
Credibility-Aware AutoML for Small Earth-Observation Constellations: The Two-to-Three-Satellite Handoff as a Search Objective

Agata M. Wijata^{1,2} Przemysław Biecek²

¹Faculty of Biomedical Engineering, Silesian University of Technology, Roosevelta 40, 44-800 Zabrze, Poland

²Centre for Credible Artificial Intelligence (CCAI), Warsaw University of Technology, Rektorska 4, 00-614 Warsaw, Poland

Abstract Earth-observation (EO) missions increasingly place learned decision-making on board, and even the smallest cooperating constellations—two or three satellites in a mothership-scout arrangement—already take autonomous actions that no operator reviews in real time. In such systems the atomic unit of a decision is not a single inference but a *handoff*: a few-hundred-byte packet in which one satellite asks another to commit an irreversible, resource-consuming action. We argue that AutoML and Neural Architecture Search (NAS) for onboard EO, which today optimise accuracy under Size, Weight, Power, and Cost (SWaP-C) budgets, should treat the *credibility* of this handoff—its calibration, cross-model agreement, explanation fidelity under a byte budget, and auditability—as a first-class search objective rather than a post-hoc add-on. We adapt the Model Science perspectives (*verify, explore, steer, refine*) to the two-to-three-satellite case, formulate credibility-aware search over the joint space of paired models and their handoff protocol, sketch a benchmark that reports credibility-cost trade-off frontiers on flight-representative hardware, and propose a machine-readable handoff schema that extends model and dataset cards toward auditable in-orbit decisions. Our position is that credibility must be co-designed with the search objective while these small constellations are still on the drawing board.

1 Motivation: The Handoff Is the Unit of Credibility

Over the past decade, EO missions have moved computation on board, from data compression toward learned, autonomous decision-making (Furano et al., 2020; Barretta et al., 2026). Between the two usual extremes—a single smart satellite and a mega-constellation of thousands of nodes—lies a case both more immediate and, for AutoML, more tractable: a *small cooperating constellation* of two or three satellites. Its canonical instance is the mothership-scout architecture (Chen et al., 2022): a wide-field *scout* screens a region and, on a candidate event, asks a higher-resolution *mothership* to re-orient and confirm; a third satellite acts as relay or second opinion. Such configurations are within reach today—onboard wildfire classification has flown on CubeSat-class hardware (Azami et al., 2022; Cratere et al., 2025)—and compress distributed autonomy into a single, analysable interaction. That interaction is a *handoff*. Within an inter-satellite window of only minutes (Leyva-Mayorga et al., 2023; Gumiel et al., 2023), the scout composes a request and the mothership adjudicates it—committing a re-tasking that a false alarm would waste irrecoverably—before any ground operator could intervene. The handoff, not the individual model, is thus the unit at which trust is established, exchanged, or lost: a scout confidence never calibrated cannot be re-calibrated downstream, and a decision that left no trace cannot be audited when a missed event later surfaces in ground archives.

Yet AutoML and NAS for onboard EO optimise the wrong object. State-of-the-art hardware-aware NAS (Cai et al., 2019, 2020; Benmeziane et al., 2021) searches single architectures for accuracy

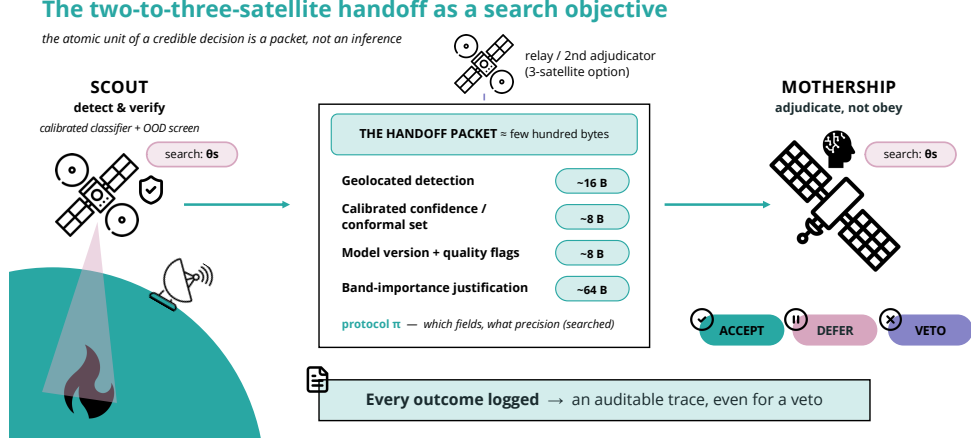


Figure 1: The handoff as the unit of credibility: the scout emits a byte-budgeted packet, the mothership *adjudicates* rather than obeys, and every outcome is logged. Models (θ_s , θ_m) and protocol (π) are all searched; a third satellite adds a relay / second-adjudicator role.

