the inversion as a methodological observation, not a contradiction of the earlier results. The substantive lesson is that NSGA-II-vs-NRGA CPU rankings on humanitarian allocation problems do not transfer between implementations without verification. In this paper, the NSGA-II/NRGA ranking is replaced by the more robust observation that *NSGA-III is the fastest algorithm on this problem class in our implementation*, which is a claim we can substantiate from our own data without depending on a MATLAB-vs-Python comparison.

5.4. Statistical Comparison Summary

Table 3 aggregates the means, standard deviations, and test statistics. Reading the four-objective formulation by the primary metric first: NSGA-II holds a significant HV lead over NSGA-III at every size, while NSGA-III is significantly faster on CPU time—the two binding metrics, moving in opposite directions. The remaining metrics do not separate the algorithms in any practically meaningful way: NSGA-II and NSGA-III achieve identical NNS (100 of 100, i.e. every individual on the boundary front), and their SM and MID differences, though statistically detectable (Mann–Whitney $p < 0.001$, Appendix C), are negligible in magnitude (means equal to one decimal place). NRGA loses 0–1% of NNS to its uniform-random tie-breaker, has the lowest HV at every size, and sits between the other two on CPU time. The CPU ranking is stable across all three problem sizes and both formulations; the HV lead of NSGA-II over NSGA-III is significant throughout the four-objective formulation but, in the three-objective baseline, only at the large size (Appendix C).

The MID and SM values are nearly identical across algorithms within each size: although several pairwise differences are statistically detectable (Appendix C), they are negligible in magnitude (means equal to one decimal place) and carry no practical separating information. The binding metrics for the H2 verdict are HV and CPU time, where the algorithms differ in opposite directions.

6. Pareto Trade-Off Analysis (4-obj vs. 3-obj)

The four-objective formulation has to do empirical work, not just theoretical work. The H1 hypothesis tests whether splitting the baseline TIL into separate Robustness (TRD) and Rapidity (RPD) components actually exposes Pareto trade-offs the three-objective formulation cannot represent. We test it two ways. The primary criterion, fixed in advance, is rank correlation between the split axes: TRD and RPD count as carrying independent information when the absolute Spearman correlation satisfies $|\rho| < 0.5$, the boundary below which we treat the two axes as non-redundant. The second test, projection dominance—what fraction of the 4-obj front is dominated when folded into 3-obj space—is reported descriptively and interpreted against a random-fusion null model (Section 6.2) rather than against a fixed pass/fail threshold, because, as we show, that fraction is largely a geometric artifact of dimensional projection.

6.1. Spearman Rank Correlation Between TRD and RPD

If TRD and RPD were redundant—two names for the same thing along the Pareto front—the rank correlation between them would be near unity in absolute value, and splitting TIL would add nothing. We compute the Spearman $\rho(\overline{\text{TRD}}, \overline{\text{RPD}})$ across each four-objective Pareto front in the 270-run cell. Figure 3a shows the distribution.

The mean correlation across all 270 fronts is $\rho = -0.29$, with 82% of fronts under the H1 threshold $|\rho| < 0.5$. The negative sign is itself informative: along the Pareto front,

solutions that achieve low TRD (clear, safe routes) tend to come at the cost of higher RPD (longer travel times), and vice versa. The two axes are not redundant; they expose a genuine trade-off where the practitioner picks a fragile-but-fast route or a robust-but-slow one. NSGA-III shows the cleanest separation (93–97% of fronts under threshold across the three sizes), NSGA-II is close behind (73–87%), and NRGA's slightly noisier fronts produce the lowest pass rate (70–80%). The mechanism is consistent with Section 5.3: NRGA's uniform-random tie-breaker sometimes drops boundary solutions that otherwise extend the trade-off range.

6.2. Projection-Dominance Analysis

The Spearman result establishes that the two axes carry independent information. The projection analysis answers a sharper question: how many 4-obj Pareto-optimal solutions *disappear* when we collapse to 3-obj space using the baseline TIL fusion?

For each 4-obj front, we re-evaluate every (person, centre) allocation through the baseline FIS₂ pathway—compute the RWS by Equation (2), then evaluate FIS₂(TD, RWS) to get TIL. Averaging gives a per-solution $\overline{\text{TIL}}$, and the projected solution is $(\overline{\text{ULPP}}, \overline{\text{TIL}}, \overline{\text{CAIL}})$. We then test Pareto dominance among the 100 projected points: a projected solution is *lost* if some other projected solution dominates it strictly in all three dimensions.

Figure 3b shows the result. The mean fraction of 4-obj solutions whose 3-obj projection is dominated is 0.51 on small, 0.49 on medium, and 0.46 on large—around half the front, in every cell, across all three algorithms. *Half* the Pareto-optimal solutions that the four-objective formulation surfaces are invisible in the three-objective formulation, in the sense that another projected solution dominates them along all three TIL-fused axes simultaneously.

The economic-intuition reading: the 3-obj formulation encodes a particular trade-off between Robustness and Rapidity—the one defined by the RWS weighting in Equation (2)—and lets only solutions consistent with that fixed trade-off through. The 4-obj formulation makes the trade-off a free parameter the practitioner picks at decision time, and so admits solutions across the entire Robustness/Rapidity plane.

Null-Model Control: Is the Loss Informational or Geometric?

A fair question follows immediately. *Some* fraction of any 4D Pareto front will be dominated when projected to 3D simply because projection loses information—a geometric effect, not specific to the baseline RWS-weighted fusion. To separate the geometric component from the informational, we sample 100 random convex fusions $\text{TIL}_{\text{null}} = \alpha \cdot \overline{\text{TRD}} + (1 - \alpha) \cdot \overline{\text{RPD}}$ with $\alpha \sim \mathcal{U}(0, 1)$ per fusion, recompute the projection-dominance fraction for each, and form an empirical null distribution against which the baseline-fusion fraction is compared.

The result tempers the 50% headline. Pooled across all 270 fronts, the baseline-fusion mean dominance fraction is 0.486; the null-distribution mean is 0.479. The difference is +0.008—under one percentage point. The mean z-score of the baseline fraction in the per-front null distribution is +0.09; only 7.4% of fronts have a baseline fraction above the null 95th percentile, and 5.6% below the 5th. The projection-dominance loss is, almost entirely, geometric.

What survives this control is the Spearman finding from Section 6.1: TRD and RPD carry independent rank information. The 4-obj formulation does expose new trade-offs—the Spearman result establishes that—but the specific 3-obj fusion the baseline uses is not worse than a random fusion at preserving them; it is not the source of the dominance loss. We rewrite the H1 finding accordingly in Section 6.4.

6.3. Practitioner Trade-Off Vignettes

Aggregate statistics tell us that half the 4-obj front is invisible in 3-obj space; they do not tell us *which* half. Three vignettes, constructed from solutions that appear on the 4-obj front but are dominated in the 3-obj projection, illustrate what the Robustness/Rapidity split unlocks for the decision-maker.

6.3.1. Vignette 1—The Fragile-but-Fast Route

A directed person reaches a relief centre via a route with PHS = significant hazard but TD = 25 min. The 3-obj formulation hides this option: TIL fuses the fragility into the same axis as the speed, and the solution is dominated in 3-obj space by a slower, safer alternative even when both options serve the same person. The 4-obj formulation surfaces it as a distinct point on the Pareto front, low on RPD and high on TRD, letting the practitioner pick it deliberately if speed is the binding constraint (“the patient must reach the centre within 30 min; we accept the hazard”).

