DOI: http://dx.doi.org/10.26483/ijarcs.v17i4.7555

ISSN No. 0976-5697

Volume 17, No. 4, July-August 2026
International Journal of Advanced Research in Computer Science
RESEARCH PAPER
Available Online at www.ijarcs.info

WEIGHT SENSITIVITY ANALYSIS IS NOT SUFFICIENT: FOUR FURTHER ANALYST-SUPPLIED INPUTS IN MULTI-CRITERIA DECISION MAKING FOR DISASTER COMMUNICATION SELECTION

Tadashi Shiroma
Faculty of Engineering
University of the Ryukyus
1 Senbaru, Nishihara, Okinawa 903-0213, Japan

Abstract: Applied multi-criteria decision making (MCDM) studies almost universally report a sensitivity analysis on the criterion weights, and rarely on the other quantities the analyst supplies. This paper contributes a reproducible computational sensitivity-audit protocol and demonstrates it on a disaster communication selection problem used as a testbed: the contribution is the protocol and its findings, not a recommendation of communication media. With fifteen media, six criteria, a non-compensatory mission screen and three aggregations under three hazard weight templates, we perturb five analyst-supplied inputs and evaluate each against the same binary endpoint, whether the recommended leading set changes. Switching the method-plus-preprocessing specification changes it in 67% of method-pair comparisons, of which normalisation alone accounts for a large share. A one-in-twenty disagreement over the performance scores changes it in 28% of draws, against 25% when every weight is randomised by up to 20% in either direction. The screening thresholds and the treatment of the ordinal rating scale as cardinal also move the recommendation. We deliberately do not rank these influences: the perturbation magnitudes are not calibrated against one another, and the ordering of weights against thresholds reverses with the perturbation size, so any ranking would be an artefact of the magnitudes chosen. What the results support is the weaker but still consequential claim that weight sensitivity alone is not an adequate robustness check. A cell-level analysis localises the sensitivity usefully: 53 of 90 score cells cannot alter any recommendation, which makes targeted re-elicitation tractable. We also report descriptively a comparison against media deployed in two Japanese disasters, and explain why the sample cannot support inferential claims.

Keywords: multi-criteria decision making; sensitivity analysis; rank reversal; aggregation method; disaster communication; non-compensatory screening

I. INTRODUCTION

An applied MCDM study asks its analyst for four distinct things: a set of alternatives, a performance score for every alternative on every criterion, a weight for every criterion, and a rule for aggregating the weighted scores into a ranking. All four are judgements. Yet the robustness section of a typical paper interrogates exactly one of them. Reviews of the field record weight sensitivity as the near-universal robustness check [1,2], and we are not aware of a study in the disaster-management MCDM literature that reports the sensitivity of its conclusions to the choice of aggregation method or to the performance scores with comparable rigour. We did not conduct a systematic search, so this is an impression from the reviews cited above rather than an established absence.

This paper asks whether that convention is an adequate robustness check, and contributes a reproducible computational protocol for answering that question on any MCDM model. We state at the outset what kind of paper this is, because it determines how the results should be read. The disaster communication problem below is used as a computational testbed: it is realistic enough that getting it wrong would matter, and small enough to perturb exhaustively. The primary contribution is the sensitivity-audit protocol and what it reveals, not a definitive recommendation of which communication media to deploy. Readers looking for the latter should treat Section VII-B as exploratory.

To answer the question we need such a testbed. We use the selection of communication infrastructure for disaster response. The problem is genuinely multi-criteria, since physical resilience, power autonomy, deployability, congestion tolerance, functionality and throughput trade off against each other, and it matters: in the 2024 Noto Peninsula earthquake (Mj 7.6, 1 January 2024) up to 839 mobile base stations were off the air, 799 of them in Ishikawa Prefecture, and the peninsular geography blocked the restoration convoys that would normally have arrived within hours [3].

Building the testbed surfaced a problem of its own, which we report because it motivates the screening stage and because it is instructive. Applying a compensatory aggregation to the full set of fifteen media ranks EnOcean, an energy-harvesting building-automation protocol that cannot carry a voice call or a text message, first under two of the three hazard templates. The mechanism is transparent: physical resilience and power autonomy together carry 55–60% of the weight immediately after a disaster, EnOcean scores 5 on both, and its categorical inability to carry human communication registers only as a four-point deduction on a criterion weighted 0.10. We therefore introduce a non-compensatory mission screen before aggregation. This is a standard conjunctive filter rather than a novel method, and we present it as part of the testbed, not as the paper's contribution.

The contribution is the sensitivity audit, whose protocol is set out in Fig. 1. We perturb each analyst-supplied input and evaluate every result against the same binary endpoint: whether the recommended leading set changes. The denominator differs between inputs and is stated with each result. We stress at the outset what this does and does not license. A common endpoint does not make the inputs commensurable: a ±20% weight perturbation and a one-in-five score perturbation are not equally plausible, and we have no empirical basis for calibrating them

against each other. We therefore report curves and refuse to rank the inputs. The findings are:

(i) Switching aggregation method changes the leading set in 67% of method-pair comparisons, and TOPSIS disagrees with PROMETHEE II in 100% of them. Decomposing this, changing only the normalisation within SAW, from raw scores to vector normalisation, already accounts for 5 of 9 cases, so "method choice" as usually reported bundles normalisation together with the aggregation operator.

(ii) Score perturbation produces comparable movement: one cell in twenty shifted by a rubric level changes the leading set in 28% of draws, one cell in five in 61%.

(iii) Weight perturbation produces 25% at $\delta = 0.2$ and 55% at $\delta = 1.0$. Whether this places weights above or below the screening thresholds (30%) depends on which δ is chosen, and that is our reason for not ranking the inputs.

(iv) Treating the ordinal 1–5 scale as cardinal is not inert. A monotone rescaling of all six criteria changes the leading set in 1 of 9 cases under a steep convex transform and 6 of 9 under a logarithmic one. Only the PROMETHEE preference threshold proved genuinely inert over its plausible range.

(v) Sensitivity is extremely localised. Of 90 score cells, 53 cannot change any recommendation under a ±1 shift, while the physical resilience of the drone-borne relay changes ten of eighteen mission–scenario perturbations on its own. This is the most directly actionable result: it identifies the fifteen judgements worth sending for expert re-elicitation.

To be explicit about the scope of these results, we separate what this paper claims from what it does not.

WHAT WE CLAIM. (a) A single MCDM model can be audited against five analyst-supplied inputs against one common binary endpoint, and the protocol for doing so is computationally inexpensive and reproducible. (b) On this testbed, four inputs other than the criterion weights each move the recommended leading set substantially under the perturbations we examined. (c) It follows that a robustness section reporting weight sensitivity alone can miss substantial instability. (d) Sensitivity is highly localised at the cell level, which makes targeted re-elicitation practical.

WHAT WE DO NOT CLAIM. (a) We do not rank the five inputs by influence. All results share a common endpoint, but the perturbation magnitudes are not calibrated against one another and we have no empirical basis for declaring them equally plausible; the relative position of weights and thresholds reverses between $\delta = 0.2$ and $\delta = 1.0$, so any ranking would be an artefact of the magnitudes chosen. (b) We do not claim the findings generalise beyond this testbed. (c) We do not claim the framework predicts historical deployment better than the model it replaces; Section VI shows it does not, detectably. (d) We do not offer the media rankings as procurement advice.