under latency and energy constraints; multi-objective NAS (Elsken et al., 2019a,b) adds model size or Floating Point Operations Per Second (FLOPS), but the objectives remain properties of one network in isolation. None optimizes whether the *confidence* a model emits is trustworthy once it crosses a satellite boundary, whether two independently searched models agree on a shared scale, or whether the byte budget of the link leaves room for a faithful justification. We adopt *Credible Artificial Intelligence* (Biecek et al., 2026) to name this composite property—interpretability, calibration, auditability, and controllability of a decision *chain*—and make three contributions: we cast the handoff as a credibility problem and adapt the Model Science capabilities to it (Section 2); we formulate credibility-aware AutoML as multi-objective search over paired models *and* their handoff protocol, with a matching benchmark (Section 3); and we propose a machine-readable handoff schema and “mission-credibility datasheet” (Section 4), closing with an agenda (Section 5).

2 The Two-to-Three-Satellite Handoff and Why It Breaks Single-Model AutoML

Consider the wildfire pipeline as a running example. The scout runs a lightweight classifier whose confidence is calibrated pre-launch (Guo et al., 2017) and whose inputs pass an Out-Of-Distribution (OOD) screen against known confounders—sun glint, cloud edges, industrial heat—before any request is issued. On a candidate detection, it emits the handoff packet of Fig. 1: geolocated detection, calibrated confidence (or a conformal set whose coverage survives the model boundary (Angelopoulos and Bates, 2023)), model lineage and input-quality flags, and a byte-budgeted justification—a band-importance vector rather than a full saliency map (Audebert et al., 2019; Selvaraju et al., 2017). The mothership then *adjudicates rather than obeys*: it weighs the received metadata against its own priors—slew budget, competing targets, and a cross-calibration map translating the scout’s confidence scale into its own—and accepts, defers, or vetoes, logging the decision either way.

The four Model Science perspectives (Biecek et al., 2026) map onto this chain—the packet is the shared interface across all four—and each imposes a requirement single-model search cannot see. *Verify*: two independently trained models must agree on a shared confidence scale that holds under orbital shift—*cross-model calibration* is a joint property of the pair, not the sum of per-model calibrations. *Explore*: the justification must stay faithful (Adebayo et al., 2018) within the link budget, so explanation fidelity is bounded by a resource the search allocates. *Steer*: the scout abstains when its OOD screen fires (Geifman and El-Yaniv, 2017) and the mothership keeps a veto—behaviours

Credibility-aware AutoML: searching the scout-mothership pair and its handoff

credibility becomes a search objective, not a post-hoc audit

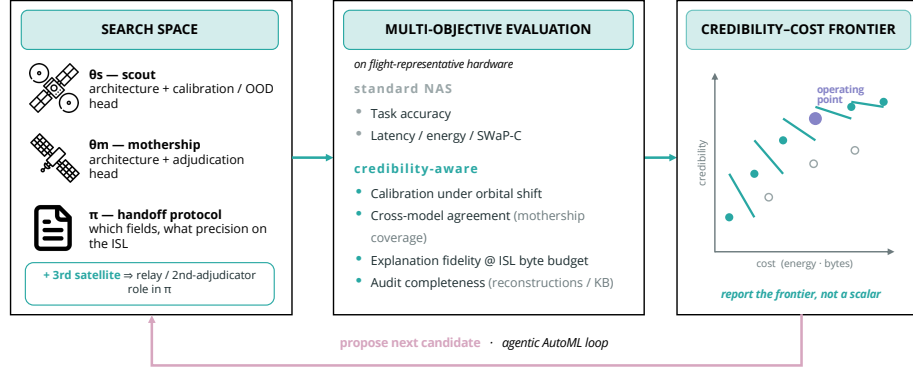


Figure 2: Credibility-aware AutoML: search over $(\theta_s, \theta_m, \pi)$ scores candidates on standard *and* credibility objectives on flight-representative hardware, yielding a credibility–cost frontier; an agentic loop proposes the next candidate.

that align the constellation without retraining, to be searched, not fixed wrappers. *Refine*: on-orbit updates change behaviour, so each decision record must cite the model version in force.

Three structural facts make this irreducible to single-model AutoML. First, the objective is a property of a *pair*: cross-model calibration and agreement do not decompose into per-model terms. Second, the models and the protocol share one budget—bytes spent on uncertainty and explanation are bytes unavailable to payload data, so the handoff schema belongs *inside* the search space, alongside the architectures. Third, the levels are coupled, and errors are irreversible: a capability absent at the node level cannot be reconstructed at the handoff, and a wrong action consumes an acquisition geometry that will not recur. Standard accuracy-under-SWaP-C search silently assumes a correction loop that a small autonomous constellation does not have.

