FEDRASR: RISK-AWARE SOFT REWEIGHTING AGAINST PERSISTENT POISONING IN FEDERATED LOAD FORECASTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Federated learning enables collaborative load forecasting without pooling private smart-meter data, but persistent poisoning is difficult to detect because small malicious updates accumulate across rounds and are obscured by regional non-IID variation. To evaluate this threat, we propose Hist-Residual, which repeatedly perturbs profile-aware targets, and Profile-LIE, which additionally shapes uploads using benign round-level coordinate statistics. Existing robust aggregation rules rely on current-round geometry, and cross-round detectors may confuse regional heterogeneity with malicious temporal behavior. To address this, we develop a Federated Risk-Aware Soft Reweighting algorithm (FedRASR), a five-stage server-side defense that forms compact update features, constructs region-calibrated residual trajectories, derives FFT-based frequency evidence, estimates risk, and softly reweights the complete updates. Under the evaluated threat setting, experiments on four regional smart-meter collections with 1,200 clients show that FedRASR attains the lowest RMSE and MAE among the compared methods under both attacks, remains competitive under clean training, and benefits from both frequency scoring and soft reweighting.

1 INTRODUCTION

Power system operators must schedule generation and reserves before actual electricity demand is observed. Therefore, the performance of forecasting may lead to direct operational consequences. Underestimation may leave insufficient reserves during demand peaks, while overestimation may lead to unnecessary generation commitments and procurement costs. In a more general perspective, load forecasts support electricity-market operation and long-term infrastructure planning (Hong & Fan, 2016), because reliable and affordable electricity is closely related to economic productivity and household well-being (Fried & Lagakos, 2023; Welsch & Biermann, 2017).

Building accurate forecasting models, however, requires data that are distributed across utilities, regions, and households. These data cannot always be centrally collected because of privacy, ownership, and regulatory constraints. To this end, federated learning (FL) has emerged to enable multiple data owners to train a shared forecasting model while keeping raw data local (McMahan et al., 2017; Zuo et al., 2026; Wang et al., 2025). However, FL remains vulnerable to compromised clients. A compromised participant may be sampled in multiple communication rounds and manipulate its local data or uploaded model update whenever it participates. Such repeated client-side perturbations can accumulate in the global forecaster and degrade the predictions used for grid operation (Liu et al., 2024). For example, false-data injection and poisoning attacks have already been studied in smart-grid measurement systems (Li et al., 2022; Zhang & Sikdar, 2026). When persistent poisoning occurs, the threat becomes harder to detect, especially in regional load forecasting. This is because regional load series are not independently and identically distributed (non-IID) across clients. Differences in scale, seasonality, and consumption patterns produce substantial variation among benign client updates. Such legitimate heterogeneity can conceal malicious changes that are small in any individual round but accumulate over time.

To study persistent poisoning in federated load forecasting, we develop two forecasting-specific attacks under different adversarial capabilities. (i) Hist-Residual performs persistent data poisoning

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

through bounded target perturbations derived from each malicious client's historical load profile; the perturbation sign is set by the global communication round whenever that client is sampled. (ii) Profile-LIE assumes stronger aggregation-aware information access. It retains the profile-based target corruption and further shapes uploaded updates using benign round-level coordinate statistics available in the evaluation threat model (Baruch et al., 2019).

Existing defenses face a difficult trade-off in this setting. Krum and robust aggregation rules (coordinate-wise median, trimmed mean, and geometric median) rely primarily on the geometry or coordinate-wise distribution of updates within the current round (Blanchard et al., 2017; Yin et al., 2018; Pillutla et al., 2019). Cross-round detectors use historical update behavior to identify temporal inconsistencies (Zhang et al., 2022; 2024). However, an unusual update may reflect legitimate regional variation rather than malicious behavior, while a persistent attacker can distribute its influence across several individually inconspicuous updates. Therefore, a practical defense must distinguish stable regional heterogeneity from evolving malicious influence by analyzing ordered cross-round trajectories rather than individual updates. Fast Fourier Transform (FFT) provides a natural framework for this purpose: it efficiently decomposes an ordered trajectory into oscillatory components at different frequencies (Cooley & Tukey, 1965; Frigo & Johnson, 2005). Gradual poisoning affects the low-frequency components, while alternating perturbations strengthen specific frequency components. However, frequency decomposition alone cannot determine how suspicious an update is or how much influence it should retain. The evidence should be further calibrated against legitimate regional variation and translated into a continuous and smooth aggregation decision. Motivated by this connection, in this paper, we propose FedRASR, a Risk-Aware Soft Reweighting defense algorithm for federated regional load forecasting. It develops regional calibrations with FFT-based cross-round analysis to mitigate persistent poisoning under non-IID data while preserving forecasting performance.

FedRASR starts from a central insight. Persistent manipulation may remain inconspicuous in any one upload, but repeated or alternating behavior can leave structured traces across communication rounds. FedRASR operates in five stages. (i) Compact Update Representation extracts a small feature vector from each upload for efficient trajectory storage, while retaining the complete update for final aggregation. (ii) Source-Calibrated Residual Trajectory Construction removes a smoothed regional reference before linking the residual features across rounds. This calibration prevents stable regional differences from being mistaken for malicious behavior. (iii) By implementing FFT, Frequency-Domain Scoring then derives three complementary statistics from the resulting spectrum. High-frequency energy reflects rapid variation, peak concentration captures a dominant repetition rate, and entropy-based periodicity indicates whether power is concentrated in a small set of frequencies. (iv) Risk Estimation compares the resulting score with regional peers and incorporates directional consistency. (v) Finally, Soft-Reweighting and Aggregation converts the accumulated risk into a continuous contribution factor and uses it to aggregate the complete updates. FedRASR can therefore suppress persistent malicious influence without relying on brittle single-round filtering or immediately removing uncertain clients.

Our contributions are summarized as follows:

- We formulate persistent poisoning in federated load forecasting as a cross-round contribution-control problem. Regional non-IID behavior makes suspicious clients difficult to identify from one update alone.
- We develop two persistent attacks tailored to federated load forecasting. Hist-Residual repeatedly poisons profile-aware targets, while Profile-LIE adds aggregation-aware update shaping.
- We develop FedRASR as a server-side defense. It calibrates compact update features against regional references and analyzes the resulting trajectories in the frequency domain. Soft reweighting then limits suspicious contributions without hard client isolation.
- Experiments on four regional smart-meter collections and 1200 clients show that FedRASR achieves the lowest RMSE and MAE among the compared methods under both attacks. Its clean performance remains competitive, and the ablations support the roles of frequency scoring and soft reweighting.

2 Related Work

2.1 Federated Time-Series Forecasting

FedAvg provides a standard framework for collaborative training, and it has been applied to residential load forecasting without exchanging raw measurements (McMahan et al., 2017; Husnoo et al., 2024). Recent studies extend this setting to multi-source and cross-regional forecasting (Wang et al., 2025; Zuo et al., 2026). Other work considers privacy-preserving reasoning for heterogeneous IoT data (Ni et al., 2026) and robust partial-label FL under client heterogeneity (Fu et al., 2026).

These methods mainly address private collaboration and data heterogeneity. FedDAP further uses regional information to learn domain-specific prototypes (Le et al., 2026), but its goal is model adaptation rather than separating legitimate regional variation from persistent poisoning. FedRASR uses regional information only to calibrate update behavior and retains a single global forecaster.

2.2 Poisoning Attacks in Federated Learning

A compromised client may alter its local objective, optimizer, or uploaded update. Model-poisoning attacks can evade norm-based checks and exploit weaknesses in robust aggregation (Bhagoji et al., 2019). LIE uses benign update statistics to keep crafted coordinates near the benign distribution (Baruch et al., 2019). Persistent attacks show that adversarial effects may extend beyond the injection round (Liu et al., 2024). Targeted smart-grid attacks can also challenge robust aggregation (Zhang & Sikdar, 2026), while FUBA demonstrates that federated unlearning may provide another path for backdoor injection (Sheng et al., 2026).

Our attacks combine profile-dependent forecasting corruption with persistence across intermittent participation. Hist-Residual applies bounded target perturbations through normal local training. Profile-LIE further shapes the resulting update around benign coordinate statistics. The two settings therefore test persistent forecasting corruption under different adversarial capabilities.

2.3 Robust Aggregation

Krum selects an update close to the other submissions (Blanchard et al., 2017). Coordinate-wise median and trimmed mean reduce the effect of extreme coordinates (Yin et al., 2018), while geometric median aggregation estimates a robust center (Pillutla et al., 2019). These approaches rely on malicious updates being sufficiently separable from benign ones. Regional non-IID data weakens this assumption because legitimate updates may already be widely dispersed.

History-aware defenses use a different signal. FoolsGold compares client histories (Fung et al., 2018), FLDetector tests cross-round consistency, and FLTracer traces poisoning provenance (Zhang et al., 2022; 2024). FreqFed applies a DCT within each current-round model and filters suspicious low-frequency clusters (Fereidooni et al., 2024). FedRDF applies an FFT across clients for each model coordinate within the same round (Campos et al., 2024). FedRASR instead follows each client's region-calibrated residual trajectory across rounds and uses the resulting risk to softly reweight its complete update.

2.4 Temporal and Frequency Analysis

Ordered client histories reveal patterns hidden in a single-round update. FedA4 uses optimization trajectories for heterogeneity-aware adaptation (Zhao et al., 2026). Adaptive Bayesian learning also improves robustness under changing IIoT topologies (Chen et al., 2025b). Frequency-domain discrepancy exposes anomalies in IIoT sequences (Chen et al., 2025a). WAFFLE uses Fourier and wavelet features to detect malicious clients offline (Licciardi et al., 2025).

These studies show the value of temporal order and frequency representations, but they do not combine regional calibration with cross-round evidence for online aggregation. FedRASR fills this gap by converting calibrated trajectory risk into soft weights instead of rejecting a client for one unusual regional update.

3 METHOD

3.1 PROBLEM FORMULATION

We study federated load forecasting with N clients. Client $i \in \{1, \ldots, N\}$ has n_i observations in its chronological training segment and constructs M_i private windows $(x_{i,m}, y_{i,m})$. The input $x_{i,m}$ contains historical load observations, while $y_{i,m} \in \mathbb{R}^2$ contains the next two load values. Let F_θ denote the forecaster parameterized by θ. Client i minimizes

$$\mathcal{L}_i(\theta) = \frac{1}{M_i} \sum_{m=1}^{M_i} \|F_\theta(x_{i,m}) - y_{i,m}\|_2^2 . \tag{1}$$

The squared ℓ_2 norm sums error over the two forecast steps, and $1/M_i$ averages it across local windows. The observation count n_i is used for aggregation. It is distinct from M_i and the number of local mini-batches.