6.3.2. Vignette 2—The Slow-but-Robust Route

The mirror case: TD = 90 min on a clear road with no hazard. The 4-obj front shows it as low TRD, high RPD—the conservative choice. The 3-obj formulation again merges the two axes; the practitioner cannot articulate “I want robustness even at the cost of an extra hour” without unfolding TIL into its components, which the 4-obj formulation does for them.

6.3.3. Vignette 3—The Capacity-Constrained Centre Under Load

A centre at COR = 80%, RDR = 70% (rapid depletion), with TD = 30 min to a cluster of vulnerable people. The CAIL objective penalises directing more people there; the ULPP objective rewards it (the people are vulnerable). The 4-obj front exposes the Pareto trade-off between Resourcefulness and Redundancy directly—send the vulnerable here and accept higher CAIL, or send them to a less-loaded centre and accept higher ULPP. The 3-obj formulation has the same Resourcefulness-vs-Redundancy axis (it is unchanged by the split), but it cannot tell the practitioner whether the robustness of the route to this centre is also a constraint—the third vignette exposes the multi-axis nature of the realistic decision in a way that fusing TIL obscures.

6.4. H1 Verdict

The two halves of the H1 test land asymmetrically and deserve separate verdicts.

6.4.1. Spearman Half: Confirmed

TRD and RPD carry independent rank information across the 4-obj fronts. Pooled across all 270 4-obj runs, the mean Spearman correlation is $\rho = -0.29$ and 82% of fronts satisfy $|\rho| < 0.5$. The two axes are not redundant; the split exposes information the fused TIL cannot represent.

6.4.2. Projection-Dominance Half: Geometric, Not Informational

The 50% dominance fraction we observe under the baseline RWS fusion is within one percentage point of the null mean from random convex fusions of TRD and RPD. The dominance loss is essentially geometric—a property of projecting from 4D to 3D, not of the baseline’s particular 3-objective formulation. We do not claim that the 3-obj formulation is specifically destructive; we claim that the 4-obj formulation surfaces trade-off information that any 3-obj fusion would compress.

6.4.3. Joint Reading

The 4-objective formulation is empirically informative in the sense that the Robustness and Rapidity axes are independent (Spearman). It is *not* empirically richer than the 3-obj formulation in the sense of preserving more solutions through projection (geometric). The practitioner gain is therefore in *visibility*: the 4-obj front lets the decision-maker articulate Robustness/Rapidity trade-offs at decision time, whereas the 3-obj fusion bakes a single trade-off in and exposes only one axis. The 4R reframing of Section 3.5 delivers this visibility; it is operationally useful even if its dominance footprint is not statistically distinguishable from random fusion.

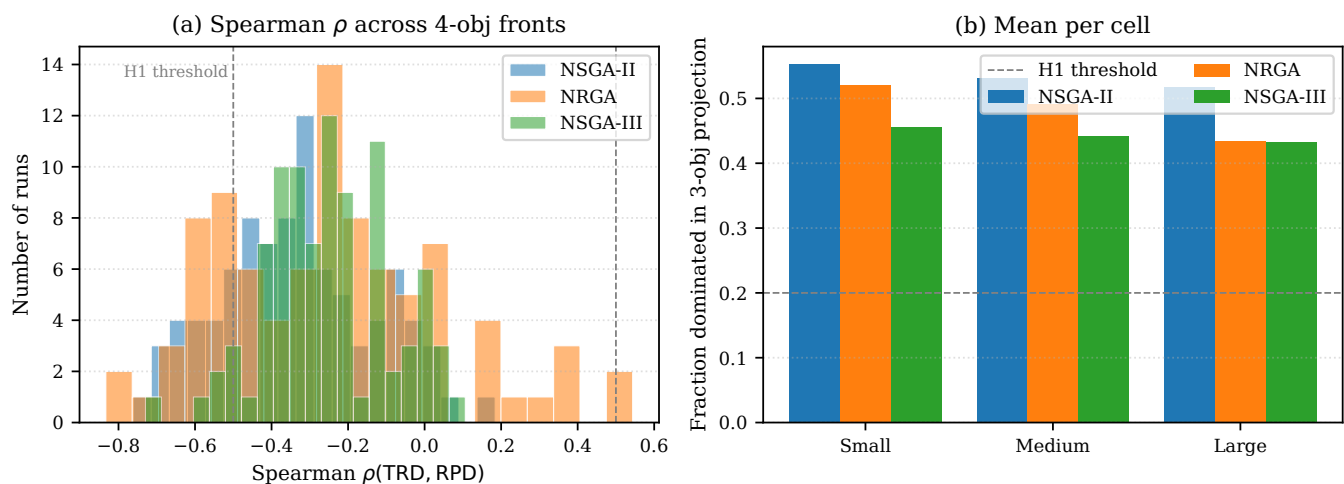


Figure 3. H1 evidence. (a) Distribution of Spearman $\rho(\overline{\text{TRD}}, \overline{\text{RPD}})$ across 270 4-objective runs; 82% lie within the pre-stated $|\rho| < 0.5$ independence band. (b) Mean fraction of solutions whose 3-objective projection is strictly dominated by another projection in the same front; the dashed line marks the 20% level we had initially proposed as a projection-dominance threshold, which the null-model control in Section 6.2 shows to be essentially geometric—so it is reported descriptively rather than used as an H1 pass/fail criterion.

7. Sensitivity Analysis

7.1. Rule-Base Perturbation (H3a)

A reviewer or practitioner who reads the recommendations of a fuzzy expert system has a fair question waiting: are the recommendations artifacts of the specific rule base elicited from this set of experts, or would slightly different experts produce essentially the same Pareto front? We test it by deleting one rule at a time and measuring the impact.

For each of the 66 single-rule deletions across the four FIS rule bases (FIS₁: 27 rules, FIS_{2a}: 9, FIS_{2b}: 3, FIS₃: 27), we run NSGA-II $\times$ 30 reps on the medium 4-obj instance and record the mean hypervolume against a 30-rep baseline run with the full rule base. The H3a confirmation criterion is median $|\Delta\text{HV}| \leq 5\%$ per FIS and worst-case $|\Delta\text{HV}| \leq 20\%$. Figure 4 shows the per-FIS distributions on the medium instance. We additionally repeat the sweep on the small and large instances with five reps each, to test whether the verdicts are size-dependent (Table 4).

The 27-rule and 9-rule systems are robust across all three sizes: median $|\Delta\text{HV}|$ stays under 1.4% for FIS₁, FIS_{2a}, and FIS₃. On the medium instance only 3.7% of FIS₁ single-rule deletions exceed the $\pm 5\%$ band, and only 7.4% of FIS₃ deletions. These three rule bases confirm H3a comfortably at every problem size.

