Flight Demonstration of On-Orbit Model Adaptation on a Nanosatellite

Anonymous AI4Space Workshop @ CVPR submission

Paper ID 7

Abstract

*AI models launched on spacecraft are often frozen at deployment, yet sensing conditions and operational demands in orbit change. In-orbit updates are hard because the standard loop (e.g. downlinking new in-orbit data, such as full-resolution images, labelling them on Earth, retraining models on Earth, uplinking new weights) is constrained by bandwidth, power, thermal limits, and short operations windows. We show that meaningful in-mission adaptation is possible under these constraints and introduce Telemetry-First In-orbit Fine-Tuning (*TFIT*), an operations-first framework for fine-tuning AI models in space. Telemetry denotes compact kilobyte-scale non-image records downlinked from the spacecraft, including logits, timestamps, geolocation, and spacecraft status.* TFIT *avoids routine image downlink: labels are produced on Earth by fusing telemetry with Earth-based context sources, while thumbnails are downlinked for human review only when label evidence is low-confidence or conflicting. Crucially, in-orbit data remains onboard and only approved labels are uplinked for bounded on-orbit updates. We demonstrate* TFIT *on a flight-operated 6U nanosatellite by executing an in-orbit update of an onboard cloud-detection model (OrbitBaseline →OrbitAdapted) and evaluating pre/post-update behavior with a same-image audit protocol. Under the recovered same-image audit, OrbitAdapted improves ROC AUC by +0.510 and F1 score by +0.871 over the pre-update baseline. These results demonstrate in-mission adaptation without routine full image downlink, widening the practical scope of AI in space.*

1. Introduction

AI models in space missions are typically trained on Earth, deployed at launch, and expected to operate for months under often changing orbital conditions. In this setting, distribution shift is persistent: illumination, viewing geometry, atmosphere, and sensor effects drift over time [16, 19]. A further mismatch is that pre-launch training data can differ substantially from mission-specific in-orbit data streams and actual sensing conditions.

The direct remedy is post-launch adaptation, but the conventional workflow is hard to achieve, i.e. downlinking newly acquired in-orbit data (often full-resolution images), labelling them on Earth, retraining models on Earth, then uplinking updated weights. Under severely constrained communication bandwidth, compute/power, thermal, and mission-window constraints, that loop is not viable. We therefore treat fine-tuning as a mission-operations problem, not just a model-training task. We address this challenge with *Telemetry-First In-orbit Fine-Tuning (*TFIT*)*. TFIT is an operations protocol for mission-time model adaptation. TFIT builds supervision without routine full-image downlinking by fusing telemetry with Earth-side context sources, escalating ambiguous cases to thumbnail-level human review, and executing bounded on-orbit updates under mission governance constraints. We demonstrate our system on a flight-operated 6U nanosatellite in low Earth orbit (LEO), with an in-orbit transition from Baseline Model to Adapted Model using only in-flight evidence. We instantiate TFIT for scenarios based around cloud detection. The workflow is transferable to tasks where weak supervision can be derived reliably from time- and location-linked Earth context at telemetry scale. For auditability, we use a *same-acquisition* protocol comparing pre- and post-update model outputs on the same images

Prior work is strong on onboard inference and selection [1, 6, 23], while update-capable demonstrations are emerging [2, 13, 20]. However, prior update-capable demonstrations typically retain a heavier ground-in-the-loop cost based on either downlinking mission imagery for Earth-side retraining and uplinking updated weights, or onboard training using pre-launch, Earth-prepared labeled datasets (rather than mission-acquired supervision) [13, 20].

We present, to our knowledge, the first publicly reported in-flight model adaptation that uses newly acquired in-orbit data while avoiding routine downlink of training imagery for supervision and avoiding the uplink of updated model weights. This is implemented as an operations-first workflow (Telemetry-First In-orbit Fine-Tuning TFIT) on a flight-operated 6U nanosatellite.

Contributions.

1. We present Telemetry-First In-orbit Fine-Tuning (TFIT) as an operations-first in-flight adaptation workflow that avoids routine training-image downlink and avoids updated-weight uplink by using telemetry-scale Earth-side supervision, bounded label uplink, and same-image audit evaluation.
2. We demonstrate TFIT in real flight on a 6U nanosatellite running a cloud-detection model trained on Earth, and demonstrate how it achieves adaptation to in-orbit conditions on newly acquired in-orbit data.

2. Related Work

Onboard autonomy and inference. Autonomous spacecraft AI has a long history in onboard planning, targeting, and mission operations [1, 4, 5, 23]. More recent small-satellite demonstrations extend this line to onboard vision inference and filtering, including cloud screening, onboard Earth observation payload processing, and federated onboard compression [6–8, 14]. Our setting starts from this flight AI context, but targets mission-time adaptation rather than inference-only deployment.

Update-capable flight systems. Approaches that include model updating in situ are rare. Joukhadar et al. [11] describe an adaptive dual-segment AI architecture for SpIRIT with onboard fine-tuning capability, but do not report end-to-end flight adaptation results. Mateo-Garcia et al. [13] present a retrainable payload where selected in-orbit images are downlinked, used for model retraining on Earth, and updated weights subsequently uplinked. Růžička et al. [20] demonstrate fast onboard training of a tiny classifier head on flight hardware, training on frozen RaVAEn latent features using a pre-prepared Sentinel-2 tile dataset with Earth-defined training and evaluation splits. Our approach is closest to this update-capable line, but differs in how supervision is generated during mission operations and in performing same-image pre/post-update validation.

Adaptation under shift in constrained operations. Dataset-shift and transfer-learning literature motivates adaptation when training and deployment distributions diverge [16, 19]. In spacecraft vision, recent domain-adaptation work for non-cooperative pose tracking addresses the synthetic-to-real gap (a form of data shift) using self-supervised pseudo-labeling pipelines [12]. In remote sensing, label scarcity makes selective sampling central to adaptation efficiency [22, 24]. Post-hoc calibration with Platt scaling provides threshold control under monotonic score mapping [17], while neural network outputs can remain miscalibrated under deployment shift [9]. For space operations, small-satellite communication and resource limits further motivate compact-supervision workflows over routine full-image transfer [18, 21, 25].