Clients have non-IID scales and temporal patterns but jointly train one forecaster without sharing raw windows. Let $\mathcal{R}$ be the set of represented regions and let $\zeta_i \in \mathcal{R}$ denote the region associated with client i. This metadata is used only for reference calibration. It is not a model input or prediction objective. This use follows the metadata precedent of FedDAP (Le et al., 2026).

At round r, the server selects $\mathcal{S}^{(r)}$, broadcasts $\theta^{(r-1)}$, and receives the complete update $\Delta_i^{(r)} = \theta_i^{(r)} - \theta^{(r-1)}$. Here, $\theta_i^{(r)}$ is obtained through local Adam optimization. Sample-count-weighted FedAvg (McMahan et al., 2017) uses $w_i^{(r)} = n_i / \sum_{j \in \mathcal{S}^{(r)}} n_j$ and updates

$$\theta_{\text{FedAvg}}^{(r)} = \theta^{(r-1)} + \sum_{i \in \mathcal{S}^{(r)}} w_i^{(r)} \Delta_i^{(r)}. \tag{2}$$

The base coefficient $w_i^{(r)}$ differs from the defense factor $a_i^{(r)}$ introduced later. Local optimization details are provided in the supplementary material.

3.2 THREAT MODEL AND CONSIDERED ATTACKS

Before training, a fixed subset $\mathcal{M} \subseteq \{1, \ldots, N\}$ is compromised. Only sampled members $\mathcal{M}^{(r)} = \mathcal{S}^{(r)} \cap \mathcal{M}$ attack at round r. This represents compromised meters, gateways, or regional data-processing clients that participate intermittently while remaining under adversarial control. Smart-grid false-data injection studies support the relevance of these targets (Li et al., 2022; Keçeci et al., 2025).

The server observes authenticated uploads and fixed regional associations, but not raw series, targets, or local batches. Attackers may corrupt local targets, optimization, or uploads. They cannot forge client identities or alter regional associations.

We consider two adversarial capabilities. Direct update manipulation alone does not represent target corruption that passes through normal local optimization (Bhagoji et al., 2019). Hist-Residual models this training-path attack by changing local targets according to the client's load profile. Profile-LIE retains the target corruption and further shapes the uploaded update using benign round-level coordinate statistics available in the evaluation setting (Baruch et al., 2019). It therefore assumes stronger information access; this does not imply that it must cause larger forecasting error.

Hist-Residual Poisoning. For malicious client i, let $y_i(t)$ be a scalar target component with historical mean $\mu_{i,y}$ and standard deviation $\sigma_{i,y}$. The attack constructs

$$\psi_i(t) = \text{clip}\left(\frac{y_i(t) - \mu_{i,y}}{\lambda_{\text{hist}} \sigma_{i,y} + \epsilon}, -1, 1 \right), \tag{3}$$

where $\text{clip}(x, l, u) = \min\{\max\{x, l\}, u\}$. The poisoned target is

$$\widetilde{y}_i^{(r)}(t) = \mathrm{clip}(y_i(t) + \chi_r \alpha \psi_i(t), 0, 1)\,. \tag{4}$$

The sign $\chi_r \in \{-1, +1\}$ alternates with the global communication-round index. Because only sampled malicious clients attack, a client's realized sign sequence is determined by the global rounds in which it participates. The parameter $\alpha > 0$ controls amplitude and λ_{hist} scales the historical deviation.

Profile-LIE Poisoning. A malicious client first trains on Hist-Residual targets and obtains $\Delta_i^{\mathrm{temp},(r)}$. For coordinate κ, the evaluation simulator supplies the exact mean $\mu_{\Delta,\kappa}^{(r)}$ and standard deviation $\sigma_{\Delta,\kappa}^{(r)}$ of the benign selected-client updates in that round. The displacement direction is $d_{i,\kappa}^{(r)} = \mathrm{sign}(\Delta_{i,\kappa}^{\mathrm{temp},(r)} - \mu_{\Delta,\kappa}^{(r)})$. It indicates whether the temporary coordinate lies below, at, or above the benign mean. The LIE-shaped coordinate is

$$\Delta_{i,\kappa}^{\mathrm{LIE},(r)} = \mu_{\Delta,\kappa}^{(r)} + z_{\mathrm{lie}} \sigma_{\Delta,\kappa}^{(r)} d_{i,\kappa}^{(r)}. \tag{5}$$

Stacking the coordinates gives $\Delta_i^{\mathrm{LIE},(r)}$. The submitted adversarial update is

$$\Delta_i^{\mathrm{adv},(r)} = (1 - \lambda_{\mathrm{mix}})\Delta_i^{\mathrm{temp},(r)} + \lambda_{\mathrm{mix}} \Delta_i^{\mathrm{LIE},(r)}. \tag{6}$$

For malicious clients, the server receives $\Delta_i^{(r)} = \Delta_i^{\mathrm{adv},(r)}$. The coefficient $\lambda_{\mathrm{mix}} \in [0, 1]$ controls the aggregation-aware component. Exact access assumptions and attack parameters are provided in the supplementary material.

3.3 Overview of FedRASR

FedRASR is a server-side defense that leaves local training unchanged and analyzes each client's update sequence across rounds. Figure 1 presents five stages. Compact Update Representation reduces the memory required for security analysis. Source-Calibrated Residual Trajectory Construction removes stable regional offsets. Frequency-Domain Scoring characterizes how residual behavior is organized across rounds. Risk Estimation compares this evidence with regional peers, and Soft-Reweighting and Aggregation controls the contribution of each complete update.

3.4 Compact Update Representation

Each selected client returns its complete update $\Delta_i^{(r)}$ and observation count n_i. Storing full updates across rounds would make trajectory memory scale with model size. The server therefore maps each update to a compact feature $f_i^{(r)} \in \mathbb{R}^p$ that combines the prediction-head update with a deterministic backbone sketch. This feature is used only for trajectory scoring. The original complete update remains unchanged for final aggregation. The construction and dimension of $f_i^{(r)}$ are given in the supplementary material.

3.5 Source-Calibrated Residual Trajectory Construction

Legitimate regional load patterns can create different update offsets, so compact features should not be compared with one global center. For region ζ, define $\mathcal{S}_\zeta^{(r)} = \{k \in \mathcal{S}^{(r)} \mid \zeta_k = \zeta\}$. The server computes $\widetilde{c}_\zeta^{(r)} = \mathrm{median}_{k \in \mathcal{S}_\zeta^{(r)}} f_k^{(r)}$ and smooths this median across rounds. Client i's residual feature is $e_i^{(r)} = f_i^{(r)} - c_{\zeta_i}^{(r)}$. This subtraction removes a shared regional offset without partitioning model training.

Let T be the maximum trajectory length and $L_r = \min\{T, r\}$. The aligned trajectory is $X_i^{(r)} = [e_i^{(r-L_r+1)}, \ldots, e_i^{(r)}] \in \mathbb{R}^{L_r \times p}$. Missing participation rounds use zero-filled positions on the global round axis. Frequency scoring starts after at least $T_{\min}$ observed residuals. The center update and missing-round handling are specified in the supplementary material.

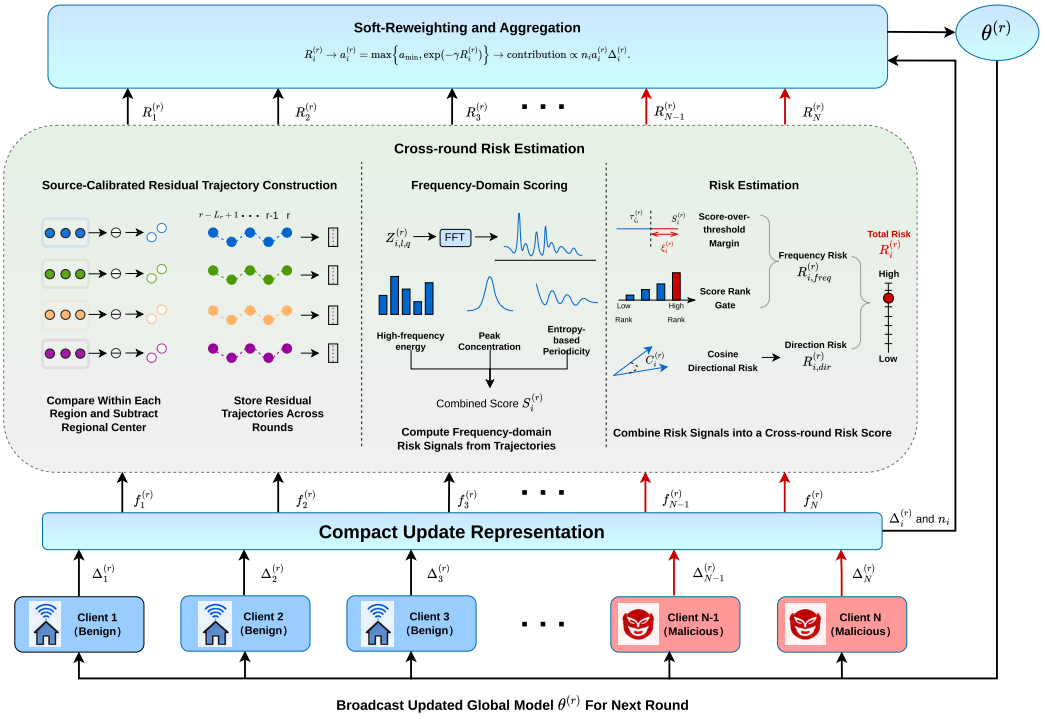

Figure 1: FedRASR and the five stages aligned with the Method subsections: Compact Update Representation, Source-Calibrated Residual Trajectory Construction, Frequency-Domain Scoring, Risk Estimation, and Soft-Reweighting and Aggregation. Trajectory colors denote the four represented regions. Red client icons are illustrative security annotations unavailable to the defense. Risk is converted to a soft weight before complete updates are aggregated.

3.6 Frequency-Domain Scoring

A single update cannot reveal how client behavior develops across rounds. The residual trajectory preserves this order, but a scalar time-domain variation score may not distinguish recurring manipulation from irregular variation. FedRASR therefore subtracts the temporal mean of each feature coordinate and applies the discrete Fourier transform along the communication-round axis (Cooley & Tukey, 1965).