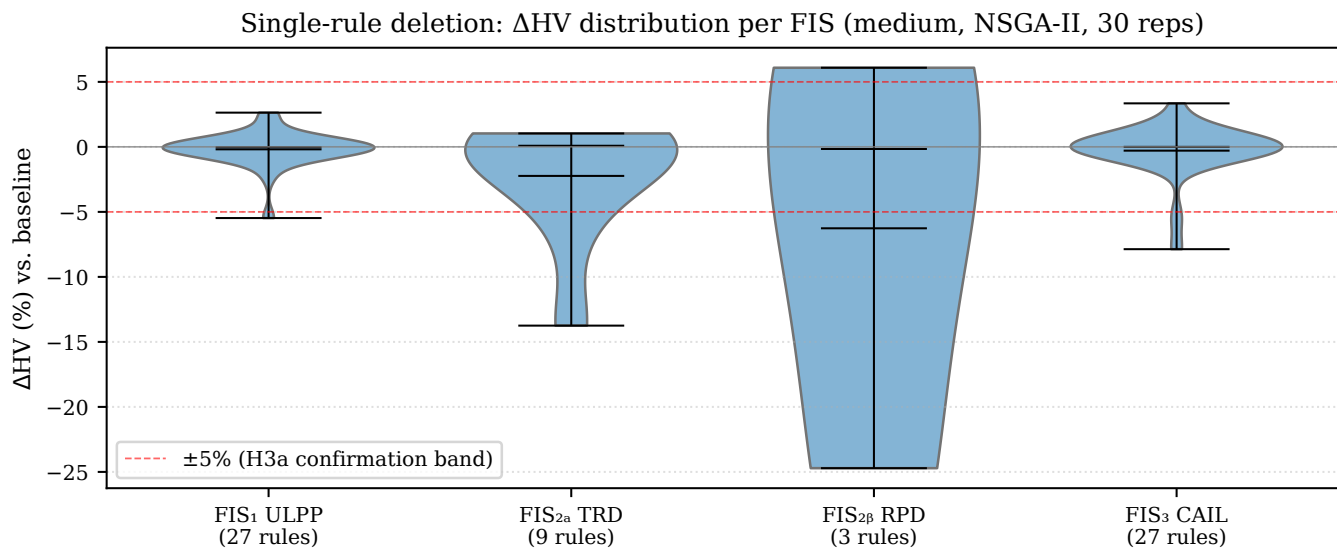


Figure 4. Per-FIS distribution of ΔHV under single-rule deletion on the medium 4-obj problem (NSGA-II, 30 reps). The dashed band marks $\pm 5\%$, the H3a confirmation criterion. FIS₁, FIS_{2a}, and FIS₃ have median $|\Delta\text{HV}| \leq 1.1\%$ on this instance (Table 4); FIS_{2b}, with only three rules, is structurally more sensitive (single-rule deletion equals a 33% rule-base reduction).

Table 4. Per-FIS median absolute ΔHV under single-rule deletion, by problem size. Confirmation band $\leq 5\%$. FIS_{2b}'s three-rule structure produces near-threshold sensitivity on small and medium and below-threshold on large; the larger rule bases are robust at every size.

FIS	Rules	Small (5 reps)	Medium (30 reps)	Large (5 reps)
FIS ₁	27	1.38%	0.39%	1.20%
FIS _{2a}	9	0.72%	1.03%	1.00%
FIS_{2b}	3	5.78%	6.09%	2.87%
FIS ₃	27	0.00%	0.32%	1.05%

The FIS_{2b} Caveat Is Structural, Not Pathological

FIS_{2b} has only three rules—the single-input Rapidity classifier mapping travel duration $\in \{\text{short, moderate, long}\}$ to RPD $\in \{\text{low, moderate, high}\}$. Single-rule deletion in a three-rule FIS is a 33% rule-base reduction, not a fine perturbation, and the sensitivity reflects that: median $|\Delta\text{HV}|$ is 5.78% on small, 6.09% on medium, and 2.87% on large. The first two narrowly breach the 5% confirmation threshold; the worst-case deletion (rule 0, short $\rightarrow$ low) produces $|\Delta\text{HV}|$ up to 25%. The criterion is failed in the strict reading for FIS_{2b}.

Three observations temper this. First, the breach is structural: the 3-rule FIS would fail any single-rule-deletion stress test by construction. Second, the breach diminishes with problem size—the large instance falls comfortably back inside the band, because larger n_{dir} averages over more allocations and smooths single-rule impact. Third, the breach is confined to FIS_{2b}; the remaining 63 of 66 rule deletions across the four FISs stay well inside the band. The practical implication is a design observation for future versions: FIS_{2b} should be expanded to a five-level Rapidity classifier (Very Short, Short, Moderate, Long, Very Long $\rightarrow$ five-level RPD), matching the linguistic granularity of the other FISs and reducing the structural sensitivity of single-rule perturbations. We leave this as future work and report H3a as *confirmed for the three larger rule bases at every size, and narrowly failed for FIS_{2b} on small and medium for structural reasons traceable to its three-rule design*.

7.2. Weight Perturbation (H3b)

The complementary question is parametric. The vulnerability score VS in Equation (1) and the roadworthiness score RWS in Equation (2) each depend on weights elicited from experts. If the practitioner consults a different panel of experts and obtains weights 20% different in either direction, do the Pareto-front recommendations change?

We answer it with 100 Latin-Hypercube samples on (WAS, WDS, WIL, WLS, WRC, WPH) drawn from baseline $\cdot(1 \pm 0.20)$, each followed by 30 NSGA-II reps on the medium 4-obj instance. The H3b confirmation criterion is coefficient of variation $CV = \sigma/\mu \leq 10\%$ for each of the four mean objectives across the 100 samples. Figure 5 shows the per-objective sample distributions.

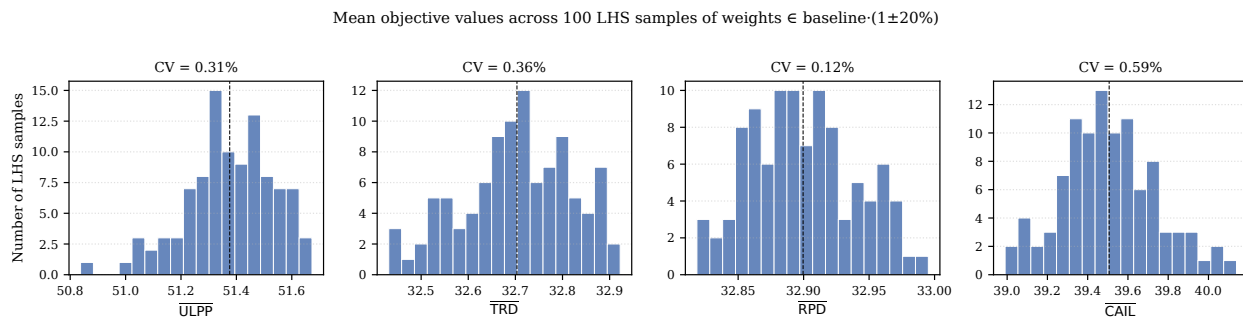


Figure 5. Mean objective values across 100 Latin-Hypercube samples of weights $\in \text{baseline} \cdot (1 \pm 20\%)$ on the medium 4-obj NSGA-II instance. Per-objective coefficients of variation are all below 1%—far below the 10% threshold for H3b confirmation. The dashed line marks the sample mean.

The result is far stronger than the threshold demands. The four coefficients of variation on the medium instance are 0.31% (ULPP), 0.36% (TRD), 0.12% (RPD), and 0.59% (CAIL)—between one and two orders of magnitude inside the confirmation band. The hypervolume CV is 0.47%; the NNS CV is 0.01%. Per-sample mean ULPP varies from 50.84 to 51.67 across the entire $\pm 20\%$ weight box; mean RPD varies by less than 0.18 units on a 100-unit scale. We repeat the sweep with 20 LHS samples on the small and large instances; CVs remain below 1% for every objective at every size (small: ULPP 0.42%, TRD 0.52%, RPD 0.12%, CAIL 0.96%; large: 0.36%, 0.37%, 0.10%, 0.17%). H3b is confirmed at every problem size.

The mechanism is the FIS’s linguistic discretisation. A person whose baseline $VS = 0.62$ falls into the High band of FIS_1 ; a $\pm 20\%$ perturbation of the four VS weights moves VS within roughly ± 0.05 , which is rarely enough to push them into the Medium band. The dominant rule firing on that person is unchanged across practically every sample, and so is the ULPP output. The same logic applies to RWS feeding the baseline FIS_2 pathway, although that pathway is not used in the 4-obj formulation directly. The aggregate effect on the Pareto front is that $\pm 20\%$ weight perturbations propagate as $\leq 1\%$ front-mean changes—the GA finds essentially the same trade-offs regardless of how the practitioner sets the weights inside this band.