Section II reviews the relevant literature. Section III defines the testbed. Section IV documents the degeneracy that motivates screening. Section V is the sensitivity audit. Section VI reports the descriptive retrospective comparison and explains its limits. Sections VII and VIII discuss and conclude.

II. Related work

A. Sensitivity analysis practice in applied MCDM

Systematic and bibliometric reviews of MCDM in disaster management [1,2,4] describe a field dominated by AHP, ANP, TOPSIS, ELECTRE, PROMETHEE and their hybrids, applied to decision objects such as shelter siting [5]. Where robustness is addressed, it is almost always weight sensitivity: one-at-a-time variation of a weight, or a Monte-Carlo sweep of the weight vector.

The methodological literature has long known that this is not the only vulnerability. Rank reversal under alternative-set changes is documented for several methods, including ELECTRE, AHP and TOPSIS [6,7,8]; the choice between normalisation schemes is known to affect TOPSIS rankings materially [9]; and comparative studies applying several methods to one dataset routinely find divergent rankings [10,11]. Zanakis et al. [10] and Zavadskas et al. [11] are explicit that method choice is consequential. What is missing, and what this paper supplies, is a study that places method choice, score uncertainty, weight uncertainty, threshold choice and scale-level assumptions on a single comparable axis for one decision problem, so that a practitioner can see which sources of instability are missed by weight sensitivity alone.

B. Compensation and non-compensatory screening

That a weighted sum permits a large surplus on one criterion to offset a disqualifying deficit on another is the classical objection to compensatory aggregation, articulated by Roy in the ELECTRE veto threshold [12], by Keeney and Raiffa in the context of preferential independence [13], by Bouyssou in a series of analyses of compensation in outranking [14], and by Munda in the ecological-economics setting [15]. Conjunctive screening, which requires a minimum on designated criteria before any trade-off is considered, is equally classical [16].

We adopt conjunctive screening rather than an ELECTRE veto for a specific reason. A veto threshold suppresses an outranking relation between a pair of alternatives; the alternative remains in the set and can still be ranked. Our requirement is stronger and is a statement about the problem rather than about preferences: a medium that cannot carry a voice call is not a candidate for a safety-confirmation mission at any score on any other criterion, and should not appear in the output at all. A veto formulation would reach a similar ranking by a less direct route while leaving infeasible media in the presented list. We make no claim of novelty for the screen itself, and Section VII notes that an ELECTRE-with-veto formulation of the same testbed would be a worthwhile comparison.

C. Disaster communication

The technical literature identifies recurring obstacles [17,18]: deployment is harder when the network is partially rather than wholly destroyed; human communication behaviour is rarely modelled in system design; and hierarchical command structures must interoperate with improvised organisations. Low-Earth-orbit constellations have changed the option set, offering 20–40 ms latency against 500–700 ms for geostationary systems, 50–500 Mbps throughput, and setup in minutes without terrestrial infrastructure [19–21], at the cost of sensitivity to space weather and slightly lower long-term availability [22]. Both KDDI and SoftBank used Starlink terminals as alternative backhaul during the Noto response [23,24].

III. The testbed

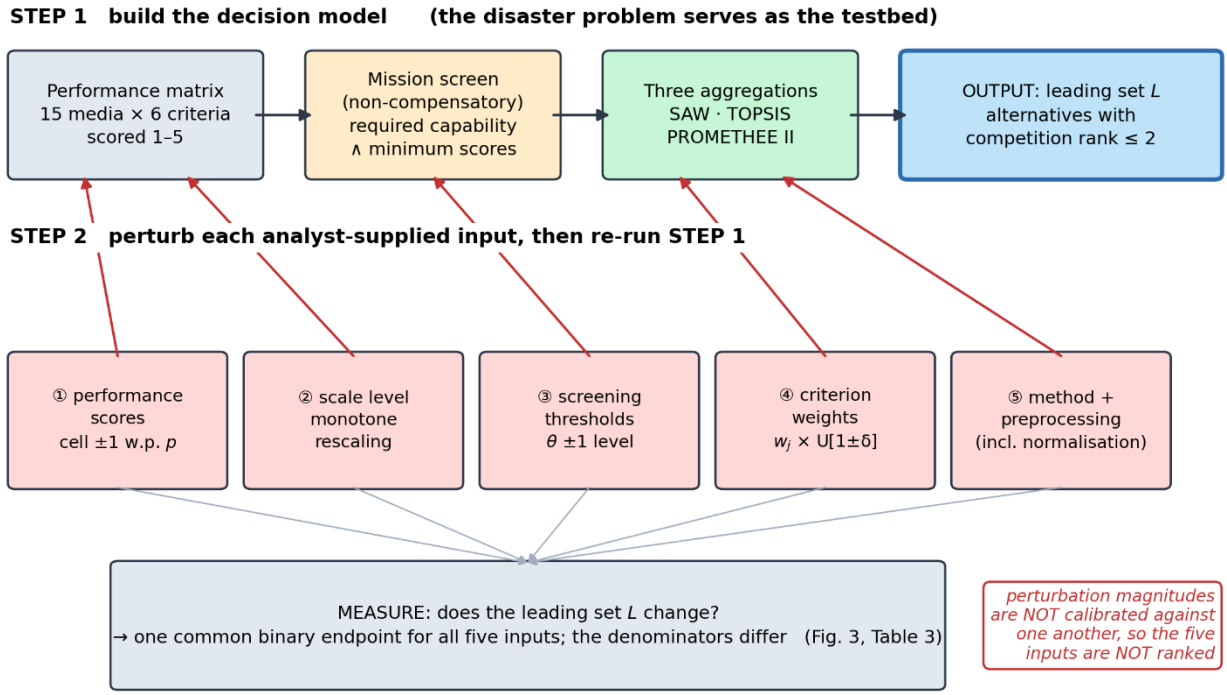

Fig. 1. The audit protocol. STEP 1 builds a conventional MCDM model and reports a leading set. STEP 2 perturbs each of the five inputs the analyst supplies, re-runs STEP 1, and records how often the leading set changes. Because all five perturbations are measured in the same unit they can be displayed together, but their magnitudes are not calibrated against one another, so the inputs are not ranked.

A. Alternatives, criteria and capability profile

The alternative set comprises fifteen bearers, chosen to span the phases of the disaster cycle and the operational purposes of alerting, safety confirmation, coordination and monitoring. The set contains bearers only. Application-layer systems, such as a disaster information sharing platform of the SIP4D type, are deliberately excluded: their physical resilience and power autonomy are properties inherited from whatever bearer carries them rather than attributes of the platform itself, so ranking such a platform against Starlink within a single set is not well defined. Selecting an application layer is a separate decision problem, and mixing the two layers in one alternative set is a formulation error we prefer to avoid rather than to caveat.

The six criteria are C1 physical resilience, C2 power autonomy, C3 usability and penetration, C4 risk tolerance, C5 functionality and C6 throughput, all benefit-type and scored 1–5 against explicit anchors. Separately we record a binary capability profile with four components: broadcast, two-way person-to-person, rich media over IP, and unattended telemetry. It records what a medium can do at all rather than how well. Table I gives the alternative set, the capability profile and the complete performance matrix in numeric form.