Let $Z_{i,\ell,q}^{(r)}$ denote the temporally centered value of feature coordinate q at trajectory position ℓ. Its DFT is

$$\widehat{X}_{i,\nu,q}^{(r)} = \sum_{\ell=0}^{L_r-1} Z_{i,\ell,q}^{(r)} \exp\left(-\frac{2\pi\sqrt{-1}\nu\ell}{L_r}\right), \tag{7}$$

where ν indexes a discrete frequency bin and $\sqrt{-1}$ is the imaginary unit. For the positive-frequency bins $\mathcal{F}^+$, normalized power is

$$\varpi_{i,\nu,q}^{(r)} = \frac{|\widehat{X}_{i,\nu,q}^{(r)}|^2}{\sum_{\nu'\in\mathcal{F}^+} |\widehat{X}_{i,\nu',q}^{(r)}|^2 + \epsilon}. \tag{8}$$

The server derives high-frequency energy $\bar{H}_i^{(r)}$, peak concentration $\bar{P}_i^{(r)}$, and entropy-based periodicity $\bar{Q}_i^{(r)}$, averaged across feature coordinates. Their combined score is

$$S_i^{(r)} = \frac{\omega_H \bar{H}_i^{(r)} + \omega_P \bar{P}_i^{(r)} + \omega_Q \bar{Q}_i^{(r)}}{\omega_H + \omega_P + \omega_Q}. \tag{9}$$

High-frequency energy responds to rapid or alternating changes. Peak concentration identifies a dominant repetition rate, while entropy-based periodicity distinguishes concentrated structure from diffuse variation. Their coordinate-level definitions and weights are provided in the supplementary.

3.7 Risk Estimation

The same frequency score may have different significance in different regions. FedRASR therefore computes a score median $\mathrm{med}_\zeta^{(r)}$ and median absolute deviation $\mathrm{MAD}_\zeta^{(r)}$ from a regional reference set. The instantaneous threshold is

$$\widetilde{\tau}_\zeta^{(r)} = \mathrm{med}_\zeta^{(r)} + \lambda_{\mathrm{mad}} \, \mathrm{MAD}_\zeta^{(r)} . \tag{10}$$

The stored threshold $\tau_\zeta^{(r)}$ is smoothed across rounds. For client i, the positive margin is $\xi_i^{(r)} = \max\{0, S_i^{(r)} - \tau_{\zeta_i}^{(r)}\}$. The server also computes a normalized within-region rank $\rho_i^{(r)}$. This rank controls a score gate $b_i^{(r)} \in [0,1]$, giving

$$R_{i,\mathrm{freq}}^{(r)} = b_i^{(r)} \frac{\xi_i^{(r)}}{s_{\mathrm{scale}} + \epsilon}. \tag{11}$$

FedRASR supplements this score with directional evidence. $C_i^{(r)}$ denotes the cosine similarity between the current residual and the most recent earlier residual of client i. A stricter regional rank gate $u_i^{(r)} \in [0,1]$ limits directional evidence to high-ranked clients:

$$R_{i,\mathrm{dir}}^{(r)} = u_i^{(r)} \max\{0, \tau_{\cos} - C_i^{(r)}\}. \tag{12}$$

Missing or stale residuals contribute no directional risk. The total risk is

$$R_i^{(r)} = \min\Big\{ R_{\max}, R_{i,\mathrm{freq}}^{(r)} + \lambda_{\mathrm{dir}} R_{i,\mathrm{dir}}^{(r)} \Big\} . \tag{13}$$

Here, λ_{mad} controls the regional threshold, s_{scale} normalizes the score margin, $\tau_{\cos}$ sets the directional threshold, λ_{dir} controls directional evidence, and $R_{\max}$ caps total risk. Threshold smoothing, rank gates, and stale-round rules are detailed in the supplementary material.

3.8 Soft-Reweighting and Aggregation

FedRASR converts risk into the defense factor

$$a_i^{(r)} = \max\{a_{\min}, \exp(-\gamma R_i^{(r)})\}. \tag{14}$$

The parameter $\gamma > 0$ controls the rate of weight decay, while $a_{\min} \in (0,1]$ prevents irreversible exclusion. The defense factor rescales the sample-count contribution and is renormalized over selected clients:

$$\theta^{(r)} = \theta^{(r-1)} + \frac{\sum_{i \in \mathcal{S}^{(r)}} n_i a_i^{(r)} \Delta_i^{(r)}}{\sum_{j \in \mathcal{S}^{(r)}} n_j a_j^{(r)}}. \tag{15}$$

This equation aggregates complete updates rather than compact features. It reduces to sample-count-weighted FedAvg when all risks are zero. Higher risk therefore reduces a client's contribution smoothly without requiring binary removal.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Datasets. We evaluate secure federated time-series prediction using smart-meter data from Portugal, Australia, London, and Switzerland. The experiment names Portugal, Ausgrid, London, and Swiss refer to the processed regional collections, not complete official dataset titles. Public records document the collections and their provenance (Trindade, 2015; Ausgrid, 2013; UK Power Networks, 2015; Sawicki, 2023). We convert each collection to 30-minute resolution and retain 300 clients per region, giving 1200 clients in total. The sources are not centrally pooled; their geographic diversity supplies the cross-region heterogeneity that motivates regional calibration, while each meter remains a federated client with fixed regional metadata.

Federated forecasting task. Each client owns one smart-meter load series. The model receives only normalized time-series windows. Regional membership and client identity support data lookup, split construction, attribution, and server-side trajectory tracking. They are not model inputs. We use 48 input steps and predict the next 2 steps. All methods use the same temporal convolutional network backbone unless stated otherwise (Bai et al., 2018). Its channels are $(32, 64)$, with kernel size 3 and dropout 0.2.

Training configuration. Training lasts for 50 communication rounds. Each round samples 20% of the 1200 clients, giving 240 participants. A selected client performs 100 local optimization steps with batch size 128 and learning rate 0.001. The main comparison uses five random seeds, $\{1, 2, 3, 4, 5\}$. Each client series is split chronologically 80/10/10. For each seed, reported metrics use a fixed region-balanced evaluation subset shared across methods; the evaluation budget and split details are provided in the supplementary material.

Attack scenarios. We evaluate clean training, Hist-Residual poisoning, and Profile-LIE poisoning. Under attack, 30% of all clients form a fixed malicious pool, giving 360 malicious clients. Only sampled members attack in a given round. Hist-Residual perturbs forecasting targets using each malicious client's historical profile. Profile-LIE combines this target poisoning with LIE-style update shaping (Baruch et al., 2019). It therefore changes both local targets and the distribution of uploaded updates. The same fixed 30% malicious-client fraction and 20% participation rate are used across all compared methods.

Compared methods. We compare FedRASR with FedAvg (McMahan et al., 2017), Krum (Blanchard et al., 2017), Trimmed Mean (Yin et al., 2018), geometric median aggregation (Pillutla et al., 2019), FoolsGold (Fung et al., 2018), WAFFLE-FT (Licciardi et al., 2025), and IF-Detector, a current-round Isolation-Forest anomaly detector (Liu et al., 2008). All methods use the same client sampling, attack schedule, architecture, and evaluation protocol.

4.2 EVALUATION METRICS AND PROTOCOL

RMSE and MAE are computed in each client's normalized space; lower is better. For each seed, metrics are averaged over the fixed region-balanced evaluation subset within each region and then macro-averaged across regions. We report mean and standard deviation over five seeds, together with malicious aggregation contribution and group mean defense weights. Full definitions are in the supplementary material.

4.3 MAIN RESULTS

Table 1 reports forecasting errors under clean training and the two attacks. FedRASR records the lowest RMSE and MAE under both attacks while remaining close to the best clean performance. Relative to the strongest non-FedRASR baseline, RMSE decreases by 9.9% under Hist-Residual and 8.1% under Profile-LIE, while MAE decreases by 8.7% and 7.0%, respectively. On clean data, FedRASR differs from the best result by only 0.0002 in RMSE and 0.0003 in MAE. Per-region results in the supplementary material show that the largest improvement occurs on Portugal while performance remains close to the strongest regional baselines elsewhere. Representative forecast traces, the GeoMed counterpart, the trace-selection protocol, and forecast-region WAPE results are also provided in the supplementary material.

Table 1: Forecasting error across methods, reported as mean ± standard deviation over five seeds. Lower is better, and bold indicates the lowest mean in each column.

	RMSE			MAE		
Method	Clean	Hist-Residual	Profile-LIE	Clean	Hist-Residual	Profile-LIE
FedRASR (ours)	0.0874±0.0015	**0.0945±0.0054**	**0.0936±0.0052**	0.0623±0.0006	**0.0727±0.0053**	**0.0707±0.0050**
FedAvg	**0.0872±0.0017**	0.1109±0.0048	0.1070±0.0049	0.0621±0.0012	0.0928±0.0036	0.0871±0.0040
Krum	0.0959±0.0138	0.1110±0.0100	0.1087±0.0140	0.0664±0.0136	0.0796±0.0097	0.0760±0.0107
Trimmed Mean	0.0887±0.0018	0.1049±0.0060	0.1065±0.0079	0.0634±0.0012	0.0816±0.0051	0.0828±0.0070
GeoMed	0.0879±0.0016	0.1049±0.0063	0.1018±0.0050	0.0623±0.0011	0.0846±0.0058	0.0798±0.0042
FoolsGold	0.0893±0.0027	0.1086±0.0049	0.1074±0.0034	0.0639±0.0021	0.0897±0.0043	0.0869±0.0031
WAFFLE-FT	0.0873±0.0017	0.1075±0.0098	0.1068±0.0072	0.0621±0.0011	0.0903±0.0108	0.0883±0.0077
IF-Detector	0.0873±0.0016	0.1124±0.0124	0.1101±0.0095	**0.0620±0.0011**	0.0892±0.0118	0.0865±0.0091

4.4 Ablation Study

Table 2: Forecasting errors in the ablation study. Results are reported as mean ± standard deviation over five seeds. Lower is better, and bold indicates the lowest mean for each metric.

Attack	Variant	RMSE	MAE
Hist-Residual	Full	**0.0945±0.0054**	**0.0727±0.0053**
Hist-Residual	Hard isolation	0.0974±0.0055	0.0765±0.0051
Hist-Residual	Time-domain	0.0991±0.0055	0.0782±0.0045
Hist-Residual	Basic frame	0.1013±0.0051	0.0810±0.0039
Profile-LIE	Full	**0.0936±0.0052**	**0.0707±0.0050**
Profile-LIE	Hard isolation	0.0959±0.0050	0.0739±0.0046
Profile-LIE	Time-domain	0.0974±0.0044	0.0755±0.0031
Profile-LIE	Basic frame	0.0998±0.0045	0.0783±0.0034

We ablate two central design choices: frequency-domain versus time-domain trajectory scoring, and soft reweighting versus hard isolation. The time-domain variant changes only the score, the hard-isolation variant changes only the aggregation decision, and the basic-frame variant combines both reduced choices. Other components are held fixed.

Table 2 shows the same ordering under both attacks. The full method gives the lowest RMSE and MAE. Replacing frequency-domain scoring with the time-domain score causes a larger increase than replacing soft reweighting with hard isolation, while the basic-frame variant produces the largest errors. These results indicate that both components contribute to robustness, with frequency-domain scoring providing the larger individual improvement in this experiment.