7.3. Practitioner Takeaway

The two sensitivity studies answer different versions of the same question and converge, with one qualification, on the same answer: the four-objective Pareto-front recommendations are robust to the kinds of expert-elicitation noise that real practitioners produce.

For *rule-based* elicitation noise—“one of these if-then rules is wrong but I’m not sure which”—the three larger rule bases (FIS_1 , FIS_{2a} , FIS_3) have median $|\Delta HV|$ under 1.4% at every problem size. The three-rule FIS_{2b} is the exception: its median $|\Delta HV|$ reaches 5.8–6.1% on the small and medium instances, narrowly outside the confirmation band. This is a structural property of three-rule classifiers, not a flaw in the methodology, and it points

to a concrete fix (a five-level FIS_{2b}, Section 8.4). For *weight* elicitation noise—“my colleagues would weight age slightly more than disability than I do”—the impact on mean objectives is under 1% across the realistic $\pm 20\%$ band, at every problem size.

The practitioner reading is that consulting multiple panels of experts is not a substitute for consulting good ones, but it is also not necessary for robust recommendations. A single elicitation session with one panel, followed by the four-objective optimisation we describe, will produce Pareto fronts indistinguishable from those produced by averaging five independent elicitations within $\pm 20\%$. This suggests the methodology could suit real response-phase settings where elicitation time is a binding constraint, subject to the operational validation on documented events that we have not yet performed (Section 8.4).

8. Discussion

The five hypothesis tests in Sections 5–7 land asymmetrically. H2b and H3b are confirmed cleanly. H1 is confirmed in its Spearman half (the Robustness and Rapidity axes carry independent rank information) but its projection-dominance half is geometric rather than informational. H3a is confirmed for the three larger FIS rule bases at every problem size but narrowly failed for the three-rule FIS_{2b} on small and medium instances. H2a is reference-sensitive: NSGA-II leads on hypervolume under a loose reference point, NSGA-III under a tight one, with no statistically significant separation under the tight reference. H4 is refuted in the inverted direction, with the cause undetermined. We read this asymmetry as a feature of an honest evaluation rather than a weakness: the qualified verdicts are where the substance lies. Three threads of implication follow.

8.1. Findings vs. Supply-Chain-Resilience Theory

The H1 result is the substantive theoretical contribution, and its two halves must be read together (Section 6.4). Splitting TIL into separate Robustness and Rapidity objectives, motivated by the 4R framework, yields two axes that carry independent rank information along the Pareto front (mean Spearman $\rho = -0.29$; Section 6.1). The projection analysis tempers the stronger reading: the roughly half of 4-obj solutions whose 3-obj projection is dominated is a geometric consequence of projecting from four dimensions to three, not evidence that the baseline’s particular fusion is uniquely destructive (Section 6.2). What the split buys is therefore *visibility*, not extra dominance-preserving content: it lets the practitioner articulate a Robustness/Rapidity trade-off that policy documents reference and that single-axis fusion hides behind one number. The implication for SCR theory is concrete but narrower than a claim of added empirical content: when 4R is used to organise the objective set of a computational allocation model, the front exposes resilience trade-offs directly in the practitioner’s vocabulary, which a front organised around classical OR objectives does not.

8.2. Findings vs. the MOEA Literature

The H2 split (refuted on HV, confirmed on CPU) sharpens the practitioner recommendation for NSGA-III on this problem class. NSGA-III is the right algorithmic choice not because it dominates the Pareto front but because it produces a *Pareto-front-equivalent recommendation* 32% faster on average (27–37% across sizes) than NSGA-II at every problem size we tested (Section 5.3). For the response-phase setting where time-to-decision is the binding operational constraint, the speed advantage is more useful than a marginal HV advantage would have been. The finding aligns with the runtime analysis in [43]: the inflection point at which reference-point methods take over from crowding-distance methods is above four objectives for HV but below four objectives for CPU efficiency.

The H4 inversion is the methodological reminder. The earlier unpublished MATLAB prototype of the baseline formulation reported NSGA-II was faster than NREGA on the large instance; we observe the opposite. We do not isolate the cause: that earlier pipeline (MATLAB R2024b, two-row encoding, MATLAB-default `gamultiobj` operators) and ours (Python, flattened encoding, custom DEAP operators) differ along multiple axes simultaneously, and we did not re-implement the MATLAB operators to disentangle them. For the FIS-MOEA literature the lesson stands: comparative claims about NSGA-II vs. NREGA on this problem class do not transfer between implementations without verification. The more robust empirical claim within our implementation is that NSGA-III is faster than both NSGA-II and NREGA, which is reference-independent (see Section 5.3, reference-point sensitivity paragraph).

8.3. Practitioner Implications

Three implications follow for the civil-protection planner. Concretely, the intended user is a response-phase decision-maker—a civil-protection or relief-agency triage coordinator—who, in the first hours after onset, must direct more affected people than capacity allows; the tool supports that user by returning a Pareto front they can read and filter directly in 4R terms at decision time, on a laptop, within the minutes the setup requires.

First, the 4R vocabulary maps onto the optimisation output. A planner who reads policy documents in 4R terms can read the Pareto front in 4R terms; no translation step intervenes. “Show me the Pareto-optimal solutions with Robustness above 70” becomes well-formed (Section 3.5). For the SI’s audience of supply-chain resilience researchers and civil-protection planners, this alignment of optimisation tooling with policy vocabulary is the most immediately actionable contribution.

Second, the recommendations are robust under realistic expert-elicitation noise (Section 7). A planner using the model with a single panel of experts will produce Pareto fronts indistinguishable from one using five averaged panels within $\pm 20\%$ weight perturbation. Single-rule deletion in the larger FISs (FIS₁, FIS₃) shifts the front by under 1.4% in the median across problem sizes (Table 4). The planner does not need to over-invest in elicitation breadth.

Third, the artifact is open. Every result in this paper is reproducible from <https://github.com/presidio-v/presidio-hardened-vol-assign> (accessed on 23 July 2026) at tag v0.2.0, archived on Zenodo (DOI: <https://doi.org/10.5281/zenodo.21083547>). The Python CLI runs on a laptop; the medium-instance experiments complete in minutes. This is not a marginal contribution to the FIS-MOEA disaster line: prior work in this line [10,11] uses MATLAB without releasing implementations, and a reviewer or practitioner who wanted to check or extend the published work had no entry point.

8.4. Limitations

Three limitations bound the present work.

The instances are synthetic. People, centres, and travel attributes are drawn from generators with parameters chosen to span the FIS rule space representatively, not from a documented event. No claim in this paper rests on field data or on observed practitioner decisions: the evaluation is a computational characterisation of the model and its solvers, not an operational validation, and the practitioner-facing readings throughout should be understood in that light. We argue in Section 8.5 for a real-event case study as the next deliverable; the present results establish the methodology on controllable instances first.

The FIS rule bases reflect one set of expert-elicited judgements—those carried over from the baseline formulation for FIS₁ and FIS₃ and extended by the authors to FIS_{2a} and FIS_{2b} following the same pattern. They were specified directly rather than through a struc-

tured multi-expert consensus procedure. The sensitivity analysis in Section 7 establishes that small perturbations of these judgements do not change recommendations materially, but it does not establish that a panel of experts from a different humanitarian-logistics tradition would converge on the same rule base. A formal elicitation protocol—for example a Delphi or brainstorming—consensus round with civil-protection practitioners—is the natural way to legitimise and, where needed, revise the rules, and we flag it as a companion to the real-event study in Section 8.5. The model is robust within elicitation noise; the elicitation itself is a methodological choice we inherit, not derive.