Two rubric anchors are worth stating because they carry most of the discriminating power. C5 level 5 is full multimedia, level 3 is standard voice plus short messaging, level 1 is a single state telegram or one-way broadcast. C6 level 5 is ≥100 Mbps, level 3 is 1–10 Mbps, level 1 is below 100 kbps. Because the C6 anchors are logarithmic in bandwidth, treating the level index as cardinal is an assumption rather than a fact; Section V-D tests it.

B. Construction of the performance matrix

The matrix was constructed for this study from the rubric anchors and the published specifications of each medium; it is not an audit of a previously published matrix. Every cell was assigned so as to be consistent both with the rubric anchors and with the medium's published specification, and no cell was allowed to imply a capability that the specification excludes. The matrix should be read as constructed here and defended on its own terms; Section V-B quantifies how much any conclusion could depend on the resulting values.

C. Mission screening

Four missions are distinguished: M1 mass alerting, M2 safety confirmation, M3 command and coordination, M4 situational and infrastructure monitoring. Each defines a feasible subset by conjunction of a required capability and minimum scores on the criteria the mission cannot trade away:

$$A_m = \{ a \in A : cap_k(a) = 1 \text{ for all } k \in K_m, \text{ and } x_{aj} \geq \theta_{mj} \text{ for all } j \in J_m \} \quad (1)$$

M2 and M3 share the two-way capability requirement and are separated by their score thresholds; the four missions are collectively exhaustive over the purposes documented in the case material but are not mutually exclusive in capability, and we do not claim that they are. Thresholds are read from the rubric anchors at the lowest level satisfying the mission definition: C5 ≥ 3 for M2 and M3 is where the rubric first admits voice plus short messaging; C4 ≥ 3 for M3 is where data first passes during voice-traffic restriction; C3 ≥ 2 for M2 excludes media requiring a licence the public does not hold; C2 ≥ 3 for M4 is where multi-day operation on portable power first becomes possible. Table II gives the resulting feasible sets.

TABLE II. MISSION SCREEN

Mission	Required capability	Minimum scores	|Am|	Feasible alternatives
M1 Mass alerting	broadcast	—	2	A1, A6
M2 Safety confirmation	two-way P2P	C5 ≥ 3, C3 ≥ 2	6	A1, A3, A4, A5, A7, A9
M3 Command & coordination	two-way P2P	C5 ≥ 3, C4 ≥ 3	5	A4, A5, A7, A8, A9
M4 Situational monitoring	unattended telemetry	C2 ≥ 3	6	A9, A11, A12, A13, A14, A15

Mission screen. Thresholds encode necessity only; all preference information is deferred to the ranking stage.

D. Ranking, weights and tie handling

Three aggregations are computed on each feasible set: simple additive weighting (SAW), $S_a = \Sigma_j w_j x_{aj}$; TOPSIS [25] with vector normalisation and ideal points defined on the feasible set; and PROMETHEE II [26,27] with a V-shape preference function, indifference threshold q = 0 and preference threshold p = 2. A consensus position is the competition rank of the mean of the three competition ranks.

We use competition ranking throughout: tied alternatives share the better rank, and the next rank is skipped. Ties are not rare here, because with integer scores and six criteria aggregate scores collide frequently, and the choice of tie rule is itself an analyst decision, so we state it explicitly and report tied positions as tied rather than breaking them arbitrarily. Scores are compared to a tolerance of 10^{-9}.

The output we track throughout is the leading set L, defined as the unordered set of alternatives whose competition rank is 2 or better. Because ties share the better rank, |L| is usually but not always two: across the nine mission–scenario cases it is two in eight and three in one (M4 under the pluvial template, where the drone-borne relay and WirelessHART tie for second). We therefore write "leading set" rather than "leading pair", and never truncate it to two members. Two perturbed outcomes are counted as different when the leading sets differ as sets; ordering within the set is not compared, since Section V-A shows that ordering is not stable across methods in the first place.

The consensus position is the competition rank of the mean of the three competition ranks, a Borda-type aggregation. This choice is not derived from any axiom and we treat it as a further analyst decision: replacing the mean by the median changes the leading set in 1 of 9 cases, and by a Copeland score in 1 of 9. The union of the three methods' individual leading sets has exactly three members in every case and their intersection exactly one, which is arguably the more honest summary and is what Section VII-A recommends reporting.

We do not describe SAW as AHP. The weights below are specified directly rather than elicited by pairwise comparison, so no analytic hierarchy is constructed. For completeness we note that reconstructing an implied pairwise matrix from each weight vector and computing its consistency ratio [28] returns CR ≤ 0.008 in all three cases, but this is a near-tautology, since a matrix built from a consistent vector is consistent, and we draw no validity conclusion from it. Weights for the immediate-response phase are defined by hazard type: wind (0.30, 0.25, 0.20, 0.10, 0.10, 0.05), seismic (0.35, 0.25, 0.20, 0.05, 0.10, 0.05) and pluvial (0.30, 0.25, 0.20, 0.05, 0.15, 0.05), the differences reflecting that seismic damage is direct structural failure, wind events produce dispersed simultaneous demand, and pluvial events isolate settlements that must be assessed remotely.

IV. THE DEGENERACY THAT MOTIVATES SCREENING

Fig. 2 shows the unscreened SAW ranking under the seismic template. EnOcean ranks first at 3.95 and WirelessHART third, tied with public-safety dedicated radio, at 3.70; four of the ten best-placed media cannot carry human communication in any form (five under the pluvial template). Because a categorical incapacity enters only through C5, weighted 0.10, no compensatory operator applied to this alternative set can avoid a variant of the result: the defect is in the problem formulation, not in the aggregation.

The strongest form of the degeneracy is, however, fragile, and we report this because it bears on how much weight the

TABLE I. ALTERNATIVE SET, BINARY CAPABILITY PROFILE AND COMPLETE PERFORMANCE MATRIX

#	Alternative	Bcast	2-way	IP	Telem.	C1	C2	C3	C4	C5	C6
A1	Cellular network (voice/SMS/data)	✓	✓	✓	–	1	2	5	1	5	4
A2	Disaster message dial (171)	–	✓	–	–	2	3	4	4	2	1
A3	Public payphone (analogue)	–	✓	–	–	1	2	4	1	3	1
A4	Satellite phone (MSS)	–	✓	–	–	5	3	2	4	3	2
A5	Starlink (LEO broadband)	–	✓	✓	✓	5	2	3	4	5	5
A6	Municipal disaster radio	✓	–	–	–	5	3	2	3	1	1
A7	Public-safety dedicated radio	–	✓	–	–	5	3	3	5	3	1
A8	Amateur radio	–	✓	–	–	5	2	1	5	3	1
A9	Drone-borne relay	–	✓	✓	✓	4	4	2	4	4	3
A10	IoT sensor network (generic)	–	–	–	✓	4	1	1	2	2	1
A11	LPWAN (LoRa / NB-IoT)	–	–	–	✓	3	4	3	3	2	1
A12	ZigBee (IEEE 802.15.4)	–	–	–	✓	3	4	3	3	2	1
A13	Sub-GHz LPWA (920 MHz)	–	–	–	✓	2	5	3	3	2	1
A14	WirelessHART	–	–	–	✓	4	5	3	4	2	1
A15	EnOcean	–	–	–	✓	5	5	3	4	1	1

Alternative set, binary capability profile and complete performance matrix. Cellular broadcast refers to ETWS / Area Mail. The matrix is reproduced numerically here, rather than only as a figure, so that the analysis can be replicated without reading values off a plot.

example can carry. EnOcean's leading position rests substantially on a single cell. Lowering its physical resilience from 5 to 4, which is arguable for an indoor building-automation component, moves it from first to fourth under the wind and seismic templates (from 3.90 and 3.95 to 3.60 in both, behind Starlink at 3.75 and 3.80 and behind public-safety radio and WirelessHART tied at 3.70), and the leader becomes Starlink. Under the pluvial template EnOcean is second rather than first even before the change, and fifth after it. What survives the change is the general pattern rather than the specific result: four to five incapable media remain among the ten best-placed under every template. It is that pattern, rather than EnOcean's first place, that motivates the screen.