Effect on aggregation. At the final round, full FedRASR reduces the malicious aggregation share from 0.1896 to 0.1037 under Hist-Residual and from 0.1931 to 0.1035 under Profile-LIE, while its average benign weights remain above 0.95. Complete mechanism statistics are reported in the supplementary material.

5 Conclusion

This paper studied persistent poisoning in heterogeneous federated smart-meter forecasting. In this setting, legitimate regional non-IID variation makes single-round update geometry unreliable. The more informative signal lies in how residualized client update features evolve across rounds. FedRASR captures this signal through frequency analysis and converts risk into soft aggregation weights. On four regional smart-meter collections, it achieves the lowest RMSE and MAE among the compared methods under both attacks while remaining competitive under clean training. Ablations support frequency-domain trajectory scoring and soft reweighting as the two central design choices studied here. Future work will extend the evaluation to broader attack schedules and participation regimes and investigate mask-aware trajectories under sparse participation.

AI USE STATEMENT

This paper only uses LLMs to polish text and not for any other purpose.

REFERENCES

Ausgrid. Solar home electricity data. Ausgrid data release, 2013. URL https://www.ausgrid.com.au/Industry/Our-Research/Data-to-share/Solar-home-electricity-data.

Shaojie Bai, J. Zico Kolter, and Vladlen Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271*, 2018.

Gilad Baruch, Moran Baruch, and Yoav Goldberg. A little is enough: Circumventing defenses for distributed learning. In *Advances in Neural Information Processing Systems*, 2019.

Arjun Nitin Bhagoji, Supriyo Chakraborty, Prateek Mittal, and Seraphin Calo. Analyzing federated learning through an adversarial lens. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 634–643, 2019.

Peva Blanchard, El Mahdi El Mhamdi, Rachid Guerraoui, and Julien Stainer. Machine learning with adversaries: Byzantine tolerant gradient descent. In *Advances in Neural Information Processing Systems*, 2017.

Enrique Mármol Campos, Aurora González Vidal, José Luis Hernández Ramos, and Antonio Skarmeta. FedRDF: A robust and dynamic aggregation function against poisoning attacks in federated learning. *arXiv preprint arXiv:2402.10082*, 2024. doi: 10.48550/arXiv.2402.10082.

Lei Chen, Xuxin Liu, Ying Zou, Jiajun Tang, Canwei Liu, Bowen Hu, and Mingyang Lv. Frequency-domain spectrum discrepancy-based fast anomaly detection for IIoT sensor time-series signals. *IEEE Transactions on Instrumentation and Measurement*, 74:1–16, 2025a. doi: 10.1109/TIM.2025.3554286.

Ning Chen, Songwei Zhang, Xiaobo Zhou, Song Cao, and Tie Qiu. Fast robustness enhancement for dynamic iiot topology with adaptive bayesian learning. *IEEE Transactions on Mobile Computing*, 24(10):10886–10899, 2025b. doi: 10.1109/TMC.2025.3571431.

James W. Cooley and John W. Tukey. An algorithm for the machine calculation of complex fourier series. *Mathematics of Computation*, 19(90):297–301, 1965.

Hossein Fereidooni, Alessandro Pegoraro, Phillip Rieger, Alexandra Dmitrienko, and Ahmad-Reza Sadeghi. FreqFed: A frequency analysis-based approach for mitigating poisoning attacks in federated learning. In *Proceedings of the Network and Distributed System Security Symposium*, 2024. doi: 10.14722/ndss.2024.23620.

Stephie Fried and David Lagakos. Electricity and firm productivity: A general-equilibrium approach. *American Economic Journal: Macroeconomics*, 15(4):67–103, 2023. doi: 10.1257/mac.20210248.

Matteo Frigo and Steven G. Johnson. The design and implementation of FFTW3. *Proceedings of the IEEE*, 93(2):216–231, 2005. doi: 10.1109/JPROC.2004.840301.

Sen Fu, Wei Bao, Zhengjie Yang, Xinyi Sheng, Yichen Guo, Yuqin Liu, and Bing Bing Zhou. Towards robust and fair partial label federated learning service. *IEEE Transactions on Services Computing*, pp. 1–12, 2026. doi: 10.1109/TSC.2026.3686458.

Clement Fung, Chris J. M. Yoon, and Ivan Beschastnikh. Mitigating sybils in federated learning poisoning. arXiv preprint arXiv:1808.04866, 2018.

Tao Hong and Shu Fan. Probabilistic electric load forecasting: A tutorial review. *International Journal of Forecasting*, 32(3):914–938, 2016. doi: 10.1016/j.ijforecast.2015.11.011.

Muhammad Akbar Husnoo, Adnan Anwar, Nasser Hosseinzadeh, Shama Naz Islam, Abdun Naser Mahmood, and Robin Doss. A secure federated learning framework for residential short-term load forecasting. *IEEE Transactions on Smart Grid*, 15(2):2044–2055, 2024.

Cihat Keçeci, Katherine R. Davis, and Erchin Serpedin. Federated learning-based distributed localization of false data injection attacks on smart grids. *IEEE Systems Journal*, 19(3):719–729, 2025. doi: 10.1109/JSYST.2025.3581524.

Huy Q. Le, Loc X. Nguyen, Yu Qiao, Seong Tae Kim, Eui-Nam Huh, and Choong Seon Hong. FedDAP: Domain-aware prototype learning for federated learning under domain shift. arXiv preprint arXiv:2604.06795, 2026. URL https://arxiv.org/abs/2604.06795.

Yang Li, Xinhao Wei, Yuanzheng Li, Zhaoyang Dong, and Mohammad Shahidehpour. Detection of false data injection attacks in smart grid: A secure federated deep learning approach. *IEEE Transactions on Smart Grid*, 13(6):4862–4872, 2022. doi: 10.1109/TSG.2022.3204796.

Alessandro Licciardi, Davide Leo, and Davide Carbone. Wavelet scattering transform and fourier representation for offline detection of malicious clients in federated learning. arXiv preprint arXiv:2506.09674, 2025.

Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation forest. In *2008 Eighth IEEE International Conference on Data Mining*, pp. 413–422. IEEE, 2008. doi: 10.1109/ICDM.2008.17.

Tao Liu, Yuhang Zhang, Zhu Feng, Zhiqin Yang, Chen Xu, Dapeng Man, and Wu Yang. Beyond traditional threats: A persistent backdoor attack on federated learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 21359–21367, 2024. doi: 10.1609/aaai.v38i19.30131.

H. Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pp. 1273–1282, 2017.

Tengjun Ni, Xin Yuan, Shenghong Li, and Wei Ni. Privacy-preserving multimodal reasoning for internet of things: A retrieval-augmented large language and vision assistant framework. *IEEE Internet of Things Magazine*, pp. 1–10, 2026. doi: 10.1109/MIOT.2026.3650753.

Krishna Pillutla, Sham M. Kakade, and Zaid Harchaoui. Robust aggregation for federated learning. arXiv preprint arXiv:1912.13445, 2019.

Benjamin Sawicki. Swiss Smart Meter Data - CKW 2021/2022 - Anonymized Individual Metering Points, February 2023. URL https://doi.org/10.5281/zenodo.7828796.

Xinyi Sheng, Wei Bao, Yichen Guo, and Sen Fu. FUBA: Backdoor federated learning via federated unlearning. *IEEE Transactions on Artificial Intelligence*, 7(5):2892–2907, 2026. doi: 10.1109/TAI.2025.3630110.

Artur Trindade. ElectricityLoadDiagrams20112014. UCI Machine Learning Repository, 2015. URL https://archive.ics.uci.edu/dataset/321/electricityloaddiagrams20112014.

UK Power Networks. Smartmeter energy consumption data in london households. London Datastore, 2015. URL https://data.london.gov.uk/dataset/smartmeter-energy-consumption-data-in-london-households-vqm0d.

Mengdi Wang, Rui Xin, Mingrui Xia, Zhifeng Zuo, Yinyin Ge, Pengfei Zhang, and Hongxing Ye. A federated LSTM network for load forecasting using multi-source data with homomorphic encryption. *AIMS Energy*, 13(2):265–289, 2025. doi: 10.3934/energy.2025011.

Heinz Welsch and Philipp Biermann. Energy affordability and subjective well-being: Evidence for european countries. *The Energy Journal*, 38(3):159–176, 2017. doi: 10.5547/01956574.38.3.hwel.

Dong Yin, Yudong Chen, Kannan Ramchandran, and Peter Bartlett. Byzantine-robust distributed learning: Towards optimal statistical rates. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, 2018.

Guihai Zhang and Biplab Sikdar. Targeted adversarial poisoning attack against robust aggregation in federated learning for smart grids. *IEEE Transactions on Smart Grid*, 17(2):1519–1530, 2026.

Xinyu Zhang, Qingyu Liu, Zhongjie Ba, Yuan Hong, Tianhang Zheng, Feng Lin, Li Lu, and Kui Ren. FLTracer: Accurate poisoning attack provenance in federated learning. *IEEE Transactions on Information Forensics and Security*, 19:9534–9549, 2024.

Zaixi Zhang, Xiaoyu Cao, Jinyuan Jia, and Neil Zhenqiang Gong. FLDetector: Defending federated learning against model poisoning attacks via detecting malicious clients. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 2545–2555. Association for Computing Machinery, 2022. doi: 10.1145/3534678.3539231.

Guanyi Zhao, Juntao Hu, Zhengjie Yang, and Dapeng Oliver Wu. FedA4: Federated learning with anti-bias aggregation and trajectory-based adaptation. *Transactions on Artificial Intelligence*, 2 (1):26–38, 2026. doi: 10.53941/tai.2026.100003.

Zhifeng Zuo, Hongxing Ye, Jie Li, and Yinyin Ge. A robust and heterogeneity-aware federated learning framework with knowledge distillation for cross-regional load forecasting. *IEEE Transactions on Smart Grid*, 2026. doi: 10.1109/TSG.2026.3660096. Early Access.

A Overview

This supplementary material provides implementation and reproducibility details for the main paper. The main paper defines the threat model, the two poisoning attacks, the proposed Risk-Aware Soft Reweighting defense, and the primary experimental results. This document expands the experimental protocol, including dataset construction, preprocessing, model and training settings, compact update features, defense hyperparameters, attack parameters, evaluation metrics, and additional mechanism statistics. An accompanying anonymous Code and Data Supplement provides the implementation code, machine-readable configurations, preprocessing and evaluation scripts, software requirements, and official data-source instructions. These materials support reproduction of the reported experiments and do not introduce claims beyond the main paper. A consolidated summary of the notation is provided in Tables S11 and S12.