3. Mission Context and Data

3.1. Platform and Operations

The deployment platform is a flight-operated 6U nanosatellite carrying six IMX219 cameras and an onboard NVIDIA Jetson Nano compute unit (128-core Maxwell GPU) capable of running onboard inference and training, operating at approximately 200 m/px ground resolution in low Earth orbit (LEO). The baseline model (OrbitBaseline) is a compact cloud-detection CNN based on MobileNetV3Small (22 layers, $\sim$2.5M parameters), selected for edge execution under flight compute constraints. We consider cloud detection as a binary image-classification task between cloudy and clear-sky images. OrbitBaseline was initialized from ImageNet-pretrained weights, trained on Earth before launch, and then deployed for in-flight operations. Pre-launch training used a curated Sentinel-2 Level-2A set of 76,800 tiles, with labels derived from Sentinel cloud-mask classes and a 20% cloudy-pixel rule per tile [3]. The data set contained 58,154 cloudy samples and 18,646 clear samples. Adaptation was constrained by severely limited communication bandwidth (about $1\,\mathrm{MB/day}$ downlink and about $100\,\mathrm{KB/day}$ uplink at 20% duty cycle, shared with other payloads) and by an operations policy that restricted routine downlink to compact non-image telemetry, including model logits (raw output scores before probability conversion), timestamps, and geolocation. Here, geolocation denotes an image-center lat/lon estimate from mission telemetry/metadata (scene-center ground intercept), not pixel-wise georeferencing. For reference, a single full-resolution image is about 6–7 MB, while a thumbnail is about $500\,\mathrm{KB}$. In nanosatellite missions, such limits routinely favor compact telemetry over raw-image transfer

Within this envelope, the conventional loop (downlink images, label them on Earth, retrain the model on Earth, uplink weights) is not practical as a routine workflow. Any update must fit communication windows, compute and power budgets, higher-priority non-AI mission tasks that can preempt onboard compute time, and other scheduled flight procedures. We therefore treat adaptation as an operations design problem, motivating Telemetry-First In-orbit Fine-Tuning (TFIT) which offers compact supervision from telemetry, bounded escalation, and same-image auditing to isolate update effects from scene-composition changes. Accordingly, mission execution imposed explicit per-run caps: 20 images per inference run and 10 images per fine-tuning run, chosen to fit available operations windows, onboard compute and power limits, and periods where higher-priority non-AI mission tasks required such resources.

3.2. Dual-Segment Operations Architecture

Operations use a dual-segment architecture. The flight segment performs image acquisition, inference, and onboard

updates under spacecraft resource constraints. The ground segment coordinates commands, tracks mission state, maintains an auditable mirror of onboard artifacts (e.g., mock model files and experiment artifacts), and manages supervision inputs for adaptation cycles. This separation enables continuous operations while preserving traceability across commands, data, and model-update events.

3.3. Mission Planner and AI Archive

To operationalize TFIT, we built a custom mission planner as the ground software foundation of the nanosatellite stack. The planner is the step-by-step execution layer for TFIT. It handles telemetry ingestion, label-generation and review workflow, update-set bookkeeping, and pre and post audit tracking, so each image remains tied to the mission context.

For Earth imaging, a capture is retained only if the spacecraft can point the camera at the intended target and hold that attitude stably during exposure. This capability is provided by a proprietary onboard ADCS (attitude determination and control subsystem), which handles attitude estimation, target alignment, and stabilization during imaging windows. In our pipeline, successful pointing is a hard validity gate for AI-use captures.

Using this workflow, we compiled the *Mission AI Image Set* (N=128): successfully pointed captures retained for AI operations. These images span latitudes from $-80.0°$ to $80.15°$ and longitudes from $-111.15°$ to $174.64°$, covering all four hemispheres and roughly 12 land-cover categories (including forest, wetland, coastline, agriculture, urban, water, desert, and frozen scenes). Each retained image was stored with mission telemetry context (time, geolocation, spacecraft status, and model outputs) in the ground archive. The result was a traceable in-orbit dataset across multiple passes and sensing conditions, which formed the basis for our adaptation workflow.

4. Method and Setup

4.1. Method Overview

We deployed *Telemetry-First In-orbit Fine-Tuning (*TFIT*)* as a mission-operations framework for in-flight adaptation (Figure 1). Here, TFIT adapts the onboard cloud-detection model on a flight-operated 6U nanosatellite. Building on the dual-segment architecture (Sec. 3.2), TFIT forms a closed-loop ground-orbit workflow: telemetry is downlinked and approved labels are uplinked to enable in-flight updates.

TFIT Round Structure. Figure 1 summarizes one full TFIT adaptation round across orbit and ground. At a high level, the stages are: **Stage 1 (Collect mission evidence):** the spacecraft acquires mission images, runs the deployed baseline model, and downlinks compact telemetry and model outputs. **Stage 2 (Create trustworthy supervision):** the ground segment converts telemetry plus Earth-context sources into approved labels, with escalation to human review when confidence is low, and calibrates baseline model scores. **Stage 3 (Select what to update on):** the ground segment identifies the most informative disagreement cases, forms a bounded fine-tuning label set, and uplinks only those labels. **Stage 4 (Execute bounded in-orbit update):** the spacecraft applies a constrained model update and runs inference with the updated model. **Stage 5 (Audit adaptation effect):** the ground segment compares baseline and updated-model outputs on the same recovered image set to isolate the impact of the adaptation. The technical implementation and quantitative settings for each stage are provided in the following subsections.

4.2. Stage 1: OrbitBaseline Inference and Downlink

Stage 1 is the operational baseline pass: the orbit segment acquires images, runs deployed OrbitBaseline inference, and downlinks telemetry and logits. Given the flight limit of 20 images per inference run, processing the full Mission AI Image Set (N=128) required seven scheduled OrbitBaseline runs. These runs were executed over multiple weeks, constrained by ground-station visibility windows, link availability, and mission-operations scheduling limits. This stage provides the mission-time evidence used in later stages.

4.3. Stage 2: Ground Labeling and Calibration

Stage 2 converts Stage-1 telemetry into approved supervision and calibrated baseline probabilities. For each acquisition, we use capture timestamp and image-center geolocation to query the cloud context from the European Centre for Medium-Range Weather Forecasts (ECMWF) fifth-generation reanalysis (ERA5) [10], accessed via the Open-Meteo Historical Weather API [15, 26]. The API returns one cloud-cover value per sampled geographic cell in percent (0 to 100). This labeling step does not use image pixels or image segmentation outputs, and does not require image downlink, hence full-resolution mission images remain onboard the nanosatellite.