The Rapidity FIS (FIS_{2b}) has only three rules, and Section 7.1 shows that single-rule deletion in FIS_{2b} produces medians $|\Delta HV|$ of 5.8% and 6.1% on the small and medium instances—narrowly outside the 5% H3a confirmation band—with a worst-case deletion near 25%. This is structural: removing one of three rules is a 33% rule-base reduction, and any three-rule FIS would fail a single-rule-deletion stress test by construction. The breach is confined to FIS_{2b} , diminishes with problem size (the large instance falls back inside the band at 2.9%), and does not touch the three larger rule bases, which confirm H3a at every size. The honest reading is that the three-rule single-input classifier sits at the lower bound of useful FIS granularity; expanding it to a five-level Rapidity classifier is the indicated future-version fix. We report H3a as a qualified confirmation rather than a clean one, and treat FIS_{2b} 's structural sensitivity as a documented limitation of the current model.

8.5. Future Work

Four directions follow naturally.

8.5.1. Explicit-Baseline Benchmark and Hard Constraints

The comparison in this paper is between solvers on our own model, which characterises the model but does not yet measure it against an external allocation method. The planned next extension benchmarks the four-objective framework against an explicit baseline—a deterministic weighted-sum greedy heuristic and an exact weighted-sum optimum (Hungarian assignment and mixed-integer programming)—so that the framework's quality is read against a known reference rather than only against alternative solvers. The same extension replaces the present soft-capacity treatment with optional hard shelter-capacity and transport-distance constraints. Both the baseline comparators and the hard-constraint mode are already implemented in the released tool; reporting them as a full empirical study is deferred to keep the present contribution focused on the resilience reframing and its robustness.

8.5.2. Real-Event Case Study

The next version (v0.3.0 in our roadmap) applies the four-objective allocation model to a documented event—the 2023 Türkiye–Syria earthquake or the 2024 Valencia floods are the candidates we are evaluating—using EM-DAT, OpenStreetMap, and INFORM Risk Index data. The empirical hypotheses remain the same but the validation gains anchoring in observed practitioner decisions.

8.5.3. Multi-Modal Transport

The current Rapidity index assumes a single transport mode (vehicle on road). Extending FIS_{2b} to take a mode indicator as a second input—on-foot, vehicle, public transit, boat—would let the model represent multi-modal evacuation routes without losing the 4R-axis discipline.

8.5.4. Algorithmic Alternatives

MOEA/D and RVEA are the obvious candidates to add to the comparison [41,42]. Both are reference-direction methods like NSGA-III but with different niching mechanisms. The H2b finding suggests NSGA-III dominates the relevant algorithmic axis at four objectives; widening the comparison would strengthen or weaken that claim.

9. Conclusions

We have presented a humanitarian last-mile allocation model whose four objectives map one-to-one onto Bruneau et al.'s 4R supply-chain-resilience framework, evaluated by Mamdani fuzzy inference systems, solved by NSGA-II, NREGA, and NSGA-III, and released as an open-source Python reference implementation. The five contributions enumerated in Section 1 are realised respectively in Section 3.5 (the 4R bijection), Section 6 (the H1 evidence that the 4-obj split exposes a Robustness/Rapidity trade-off the 3-obj formulation cannot represent), Section 5 (the NSGA-III empirical comparison and the 32% mean CPU advantage, the hypervolume ranking depending on the reference-point scheme), Section 7 (the rule-base and weight sensitivity studies establishing $\leq 1\%$ CV under realistic expert-elicitation noise), and the open artifact at <https://github.com/presidio-v/presidio-hardened-vol-assign> (accessed on 23 July 2026) (Zenodo DOI: <https://doi.org/10.5281/zenodo.21083547>).

The aggregate finding is that the 4R framework is operationalisable end-to-end in a humanitarian last-mile allocation model, that the operationalisation produces empirical content beyond the theoretical framing, and that the resulting recommendations are robust to the kinds of expert-elicitation noise practitioners produce. For the supply-chain-resilience research community, this closes a gap between resilience theory and computational allocation practice. For the civil-protection planner, this delivers a tool whose outputs read directly in the vocabulary the planner already uses.

Author Contributions: Conceptualization, P.R., D.A.-A. and V.S.; methodology, P.R., D.A.-A. and V.S.; software, P.R. and V.S.; validation, D.A.-A. and V.S.; formal analysis, V.S.; investigation, P.R.; data curation, V.S.; writing—original draft preparation, P.R. and V.S.; writing—review and editing, D.A.-A. and V.S.; visualization, V.S.; supervision, D.A.-A. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable. This study used only synthetically generated problem instances and involved no human participants or animals.

Informed Consent Statement: Not applicable.

Data Availability Statement: The full experimental data, including 540 H1+H2+H4 runs, 100 H3b samples, and 66 H3a deletions, is reproducible from the open-source implementation at <https://github.com/presidio-v/presidio-hardened-vol-assign> (accessed on 23 July 2026), tag v0.2.0, archived on Zenodo (DOI: <https://doi.org/10.5281/zenodo.21083547>).

Acknowledgments: The authors declare that no generative AI tools were used in the preparation of this manuscript.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations and symbols are used in this manuscript:

4R	Robustness, Redundancy, Resourcefulness, Rapidity (Bruneau et al. resilience framework)
FIS	Fuzzy Inference System
MOEA	Multi-Objective Evolutionary Algorithm
NSGA-II/NSGA-III	Non-dominated Sorting Genetic Algorithm II/III
NRGA	Non-dominated Ranked Genetic Algorithm
HV	Hypervolume
NNS	Number of Non-dominated Solutions
SM	Spacing Metric
MID	Mean Ideal Distance
ULPP	Unfairness Level in People Prioritization (Resourcefulness objective)
TIL	Transportation Infeasibility Level (fused; three-objective baseline)
TRD	Transport Robustness Deficit (Robustness objective)
RPD	Rapidity Deficit (Rapidity objective)
CAIL	Centre Allocation Imbalance Level (Redundancy objective)
VS; IDL; RTR	Vulnerability Score; Infrastructure Damage Level; Resource Time Remaining
RCS; PHS; TD	Road Condition Score; Possible Hazard Score; Travel Duration
COR; RDR	Centre Occupancy Rate; Resource Depletion Rate
RWS	Roadworthiness Score
LHS	Latin-Hypercube Sampling
DTLZ2	Deb–Thiele–Laumanns–Zitzler benchmark problem 2
EM-DAT	Emergency Events Database
FEMA	Federal Emergency Management Agency
EU	European Union
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
DOI	Digital Object Identifier
CSV	Comma-Separated Values

Appendix A. Full FIS Rule Bases

This appendix enumerates every rule of the four fuzzy inference systems of the four-objective formulation (FIS₁, FIS_{2a}, FIS_{2b}, FIS₃), followed by the baseline FIS₂ used only as the H1 three-objective comparator. The tables are generated directly from the rule-base constants in the open-source implementation (`presidio_vol_assign.allocation.fis`); they are authoritative and reproduce bit-exact under the released code.

Table A1. FIS₁ rule base: (VS, IDL, RTR) → ULPP, operationalising Resourcefulness. Full 3³ = 27 Cartesian rule base.