B Related Work

B.1 Federated Forecasting under Data Heterogeneity

FedAvg provides the basic parameter-aggregation framework used by many federated forecasting systems (McMahan et al., 2017). Residential load forecasting has adopted FL to train shared predictors without directly exchanging household measurements (Husnoo et al., 2024). Multi-source and cross-regional methods further allow geographically distributed owners to retain their local data while learning a common forecasting model (Wang et al., 2025; Zuo et al., 2026).

Related studies address other forms of heterogeneity. Privacy-preserving reasoning has been studied for heterogeneous IoT data (Ni et al., 2026), while robust partial-label FL considers clients with incomplete or inconsistent supervision (Fu et al., 2026). FedDAP uses regional information to learn domain-specific prototypes under domain shift (Le et al., 2026). These methods improve collaboration or adaptation, but they do not treat regional information as a security reference for distinguishing legitimate update variation from persistent malicious behavior. FedRASR retains one global forecaster and uses regional context only when calibrating client update trajectories.

B.2 Poisoning Attacks

A malicious FL client may interfere with training at several points. It can corrupt local examples or targets, modify the local objective, alter the optimizer, or directly replace the uploaded update. Direct model-poisoning attacks demonstrate that crafted updates may evade simple norm checks and exploit assumptions made by robust aggregation (Bhagoji et al., 2019). LIE constructs malicious coordinates from estimates of benign round-level statistics. This keeps individual coordinates near the benign distribution while shifting the aggregate (Baruch et al., 2019).

Persistent poisoning extends the threat beyond a single injection round (Liu et al., 2024). Its effect may accumulate through several uploads even when no individual upload is extreme. In smart-grid settings, TAPA shows that targeted update construction can undermine robust aggregation (Zhang & Sikdar, 2026). FUBA identifies federated unlearning as another possible route for backdoor injection (Sheng et al., 2026).

Hist-Residual and Profile-LIE represent two adversarial capabilities in our forecasting setting. Hist-Residual corrupts forecasting targets using bounded perturbations derived from the malicious client's load profile. The corruption passes through normal local optimization and persists across intermittent participation. Profile-LIE retains this target corruption and further shapes the submitted update using exact benign selected-client coordinate statistics supplied by the evaluation simulator. The first attack therefore models persistent data poisoning through an ordinary training path, whereas the second also grants the attacker control over the final uploaded update.

B.3 Robust and History-Aware Defenses

Krum selects a submitted update that remains close to other updates (Blanchard et al., 2017). Coordinate-wise median and trimmed mean reduce the influence of extreme values independently in each model coordinate (Yin et al., 2018). Geometric median aggregation instead estimates a robust

center of the submitted updates (Pillutla et al., 2019). These defenses work best when malicious updates are geometrically distinct from a sufficiently concentrated benign majority. Regional non-IID forecasting can violate this condition. Legitimate updates may be dispersed because clients observe different demand patterns, while a shaped malicious update may remain inside that dispersed set.

History-aware methods use evidence from more than one communication round. FoolsGold assigns weights according to similarity among client histories (Fung et al., 2018). FLDetector evaluates cross-round consistency, and FLTracer analyzes the provenance of poisoning behavior (Zhang et al., 2022; 2024). Such methods establish the value of client history, although their evidence and decision rules differ from the region-calibrated trajectory scoring used by FedRASR. FLDetector is therefore retained as a representative cross-round detector in related work, while the experimental IF-Detector is a separate current-round Isolation-Forest anomaly-detection baseline (Liu et al., 2008).

FreqFed and FedRDF are particularly relevant because both use frequency-related representations in FL. FreqFed applies a DCT within each client's current-round model weights. It clusters the resulting low-frequency representations and filters suspicious clusters before aggregation (Fereidooni et al., 2024). Its transformed dimension lies inside one uploaded model.

FedRDF applies an FFT to the current-round values of each model coordinate across clients (Campos et al., 2024). The transform is therefore taken along the client axis within one round. The resulting representation is used to estimate where coordinate values concentrate and to exclude values considered malicious.

FedRASR transforms a different object. It first constructs a region-calibrated residual feature trajectory for each client and then applies the FFT along the communication-round axis. The resulting score represents cross-round client risk rather than within-model structure or current-round cross-client density. FedRASR also retains the original complete update. Risk modifies its aggregation contribution continuously instead of directly discarding the client or individual coordinate values.

B.4 Temporal and Frequency Representations

FedA4 uses client optimization trajectories instead of relying only on final model snapshots (Zhao et al., 2026). This supports heterogeneity-aware adaptation rather than poisoning defense. Related IIoT research applies adaptive Bayesian learning to improve robustness when the network topology changes (Chen et al., 2025b). These studies show that an ordered history can preserve behavior that is absent from one isolated observation.

Frequency analysis offers another view of an ordered sequence. Frequency-domain discrepancy has been used to detect sequence-level anomalies in IIoT sensor signals (Chen et al., 2025a). WAFFLE studies Fourier and wavelet representations for offline malicious-client detection (Licciardi et al., 2025). FedRASR differs by applying frequency analysis online to each client's cross-round residual trajectory and by using the resulting evidence in soft aggregation.

FreqFed and FedRDF are included as conceptual comparisons because their published frequency transforms operate on representation axes different from FedRASR. They are therefore not included in the shared numerical comparison, which focuses on methods evaluated under the common client-sampling, attack, forecasting, and aggregation protocol.

C Dataset Construction and Preprocessing

C.1 Regional Data Sources

We use four processed smart-meter sources: Portugal, Ausgrid, London, and Swiss. Each source is balanced to 300 clients, giving 1200 clients in total. Using geographically distinct sources creates cross-region heterogeneity in scale, seasonality, and consumption patterns that motivates regional calibration while preserving a single federated forecasting task. Table S1 summarizes the processed files.

The complete raw and processed datasets are not included in the Code and Data Supplement because their combined size exceeds the submission limit and the original sources have their own access and redistribution terms. Instead, the supplement provides the official source references, expected

Table S1: Processed data sources used in the experiments. All sources are converted to 30-minute resolution and balanced to 300 clients.

Region	Processed file	Clients used
Portugal	`portugal_2012_2014_30min_top300_LD2011_2014.csv`	300
Ausgrid	`ausgrid_2010_2013_GC_30min_top300.csv`	300
London	`london_2012_2014_30min_top300_LCL.csv`	300
Swiss	`swiss_2021_2022_30min_top300_CKW.csv`	300

filenames, preprocessing scripts, and instructions needed to reconstruct the processed experimental inputs after the original data have been obtained through their official channels. Missing values are handled during client construction by forward filling within each client series after the valid starting point is identified.

C.2 Client Construction, Splits, and Normalization

Each column in a processed CSV file corresponds to one client time series. For each client, we remove the leading all-zero region by starting from the first finite nonzero observation. The remaining sequence is split chronologically into training, validation, and test segments using an 80/10/10 split:

$$n_{\text{train}} = \lfloor 0.8n \rfloor,\ n_{\text{val}} = \lfloor 0.1n \rfloor,\ n_{\text{test}} = n - n_{\text{train}} - n_{\text{val}}.$$

The reported evaluation uses the validation segment of this split. Each client sequence is normalized by min–max scaling fitted only on that client's training segment. The model receives normalized time-series windows. Dataset and client identifiers are used only for data lookup, split construction, attribution, and server-side trajectory tracking; they are not provided as model inputs.

C.3 Client Identity and Regional Metadata

Let $\mathcal{R}$ denote the set of represented regions. Each client is constructed with a stable key (ζ_i, id_i), where id_i identifies the client time series and $\zeta_i \in \mathcal{R}$ denotes its region. The server-side scheduler stores this association before training and uses it to match each returned update to the selected client. Equivalently, a deployment may attach authenticated client and regional metadata to the contribution.

Regional membership is not inferred from the update tensor and is not provided to the forecasting model. It is used only to construct server-side reference statistics. We assume that client identities and their regional associations cannot be spoofed. Identity spoofing and Sybil attacks are outside the present threat model.

C.4 Forecasting Windows

Each example consists of an input window of 48 time steps and a forecast horizon of 2 time steps. Since all sources are converted to 30-minute resolution, the input window covers 24 hours and the model predicts the next hour. During local training, mini-batches are sampled from random windows of each selected client's normalized training series. During evaluation and plotting, fixed chronological windows from the evaluation segment are used.

D Federated Training and Model Configuration

D.1 Federated Protocol

The federated training configuration is summarized in Table S2. All compared methods use the same client sampling schedule, model architecture, local optimizer, attack schedule, and evaluation protocol.

Each selected client initializes its local model from the broadcast global model,

$$\theta_i^{(r,0)} = \theta^{(r-1)}. \tag{16}$$

Table S2: Federated training configuration used for the main comparison and ablation experiments.

Item	Value
Number of clients	1200
Clients per region	300
Communication rounds	50
Client sampling rate	20%
Approx. clients per round	240
Local optimization steps	100
Batch size	128
Evaluation batch size	512
Evaluation clients per region	10
Evaluation window cap per client	5000
Optimizer	Adam
Client learning rate	0.001
Server learning rate	none
Random seeds	1, 2, 3, 4, 5

It then performs E local optimization steps with Adam on mini-batches from its local training data:

$$\theta_i^{(r,h+1)} = \mathrm{AdamStep}\Big(\theta_i^{(r,h)}; \mathcal{B}_{i,h}, \eta_c\Big), \; h = 0, \ldots, E-1, \tag{17}$$

where h indexes the local optimization step, $\mathcal{B}_{i,h}$ is the corresponding mini-batch, and $\eta_c = 0.001$ is the client-side learning rate. The locally trained model is

$$\theta_i^{(r)} = \theta_i^{(r,E)}. \tag{18}$$

The uploaded model update is then defined as the parameter difference

$$\Delta_i^{(r)} = \theta_i^{(r)} - \theta^{(r-1)}. \tag{19}$$

Together with this complete update, the client returns $n_i = \max\{1, \texttt{client.n_train}\}$. Here, `client.n_train` is computed from the valid sequence length and chronological training split and counts observations in that training segment. It is not the number of local mini-batches or the number of sliding-window examples. Thus, the local regression objective in Eq. (1) of the main paper is minimized through these Adam steps. The resulting local model is then used to construct the parameter-difference update $\Delta_i^{(r)}$; Eq. (1) is the regression loss itself, not the parameter-difference definition.