We sample a local 6×6 grid centered at the image geolocation ($\pm 1°$ in latitude and longitude) to absorb pointing/geolocation uncertainty. Let $c_i \in [0, 100]$ denote cloud cover (%) for cell i for $i = 1, \ldots, 36$ we compute:

$$\mu = \frac{1}{36}\sum_{i=1}^{36} c_i, \qquad \sigma = \sqrt{\frac{1}{36}\sum_{i=1}^{36}(c_i - \mu)^2}.$$

The weak-label score is $w = \mu/100 \in [0, 1]$, and the binary weather weak label is $y_{\text{weather}} = \mathbf{1}[\mu > 20]$. Dividing by 100 is a unit conversion from percent to $[0, 1]$, so w is on the same scale as model probabilities. The 20% cutoff is aligned with the pre-launch training-label rule (the image is regarded as cloudy if cloud coverage exceeds 20%).

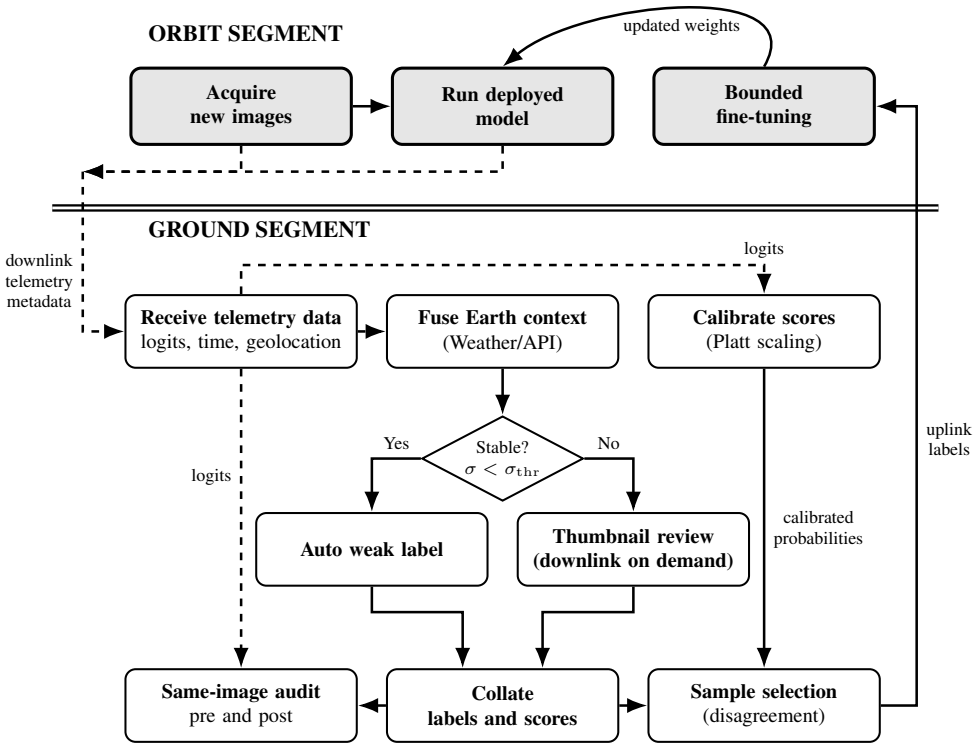


Figure 1. Telemetry-First In-orbit Fine-Tuning (TFIT). The orbit segment acquires new images, runs inference, and performs bounded fine-tuning. The ground segment converts downlinked telemetry into approved labels via calibration, Earth-context fusion, and stability-gated labeling, with thumbnail downlink only for escalated human-review cases, then selects and uplinks labels for update. A same-image audit on the ground compares pre- and post-update outputs to isolate adaptation effects.

Stability gate. A case is *stable* when nearby weather cells agree, and *unstable* when local weather is patchy or conflicting. We quantify this using the grid-level variability σ (in percentage points) and set $\sigma_{\text{thr}} = 20$ as a fixed operating threshold chosen empirically to separate the stable/unstable regimes illustrated in Figure 2.

Stable cases ($\sigma < 20$) are auto-labeled, while unstable cases ($\sigma \geq 20$) are escalated to thumbnail-based human review, with thumbnails downlinked only for these cases [22]. For escalated cases, the mission planner shows a simple thumbnail bright-pixel-fraction cue (annotation aid only, not a cloud detector or label), and the human assigns the final binary label using a visual rule aligned with pre-training semantics ("cloudy" if coverage appears roughly $\geq 20\%$). Approved labels are collated into the Stage-2.

We then apply Platt scaling for OrbitBaseline:

$$p_{\text{cal}} = \frac{1}{1 + \exp\left(-\frac{\ell - B}{T}\right)}, \tag{1}$$

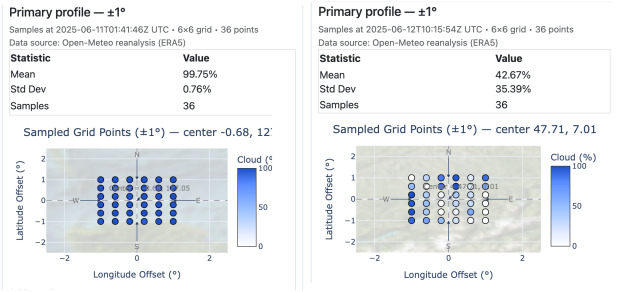
where ℓ is the downlinked logit and (T, B) are fitted calibration parameters [17].

Throughout the paper, *uncalibrated (uncal.)* denotes the pre-launch operating point: an identity mapping (T=1, B=0) with the default decision threshold θ=0.5.