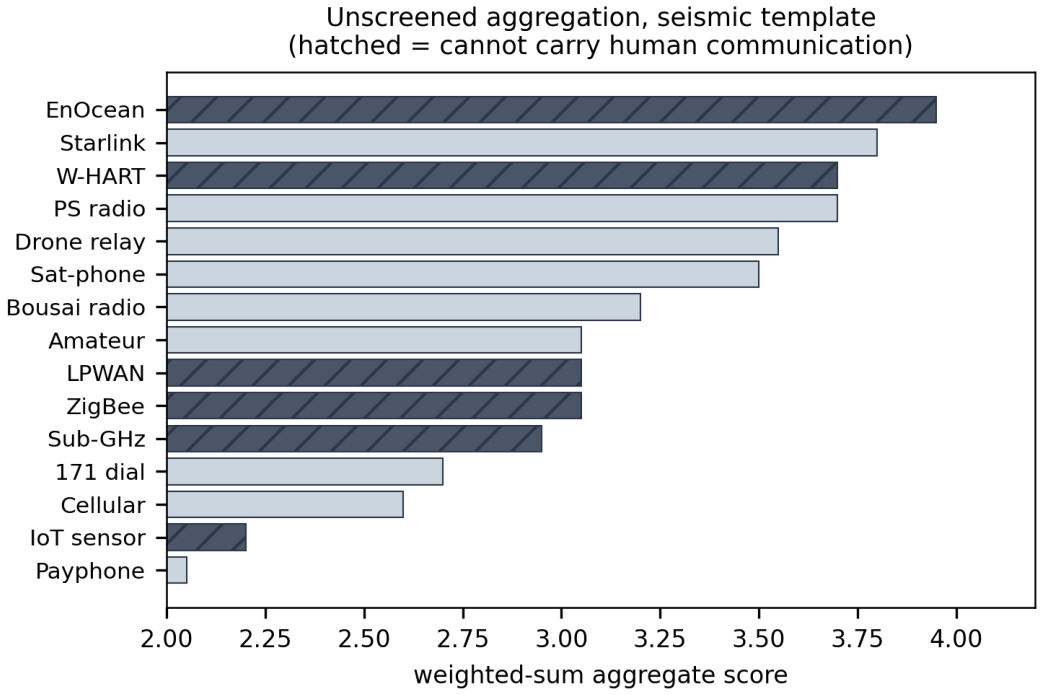

Fig. 2. Unscreened SAW ranking under the seismic template. Hatched bars mark media that cannot carry human communication.

V. THE SENSITIVITY AUDIT

Every result in this section is evaluated against one binary endpoint: whether the leading set changes. The denominator is not the same for every input and is stated with each result. Monte-Carlo entries are proportions of draws pooled over the nine mission–scenario cases; the method comparison is over method-pair by case comparisons; the threshold comparison is over threshold-move by scenario comparisons. Nine mission–scenario cases have three or more feasible alternatives and are used throughout; M1 with two alternatives is excluded because its leading set is the whole feasible set. Monte-Carlo estimates use 3,000 draws per case, generated with NumPy's default_rng (PCG64) under seed 21 throughout; the Monte-Carlo error on

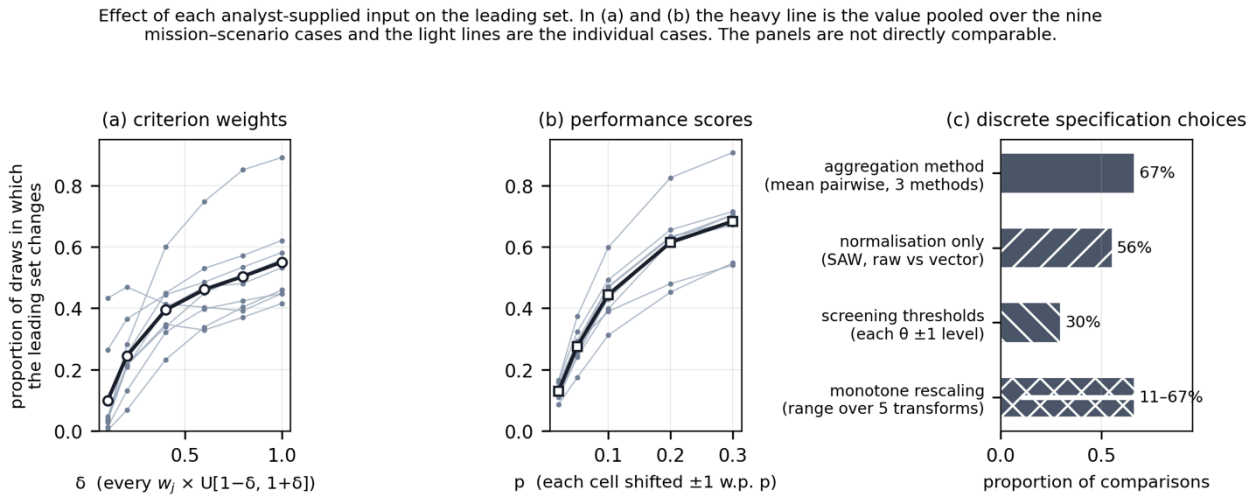

Fig. 3. Effect of each analyst-supplied input on the leading set. In (a) and (b) the heavy line is the proportion pooled over the nine mission–scenario cases and the light lines are the nine cases individually; the spread between them is dispersion in the quantity itself, not sampling error, and no inferential interval is attached, for the reason given at the start of Section V. (c) reports discrete specification choices. The panels share a vertical axis, but their denominators differ and the perturbation magnitudes are not calibrated against one another, so they must not be read as a ranking.

any single case is about ±0.017 at 95%, but the pooled figures reported below average over nine cases that share one performance matrix and three similar weight vectors and are therefore not independent. We report them as descriptive of this testbed and attach no interval to them. For the same reason we write "proportion under the specified perturbation distribution" rather than "probability".

A common endpoint does not make the inputs comparable. The perturbations below differ in kind: a discrete switch between three methods, a simultaneous ±1 shift of an expected 18 of 90 cells, a ±100% randomisation of every weight, a one-level move of a threshold. We have no empirical basis on which to declare them equally plausible. Fig. 3 therefore presents them in three panels rather than on one axis, and we do not rank them. That the ranking would be arbitrary is easy to demonstrate: weights at $\delta = 0.2$ give 25% and thresholds give 30%, so weights rank below thresholds; weights at $\delta = 1.0$ give 55%, so they rank above. Nothing in the data selects between these.