3 Credibility-Aware Search: Objectives, Space, and Benchmark

Search space. We search jointly over a triple $(\theta_s, \theta_m, \pi)$: the scout architecture with its calibration/OOD head, the mothership architecture with its adjudication head, and the handoff protocol π fixing which credibility fields are sent and at what precision. A third satellite extends π with a relay or second-adjudicator role without changing the formulation. Crucially π is a genuine search dimension, trading link bytes against the calibration and explanation the mothership receives.

Objectives. Beyond task accuracy and SWaP-C/latency measured on flight-representative hardware, credibility-aware search evaluates calibration of the transmitted confidence *under orbital shift* (Ovadia et al., 2019), cross-model agreement (the coverage the mothership obtains when it consumes the scout’s uncertainty), explanation fidelity at the packet’s byte budget (Adebayo et al., 2018), and audit completeness (decisions reconstructible per kilobyte of record); Table 1 and Fig. 2 place these in the loop. They are multi-objective terms in the sense familiar to NAS (Elsken et al., 2019a), but defined over the pair-plus-protocol rather than a single network, and the operationally relevant output is a Pareto frontier, not a scalar: never “which model is most calibrated” but “most calibrated *at this energy and this link budget*”.

Benchmark. Realizing this requires a shared benchmark. Minimally, it provides paired scout/mothership tracks over a common EO task (wildfire or flood onset), a simulated inter-satellite link with a configurable byte budget, and shift-inducing test splits spanning a mission lifetime. Crucially, it scores *credibility cost*: the SWaP-C overhead of each uncertainty and explanation method

Table 1: Standard hardware-aware NAS optimizes properties of a single network; credibility-aware search adds objectives defined over the scout–mothership pair and the handoff protocol π . ISL: inter-satellite link.

Objective	Scope	Measured on
Task accuracy	single model	held-out + shifted splits
Latency / energy / SWaP-C	single model	flight-representative HW
Calibration under shift	per model	seasonal/geographic shift
Cross-model agreement	pair	mothership-side coverage
Explanation fidelity	pair + π	at ISL byte budget
Audit completeness	pair + π	reconstructions per KB

on flight-representative hardware (Barretta et al., 2026), so that “lightweight” becomes a number. Single-node tracks are necessary but not sufficient—handoff miscalibration and adjudication error appear only at the inter-node level—so a simulation track is required alongside any flight data. Reporting frontiers rather than leaderboard scalars is what lets such a benchmark shape the field instead of decorating it.

4 Standardized, Machine-Readable Credibility Metadata

Credibility-aware search produces artifacts that must stay interpretable long after the pass that generated them, making standardized metadata part of the method, not an afterthought. We propose two FAIR-aligned descriptions. A *handoff schema* specifies the packet’s credibility fields (Fig. 1) so a mothership from one vendor can consume a scout from another and a decision record can be replayed. A *mission-credibility datasheet* extends model cards and datasheets (Mitchell et al., 2019; Gebru et al., 2021) with what an autonomous EO handoff needs: calibration under documented orbital shifts, the OOD confounders screened, the byte budget under which explanations were validated, and the abstention/veto policy. Together these make the outputs of search auditable, reproducible, and interoperable—serving the workshop’s interest in trustworthy evaluation and semantic interoperability across EO repositories—and let agentic AutoML automate safely: an agent proposing protocols must read and write exactly these fields.

5 Research Agenda and Conclusion

Three open problems follow directly. *Heterogeneous-sensor uncertainty handoff*: scout and mothership observe at different resolutions and modalities, so the cross-calibration map is itself a learned, searchable object. *Joint budgeting under one SWaP-C envelope*: as a rate–distortion problem, what is the minimal representation of “why” that preserves the mothership’s veto-relevant information? *Search efficiency*: scoring credibility under shift is costlier than scoring accuracy, inviting surrogates and multi-fidelity strategies. We scope these to the two-to-three-satellite case—concrete and measurable—before the constellation-scale problems (Wang, 2026; Thangavel et al., 2024) that inherit them.

Our position is simple: for small autonomous EO constellations, credibility is an architectural property of the handoff, not of either model, and cannot be retrofitted onto architectures searched without it. AutoML for onboard EO already treats latency and power as first-class objectives; credibility belongs on that list, while these constellations are still on the drawing board.

Acknowledgements. Work on this project is financially supported by the Foundation for Polish Science (FNP) grant ‘Centre for Credible AI’ No. FENG.02.01-IP.05-0058/24.