We fit (T, B) on a pre-update labeled calibration cohort using an out-of-fold (OOF) protocol. A fit subset is used to estimate (T, B) from the downlinked logits and approved Stage-2 labels, and a disjoint held-out subset is used only for verification. Operationally, this is a one-dimensional logistic regression on the downlinked logit ℓ (Platt scaling), with (T, B) estimated by minimizing binary cross-entropy (negative log-likelihood) on the OOF fit subset. After fitting (T, B), we choose the calibrated decision threshold θ on the OOF calibration data by threshold search over calibrated probabilities using a criterion that balances F1 with a specificity constraint (maximize F1 subject to a target specificity). Calibration changes only score-to-probability mapping (operating point), not the model weights. Because the mapping is monotonic, ranking (AUC) is preserved while thresholded metrics can improve.

4.4. Stage 3: Disagreement-Based Selection

This stage converts the approved Stage 2 label set into a bounded uplink set for in-orbit update. To prioritize informative samples, we compare the Stage 2 weak-label score w with calibrated OrbitBaseline output using the model–



Stable case: μ=99.75%, σ=0.76% Unstable case: μ=42.67%, σ=35.39%

Figure 2. Screenshots from the mission planner app showing weather-grid stability examples from mission acquisitions. Low-σ cases are auto-labeled; high-σ cases trigger thumbnail-human escalation.

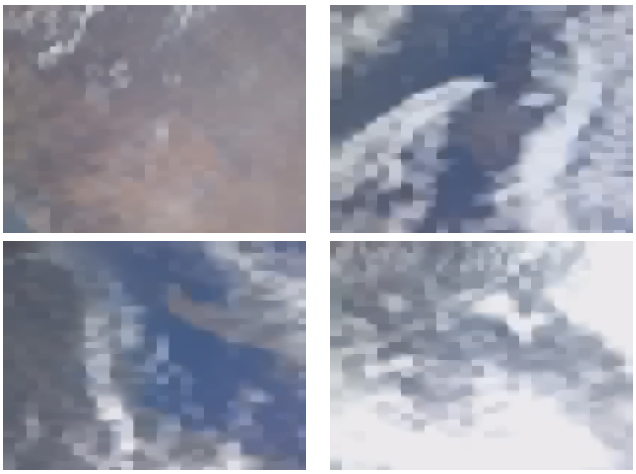


Figure 3. Examples of thumbnail downlinked data from the Mission AI Image Set. These compact previews are downlinked on demand only when the weather weak label is unstable (high-σ) and human annotation is required (Section 4.3), avoiding routine full-image downlink.

weather conflict score $\delta = |p_{\text{cal}} - w|$. Higher δ indicates stronger conflict between model belief and context signal, so higher-δ samples are ranked earlier. Only a bounded top-ranked subset is selected and uplinked for Stage 4 update execution. For the OrbitBaseline →OrbitAdapted update in this study, we ranked the full Mission AI Image Set (N=128) by δ and selected the top 10 samples for uplink. This fine-tuning set size is bounded for flight execution.

4.5. Stage 4: In-Orbit Update Execution

This stage begins once the selected label set is uplinked. To keep updates bounded under onboard compute limits, adaptation is restricted to last-layer fine-tuning. Let $f_\phi(x)$ be the frozen feature extractor and ℓ the model logit:

$$\ell = w_{\text{fc}}^\top f_\phi(x) + b, \qquad p = \frac{1}{1+\exp(-\ell)}. \tag{2}$$

Only the final fully connected parameters (w_{fc}, b) are updated in-orbit and ϕ remains fixed. Optimization is performed on the onboard cloud and clear labels with binary cross-entropy with logits (BCEWithLogitsLoss) and AdamW over the unfrozen parameters, using learning rate 10^{-3}, weight decay 10^{-2}, and a fixed budget of 5 epochs. The update is executed as image-wise steps (effective batch size 1) over the bounded 10-image fine-tuning set. After the OrbitBaseline →OrbitAdapted update, OrbitAdapted inference is scheduled over the non-training remainder of the Mission AI Image Set (N=118, i.e., 128 total captures minus the 10-image fine-tuning set). Because flight operations limit each inference experiment to 20 images per run, the audit required multiple scheduled runs. In practice, only one 20-image OrbitAdapted run was executed and successfully downlinked. Shortly afterward, spacecraft anomalies forced the nanosatellite into end-of-life operations, with intermittent/lost communications and no reliable access to the onboard computing unit; as a result, the remaining scheduled OrbitAdapted inference runs could not be recovered and no further fine-tuning rounds could be scheduled. This recovered batch is the Stage 4 output carried into Stage 5.

4.6. Stage 5: Same-Image Audit for Adapted-Model Performance

Stage 5 is executed on the ground after downlink and evaluates pre- vs post-update behavior using only downlinked model outputs (no image data are used in this step). The full images remain on the nanosatellite; the audit compares OrbitBaseline and OrbitAdapted outputs for the same recovered 20 image captures from Stage 4, using the corresponding labels established in Stage 2.

20-image audit set. For these same 20 images, pre-update OrbitBaseline outputs were already available from Stage 1, so the audit set contains complete pre- and post-update model outputs on identical captures. For each image in this batch, we compare OrbitBaseline and OrbitAdapted on the exact same capture and record OrbitBaseline uncalibrated outputs (logit and uncalibrated probability), OrbitBaseline calibrated probability (p_{cal}), and OrbitAdapted uncalibrated outputs (logit and uncalibrated probability). Using identical captures removes scene-selection confounds and isolates the effect of adaptation. In this 20-image same-image set, ground-truth sources comprise 4 human annotations and 16 weather-derived auto weak labels, with a class balance of 17 cloudy and 3 clear images.

Post-update reporting choices. Results (Section 5) report three same-image settings: OrbitBaseline uncalibrated, OrbitBaseline calibrated, and OrbitAdapted uncalibrated. We do not apply post-update calibration to OrbitAdapted because only this single 20-image same-image run has recovered OrbitAdapted outputs. This sample is too small for reliable out-of-fold calibration.

Table 1. Metric definitions (TP, TN, FP, FN; $N{=}TP{+}TN{+}FP{+}FN$). For AUC, $s(\cdot)$ is the model score, x^+ is a positive (cloudy) sample, and x^- is a negative (clear) sample.