A. Aggregation method

Taking each pair of methods in turn and asking how often they disagree about the leading set gives: SAW against PROMETHEE II, 11%; SAW against TOPSIS, 89%; PROMETHEE II against TOPSIS, 100%. Pooled over the three pairs and the nine cases, 18 of 27 method-pair by case comparisons disagree, that is 67%. In not one of the nine mission–scenario pairs do all three methods agree on the leading set.

The SAW–PROMETHEE agreement is unsurprising and should not be read as corroboration: with a V-shape preference function and no alternative reaching the preference threshold on most criteria, PROMETHEE's net flow is close to a linear function of score differences, as is SAW. The informative comparison is against TOPSIS, whose vector normalisation and distance-to-ideal construction reward alternatives balanced across criteria over alternatives extreme on the heavily weighted ones. The two constructions encode genuinely different notions of what a good compromise is, and on this problem they almost never agree at the top.

The label "aggregation method" is, however, too coarse for what is being measured, and we decompose it. SAW applied to raw scores and SAW applied to vector-normalised scores, which share an aggregation operator but differ in normalisation, already give different leading sets in 5 of 9 cases; sum normalisation likewise gives 5 of 9, min–max 3 of 9, and max normalisation 1 of 9. Holding normalisation fixed at vector and changing only the operator, SAW against TOPSIS gives 6 of 9. Normalisation and operator therefore contribute comparably, and the 67% figure should be read as the effect of the whole method-plus-preprocessing specification, not of the aggregation operator alone. Reported this way it is arguably more troubling, since normalisation is a choice most applied papers do not report at all.

The consequence for the reporting convention is that a study which runs one method on one normalisation and then perturbs its weights is measuring variation within a specification whose own selection contributes variation of a similar magnitude.

B. Performance scores

Lacking a second rater, we simulate one: each cell is independently shifted by one rubric level, up or down with equal probability, with probability p, and the screen is re-run on the perturbed matrix so that feasibility as well as ordering can change. This represents a second analyst who reads the rubric differently on a proportion p of cells. Perturbed scores were clipped to the rubric range [1, 5], so cells already at 1 or 5 move in one direction only and are correspondingly less mobile than interior cells.

At $p = 0.02$ the leading set changes in 13% of draws; at $p = 0.05$, 28%; at $p = 0.10$, 44%; at $p = 0.20$, 61%; at $p = 0.30$, 68%. A disagreement rate of one cell in five thus reproduces the instability caused by switching aggregation method. Because the perturbation is applied before screening, part of this is feasibility churn rather than reordering; repeating the exercise with the feasible set held fixed changes the figures by under three percentage points, so the effect is in the ordering.

C. Weights

Each weight is multiplied by an independent uniform factor on $[1-\delta, 1+\delta]$ and the vector renormalised. At $\delta = 0.1$ the leading set changes in 10% of draws; at $\delta = 0.2$, 25%; at $\delta = 0.4$, 40%; at $\delta = 0.6$, 46%; at $\delta = 0.8$, 50%; and at $\delta = 1.0$, with every weight randomised by up to ±100%, 55%. No author would describe the last as a conservative check. The curve is concave and does not reach the 67% produced by a change of aggregation method anywhere in the plausible range.

Comparing the curves gives the exchange rate. A one-in-twenty scoring disagreement ($p = 0.05$, 28% instability) requires roughly $\delta = 0.25$ to match. A one-in-ten scoring disagreement (44%) requires roughly $\delta = 0.55$. These comparisons show that modest score disagreement can produce changes comparable to substantial weight perturbation, but they do not establish a general ordering of influence: whether the weight curve lies above or below any given discrete choice depends on where it is read.

D. Scale level and preference parameters

The rubric anchors are not equally spaced, the C6 anchors being logarithmic in bandwidth, so treating the level index as cardinal is an assumption. We test it in two ways, and the contrast between them is instructive. A single-criterion test is too weak to be informative: replacing only C6 by the normalised log of its band midpoint changes no leading set, but C6 carries a weight of 0.05 in every template and its log-midpoint transform is close to linear in the level index, so the test is near-vacuous.

Applying a monotone rescaling to all six criteria and renormalising each to [1, 5] gives a very different picture. A steep convex transform (x^3) changes the leading set in 1 of 9 cases, an exponential transform (2^x) in 1 of 9, a convex transform (x^2) in 2 of 9, a concave square-root transform in 4 of 9, and a logarithmic transform in 6 of 9. The cardinality assumption is therefore not inert, and concave transforms matter most, since they compress the top of the scale and so reduce the advantage of a 5 over a 4. We report the single-criterion result alongside this one because the contrast is the point: a sensitivity test confined to one lightly weighted criterion can pass while the assumption it purports to test does not hold.

Varying the PROMETHEE preference threshold p over {1, 2, 3, 4} changes no leading set in any case. This is the one input we tested that was genuinely inert over its plausible range.

E. Screening thresholds

Each threshold was moved by one rubric level in each direction, the screen re-run and the leading set recomputed. Three of the ten perturbations change the leading set. Because each perturbation was evaluated under all three hazard templates, that is 9 of 30 threshold-move by scenario comparisons, or 30%. They do so by two different mechanisms. Tightening C5 to ≥ 4 for M2 and for M3 removes leading alternatives from the feasible set, shrinking it to 3 and 2 members respectively. Loosening C2 to ≥ 2 for M4 has the opposite mechanism: it admits Starlink, which carries the unattended-telemetry capability but scores 2 on power autonomy, and Starlink then

enters the leading set in all three scenarios. Tightening C2 to ≥ 4 changes nothing, because no currently feasible telemetry alternative has C2 = 3. The remaining seven perturbations leave the leading set intact while varying the feasible set size between 4 and 7. Thresholds therefore govern how many alternatives are carried forward more often than which of them leads, but they can do either, and the direction of the effect is not predictable from the direction of the move.

F. Cell criticality

Aggregate perturbation conceals where sensitivity lives. Shifting each of the 90 cells by ±1 in turn, and counting how many of the eighteen mission–scenario perturbations change their leading set, gives Fig. 4. Fifty-three cells (59%) cannot change any recommendation. The influence concentrates in the rows of the drone-borne relay and Starlink, media that sit near the top of several feasible sets and are therefore contestable, and specifically in C1 of the drone-borne relay, which alone changes ten of eighteen.

This is a practically useful diagnostic. It tells an analyst which handful of judgements to send for expert review, and it tells a reader which cells to argue about. It also reframes the demand for multi-rater scoring: full re-elicitation of a 90-cell matrix is expensive, whereas re-elicitation of the fifteen cells with criticality of 4 or more is not.