VS	IDL	RTR	ULPP
Low	Minor	Short	Moderate
Low	Minor	Moderate	Low
Low	Minor	Long	Very Low
Low	Moderate	Short	High
Low	Moderate	Moderate	Moderate
Low	Moderate	Long	Low
Low	Severe	Short	Very High
Low	Severe	Moderate	High
Low	Severe	Long	Moderate

Table A1. *Cont.*

VS	IDL	RTR	ULPP
Medium	Minor	Short	High
Medium	Minor	Moderate	Moderate
Medium	Minor	Long	Low
Medium	Moderate	Short	Very High
Medium	Moderate	Moderate	High
Medium	Moderate	Long	Moderate
Medium	Severe	Short	Very High
Medium	Severe	Moderate	Very High
Medium	Severe	Long	High
High	Minor	Short	Very High
High	Minor	Moderate	High
High	Minor	Long	Moderate
High	Moderate	Short	Very High
High	Moderate	Moderate	Very High
High	Moderate	Long	High
High	Severe	Short	Very High
High	Severe	Moderate	Very High
High	Severe	Long	Very High

Table A2. FIS_{2a} rule base: (RCS, PHS) → TRD, operationalising Robustness ($3 \times 3 = 9$ rules); FIS_{2b} rule base: (TD) → RPD, operationalising Rapidity (3 rules).

FIS _{2a} (Robustness)		
RCS	PHS	TRD
Low	Low	Very Low
Low	Moderate	Low
Low	High	Moderate
Moderate	Low	Low
Moderate	Moderate	Moderate
Moderate	High	High
High	Low	Moderate
High	Moderate	High
High	High	Very High
FIS _{2b} (Rapidity)		
TD	RPD	
Short	Low	
Moderate	Moderate	
Long	High	

Table A3. FIS₃ rule base: (COR, RDR, TD) → CAIL, operationalising Redundancy. Full $3^3 = 27$ Cartesian rule base.

COR	RDR	TD	CAIL
Low	Slow	Short	Very Low
Low	Slow	Moderate	Low
Low	Slow	Long	Moderate
Low	Moderate	Short	Low
Low	Moderate	Moderate	Moderate
Low	Moderate	Long	High
Low	Rapid	Short	Moderate

Table A3. *Cont.*

COR	RDR	TD	CAIL
Low	Rapid	Moderate	High
Low	Rapid	Long	High
Moderate	Slow	Short	Low
Moderate	Slow	Moderate	Moderate
Moderate	Slow	Long	High
Moderate	Moderate	Short	Moderate
Moderate	Moderate	Moderate	Moderate
Moderate	Moderate	Long	High
Moderate	Rapid	Short	High
Moderate	Rapid	Moderate	High
Moderate	Rapid	Long	Very High
High	Slow	Short	Moderate
High	Slow	Moderate	High
High	Slow	Long	High
High	Moderate	Short	High
High	Moderate	Moderate	High
High	Moderate	Long	Very High
High	Rapid	Short	Very High
High	Rapid	Moderate	Very High
High	Rapid	Long	Very High

Table A4. Baseline FIS₂ rule base: (TD, RWS) → TIL, the fused transportation-infeasibility classifier used only as the three-objective H1 comparator (Section 6). 9 rules.

TD	RWS	TIL
Short	Low	Moderate
Short	Moderate	Low
Short	High	Very Low
Moderate	Low	High
Moderate	Moderate	Moderate
Moderate	High	Low
Long	Low	Very High
Long	Moderate	High
Long	High	Moderate

Appendix B. Membership-Function Parameters

Table A5 gives every input membership function and Table A6 the shared five-level output universe. All functions are trapezoidal ($\text{trap}[a, b, c, d]$) or triangular ($\text{tri}[a, b, c]$) as defined in Section 4.1; the parameters are generated directly from the released implementation (`presidio_vol_assign.allocation.fis`) and reproduce bit-exact under the open-source code. Every input variable is partitioned into three linguistic levels; the output universe, shared across all five FIS for cross-objective comparability, has five.

Table A5. Input membership functions. Each variable spans the universe shown and is partitioned into three linguistic levels. RCS, PHS, RWS, and VS are scored on $[0, 1]$; RWS feeds only the baseline FIS₂ (three-objective comparator).

Variable (Universe)	Level	Shape	Parameters
VS $\in [0, 1]$	Low	trap	$[0, 0, 0.25, 0.5]$
	Medium	tri	$[0.25, 0.5, 0.75]$
	High	trap	$[0.5, 0.75, 1, 1]$
IDL $\in [0, 100]$	Minor	trap	$[0, 0, 25, 50]$
	Moderate	tri	$[25, 50, 75]$
	Severe	trap	$[50, 75, 100, 100]$
RTR $\in [0, 48]$ h	Short	trap	$[0, 0, 12, 24]$
	Moderate	tri	$[12, 24, 36]$
	Long	trap	$[24, 36, 48, 48]$
TD $\in [0, 180]$ min	Short	trap	$[0, 0, 30, 60]$
	Moderate	tri	$[30, 60, 90]$
	Long	trap	$[60, 90, 180, 180]$
RCS $\in [0, 1]$	Low (clear)	trap	$[0, 0, 0.2, 0.5]$
	Moderate	tri	$[0.2, 0.5, 0.8]$
	High (blocked)	trap	$[0.5, 0.8, 1, 1]$
PHS $\in [0, 1]$	Low	trap	$[0, 0, 0.2, 0.4]$
	Moderate	tri	$[0.2, 0.5, 0.8]$
	High	trap	$[0.6, 0.8, 1, 1]$
RWS $\in [0, 1]$	Low	trap	$[0, 0, 0.25, 0.5]$
	Moderate	tri	$[0.25, 0.5, 0.75]$
	High	trap	$[0.5, 0.75, 1, 1]$
COR $\in [0, 100]$	Low	trap	$[0, 0, 25, 50]$
	Moderate	tri	$[25, 50, 75]$
	High	trap	$[50, 75, 100, 100]$
RDR $\in [0, 100]$	Slow	trap	$[0, 0, 25, 50]$
	Moderate	tri	$[25, 50, 75]$
	Rapid	trap	$[50, 75, 100, 100]$

Table A6. Shared output universe on $[0, 100]$ (ULPP, TIL, TRD, RPD, CAIL); five levels, identical across all five FIS.

Level	Shape	Parameters
Very Low	trap	$[0, 0, 10, 30]$
Low	tri	$[10, 30, 50]$
Moderate	tri	$[30, 50, 70]$
High	tri	$[50, 70, 90]$
Very High	trap	$[70, 90, 100, 100]$

Appendix C. Statistical Test Outputs (Extended)

This appendix gives the complete pairwise Mann–Whitney U two-sided test outputs underlying the summary in Section 5.4. The Mann–Whitney U test is identical to the Wilcoxon rank-sum test, the significance test most multi-objective optimisation studies report; we use the U formulation throughout for consistency with the normality-gated test selection in Section 5.1. For each problem size and each of the five quality metrics (NNS, HV, SM, CPU time, and the MID diagnostic), we report the p -value of every algorithm pair (NSGA-II, NPGA, NSGA-III), over 30 repetitions per cell. Tables A7 and A8 cover the four-objective and the three-objective baseline formulations respectively. The values are computed directly from the released run manifest (experiments/results/h1_h2_h4/manifest.csv) and reproduce under the open-source code.

Table A7. Pairwise Mann–Whitney U two-sided p -values per (size, metric) on the **four-objective** formulation, 30 reps per cell. * marks $p < 0.05$. “—” denotes a degenerate comparison (both samples constant, e.g., NNS at the 100-individual ceiling).