Accuracy	$\mathrm{Acc} = \frac{TP+TN}{N}$	Specificity	$\mathrm{Spec} = \frac{TN}{TN+FP}$
Precision	$\mathrm{Prec} = \frac{TP}{TP+FP}$	Recall	$\mathrm{Rec} = \frac{TP}{TP+FN}$
F1	$F1 = \frac{2TP}{2TP+FP+FN}$	Balanced accuracy	$\mathrm{BAcc} = \frac{\mathrm{Rec}+\mathrm{Spec}}{2}$
MCC	$\mathrm{MCC} = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$		
AUC	$\mathrm{AUC} = \mathbb{P}\big(s(x^+) > s(x^-)\big)$		

Metrics. We report accuracy (Acc), specificity (Spec), precision (Prec), recall (Rec), F1, balanced accuracy (BAcc), Matthews correlation coefficient (MCC), and AUC. Throughout the paper, precision and recall use the cloudy class as positive (clear is negative). Let TP (true positive), TN (true negative), FP (false positive), and FN (false negative), and let $N = TP{+}TN{+}FP{+}FN$. Metric definitions are given in Table 1.

4.7. Robustness Analyses

Given the small cohort, we assess stability via threshold sweeps, bootstrap metrics, and per-image analysis.

Operating-point robustness. We sweep the OrbitAdapted threshold $\theta \in [0.01, 0.99]$ and track F1, testing whether gains persist across operating points.

Bootstrap resampling metric stability. The same-image cohort is small ($N{=}20$) and highly imbalanced (17 cloudy, 3 clear), so single-point estimates of AUC and F1 can be sensitive to which images are included. We therefore assess metric stability using a stratified bootstrap on the Stage 5 same-image audit outputs for each reported model setting, using the same 20 prediction–label pairs from this cohort (no additional model runs). Each bootstrap replicate is constructed by sampling *with replacement* within each class while preserving the observed class balance (17 cloudy, 3 clear), and we recompute AUC and F1. We repeat this procedure 5000 times and report 95% bootstrap percentile intervals (2.5th and 97.5th percentiles) as a resampling-based metric-stability check for this audit cohort, not as population-level uncertainty intervals.

Case-level diagnostics. We inspect image-level prediction changes from OrbitBaseline to OrbitAdapted, with emphasis on boundary-near samples. This identifies whether residual errors are more consistent with model-update limits or with label ambiguity/noise.

5. Results

We present stage-wise evidence for Stages 1–5 of the pipeline (Section 4), then stress-test the results with robustness and operational analyses. Telemetry-First In-orbit Fine-Tuning (TFIT) defines the operating architecture, and outcomes are reported in execution order.

Table 2. Mission AI Image Set ($N{=}128$) labeling composition from Stage 2.

Label source	Total	Cloudy	Clear
All labels	128	95	33
Human-labeled	27	18	9
Weather weak-labeled	101	77	24

Table 3. OrbitBaseline calibration performance on the pre-update out-of-fold (OOF) split. Precision and recall are reported for the cloudy class. Arrows indicate whether higher (↑) is better. Best value per slice and metric is in bold font.

Slice	Setting	θ	Acc↑	Spec↑	Prec↑	Rec↑	F1↑
OOF fit set ($N{=}58$)	Uncal.	0.50	0.379	**0.786**	0.786	0.250	0.379
OOF fit set ($N{=}58$)	Cal.	0.685	**0.517**	0.714	**0.833**	**0.455**	**0.588**
OOF held-out ($N{=}35$)	Uncal.	0.50	0.429	**1.000**	**1.000**	0.048	0.091
OOF held-out ($N{=}35$)	Cal.	0.685	**0.657**	0.857	0.846	**0.524**	**0.647**

Table 4. Full Mission AI Image Set ($N{=}128$) OrbitBaseline results before and after Stage 2 calibration. Precision and recall are reported for the cloudy class. Arrows indicate whether higher (↑) or lower (↓) is better. The best value per metric is in bold font.

Setting	θ	AUC↑	Acc↑	Spec↑	Prec↑	Rec↑	F1↑	BAcc↑
Uncalibrated (uncal.)	0.50	**0.555**	0.418	**0.879**	0.750	0.185	0.296	0.532
Calibrated (cal.)	0.685	**0.555**	**0.571**	0.758	**0.795**	**0.477**	**0.596**	**0.617**

5.1. Stage 1 Results: OrbitBaseline Inference and Downlink

We establish execution evidence for the pre-update baseline pass. OrbitBaseline ran over the Mission AI Image Set ($N{=}128$), and telemetry/logits were downlinked from orbit to ground for Stage 2 labeling and calibration.

5.2. Stage 2 Results: Ground labeling and Calibration

First, Stage 2 labeling produces the mission label composition summarized in Table 2. Second, we fit calibration and verify it on disjoint out-of-fold splits (Table 3). The calibrated operating threshold is then selected on the OOF calibration data by threshold search over calibrated probabilities; the resulting deployed calibrated threshold is $\theta{=}0.685$, chosen by maximizing F1 subject to a specificity constraint. Third, we apply calibration to the full Mission AI Image Set and report uncalibrated versus calibrated OrbitBaseline metrics in Table 4. Across Tables 3 and 4, calibration improves operating-point behavior, as expected, because it changes the score-to-decision mapping without changing score ordering (and thus does not improve ranking quality AUC). The remaining ranking error is therefore the part that calibration cannot fix, which motivates weight updates.

Table 5. Stage 3 selection summary for the 10-image fine-tuning set, including score statistics (w, p_{cal}, δ) and the label–prediction conflict pattern that determine uplink selection.

Score statistics over the 10 selected images			
Statistic	Mean	Min	Max
Weather weak-label score w	0.052	0.000	0.121
OrbitBaseline calibrated probability p_{cal}	0.712	0.685	0.719
Conflict $\delta = \lvert p_{\text{cal}} - w \rvert$	0.659	0.564	0.719
Labels and OrbitBaseline predictions on the 10 selected images			
	Cloudy	Clear	Total
Ground Truth labels	0	10	10
OrbitBaseline prediction (calibrated)	10	0	10

5.3. Stage 3 Results: Disagreement-Based Selection and Uplink

The aim of Stage 3 is to form a small, highly informative fine-tuning set under the mission cap of 10 images per fine-tuning run (Section 3). We therefore compute disagreement δ for all samples in the Mission AI Image Set (N=128) and select the top 10 to form the in-orbit fine-tuning set for the OrbitBaseline →OrbitAdapted update. These selected images form a high-conflict subset: weather weak-label scores are low (mean $w = 0.052$), calibrated OrbitBaseline probabilities are high (mean $p_{\text{cal}} = 0.712$), and the resulting disagreement is large (mean $\delta = 0.659$). Table 5 shows the mismatch clearly: under the Stage-2 ground-truth labels, all 10 selected images are clear-sky (10 clear, 0 cloudy). By contrast, calibrated OrbitBaseline predicts cloudy for every image (10 cloudy, 0 clear). In this stage, we uplink the selected labels to the nanosatellite as defined by TFIT (Figure 1) emphasising again that the corresponding full-resolution images are not downlinked and remain onboard. This makes Stage 3 a targeted correction step, where uplinked supervision is concentrated on the model's strongest label conflicts rather than on a manually-balanced sample.