No separate server optimizer or tuned server learning rate is used. The server directly applies the aggregated update produced by the selected aggregation rule,

$$\theta^{(r)} = \theta^{(r-1)} + \Delta_{\mathrm{agg}}^{(r)}. \tag{20}$$

Here $\Delta_{\mathrm{agg}}^{(r)}$ denotes the server-side aggregated update. For standard FedAvg this is $\Delta_{\mathrm{FedAvg}}^{(r)}$, while for the proposed method it is the defense-weighted update $\Delta_{\mathrm{def}}^{(r)}$. For consistency with the main paper, the standard FedAvg coefficient is

$$w_i^{(r)} = \frac{n_i}{\sum_{j \in \mathcal{S}^{(r)}} n_j}, \tag{21}$$

and $w_i^{(r)}$ is distinct from the defense soft weight $a_i^{(r)}$ defined later.

D.2 Software and Hardware Environment

The reported experiments were run on a Windows 11 workstation with an AMD Ryzen 9 8945HX processor and 32.0 GB RAM. The GPU was an NVIDIA GeForce RTX 5060 Laptop GPU with 8 GB memory, reported as 8151 MiB by the CUDA driver. The GPU driver version was 582.05. The operating system was Windows 11 Home, version 25H2, OS build 26200.8655. The software environment used Python 3.14.2, PyTorch 2.11.0 with CUDA 12.8 support, NumPy 2.4.4, and pandas 3.0.2. CUDA execution was enabled for the reported runs.

Table S3: Forecasting model configuration used by all methods.

Item	Value
Model type	TCN
Input channels	1
Input length	48
Output horizon	2
TCN channels	$(32, 64)$
Kernel size	3
Dropout	0.2
Prediction head	linear layer from 64 to 2

All experiments use fixed random seeds and write per-seed CSV logs, metadata files, and aggregated result tables. To reduce avoidable nondeterminism, the number of data-loading workers is set to zero. Each metadata file records the config path, dataset files, split ratios, model configuration, training parameters, attack parameters, defense parameters, selected methods, runtime seed, and device setting. The accompanying Code and Data Supplement includes the software requirements and machine-readable configurations used in the reported experiments.

D.3 Forecasting Model

All methods use the same temporal convolutional network (TCN) backbone. The model configuration is shown in Table S3. The TCN has two causal temporal blocks. Each block contains one-dimensional convolution layers, ReLU activations, dropout, and a residual connection. The final hidden state at the last input position is passed to a linear prediction head to produce the two-step forecast.

D.4 Baseline Evaluation Protocol

All baselines operate on client updates generated under the same federated training protocol. FedAvg uses sample-count-weighted averaging. Krum selects a single robust update according to pairwise update distances. Trimmed Mean applies coordinate-wise trimming before averaging. Geometric median aggregation computes a robust center of the uploaded updates. FoolsGold uses historical-similarity weighting; WAFFLE-FT is the Fourier-only current-round adaptation implemented in our code; and IF-Detector is an Isolation-Forest detector on normalized flattened updates. For all methods, the reported results use the same selected clients, malicious pool, attack schedule, model architecture, random seeds, and evaluation windows. This protocol isolates the server-side aggregation or filtering rule from differences in training setup.

D.5 Baseline Hyperparameters

All baseline methods are evaluated under the same client sampling, attack schedule, model architecture, local optimizer, training rounds, and evaluation protocol as the proposed method. The only difference is the server-side aggregation or filtering rule. Table S4 summarizes the baseline-specific settings used in our implementation.

E FedRASR Method Details

E.1 Compact Update Representation

The server converts each complete update $\Delta_i^{(r)}$ into a compact feature using the `head_plus_backbone` mode. The compact feature is used only by the defense scoring path. The original complete update remains unchanged for final aggregation.

The feature contains the flattened updates of `head.weight` and `head.bias`. For each non-head update tensor U, the implementation applies one-dimensional adaptive average pooling to produce

Table S4: Baseline-specific hyperparameters and implementation choices. All other training, attack, and evaluation settings are shared across methods.

Baseline	Implementation details
FedAvg	Aggregates client updates using training-segment observation counts.
Krum	Uses the standard squared-distance Krum score. The assumed Byzantine count is set automatically to $f_B = \lfloor \rho_{\mathrm{mal}} \lvert\mathcal{S}^{(r)}\rvert \rfloor$ in poisoning scenarios, where ρ_{mal} is the malicious-client ratio, and to 0 in the clean scenario.
Trimmed Mean	Applies coordinate-wise trimmed mean. The trimming parameter f_B is set using the same automatic rule as Krum. If $2f_B \geq \lvert\mathcal{S}^{(r)}\rvert$, the implementation falls back to the coordinate-wise mean.
GeoMed	Uses smoothed Weiszfeld geometric median aggregation with training-segment observation-count weights, maximum 8 iterations, and numerical floor $\epsilon = 10^{-6}$.
FoolsGold	Maintains a cumulative historical update vector for each client and assigns soft weights based on adjusted cosine similarity between historical update directions. Weights are normalized to $[10^{-3}, 1]$ before weighted aggregation.
WAFFLE-FT	Fourier-only current-round comparator motivated by frequency-based malicious-client detection; it computes the high-frequency energy ratio of each flattened current-round update using a real FFT and removes the top-f_B clients. The high-frequency fraction is 0.33, and f_B follows the same automatic rule as Krum.
IF-Detector	Uses Isolation Forest on normalized flattened current-round updates, with contamination set to $f_B/\lvert\mathcal{S}^{(r)}\rvert$, 200 trees, and a seed-specific random state. If more than f_B clients are flagged, only the top-f_B anomaly scores are removed.

four values and appends its root-mean-square magnitude

$$\frac{\|\operatorname{vec}(U)\|_2}{\sqrt{\operatorname{numel}(U)}}. \tag{22}$$

This construction is deterministic and does not use random projection. The TCN contains 130 prediction-head coordinates and 12 non-head tensors. Each non-head tensor contributes five values, producing $12 \times 5 = 60$ backbone coordinates. The final compact-feature dimension is therefore $p = 190$.

E.2 Source-Calibrated Residual Trajectory Construction

For represented region ζ, define

$$\mathcal{S}_\zeta^{(r)} = \{k \in \mathcal{S}^{(r)} : \zeta_k = \zeta\}. \tag{23}$$

The instantaneous coordinate-wise regional median is

$$\widetilde{c}_\zeta^{(r)} = \operatorname{median}_{k \in \mathcal{S}_\zeta^{(r)}} f_k^{(r)}. \tag{24}$$

The first observed median for region ζ initializes its stored center, so $c_\zeta^{(r)} = \widetilde{c}_\zeta^{(r)}$ when no earlier center exists. On later rounds, the implementation applies

$$c_\zeta^{(r)} = \beta_{\mathrm{center}} c_\zeta^{(r-1)} + (1 - \beta_{\mathrm{center}}) \widetilde{c}_\zeta^{(r)}, \tag{25}$$

where $\beta_{\mathrm{center}} = 0.80$. Client i's residual is

$$e_i^{(r)} = f_i^{(r)} - c_{\zeta_i}^{(r)}. \tag{26}$$

This subtraction removes stable regional offsets before security scoring. It does not divide model training into regional models.

The maximum trajectory length is $T = 40$, and the available global-round window is

$$L_r = \min\{T, r\}. \tag{27}$$

The aligned residual trajectory is

$$X_i^{(r)} = \left[e_i^{(r-L_r+1)}, \ldots, e_i^{(r)}\right] \in \mathbb{R}^{L_r \times p}. \tag{28}$$

A missing participation round is represented by a zero vector at its global-round position. This is a bookkeeping encoding of no upload in that round, not a locally computed residual. Frequency scoring starts only after client i has accumulated at least $T_{\min} = 5$ observed residuals.

E.3 Frequency-Domain Scoring

For $X_i^{(r)} \in \mathbb{R}^{L_r \times p}$, the temporal mean of each feature coordinate is removed:

$$Z_{i,\ell,q}^{(r)} = X_{i,\ell,q}^{(r)} - \frac{1}{L_r} \sum_{\ell'=0}^{L_r-1} X_{i,\ell',q}^{(r)}. \tag{29}$$

The discrete Fourier transform along the global-round axis is

$$\widehat{X}_{i,\nu,q}^{(r)} = \sum_{\ell=0}^{L_r-1} Z_{i,\ell,q}^{(r)} \exp\left(-\frac{2\pi\sqrt{-1}\nu\ell}{L_r}\right). \tag{30}$$

Here, ℓ indexes a trajectory position, q indexes a compact feature coordinate, and ν indexes a discrete frequency bin. The implementation uses a real FFT, removes the zero-frequency component, and retains the non-redundant positive-frequency bins $\mathcal{F}^+$. Normalized power is

$$\varpi_{i,\nu,q}^{(r)} = \frac{|\widehat{X}_{i,\nu,q}^{(r)}|^2}{\sum_{\nu' \in \mathcal{F}^+} |\widehat{X}_{i,\nu',q}^{(r)}|^2 + \epsilon}. \tag{31}$$

Let $\mathcal{F}^h \subseteq \mathcal{F}^+$ contain the highest 0.33 fraction of positive-frequency bins, with at least one bin. The coordinate-level high-frequency energy is

$$H_{i,q}^{(r)} = \sum_{\nu \in \mathcal{F}^h} \varpi_{i,\nu,q}^{(r)}. \tag{32}$$

Peak concentration is

$$P_{i,q}^{(r)} = \max_{\nu \in \mathcal{F}^+} \varpi_{i,\nu,q}^{(r)}. \tag{33}$$

Entropy-based periodicity is

$$Q_{i,q}^{(r)} = 1 + \frac{\sum_{\nu \in \mathcal{F}^+} \varpi_{i,\nu,q}^{(r)} \log\left(\varpi_{i,\nu,q}^{(r)} + \epsilon\right)}{\log|\mathcal{F}^+|}. \tag{34}$$

The server averages these values over the p feature coordinates to obtain $\bar{H}_i^{(r)}$, $\bar{P}_i^{(r)}$, and $\bar{Q}_i^{(r)}$. Their combined score is

$$S_i^{(r)} = \frac{\omega_H \bar{H}_i^{(r)} + \omega_P \bar{P}_i^{(r)} + \omega_Q \bar{Q}_i^{(r)}}{\omega_H + \omega_P + \omega_Q}. \tag{35}$$

The reported configuration uses $\omega_H = 1.0$, $\omega_P = 0.7$, and $\omega_Q = 0.3$.

E.4 Risk Estimation

The regional reference set first includes selected clients whose stored soft weight is at least 0.85. If at least 12 such clients are available in region ζ, their scores form $\mathcal{I}_\zeta^{(r)}$. Otherwise, $\mathcal{I}_\zeta^{(r)} = \mathcal{S}_\zeta^{(r)}$.