Size	Pair	NNS	HV	SM	CPU	MID [†]
small	NSGA-II vs. NSGA-III	—	0.027 *	<0.001 *	<0.001 *	<0.001 *
	NSGA-II vs. NRGa	0.334	<0.001 *	<0.001 *	<0.001 *	<0.001 *
	NSGA-III vs. NRGa	0.334	0.006 *	0.066	<0.001 *	<0.001 *
medium	NSGA-II vs. NSGA-III	—	0.001 *	<0.001 *	<0.001 *	<0.001 *
	NSGA-II vs. NRGa	0.334	<0.001 *	<0.001 *	<0.001 *	<0.001 *
	NSGA-III vs. NRGa	0.334	<0.001 *	0.003 *	<0.001 *	0.112
large	NSGA-II vs. NSGA-III	—	<0.001 *	<0.001 *	<0.001 *	<0.001 *
	NSGA-II vs. NRGa	0.334	<0.001 *	<0.001 *	<0.001 *	<0.001 *
	NSGA-III vs. NRGa	0.334	<0.001 *	0.652	<0.001 *	0.181

[†] MID is a diagnostic, not a ranking metric (see Table 3); its column is placed last accordingly.

Table A8. Pairwise Mann–Whitney U two-sided p -values per (size, metric) on the **three-objective** baseline formulation, 30 reps per cell. * marks $p < 0.05$.

Size	Pair	NNS	HV	SM	CPU	MID [†]
small	NSGA-II vs. NSGA-III	0.334	0.348	<0.001 *	<0.001 *	0.044 *
	NSGA-II vs. NRGa	0.021 *	<0.001 *	<0.001 *	<0.001 *	<0.001 *
	NSGA-III vs. NRGa	0.006 *	<0.001 *	0.002 *	<0.001 *	0.158
medium	NSGA-II vs. NSGA-III	—	0.284	<0.001 *	<0.001 *	0.041 *
	NSGA-II vs. NRGa	0.082	<0.001 *	<0.001 *	<0.001 *	0.042 *
	NSGA-III vs. NRGa	0.082	<0.001 *	<0.001 *	<0.001 *	0.784
large	NSGA-II vs. NSGA-III	—	0.018 *	<0.001 *	<0.001 *	0.853
	NSGA-II vs. NRGa	0.022 *	<0.001 *	<0.001 *	<0.001 *	<0.001 *
	NSGA-III vs. NRGa	0.022 *	<0.001 *	0.007 *	<0.001 *	<0.001 *

[†] MID is a diagnostic, not a ranking metric (see Table 3); its column is placed last accordingly.

Appendix D. Reproducibility

Every result in this paper regenerates from the archived artifact (tag v0.2.0, Zenodo DOI 10.5281/zenodo.21083547) with its pinned dependency set (uv.lock). After uv sync, the three problem instances are materialised bit-exact from their seed via the released generator experiments/generate_instances.py (Section 5.1), and a single allocation is produced from the command line, for example:

```
pva allocate-people -people people.csv -centers centers.csv \
-solver nsga3 -seed 1007 -pop-size 100 -generations 200
```

The full experiment matrix—the 540 H1/H2/H4 runs, the 66 H3a rule deletions, and the 100 H3b weight samples—is orchestrated from the experiments/ drivers; each writes a run manifest under experiments/results/ (for example h1_h2_h4/manifest.csv) from which Tables 3, 4, A7 and A8 and Figures 1–5 are regenerated. On the single-machine setup of Section 5.1 the full matrix completes in about 13 min; a continuous-integration workflow re-runs a reduced smoke matrix on every commit, and a cross-environment check (experiments/results/repro/) confirms bit-for-bit identical fronts across machines.

References

1. CRED. 2023: *Disasters in Numbers*; CRED: Brussels, Belgium, 2024.
2. Wirtz, A.; Kron, W.; Löw, P.; Steuer, M. The Need for Data: Natural Disasters and the Challenges of Database Management. *Nat. Hazards* **2014**, *70*, 135–157. [CrossRef]
3. Avishan, F.; Elyasi, M.; Yanikoğlu, İ.; Ekici, A.; Özener, O.Ö. Humanitarian Relief Distribution Problem: An Adjustable Robust Optimization Approach. *Transp. Sci.* **2023**, *57*, 1096–1114. [CrossRef]
4. Krishna, R.N.; Ronan, K.; Spencer, C.; Alisic, E. The Lived Experience of Disadvantaged Communities Affected by the 2015 South Indian Floods: Implications for Disaster Risk Reduction Dialogue. *Int. J. Disaster Risk Reduct.* **2021**, *54*, 102046. [CrossRef]