5.4. Stage 4 Results: On-Orbit Update Execution

Using the 10 labels uplinked in Stage 3, bounded last-layer fine-tuning was executed in orbit to obtain OrbitAdapted from OrbitBaseline. Stage 4 then scheduled OrbitAdapted inference over the non-training remainder of the Mission AI Image Set (N=118), with a mission cap of 20 images per inference run. Thus, Stage 4 defines the planned N=118 inference campaign, while Stage 5 uses the single recovered 20-image run as the available audited evidence.

5.5. Stage 5 Results: Same-Image Audit (N=20)

Stage 5 reports the final audit outcomes on the recovered 20-image subset of the Stage 4 inference campaign. For each of these images, we compare three outputs: OrbitBaseline uncalibrated, OrbitBaseline calibrated, and OrbitAdapted uncalibrated. We follow the same-image audit protocol defined in Section 4. This 20-image audit subset is disjoint from the 10-image fine-tuning set used in Stage 4. The audit cohort contains: 17 cloudy, 3 clear images and labels from 4 human annotations and 16 weather-derived weak labels.

Table 6. Performance summary on the 20-image audit set. Precision and recall are reported for the cloudy class. All AUC values in this table are computed only on these 20 audit images. For context, OrbitBaseline AUC on the full pre-update Mission AI Image Set is 0.55 (Table 4). Arrows indicate whether higher (↑) or lower (↓) is better. The best value per metric is in bold font.

Model / setting	θ	AUC↑	Acc↑	Spec↑	Prec↑	Rec↑	F1↑	BAcc↑	MCC↑
OrbitBaseline uncal.	0.5	0.284	0.100	0.333	0.333	0.059	0.100	0.196	-0.608
OrbitBaseline cal.	0.685	0.284	0.400	0.333	0.778	0.412	0.538	0.373	-0.183
OrbitAdapted uncal.	0.5	**0.794**	**0.950**	**0.667**	**0.944**	**1.000**	**0.971**	**0.833**	**0.793**

Table 7. Confusion matrices on the same-image 20-image audit cohort (positive class: cloudy; negative class: clear). Rows are predictions; columns are ground truth.

Setting	Prediction	GT cloudy	GT clear
OrbitBaseline uncal.	Pred cloudy	1	2
	Pred clear	16	1
OrbitBaseline cal.	Pred cloudy	7	2
	Pred clear	10	1
OrbitAdapted uncal.	Pred cloudy	17	1
	Pred clear	0	2

Table 6 shows the primary Stage 5 comparison. Relative to OrbitBaseline (uncalibrated), OrbitAdapted improves F1 from 0.100 to 0.971 (+0.871), which corresponds to a strong operational recovery in cloud-classification decisions. On this same-image 20-image audit cohort, AUC and F1 provide complementary evidence (threshold-free ranking/separability and thresholded operating behavior), and both should be considered in the context of small-sample audit evidence under the 17-cloudy/3-clear class split. Table 7 makes the error shift explicit: missed-cloud errors (FN) drop from 16 (uncalibrated OrbitBaseline) and 10 (calibrated OrbitBaseline) to 0 (OrbitAdapted), while total correct decisions increase from 2/20 (uncalibrated OrbitBaseline) to 19/20 (OrbitAdapted).

Case-by-Case Analysis. Figure 4 makes Stage 5 behavior explicit across all 20 images. Two representative corrections are: Image-02 (GT clear) shifts from 0.552 under uncalibrated OrbitBaseline (wrong; predicted cloudy) to 0.324 under uncalibrated OrbitAdapted (correct; predicted clear), and Image-04 (GT cloudy) shifts from 0.456 under uncalibrated OrbitBaseline (wrong; predicted clear) to 1.000 under uncalibrated OrbitAdapted (correct; predicted cloudy). At θ=0.5, only one OrbitAdapted error remains (Image-01; GT clear, score 1.000). The figure also shows a threshold-distance pattern under the reported settings: calibrated Or-

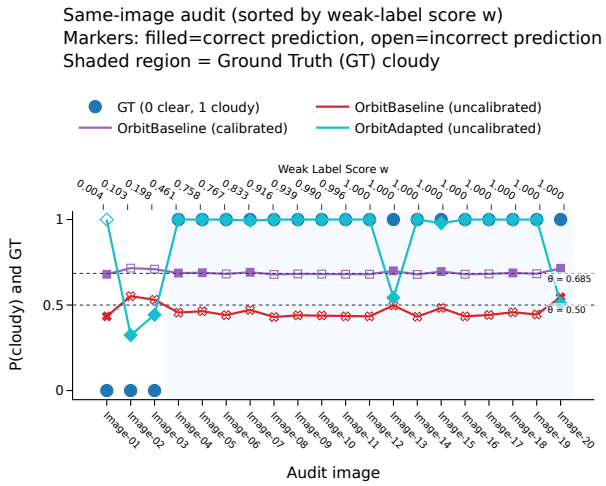


Figure 4. All 20 Stage-5 same-image audit cases (scores are positive-class probabilities). Dotted threshold lines mark the uncalibrated default decision threshold ($\theta = 0.5$) and the calibrated OrbitBaseline threshold ($\theta = 0.685$).

Table 8. Stratified-bootstrap 95% resampling intervals for AUC and F1 on the Stage 5 same-image 20-image audit cohort (5000 resamples; fixed outputs, no additional model runs), preserving the observed class balance (17 cloudy, 3 clear) in each redraw. Intervals are reported as a resampling-based metric-stability check for this audit cohort (not population-level uncertainty intervals).