The regional score median is

$$\mathrm{med}_\zeta^{(r)} = \mathrm{median}_{k \in \mathcal{I}_\zeta^{(r)}} S_k^{(r)}. \tag{36}$$

The median absolute deviation is

$$\mathrm{MAD}_\zeta^{(r)} = \mathrm{median}_{k \in \mathcal{I}_\zeta^{(r)}} \left| S_k^{(r)} - \mathrm{med}_\zeta^{(r)} \right|. \tag{37}$$

The instantaneous threshold is

$$\widetilde{\tau}_\zeta^{(r)} = \mathrm{med}_\zeta^{(r)} + \lambda_{\mathrm{mad}} \, \mathrm{MAD}_\zeta^{(r)}. \tag{38}$$

When a region is first observed, its stored threshold is initialized from $\widetilde{\tau}_\zeta^{(r)}$. Client risk in round r uses the stored threshold $\tau_\zeta^{(r)}$. The positive score margin is

$$\xi_i^{(r)} = \max\{0, S_i^{(r)} - \tau_{\zeta_i}^{(r)}\}. \tag{39}$$

After clients have been scored, the threshold for the next round is

$$\tau_\zeta^{(r+1)} = \beta_\tau \tau_\zeta^{(r)} + (1 - \beta_\tau) \widetilde{\tau}_\zeta^{(r)}. \tag{40}$$

The zero-based ascending regional rank is

$$\rho_i^{(r)} = \frac{\mathrm{rank}_{\zeta_i}(S_i^{(r)})}{\max\{|\mathcal{S}_{\zeta_i}^{(r)}| - 1, 1\}}. \tag{41}$$

The score-rank gate is

$$b_i^{(r)} = \mathrm{clip}\left(\frac{\rho_i^{(r)} - \rho_b^{\mathrm{lo}}}{\rho_b^{\mathrm{hi}} - \rho_b^{\mathrm{lo}} + \epsilon}, 0, 1 \right). \tag{42}$$

The frequency-risk component is

$$R_{i,\mathrm{freq}}^{(r)} = b_i^{(r)} \frac{\xi_i^{(r)}}{s_{\mathrm{scale}} + \epsilon}. \tag{43}$$

Let $r_i^- < r$ denote the latest earlier round in which client i participated. The cosine directional signal is

$$C_i^{(r)} = \frac{\langle e_i^{(r)}, e_i^{(r_i^-)} \rangle}{\|e_i^{(r)}\|_2 \|e_i^{(r_i^-)}\|_2 + \epsilon}. \tag{44}$$

The stricter directional gate is

$$u_i^{(r)} = \mathrm{clip}\left(\frac{\rho_i^{(r)} - \rho_u^{\mathrm{lo}}}{\rho_u^{\mathrm{hi}} - \rho_u^{\mathrm{lo}} + \epsilon}, 0, 1\right). \tag{45}$$

Directional risk is

$$R_{i,\mathrm{dir}}^{(r)} = u_i^{(r)} \max\{0, \tau_{\cos} - C_i^{(r)}\}. \tag{46}$$

Directional risk is set to zero when no previous residual exists or when $r - r_i^- > L_{\mathrm{stale}}$, where $L_{\mathrm{stale}} = 10$. If either residual norm is below the numerical tolerance, the implementation sets $C_i^{(r)} = 1$.

The capped total risk is

$$R_i^{(r)} = \min\left\{R_{\max}, R_{i,\mathrm{freq}}^{(r)} + \lambda_{\mathrm{dir}} R_{i,\mathrm{dir}}^{(r)}\right\}. \tag{47}$$

E.5 Soft-Reweighting and Aggregation

The defense weight is

$$a_i^{(r)} = \max\{a_{\min}, \exp(-\gamma R_i^{(r)})\}, \tag{48}$$

where $\gamma = 1.8$, $a_{\min} = 10^{-4}$, and $R_{\max} = 15.0$. The defense-weighted aggregated update is

$$\Delta_{\mathrm{def}}^{(r)} = \frac{\sum_{i \in \mathcal{S}^{(r)}} n_i a_i^{(r)} \Delta_i^{(r)}}{\sum_{j \in \mathcal{S}^{(r)}} n_j a_j^{(r)}}. \tag{49}$$

No separate server optimizer is used, so

$$\theta^{(r)} = \theta^{(r-1)} + \Delta_{\mathrm{def}}^{(r)}. \tag{50}$$

The update $\Delta_i^{(r)}$ is the complete client update. The compact feature $f_i^{(r)}$ affects only risk estimation. When every $a_i^{(r)} = 1$, the method reduces to sample-count-weighted FedAvg.

E.6 Computational Overhead

For each tracked client, FedRASR stores a trajectory containing $T \times p$ floating-point values. The reported configuration uses $T = 40$ and $p = 190$. This is substantially smaller than retaining complete model updates across rounds. The remaining operations consist of regional coordinate-wise statistics, an FFT over the trajectory window, risk evaluation, and scalar reweighting.

E.7 End-to-End FedRASR Procedure

Algorithm 1 summarizes the server-side procedure. Local training is unchanged. The symbol ζ_i denotes client i's region, while j is used only as a summation index.

F Attack Implementation Details

F.1 Attack Motivation and Scope

The attack design starts from a persistent threat in federated smart-meter forecasting rather than from the defense components. In large-scale smart-meter or industrial IoT settings, a compromised

Algorithm 1 FedRASR server update at round r

Require: Global model $\theta^{(r-1)}$, selected clients $\mathcal{S}^{(r)}$, regional associations $\{\zeta_i\}_{i=1}^{N}$, trajectory length T, and minimum observation count $T_{\min}$

for each $i \in \mathcal{S}^{(r)}$ in parallel **do**
  Receive complete update $\Delta_i^{(r)}$ and count n_i
  Extract compact feature $f_i^{(r)}$
end for
for each represented region $\zeta \in \mathcal{R}$ **do**
  Compute $\widetilde{c}_\zeta^{(r)}$
  Update the smoothed center $c_\zeta^{(r)}$
end for
for each $i \in \mathcal{S}^{(r)}$ **do**
  Compute $e_i^{(r)} = f_i^{(r)} - c_{\zeta_i}^{(r)}$
  Update the aligned trajectory $X_i^{(r)}$
  if client i has at least $T_{\min}$ observed residuals **then**
    Compute frequency score $S_i^{(r)}$
  else
    Set $S_i^{(r)} \leftarrow 0$
  end if
end for
for each represented region $\zeta \in \mathcal{R}$ **do**
  Form the previous-weight reference set
  Compute the score median, MAD, and rank gates
  Compute $\widetilde{\tau}_\zeta^{(r)}$
  Initialize an unseen stored threshold if required
end for
for each $i \in \mathcal{S}^{(r)}$ **do**
  Compute $R_{i,\mathrm{freq}}^{(r)}$
  Compute $R_{i,\mathrm{dir}}^{(r)}$
  Compute total risk $R_i^{(r)}$
  Set $a_i^{(r)} = \max\{a_{\min}, \exp(-\gamma R_i^{(r)})\}$
end for
for each represented region $\zeta \in \mathcal{R}$ **do**
  Update $\tau_\zeta^{(r+1)}$
end for

$$\theta^{(r)} \leftarrow \theta^{(r-1)} + \frac{\sum_{i \in \mathcal{S}^{(r)}} n_i a_i^{(r)} \Delta_i^{(r)}}{\sum_{j \in \mathcal{S}^{(r)}} n_j a_j^{(r)}}$$

return $\theta^{(r)}$

meter, gateway, or regional data-processing endpoint may participate intermittently while remaining under adversarial control across many communication rounds. It can repeatedly corrupt its own multi-step forecasting targets in a way that follows its historical load profile. Each resulting update may remain bounded in one round, while the accumulated contribution biases the global forecaster over repeated participation. Here, persistent refers to the continued control and repeated activity of a fixed malicious pool, not to a claim that the effect necessarily remains after the attack stops.

Hist-Residual poisoning represents this client-local temporal manipulation by perturbing normalized forecasting targets according to each malicious client's centered load-deviation profile. Profile-LIE assumes a stronger information-access capability: it combines the same temporal target manipulation with update shaping based on benign round-level statistics supplied by the simulation environment. The term "stronger" refers to adversarial access, not guaranteed forecasting damage; blending toward benign coordinate statistics can also constrain the submitted perturbation. Profile-

LIE is therefore interpreted as an aggregation-aware stress-test setting with stronger information access.

These threat capabilities motivate the defense components. Regional calibration prevents regional heterogeneity from dominating the reference score, cross-round trajectory scoring captures repeated behavior, and soft reweighting converts the resulting evidence into continuous contribution control. Together, the two attacks cover client-local target corruption and an aggregation-aware update-shaping extension under the defined threat model.

F.2 Persistent Malicious Pool

Under attack, 30% of all clients are assigned to a persistent malicious pool. With 1200 total clients, this gives 360 malicious clients. Malicious identities persist across communication rounds, but a malicious client is active only when sampled in the current round. This fixed malicious-client fraction is used consistently across attacks and compared methods. The clean setting uses the same federated training protocol but no active malicious clients.

F.3 Hist-Residual Poisoning

Hist-Residual perturbs the local forecasting targets of malicious clients. For malicious client i, let $y_i(t)$ denote a scalar target component from its normalized forecasting windows. Its historical target mean and standard deviation are denoted by $\mu_{i,y}$ and $\sigma_{i,y}$. The attack constructs

$$\psi_i(t) = \mathrm{clip}\left(\frac{y_i(t) - \mu_{i,y}}{\lambda_{\mathrm{hist}} \sigma_{i,y} + \epsilon}, -1, 1 \right). \tag{51}$$

Here, $\mathrm{clip}(x, l, u) = \min\{\max\{x, l\}, u\}$ restricts x to $[l, u]$. The poisoned target is

$$\widetilde{y}_i^{(r)}(t) = \mathrm{clip}(y_i(t) + \chi_r \alpha \psi_i(t), 0, 1). \tag{52}$$

The reported configuration uses $\alpha = 0.25$ and $\lambda_{\mathrm{hist}} = 1.5$. The sign $\chi_r \in \{-1, +1\}$ is determined by the global communication-round index (odd versus even round). A malicious client applies the sign of the round only when it is sampled, so the realized sequence for an intermittently participating client is a subsequence of the global-round schedule.