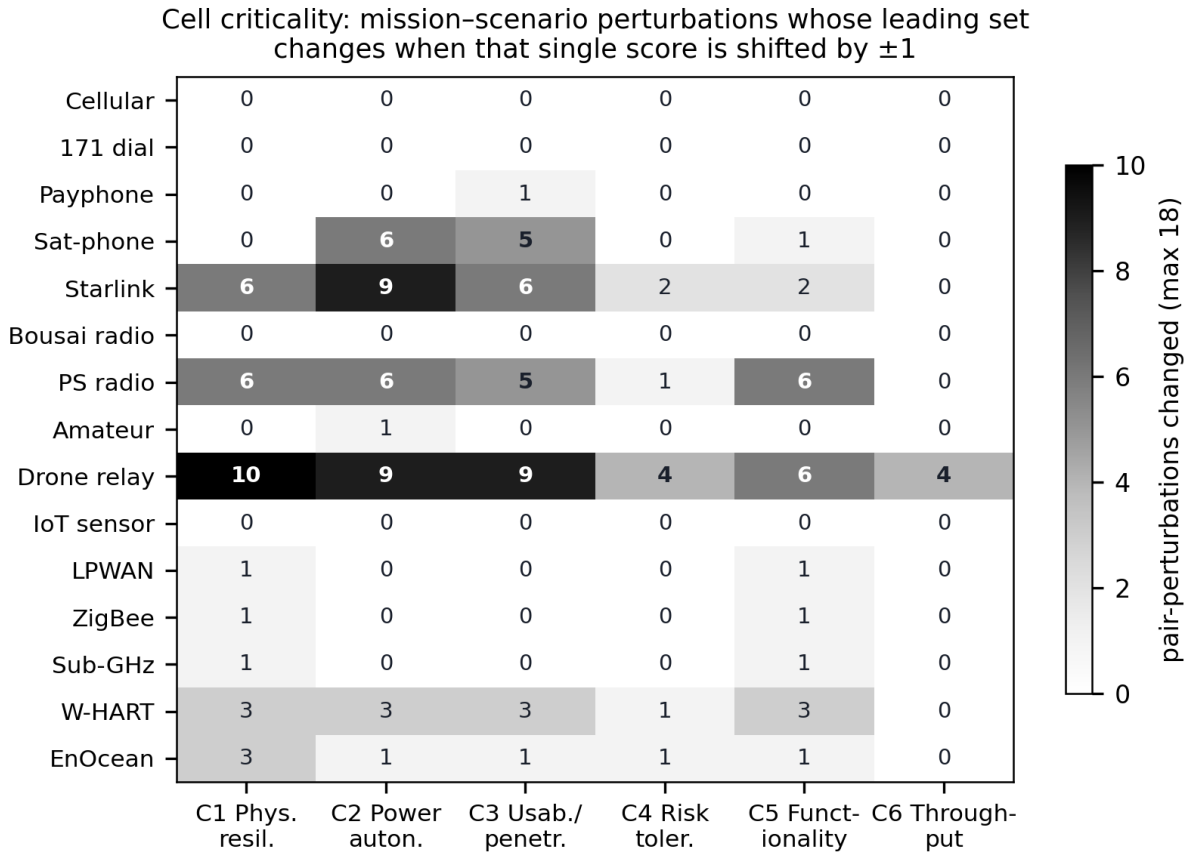

Fig. 4. Cell criticality. Entries count how many of the eighteen mission–scenario perturbations change their leading set when that single cell is shifted by ±1. Bold marks cells with criticality 4 or more, the fifteen judgements worth re-eliciting.

G. Summary

Table III collects the results of this section.

TABLE III. EFFECT OF EACH ANALYST-SUPPLIED INPUT ON THE LEADING SET

Input	Perturbation	Leading set changes in
Method-plus-preprocessing	SAW / TOPSIS / PROMETHEE II, mean pairwise	18 of 27 method-pair by case comparisons (67%)
– normalisation alone	SAW, raw vs vector normalisation	5 of 9 cases
– operator alone	SAW vs TOPSIS, both vector-normalised	6 of 9 cases
Performance scores	one cell in twenty shifted ±1 level	28% of draws
Performance scores	one cell in five shifted ±1 level	61% of draws
Screening thresholds	each θ moved ±1 level	9 of 30 move by scenario comparisons (30%)
Criterion weights	every weight randomised ±20%	25% of draws
Criterion weights	every weight randomised ±100%	55% of draws
Scale level	monotone rescaling of all six criteria	1 to 6 of 9 cases by transform
Rank-aggregation rule	mean vs median vs Copeland	1 of 9 cases

Preference parameter	PROMETHEE p ∈ {1,2,3,4}	0 of 9 cases

Effect of each analyst-supplied input on the leading set. Monte-Carlo entries pool the nine mission–scenario cases, each simulated with 3,000 draws under seed 21, and are descriptive rather than estimates of a population quantity. The rows are NOT ordered by influence and must not be read as a ranking: the perturbation magnitudes are not calibrated against one another, and the relative position of weights and thresholds reverses with the choice of δ.

TABLE IV. DESCRIPTIVE RETROSPECTIVE COMPARISON

Case	Medium deployed	|Am|	Consensus rank	Normalised	Unscreened
Noto 2024 / M1	Municipal disaster radio	2	1	0.00	0.43
Noto 2024 / M2	Starlink	6	1	0.00	0.14
Noto 2024 / M3	Starlink	5	1	0.00	0.14
Kyushu 2017 / M1	Municipal disaster radio	2	1	0.00	0.57
Kyushu 2017 / M3	Satellite comms vehicle (MSS)	5	4	0.75	0.36
Mean, all five				0.150	0.328
Mean, excluding the two M1 cases				0.250	0.213
Mean, de-duplicated and excluding M1 (n = 2)				0.375	0.250

Descriptive retrospective comparison. Normalised rank is (rank − 1)/(|Am| − 1); 0 is best in the feasible set, 1 is worst. All ranks are consensus ranks.

VI. DESCRIPTIVE RETROSPECTIVE COMPARISON, AND WHY IT CANNOT BE INFERENTIAL

Comparison against media actually deployed is the customary check in this literature, and we report it as a descriptive retrospective comparison. We use that phrase throughout, in preference to "validation", because the sample cannot bear the inferential weight that the word implies.

In the Noto response KDDI introduced Starlink from the first days for areas unreachable by road, and both KDDI and SoftBank used Starlink terminals as alternative backhaul, SoftBank supplying 100 Starlink Business kits to administrative bodies in Ishikawa Prefecture [23,24]. Municipal public-address radio carried evacuation information. In the 2017 Kyushu event, which brought 765 mm of rain in 24 hours at Toho Village and debris flows at 275 sites across Fukuoka and Oita Prefectures [29,30], the Kyushu Regional Development Bureau dispatched a satellite communication vehicle to Toho Village and Asakura City, and municipal radio again carried evacuation information. Unmanned aircraft were also flown, but for aerial survey rather than as a communication relay; since the alternative set contains bearers only, that is not a deployment of A9 and we do not count it. Table IV sets the five resulting cases against the framework's consensus ranks and against the unscreened ranking.

Read naively, the five cases favour the framework: mean normalised rank 0.150 against 0.328 unscreened and 0.500 for a random ranking. Three features of the sample forbid that reading.

First, two of the five cases are M1, whose feasible set has two members, so the normalised rank can only be 0 or 1. Two of the four zeros are therefore successful coin flips. Excluding them, the framework scores 0.250 and the unscreened ranking 0.213, so on the three cases that remain the unscreened ranking is nominally the better of the two.

Second, the cases are not independent. Starlink appears twice for the same event and municipal radio twice across events. De-duplicating and excluding the binary cases leaves two cases, at which point no inference is possible; the mean is 0.375 against the unscreened 0.250, and this comparison is itself meaningless at n = 2.