Model / setting	AUC (95% CI)↑	F1 (95% CI)↑
OrbitBaseline uncal.	0.284 [0.000, 0.765]	0.100 [0.000, 0.286]
OrbitBaseline cal.	0.284 [0.000, 0.765]	0.538 [0.273, 0.733]
OrbitAdapted uncal.	**0.794** [0.412, 1.000]	**0.971** [0.919, 1.000]

bitBaseline scores cluster tightly near their threshold, while OrbitAdapted scores are farther from threshold on most images. Quantitatively, 17/20 calibrated OrbitBaseline scores lie within ± 0.02 of threshold versus 1/20 for OrbitAdapted; we report this as a descriptive visualization of fewer borderline calls in this audit, noting that it compares calibrated OrbitBaseline outputs against uncalibrated OrbitAdapted outputs. Overall, the case-level audit matches the cohort-level result: the update corrects most previously wrong decisions while retaining a small clear-class edge-case risk.

5.6. Robustness Evidence

Table 8 reports stratified-bootstrap resampling intervals for the Stage 5 same-image 20-image audit set (17 cloudy, 3 clear), using 5000 redraws of the same fixed 20 prediction–label pairs and no additional model runs while preserving the observed class balance in each redraw. Because this cohort is small and imbalanced, both AUC and F1 should be interpreted cautiously as small-sample resampling stability checks rather than population-level uncertainty intervals. Nevertheless AUC remains important because it measures ranking/separability (which calibration cannot improve), but its intervals are still wide on this cohort (e.g., 0.000–0.765 for OrbitBaseline and 0.412–1.000 for OrbitAdapted). F1 captures the threshold operating behavior at the reported decision thresholds; Table 8 shows improved F1 resampling intervals for OrbitAdapted relative to both OrbitBaseline settings. We also evaluated **operating-point robustness** by sweeping the OrbitAdapted decision threshold over $\theta \in [0.01, 0.99]$ on the 20-image audit cohort. The sweep shows that OrbitAdapted maintains F1$>$ 0.90 for $\theta \in [0.01, 0.97]$ and exceeds the calibrated OrbitBaseline F1 on the same 20-image audit cohort (Table 6; F1$=$ 0.538 at θ=0.685) across the full sweep, indicating that F1 gain is not an artifact of a single threshold choice.

6. Discussion and Limitations

The central innovation is operational: Telemetry-First In-orbit Fine-Tuning (TFIT) treats on-orbit adaptation as a mission workflow, keeping full-resolution images onboard while generating Earth-side supervision from telemetry-plus-context and uplinking only selected labels for bounded updates. Although demonstrated here for cloud detection, the same approach can extend to other tasks such as ship detection by replacing the context-to-label module; for sparse targets, weaker pointing or geolocation tolerance makes label reliability more sensitive and requires stricter confidence gating and more human escalation.

These findings are still bounded by flight constraints. Only one inference run after the on-orbit update was recovered for audit (N=20), with class imbalance (17 cloudy, 3 clear) and mostly weather-derived labels (16/20, 4 human), so uncertainty is higher for clear-class behavior. Accordingly, OrbitAdapted reporting remains uncalibrated (small audit set). The analyzed 10-sample update was the first bounded fine tuning of a planned multi-round campaign with mission end-of-life preventing additional rounds.

7. Conclusion

In this paper we demonstrate in-flight model adaptation on newly acquired in-orbit data using an operations-first workflow (Telemetry-First In-orbit Fine-Tuning TFIT) on a flight-operated 6U nanosatellite. The workflow avoids routine downlinking of training imagery and avoids uplinking of updated model weights by generating Earth-side supervision from telemetry-scale mission records and context sources, then uplinking only highly informative selected labels for bounded onboard updates. In this campaign, a bounded 10-image update improved same-image audit performance by +0.510 AUC and +0.871 F1, and this operations-first pattern can extend beyond cloud detection when context-derived supervision is reliable.

References

[1] Steve Chien, Rob Sherwood, Daniel Tran, Benjamin Cichy, Gregg Rabideau, Rebecca Castaño, Ashley Davies, Dan Mandl, Stuart Frye, Bruce Trout, Jeff D'Agostino, Seth Shulman, Darrell Boyer, Sandra Hayden, Adam Sweet, and Scott Christa. Lessons learned from autonomous sciencecraft experiment. In *Proceedings of the Fourth International Joint Conference on Autonomous Agents and Multiagent Systems*, page 11–18, New York, NY, USA, 2005. Association for Computing Machinery. 1, 2

[2] Miguel Ortiz Del Castillo, Jonathan Morgan, Jack McRobbie, Clint Therakam, Zaher Joukhadar, Robert Mearns, Simon Barraclough, Richard Sinnott, Andrew Woods, Chris Bayliss, Kris Ehinger, Ben Rubinstein, James Bailey, Airlie Chapman, and Michele Trenti. Mitigating challenges of the space environment for onboard Artificial Intelligence: Design overview of the imaging payload on SpIRIT. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 6789–6798, 2024. 1

[3] M. Drusch, U. Del Bello, S. Carlier, O. Colin, V. Fernandez, F. Gascon, B. Hoersch, C. Isola, P. Laberinti, P. Martimort, A. Meygret, F. Spoto, O. Sy, F. Marchese, and P. Bargellini. Sentinel-2: Esa's optical high-resolution mission for gmes operational services. *Remote Sensing of Environment*, 120: 25–36, 2012. The Sentinel Missions - New Opportunities for Science. 2

[4] Tara A. Estlin, Benjamin J. Bornstein, Daniel M. Gaines, Robert C. Anderson, David R. Thompson, Michael Burl, Rebecca Castaño, and Michele Judd. Aegis automated science targeting for the MER opportunity rover. *ACM Trans. Intell. Syst. Technol.*, 3(3), 2012. 2

[5] Raymond Francis, Tara Estlin, Stephen Johnstone, Laurent Peret, Valerie Mousset, Gary Doran, Daniel Gaines, Suzanne Montaño, Olivier Gasnault, Jens Frydenvang, Roger Wiens, Steven Schaffer, Betina Pavri, Vandana Verma, Debarati Chattopadhyay, Benjamin Bornstein, Nimisha Mittal, and Lauren DeFlores. *Incorporating AEGIS autonomous science into Mars Science Laboratory rover mission operations*. 2

[6] Gianluca Giuffrida, Lorenzo Diana, Francesco de Gioia, Gionata Benelli, Gabriele Meoni, Massimiliano Donati, and Luca Fanucci. CloudScout: A deep neural network for on-board cloud detection on hyperspectral images. *Remote Sensing*, 12(14), 2020. 1, 2

[7] Gianluca Giuffrida, Luca Fanucci, Gabriele Meoni, Matej Batič, Léonie Buckley, Aubrey Dunne, Chris van Dijk, Marco Esposito, John Hefele, Nathan Vercruyssen, Gianluca Furano, Massimiliano Pastena, and Josef Aschbacher. The Phi-Sat-1 mission: The first on-board deep neural network demonstrator for satellite earth observation. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–14, 2022.