5. Behl, A.; Dutta, P. Humanitarian Supply Chain Management: A Thematic Literature Review and Future Directions of Research. *Ann. Oper. Res.* **2019**, *283*, 1001–1044. [\[CrossRef\]](#)
6. Christopher, M.; Peck, H. Building the Resilient Supply Chain. *Int. J. Logist. Manag.* **2004**, *15*, 1–14. [\[CrossRef\]](#)
7. Tukamuhabwa, B.R.; Stevenson, M.; Busby, J.; Zorzini, M. Supply Chain Resilience: Definition, Review and Theoretical Foundations for Further Study. *Int. J. Prod. Res.* **2015**, *53*, 5592–5623. [\[CrossRef\]](#)
8. Bruneau, M.; Chang, S.E.; Eguchi, R.T.; Lee, G.C.; O'Rourke, T.D.; Reinhorn, A.M.; Shinozuka, M.; Tierney, K.; Wallace, W.A.; von Winterfeldt, D. A Framework to Quantitatively Assess and Enhance the Seismic Resilience of Communities. *Earthq. Spectra* **2003**, *19*, 733–752. [\[CrossRef\]](#)
9. Hosseini, S.; Barker, K.; Ramirez-Marquez, J.E. A Review of Definitions and Measures of System Resilience. *Reliab. Eng. Syst. Saf.* **2016**, *145*, 47–61. [\[CrossRef\]](#)
10. Rabiei, P.; Arias-Aranda, D. Introducing a Novel Multi-Objective Optimization Model for Vehicle Routing and Relief Supply Distribution in Post-Disaster Phase. In Proceedings of the 2021 6th International Conference on Transportation Information and Safety (ICTIS), Wuhan, China, 22–24 October 2021; pp. 1226–1243.
11. Rabiei, P.; Arias-Aranda, D.; Stantchev, V. Introducing a Novel Multi-Objective Optimization Model for Volunteer Assignment in the Post-Disaster Phase: Combining Fuzzy Inference Systems with NSGA-II and NREGA. *Expert Syst. Appl.* **2023**, *226*, 120142. [\[CrossRef\]](#)
12. Pettit, T.J.; Fiksel, J.; Croxton, K.L. Ensuring Supply Chain Resilience: Development of a Conceptual Framework. *J. Bus. Logist.* **2010**, *31*, 1–21. [\[CrossRef\]](#)
13. Pettit, T.J.; Croxton, K.L.; Fiksel, J. Ensuring Supply Chain Resilience: Development and Implementation of an Assessment Tool. *J. Bus. Logist.* **2013**, *34*, 46–76. [\[CrossRef\]](#)
14. Kamalahmadi, M.; Parast, M.M. A Review of the Literature on the Principles of Enterprise and Supply Chain Resilience: Major Findings and Directions for Future Research. *Int. J. Prod. Econ.* **2016**, *171*, 116–133. [\[CrossRef\]](#)
15. Wieland, A. Dancing the Supply Chain: Toward Transformative Supply Chain Management. *J. Supply Chain Manag.* **2021**, *57*, 58–73. [\[CrossRef\]](#)
16. Altay, N.; Green, W.G. OR/MS Research in Disaster Operations Management. *Eur. J. Oper. Res.* **2006**, *175*, 475–493. [\[CrossRef\]](#)
17. Galindo, G.; Batta, R. Review of Recent Developments in OR/MS Research in Disaster Operations Management. *Eur. J. Oper. Res.* **2013**, *230*, 201–211. [\[CrossRef\]](#)
18. Caunhye, A.M.; Nie, X.; Pokharel, S. Optimization Models in Emergency Logistics: A Literature Review. *Socio-Econ. Plan. Sci.* **2012**, *46*, 4–13. [\[CrossRef\]](#)
19. Anaya-Arenas, A.M.; Renaud, J.; Ruiz, A. Relief Distribution Networks: A Systematic Review. *Ann. Oper. Res.* **2014**, *223*, 53–79. [\[CrossRef\]](#)
20. Tzeng, G.H.; Cheng, H.J.; Huang, T.D. Multi-Objective Optimal Planning for Designing Relief Delivery Systems. *Transp. Res. Part E Logist. Transp. Rev.* **2007**, *43*, 673–686. [\[CrossRef\]](#)
21. Sheu, J.B. Challenges of Emergency Logistics Management. *Transp. Res. Part E Logist. Transp. Rev.* **2007**, *43*, 655–659. [\[CrossRef\]](#)
22. Abounacer, R.; Rekik, M.; Renaud, J. An Exact Solution Approach for Multi-Objective Location-Transportation Problem for Disaster Response. *Comput. Oper. Res.* **2014**, *41*, 83–93. [\[CrossRef\]](#)
23. Gama, M.; Santos, B.F.; Scaparra, M.P. A Multi-Period Shelter Location-Allocation Model with Evacuation Orders for Flood Disasters. *EURO J. Comput. Optim.* **2016**, *4*, 299–323. [\[CrossRef\]](#)
24. Zhao, L.; Li, H.; Sun, Y.; Huang, R.; Hu, Q.; Wang, J.; Gao, F. Planning Emergency Shelters for Urban Disaster Resilience: An Integrated Location-Allocation Modeling Approach. *Sustainability* **2017**, *9*, 2098. [\[CrossRef\]](#)
25. Balcik, B.; Beamon, B.M. Facility Location in Humanitarian Relief. *Int. J. Logist. Res. Appl.* **2008**, *11*, 101–121. [\[CrossRef\]](#)
26. Pedraza-Martinez, A.J.; Stapleton, O.; Van Wassenhove, L.N. Field Vehicle Fleet Management in Humanitarian Operations: A Case-Based Approach. *J. Oper. Manag.* **2011**, *29*, 404–421. [\[CrossRef\]](#)
27. Holguín-Veras, J.; Jaller, M.; Van Wassenhove, L.N.; Pérez, N.; Wachtendorf, T. On the Unique Features of Post-Disaster Humanitarian Logistics. *J. Oper. Manag.* **2012**, *30*, 494–506. [\[CrossRef\]](#)
28. Sahebjamnia, N.; Torabi, S.A.; Mansouri, S.A. A Hybrid Decision Support System for Managing Humanitarian Relief Chains. *Decis. Support Syst.* **2017**, *95*, 12–26. [\[CrossRef\]](#)
29. Goldschmidt, K.H.; Kumar, S. Humanitarian Operations and Crisis/Disaster Management: A Retrospective Review of the Literature and Framework for Development. *Int. J. Disaster Risk Reduct.* **2016**, *20*, 1–13. [\[CrossRef\]](#)
30. Yazdani, M.; Haghani, M. A Dynamic Emergency Planning System for Relocating Vulnerable People to Safe Shelters in Response to Heat Waves. *Expert Syst. Appl.* **2023**, *228*, 120224. [\[CrossRef\]](#)
31. Soghrati Ghasbeh, S.; Pourmohammadzia, N.; Rabbani, M. Equitable Post-Disaster Relief Distribution: A Robust Multi-Objective Multi-Stage Optimization Approach. *J. Humanit. Logist. Supply Chain Manag.* **2022**, *12*, 618–651. [\[CrossRef\]](#)
32. Yazdani, M.; Matinrad, N.; Haghani, M. Optimizing Community-Led Post-Disaster Recovery in Flood-Affected Areas: A Resilience Model for Volunteer Deployment. *Reliab. Eng. Syst. Saf.* **2026**, *272*, 112373. [\[CrossRef\]](#)

33. Mamdani, E.H.; Assilian, S. An Experiment in Linguistic Synthesis with a Fuzzy Logic Controller. *Int. J. Man-Mach. Stud.* **1975**, *7*, 1–13. [\[CrossRef\]](#)
34. Bahmani, V.; Adibi, M.A.; Mehdizadeh, E. A Fuzzy Multi-Objective Routing Model for Disaster Relief Distribution Considering Multi-Period Demands and Damaged Network. *Soft Comput.* **2018**, *22*, 6927–6950. [\[CrossRef\]](#)
35. Moreno, A.; Alem, D.; Ferreira, D.; Clark, A. An Effective Two-Stage Stochastic Multi-Trip Location-Transportation Model with Social Concerns in Relief Supply Chains. *Eur. J. Oper. Res.* **2018**, *269*, 1050–1071. [\[CrossRef\]](#)
36. Deb, K.; Pratap, A.; Agarwal, S.; Meyarivan, T. A Fast and Elitist Multiobjective Genetic Algorithm: NSGA-II. *IEEE Trans. Evol. Comput.* **2002**, *6*, 182–197. [\[CrossRef\]](#)
37. Ishibuchi, H.; Tsukamoto, N.; Nojima, Y. Evolutionary Many-Objective Optimization: A Short Review. In Proceedings of the IEEE Congress on Evolutionary Computation (CEC), Hong Kong, China, 1–6 June 2008; pp. 2419–2426. [\[CrossRef\]](#)
38. Li, B.; Li, J.; Tang, K.; Yao, X. Many-Objective Evolutionary Algorithms: A Survey. *ACM Comput. Surv.* **2015**, *48*, 1–35. [\[CrossRef\]](#) [\[PubMed\]](#)
39. Coello Coello, C.A.; Lamont, G.B.; Van Veldhuizen, D.A. *Evolutionary Algorithms for Solving Multi-Objective Problems*, 2nd ed.; Springer: Berlin/Heidelberg, Germany, 2007. [\[CrossRef\]](#)
40. Deb, K.; Jain, H. An Evolutionary Many-Objective Optimization Algorithm Using Reference-Point-Based Nondominated Sorting Approach, Part I: Solving Problems with Box Constraints. *IEEE Trans. Evol. Comput.* **2014**, *18*, 577–601. [\[CrossRef\]](#)
41. Zhang, Q.; Li, H. MOEA/D: A Multiobjective Evolutionary Algorithm Based on Decomposition. *IEEE Trans. Evol. Comput.* **2007**, *11*, 712–731. [\[CrossRef\]](#)
42. Cheng, R.; Jin, Y.; Olhofer, M.; Sendhoff, B. A Reference Vector Guided Evolutionary Algorithm for Many-Objective Optimization. *IEEE Trans. Evol. Comput.* **2016**, *20*, 773–791. [\[CrossRef\]](#)
43. Zheng, W.; Doerr, B. Runtime Analysis for the NSGA-II: Proving, Quantifying, and Explaining the Inefficiency for Many Objectives. *IEEE Trans. Evol. Comput.* **2024**, *28*, 1442–1454. [\[CrossRef\]](#)
44. Al Jadaan, O.; Rajamani, L.; Rao, C.R. Non-Dominated Ranked Genetic Algorithm for Solving Multi-Objective Optimization Problems: NRGGA. *J. Theor. Appl. Inf. Technol.* **2008**, *4*, 60–67. [\[CrossRef\]](#)

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.