Third, the two rankings are not commensurable. Normalised rank on a set of five or six alternatives and on a set of fifteen have different supports and different discretisations, and the unscreened value is drawn towards the middle by construction.

A fourth problem is not statistical. The screen was designed by an author who already knew what had been deployed, and in all five cases the deployed medium survives screening. That is unlikely to be a coincidence, and it means "the deployed medium is feasible" cannot serve as an independent check.

The one clear discrepancy is worth naming. The satellite communication vehicle used in Kyushu ranks fourth of five, in the bottom half. Two of the three media above it, Starlink and the drone-borne relay, were not deployable in Japan in July 2017; restricting the set to what existed then moves it to second of three. We report both figures and note that this correction was applied after the discrepancy was observed, which is exactly the manoeuvre that makes retrospective validation unreliable. We draw no conclusion from it.

Our position is therefore that this comparison is descriptive only. It shows the framework is not obviously wrong. It cannot show that it is better than the alternative it replaces, and we do not claim so. Given that this is the standard design used to support validation claims in the applied MCDM literature, the diagnosis generalises: five historical cases, two of them binary and two duplicating a medium already counted, cannot distinguish two decision procedures, and studies that report such comparisons as validation are over-reading them.

VII. DISCUSSION

A. What an applied MCDM study should report

If four inputs other than the weights can also move the recommendation substantially under the perturbations examined, then a robustness section that inspects the weights alone can miss that instability, whether or not the weights turn out to be the least influential, which our data cannot establish. We suggest four changes, none of them costly. Report at least two structurally dissimilar aggregations, one additive and one distance- or outranking-based, and present the union or intersection of their leading sets rather than one method's winner. Report the normalisation used and the effect of changing it, since Section V-A shows normalisation alone accounts for a large share of what is usually attributed to method choice. Report a score-perturbation curve alongside the weight curve, using simulated rater disagreement if a second rater is unavailable. Report a cell-criticality map, which costs one pass over the matrix and tells the reader precisely which judgements the conclusion rests on.

We would also encourage reporting a set-valued output rather than a winner, and defining it before the analysis rather than after. On this problem the three methods' individual leading sets have a union of exactly three alternatives and an intersection of exactly one in every case, so reporting the union, three of five or six alternatives, is honest but weak, while reporting the intersection produces a very narrow recommendation which,

although membership in all three leading sets does carry some evidence of robustness to method choice, resolves neither the uncertainty about ordering nor the uncertainty about the scores themselves. Neither option is satisfactory, and we think that is the substantive point rather than a presentational difficulty: on a problem of this size the data do not identify a single best alternative, so a study that names one is largely reporting its choice of method.

B. Substantive observations, offered tentatively

It is tempting to read the results as showing that LEO broadband occupies the leading set for M2 and M3 under all three hazard templates and all three methods, and to argue from that for pre-positioning terminals. The data support neither step. Starlink is in the leading set of all three methods in only one of the six relevant mission–scenario combinations: under TOPSIS it is third in M2 under the wind and seismic templates, behind public-safety radio and satellite telephony, and third in M3 under all three templates, behind the drone-borne relay and public-safety radio. This is consistent with the 67% method disagreement reported in Section V-A. Nor are Starlink's cells uncritical: its power-autonomy score is the second most critical cell in the entire matrix (Fig. 4, criticality 9), with physical resilience and usability at 6 each.

What can be said is weaker. Starlink is in the leading set under SAW and PROMETHEE II in all six combinations, is in the top three under all three methods in all six, and was in fact deployed in Noto. That is suggestive, not a basis for a procurement recommendation, particularly since the criterion set models neither deployment lead time, nor stock on hand, nor transportability, nor power supply at the site, which are the quantities a pre-positioning decision turns on. We therefore offer the observation as exploratory and do not present it as procurement advice.

One observation does seem robust. Only two of fifteen media can perform mass alerting at all, and they fail in different ways: public-address radio is vulnerable to wind and cannot confirm receipt, cellular broadcast fails with the base station. A two-legged alerting stack in which both legs failed partially in one event is thin. This conclusion follows from the capability profile rather than from any score, weight or aggregation, and is therefore untouched by everything in Section V.

C. Limitations

The perturbation magnitudes are not calibrated against empirical data on how much analysts actually disagree, how imprecise elicited weights actually are, or how much experts differ over threshold placement. Until they are, the curves in Fig. 3 can support the claim that weight sensitivity alone is insufficient but cannot support any ordering among the inputs. Calibrating them, from repeated elicitation studies or from dispersion reported in published elicitation exercises, is the single change that would most strengthen this work.

The performance scores are ordinal judgements by a single analyst, assigned against the rubric anchors of Section III-B; the reasoning behind individual cells is not reproduced here. Section V-F shows that only fifteen cells matter much, which makes targeted multi-rater elicitation tractable, but it has not been done, and until it is the ordering within the leading set should be treated as unresolved. The criterion definitions are also imperfect: C3 bundles ease of use with prevalence, which can diverge; C4 names a construct that its rubric anchors define only operationally; and the binary capability profile partly overlaps C5.

The alternative set is not a set of mutually exclusive choices. NB-IoT depends on the cellular network; Starlink is both an endpoint and a backhaul for other bearers; a drone-borne relay is a platform that carries some other radio; and the generic IoT sensor network overlaps LPWAN, ZigBee and WirelessHART. The problem is closer to selecting a complementary portfolio than to choosing one medium, and a portfolio formulation with explicit dependencies would be a better model of the underlying decision. We keep the single-choice framing because it is the framing in which the sensitivity question arises, but the substantive rankings should be read with this in mind. The criterion set omits spatial coverage, capacity in simultaneous users, procurement and maintenance cost, interoperability, regulatory availability, and deployment lead time; the absence of coverage in particular means M4's rankings should not be relied on, and we predict, as a falsifiable claim, that adding a coverage criterion would demote EnOcean and promote the drone-borne relay in M4. The framework ranks media by desirability conditional on availability and does not model whether a medium can physically be brought to the site, which was the binding constraint in Kyushu. The analysis covers the immediate-response phase only. And the findings of Section V are established on one author-constructed testbed with fifteen alternatives, six criteria and nine mission–scenario cases. They show that on this problem four inputs other than the weights move the recommendation substantially; they do not show that this holds generally, and still less that any particular ordering of influence does. Applying the same audit protocol to several published MCDM datasets is the natural next study and the only way to establish whether the pattern generalises.

VIII. CONCLUSION

We built a disaster communication selection problem as a testbed and perturbed five inputs its analyst supplies, reporting each as the proportion of mission–scenario cases in which the recommended leading set changes. Switching the method-plus-preprocessing specification changes it in 67% of method-pair comparisons, of which normalisation alone accounts for 5 of 9; shifting one score cell in twenty by a rubric level changes it in 28% of draws and one cell in five in 61%; randomising every weight by ±20% changes it in 25% and by ±100% in 55%; moving screening thresholds by one level changes it in 3 of 10 perturbations; and a monotone rescaling of all criteria changes it in 1 to 6 of 9 cases depending on the transform. Sensitivity is highly localised: 53 of 90 score cells cannot change any recommendation, and one cell changes ten of eighteen.