[8] Pablo Gómez and Gabriele Meoni. Tackling the satellite downlink bottleneck with federated onboard learning of image compression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 6809–6818, 2024. 2

[9] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, pages 1321–1330. PMLR, 2017. 2

[10] Hans Hersbach, Bill Bell, Paul Berrisford, Shoji Hirahara, András Horányi, Joaquín Muñoz-Sabater, Julien Nicolas, Carole Peubey, Raluca Radu, Dinand Schepers, Adrian Simmons, Cornel Soci, Saleh Abdalla, Xavier Abellan, Gianpaolo Balsamo, Peter Bechtold, Gionata Biavati, Jean Bidlot, Massimo Bonavita, Giovanna De Chiara, Per Dahlgren, Dick Dee, Michail Diamantakis, Rossana Dragani, Johannes Flemming, Richard Forbes, Manuel Fuentes, Alan Geer, Leo Haimberger, Sean Healy, Robin J. Hogan, Elías Hólm, Marta Janisková, Sarah Keeley, Patrick Laloyaux, Philippe Lopez, Cristina Lupu, Gabor Radnoti, Patricia de Rosnay, Iryna Rozum, Freja Vamborg, Sebastien Villaume, and Jean-Noël Thépaut. The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730):1999–2049, 2020. 3

[11] Zaher Joukhadar, Jonathan Morgan, Christopher Bayliss, Miguel Ortiz del Castillo, Jack McRobbie, Robert Mearns, Krista A. Ehinger, Benjamin I. P. Rubinstein, Richard O. Sinnott, Michele Trenti, and James Bailey. Designing an adaptive AI system for operation on board the SpIRIT nanosatellite. In *AI 2024: Advances in Artificial Intelligence*, pages 329–341, Singapore, 2025. Springer Nature Singapore. 2

[12] Kun Liu and Yongjun Yu. Revisiting the domain gap issue in non-cooperative spacecraft pose tracking. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 6864–6873, 2024. 2

[13] Gonzalo Mateo-Garcia, Josh Veitch-Michaelis, Cormac Purcell, Nicolas Longepe, Simon Reid, Alice Anlind, Fredrik Bruhn, James Parr, and Pierre Philippe Mathieu. In-orbit demonstration of a re-trainable machine learning payload for processing optical imagery. *Scientific Reports*, 13(1):10391, 2023. 1, 2

[14] Gabriele Meoni, Marcus Märtens, Dawa Derksen, Kenneth See, Toby Lightheart, Anthony Sécher, Arnaud Martin, David Rijlaarsdam, Vincenzo Fanizza, and Dario Izzo. The OPS-SAT case: A data-centric competition for onboard satellite image classification. *Astrodynamics*, 8(4):507–528, 2024. 2

[15] Open-Meteo. Historical weather api. https://open-meteo.com/en/docs/historical-weather-api, 2026. Accessed: 2026-02-22. 3

[16] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2010. 1, 2

[17] John C. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in Large Margin Classifiers*, pages 61–74. MIT Press, Cambridge, MA, 1999. 2, 4

[18] Armen Poghosyan and Alessandro Golkar. CubeSat evolution: Analyzing CubeSat capabilities for conducting science missions. *Progress in Aerospace Sciences*, 88:59–83, 2017. 2

[19] Joaquin Quiñonero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D. Lawrence. *Dataset Shift in Machine Learning*. The MIT Press, 2008. 1, 2

[20] Vít Růžička, Gonzalo Mateo-García, Chris Bridges, Chris Brunskill, Cormac Purcell, Nicolas Longépé, and Andrew Markham. Fast model inference and training on-board of satellites. In *IGARSS 2023 - 2023 IEEE International Geoscience and Remote Sensing Symposium*, pages 2002–2005, 2023. 1, 2

[21] Daniel Selva and David Krejci. A survey and assessment of the capabilities of cubesats for earth observation. *Acta Astronautica*, 74:50–68, 2012. 2

[22] Burr Settles. Active learning literature survey. Technical Report TR1648, University of Wisconsin–Madison, Department of Computer Sciences, 2009. 2, 4

[23] Rob Sherwood, Steve Chien, Daniel Tran, Benjamin Cichy, Rebecca Castano, Ashley Davies, and Gregg Rabideau. Safe agents in space: Lessons from the autonomous sciencecraft experiment. In *AI 2004: Advances in Artificial Intelligence*, pages 51–63, Berlin, Heidelberg, 2005. Springer Berlin Heidelberg. 1, 2

[24] Devis Tuia, FrÉdÉric Ratle, Fabio Pacifici, Mikhail F. Kanevski, and William J. Emery. Active learning methods for remote sensing image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 47(7):2218–2232, 2009. 2

[25] Sasha V. Weston, Craig D. Burkhard, Jan M. Stupl, Rachel L. Ticknor, Bruce D. Yost, Rebekah A. Austin, Pavel Galchenko, Lauri K. Newman, and Luis Santos Soto. State-of-the-Art small spacecraft technology. NASA Technical Publication NASA/TP—20250000142, National Aeronautics and Space Administration (NASA), 2025. PDF: https://www.nasa.gov/wp-content/uploads/2025/02/soa-2024.pdf. 2

[26] Patrick Zippenfenig. Open-Meteo.com Weather API, 2024. Software. Repository: https://github.com/open-meteo/open-meteo. 3