References

- Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I. J., Hardt, M., and Kim, B. (2018). Sanity checks for saliency maps. In *Neural Information Processing Systems*.
- Angelopoulos, A. N. and Bates, S. (2023). Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4):494–591.
- Audebert, N., Le Saux, B., and Lefevre, S. (2019). Deep learning for classification of hyperspectral data: A comparative review. *IEEE Geoscience and Remote Sensing Magazine*, 7(2):159–173.
- Azami, M. H. b., Orger, N. C., Schulz, V. H., Oshiro, T., and Cho, M. (2022). Earth observation mission of a 6U CubeSat with a 5-meter resolution for wildfire image classification using convolution neural network approach. *Remote Sensing*, 14(8).
- Barretta, D., Marini, L., Razzano, F., Guerrisi, G., Sandberg, S., del Prete, R., Zancan, V., Mariotti, E., Furano, G., Manovi, L., Guo, J., Ullo, S. L., Longép  , N., and Meoni, G. (2026). Toward real-time Earth system intelligence through advancements in onboard AI processing: A review. *IEEE Geoscience and Remote Sensing Magazine*, 14(3):392–441.
- Benmeziane, H., El Maghraoui, K., Ouarnoughi, H., Niar, S., Wistuba, M., and Wang, N. (2021). Hardware-aware neural architecture search: Survey and taxonomy. In Zhou, Z.-H., editor, *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 4322–4329. International Joint Conferences on Artificial Intelligence Organization. Survey Track.
- Biecek, P., Longo, L., Zhou, J., Fel, T., Holzinger, A., and Samek, W. (2026). The case for model science: Verify, explore, steer, refine.
- Cai, H., Gan, C., Wang, T., Zhang, Z., and Han, S. (2020). Once for all: Train one network and specialize it for efficient deployment. In *International Conference on Learning Representations (ICLR)*.
- Cai, H., Zhu, L., and Han, S. (2019). ProxylessNAS: Direct neural architecture search on target task and hardware. In *International Conference on Learning Representations (ICLR)*.
- Chen, H., Rambaux, N., Lainey, V., and Hestroffer, D. (2022). Mothership-cubesat radioscience for phobos geodesy and autonomous navigation. *Remote Sensing*, 14(7).
- Cratere, A., Carbone, A., Cannizzaro, I., Asciola, M., Spiller, D., and Dell’Olio, F. (2025). Enabling real-time wildfire detection through edge AI onboard CubeSat platforms. In Bruzzone, L., Bovolo, F., and Bovenga, F., editors, *Artificial Intelligence and Image and Signal Processing for Remote Sensing XXXI*, volume 13670, page 136700S. International Society for Optics and Photonics, SPIE.
- Elsken, T., Metzen, J. H., and Hutter, F. (2019a). Efficient multi-objective neural architecture search via lamarckian evolution. In *International Conference on Learning Representations (ICLR)*.
- Elsken, T., Metzen, J. H., and Hutter, F. (2019b). Neural architecture search: A survey. *Journal of Machine Learning Research*, 20(55):1–21.
- Furano, G., Meoni, G., Dunne, A., Moloney, D., Ferlet-Cavrois, V., Tavoularis, A., Byrne, J., Buckley, L., Psarakis, M., Voss, K.-O., and Fanucci, L. (2020). Towards the use of artificial intelligence on the edge in space systems: Challenges and opportunities. *IEEE Aerospace and Electronic Systems Magazine*, 35(12):44–56.
- Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., III, H. D., and Crawford, K. (2021). Datasheets for datasets. *Commun. ACM*, 64(12):86–92.

- Geifman, Y. and El-Yaniv, R. (2017). Selective classification for deep neural networks. In *Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17*, page 4885–4894, Red Hook, NY, USA. Curran Associates Inc.
- Gumiela, M., Musiał, A., Wijata, A. M., Lazaj, D., Weiss, J., Sagmeister, P., Morf, T., Schmatz, M., and Nalepa, J. (2023). Benchmarking space-based data center architectures. In *IGARSS 2023 - 2023 IEEE International Geoscience and Remote Sensing Symposium*, pages 4970–4973.
- Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. (2017). On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70, ICML'17*, page 1321–1330. JMLR.org.
- Leyva-Mayorga, I., Martinez-Gost, M., Moretti, M., Perez-Neira, A., Vazquez, M. A., Popovski, P., and Soret, B. (2023). Satellite edge computing for real-time and very-high resolution Earth observation. *IEEE Transactions on Communications*, 71(10):6180–6194.
- Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., and Gebru, T. (2019). Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* '19*, page 220–229, New York, NY, USA. Association for Computing Machinery.
- Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J. V., Lakshminarayanan, B., and Snoek, J. (2019). *Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift*. Curran Associates Inc., Red Hook, NY, USA.
- Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 618–626.
- Thangavel, K., Sabatini, R., Gardi, A., Ranasinghe, K., Hilton, S., Servidia, P., and Spiller, D. (2024). Artificial intelligence for trusted autonomous satellite operations. *Progress in Aerospace Sciences*, 144:100960.
- Wang, Z. (2026). Space AI: Leveraging artificial intelligence for space to improve life on Earth.