We do not rank these influences and we caution against doing so. The perturbation magnitudes are not calibrated against one another, and the relative position of weights and thresholds reverses between $\delta = 0.2$ and $\delta = 1.0$. What the results support is the weaker claim that reporting weight sensitivity alone is not an adequate robustness check, since weight sensitivity alone can miss substantial instability associated with the scores, the screening thresholds, the scale transformation and the method-plus-preprocessing specification. We recommend that applied MCDM studies report method and normalisation sensitivity, report a score-perturbation curve, define their set-valued output before analysing, and publish a cell-criticality map so readers can see which of the analyst's judgements the conclusions rest on. The last is the least costly and, on this problem, the most useful: it reduced the set of judgements worth re-eliciting from ninety to fifteen.

We also report a descriptive retrospective comparison against two documented Japanese disasters and argue that it cannot be inferential: five cases, two of them binary and two duplicating a medium already counted, cannot distinguish two decision procedures. Since that design is standard in this literature, the caution generalises.

IX. ACKNOWLEDGMENT

The author thanks Shizuto Nakatani for helpful discussions on disaster communication selection and for contributions to the development of the alternative set and the criterion definitions used in this testbed.

X. REFERENCES

[1] A. Mardani, A. Jusoh, and E. K. Zavadskas, "Fuzzy multiple criteria decision-making techniques and applications: two decades review from 1994 to 2014," Expert Syst. Appl., vol. 42, pp. 4126–4148, 2015.

[2] E. Ilbahar, S. Cebi, and C. Kahraman, "A bibliometric analysis of multi-criteria decision-making techniques in disaster management and transportation in emergencies: towards sustainable solutions," Sustainability, vol. 17, art. 2644, 2025.

[3] Ministry of Internal Affairs and Communications, "Damage and response concerning the 2024 Noto Peninsula earthquake, report no. 13," MIC, Tokyo, Japan, Jan. 3, 2024; and White Paper on Information and Communications in Japan 2024, MIC, Tokyo, Japan, 2024.

[4] C. Kahraman, S. C. Onar, and B. Oztaysi, "Fuzzy multicriteria decision-making: a literature review," Int. J. Comput. Intell. Syst., vol. 8, pp. 637–666, 2015.

[5] F. Kılcı, B. Y. Kara, and B. Bozkaya, "Locating temporary shelter areas after an earthquake: a case for Turkey," Eur. J. Oper. Res., vol. 243, pp. 323–332, 2015.

[6] E. Triantaphyllou, Multi-Criteria Decision Making Methods: A Comparative Study. Dordrecht, The Netherlands: Kluwer, 2000.

[7] X. Wang and E. Triantaphyllou, "Ranking irregularities when evaluating alternatives by using some ELECTRE methods," Omega, vol. 36, pp. 45–63, 2008.

[8] R. F. F. Aires and L. Ferreira, "The rank reversal problem in multi-criteria decision making: a literature review," Pesqui. Oper., vol. 38, pp. 331–362, 2018.

[9] S. Chakraborty and C.-H. Yeh, "A simulation comparison of normalization procedures for TOPSIS," in Proc. 39th Int. Conf. Computers and Industrial Engineering (CIE39), 2009, pp. 1815–1820.

[10] S. H. Zanakis, A. Solomon, N. Wishart, and S. Dublish, "Multi-attribute decision making: a simulation comparison of select methods," Eur. J. Oper. Res., vol. 107, pp. 507–529, 1998.

[11] E. K. Zavadskas, Z. Turskis, and S. Kildienė, "State of art surveys of overviews on MCDM/MADM methods," Technol. Econ. Dev. Econ., vol. 20, pp. 165–179, 2014.

[12] B. Roy, "Classement et choix en présence de points de vue multiples: la méthode ELECTRE," RIRO, vol. 8, pp. 57–75, 1968.

[13] R. L. Keeney and H. Raiffa, Decisions with Multiple Objectives: Preferences and Value Tradeoffs. New York, NY, USA: Wiley, 1976.

[14] D. Bouyssou and M. Pirlot, "Nontransitive decomposable conjoint measurement," J. Math. Psychol., vol. 46, pp. 677–703, 2002.

[15] G. Munda, Social Multi-Criteria Evaluation for a Sustainable Economy. Berlin, Germany: Springer, 2008.

[16] J. W. Payne, J. R. Bettman, and E. J. Johnson, The Adaptive Decision Maker. Cambridge, U.K.: Cambridge Univ. Press, 1993.

[17] M. Erdelj, E. Natalizio, K. R. Chowdhury, and I. F. Akyildiz, "Help from the sky: leveraging UAVs for disaster management," IEEE Pervasive Comput., vol. 16, pp. 24–32, 2017.

[18] T. Sakano et al., "Bringing movable and deployable networks to disaster areas: development and field test of MDRU," IEEE Netw., vol. 30, pp. 86–91, 2016.

[19] I. del Portillo, B. G. Cameron, and E. F. Crawley, "A technical comparison of three low earth orbit satellite constellation systems to provide global broadband," Acta Astronaut., vol. 159, pp. 123–135, 2019.

[20] M. M. Kassem, A. Raman, D. Perino, and N. Sastry, "A browser-side view of Starlink connectivity," in Proc. ACM Internet Measurement Conf., 2022, pp. 151–158.

[21] S. Kassing, D. Bhattacherjee, A. B. Águas, J. E. Saethre, and A. Singla, "Exploring the 'Internet from space' with Hypatia," in Proc. ACM Internet Measurement Conf., 2020, pp. 214–229.

[22] H. Al-Hraishawi, H. Chougrani, S. Kisseleff, E. Lagunas, and S. Chatzinotas, "A survey on nongeostationary satellite systems: the communication perspective," IEEE Commun. Surveys Tuts., vol. 25, pp. 101–132, 2023.

[23] KDDI Corporation, "Deployment of Starlink to support restoration following the 2024 Noto Peninsula earthquake," press release, Jan. 2024.

[24] SoftBank Corp., "Behind the scenes: SoftBank's initiatives to restore communications on the Noto Peninsula," SoftBank News, Jan. 19, 2024.

[25] C.-L. Hwang and K. Yoon, Multiple Attribute Decision Making: Methods and Applications. Berlin, Germany: Springer, 1981.

[26] J. P. Brans and P. Vincke, "A preference ranking organisation method: the PROMETHEE method for multiple criteria decision-making," Manage. Sci., vol. 31, pp. 647–656, 1985.

[27] M. Behzadian, R. B. Kazemzadeh, A. Albadvi, and M. Aghdasi, "PROMETHEE: a comprehensive literature review on methodologies and applications," Eur. J. Oper. Res., vol. 200, pp. 198–215, 2010.

[28] T. L. Saaty, The Analytic Hierarchy Process. New York, NY, USA: McGraw-Hill, 1980.

[29] A. Nonomura, K. Fujisawa, M. Takahashi, H. Matsumoto, and S. Hasegawa, "Analysis of the actions and motivations of a community during the 2017 torrential rains in northern Kyushu, Japan," Int. J. Environ. Res. Public Health, vol. 17, art. 2424, 2020.

[30] Disaster Prevention Research Institute, Kyoto University, "Disaster survey report on the July 2017 heavy rain in northern Kyushu," DPRI, Kyoto, Japan, 2017.