F.4 Profile-LIE Poisoning

Profile-LIE retains Hist-Residual target poisoning and further shapes the resulting model update. A malicious client first trains on the poisoned targets and obtains the temporary full-model update $\Delta_i^{\mathrm{temp},(r)}$. The simulator supplies the coordinate-wise mean and standard deviation of benign round-level updates, denoted by $\mu_{\Delta,\kappa}^{(r)}$ and $\sigma_{\Delta,\kappa}^{(r)}$. The attack computes

$$d_{i,\kappa}^{(r)} = \mathrm{sign}\Big(\Delta_{i,\kappa}^{\mathrm{temp},(r)} - \mu_{\Delta,\kappa}^{(r)} \Big). \tag{53}$$

When the difference is exactly zero, the implementation deterministically breaks the tie using the sign of the temporary coordinate, falling back to +1 if that coordinate is also zero. The sign function returns -1, 0, or $+1$ according to the sign of its argument. It determines the displacement direction relative to the benign coordinate mean. The shaped coordinate is

$$\Delta_{i,\kappa}^{\mathrm{LIE},(r)} = \mu_{\Delta,\kappa}^{(r)} + z_{\mathrm{lie}} \sigma_{\Delta,\kappa}^{(r)} d_{i,\kappa}^{(r)}. \tag{54}$$

Stacking these coordinates gives the complete shaped update $\Delta_i^{\mathrm{LIE},(r)}$. The attack then forms

$$\Delta_i^{\mathrm{adv},(r)} = (1 - \lambda_{\mathrm{mix}}) \Delta_i^{\mathrm{temp},(r)} + \lambda_{\mathrm{mix}} \Delta_i^{\mathrm{LIE},(r)}. \tag{55}$$

For a malicious client, the server receives $\Delta_i^{(r)} = \Delta_i^{\mathrm{adv},(r)}$. The reported configuration is

$$z_{\mathrm{lie}} = 2.0, \qquad \lambda_{\mathrm{mix}} = 0.42, \qquad \text{norm clipping disabled (scale = 0.0).}$$

In the released implementation, these statistics are computed from the benign selected-client updates in the current round and supplied to the attack generator by the simulation environment. This represents simulation-level aggregation-aware access beyond ordinary client-local knowledge. Because the final update is blended with a benign-statistics-shaped update ($\lambda_{\mathrm{mix}} = 0.42$), it need not produce larger forecasting degradation than Hist-Residual for every baseline.

Table S5: Active FedRASR and attack hyperparameters used in the reported experiments. The machine-readable configurations are included in the Code and Data Supplement.

Group	Parameter	Value
Trajectory	maximum window length T	40
Trajectory	minimum observed updates $T_{\min}$	5
Trajectory	stale-round limit L_{stale}	10
Feature	feature mode	head_plus_backbone
Feature	backbone sketch dimension	4
Feature	compact feature dimension p	190
Centering	EMA coefficient β_{center}	0.80
FFT	high-frequency fraction	0.33
FFT	coordinate aggregation	mean
Score weights	$(\omega_H, \omega_P, \omega_Q)$	$(1.0, 0.7, 0.3)$
Reference set	previous soft-weight cutoff	0.85
Reference set	minimum reference clients	12
Robust statistics	threshold scale λ_{mad}	4.5
Soft risk	score scale s_{scale}	0.06
Soft risk	frequency-rank gate	0.68–0.86
Soft risk	directional-rank gate	0.50–0.80
Soft risk	cosine threshold $\tau_{\cos}$	0.0
Soft risk	directional weight λ_{dir}	1.5
Threshold EMA	coefficient β_τ	0.95
Soft weight	decay coefficient γ	1.8
Soft weight	risk cap $R_{\max}$	15.0
Soft weight	floor $a_{\min}$	10^{-4}
Attack	malicious-client ratio	0.30
Attack	persistent malicious clients	360
Hist-Residual	α	0.25
Hist-Residual	λ_{hist}	1.5
Profile-LIE	z_{lie}	2.0
Profile-LIE	λ_{mix}	0.42
Profile-LIE norm clipping disabled (scale = 0.0)		

G HYPERPARAMETERS

Table S5 reports the active defense and attack parameters used in the experiments. Diagnostic-only quantities that do not affect the soft-reweighted aggregation path are retained in the machine-readable logs but are not listed as method parameters.

H EVALUATION METRICS

The primary forecasting metrics are root mean squared error (RMSE) and mean absolute error (MAE). All reported RMSE and MAE values are computed in each client's normalized space, using min–max parameters fitted exclusively on that client's training segment. For N_{eval} evaluated scalar target values, they are

$$\text{RMSE} = \sqrt{\frac{1}{N_{\text{eval}}} \sum_{m=1}^{N_{\text{eval}}} (\widehat{y}_m - y_m)^2}, \tag{56}$$

$$\text{MAE} = \frac{1}{N_{\text{eval}}} \sum_{m=1}^{N_{\text{eval}}} |\widehat{y}_m - y_m| . \tag{57}$$

Here, y_m and $\widehat{y}_m$ are the observed and predicted target values. RMSE gives more influence to large forecasting errors. MAE measures the average absolute deviation and is less sensitive to isolated large errors. The two metrics provide complementary views of forecasting quality.

For each method and scenario, both metrics are computed separately for each of the four regional data sources and then macro-averaged with equal source weights. The reported mean and standard deviation are computed over five random seeds. This prevents a source with a larger scale or more clients from dominating the result.

For mechanism analysis, the effective malicious aggregation contribution is

$$\mathrm{Share}_{\mathrm{mal}}^{(r)} = \frac{\sum_{i \in \mathcal{S}^{(r)} \cap \mathcal{M}} n_i a_i^{(r)}}{\sum_{j \in \mathcal{S}^{(r)}} n_j a_j^{(r)}}. \tag{58}$$

Here, n_i is the number of observations in client i's chronological training segment and $a_i^{(r)}$ is its defense factor. The metric sums the final normalized aggregation coefficients assigned to all selected malicious clients in round r. It is not a per-client score. This quantity is reported only as a post-hoc mechanism diagnostic and does not affect risk estimation, soft weights, or model updates.

For a nonempty selected group, the round-level mean defense factors assigned to malicious and benign clients are

$$\overline{a}_{\mathrm{mal}}^{(r)} = \frac{1}{\left|\mathcal{S}^{(r)} \cap \mathcal{M}\right|} \sum_{i \in \mathcal{S}^{(r)} \cap \mathcal{M}} a_i^{(r)}, \tag{59}$$

$$\overline{a}_{\mathrm{ben}}^{(r)} = \frac{1}{\left|\mathcal{S}^{(r)} \setminus \mathcal{M}\right|} \sum_{i \in \mathcal{S}^{(r)} \setminus \mathcal{M}} a_i^{(r)}. \tag{60}$$

These quantities average the client-level defense factors before sample-count weighting and normalization. They therefore differ from $\mathrm{Share}_{\mathrm{mal}}^{(r)}$, which measures the malicious group's final normalized aggregation contribution.

Table S9 reports these three mechanism statistics at the final communication round only. Let $R = 50$ denote the final round and $K = 5$ the number of random seeds. For

$$x \in \{\mathrm{Share}_{\mathrm{mal}}, \overline{a}_{\mathrm{mal}}, \overline{a}_{\mathrm{ben}}\},$$

Let $x_s^{(R)}$ denote the corresponding final-round value from seed s. The reported mean and sample standard deviation, with $K - 1$ used in the denominator of the latter, are

$$\overline{x} = \frac{1}{K} \sum_{s=1}^{K} x_s^{(R)}, \tag{61}$$

$$\mathrm{SD}(x) = \sqrt{\frac{1}{K-1} \sum_{s=1}^{K} \left(x_s^{(R)} - \overline{x}\right)^2}. \tag{62}$$

For the representative forecast panels, we additionally report forecast-region weighted absolute percentage error:

$$\mathrm{WAPE} = 100\% \frac{\sum_{t \in \mathcal{T}_{\mathrm{pred}}} |\widehat{y}_t - y_t|}{\sum_{t \in \mathcal{T}_{\mathrm{pred}}} |y_t|}. \tag{63}$$

WAPE is used for the displayed forecast regions because pointwise percentage errors become unstable when observed loads approach zero.

I Additional Experimental Results

I.1 Per-Source RMSE

Tables S6–S8 report the seed-averaged values visualized in Fig. S1. They complement the main paper's macro-average by showing whether robustness gains are concentrated in a single source. Values in Tables S6–S8 are rounded to three decimal places for display; the macro-averages reported in the main paper are computed from the unrounded per-seed regional metrics.

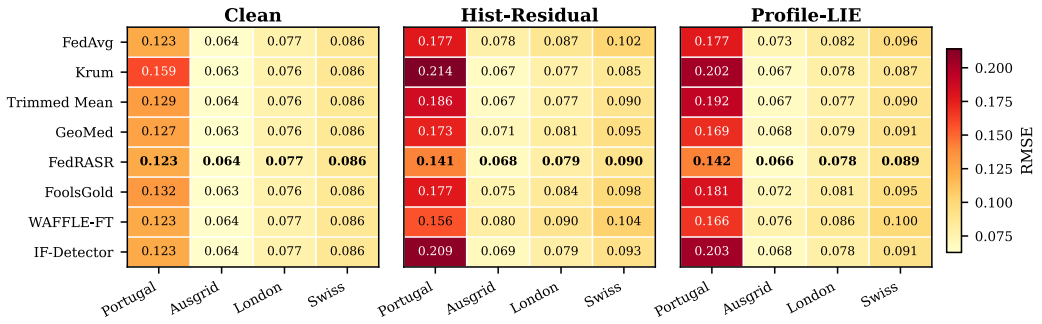

Figure S1: Per-source RMSE under clean training, Hist-Residual poisoning, and Profile-LIE poisoning. The common color scale exposes regional variation hidden by the macro-average in the main results table; seed-averaged values are reported in Tables S6–S8. Bold text highlights FedRASR for visual identification and does not denote the minimum value in each cell.

Table S6: Seed-averaged per-source RMSE under clean training. Lower is better.

Method	Portugal	Ausgrid	London	Swiss
FedAvg	0.123	0.064	0.077	0.086
Krum	0.159	0.063	0.076	0.086
Trimmed Mean	0.129	0.064	0.076	0.086
GeoMed	0.127	0.063	0.076	0.086
FedRASR	0.123	0.064	0.077	0.086
FoolsGold	0.132	0.063	0.076	0.086
WAFFLE-FT	0.123	0.064	0.077	0.086
IF-Detector	0.123	0.064	0.077	0.086

Table S7: Seed-averaged per-source RMSE under Hist-Residual poisoning. Lower is better.

Method	Portugal	Ausgrid	London	Swiss
FedAvg	0.177	0.078	0.087	0.102
Krum	0.214	0.067	0.077	0.085
Trimmed Mean	0.186	0.067	0.077	0.090
GeoMed	0.173	0.071	0.081	0.095
FedRASR	0.141	0.068	0.079	0.090
FoolsGold	0.177	0.075	0.084	0.098
WAFFLE-FT	0.156	0.080	0.090	0.104
IF-Detector	0.209	0.069	0.079	0.093

Table S8: Seed-averaged per-source RMSE under Profile-LIE poisoning. Lower is better.

Method	Portugal	Ausgrid	London	Swiss
FedAvg	0.177	0.073	0.082	0.096
Krum	0.202	0.067	0.078	0.087
Trimmed Mean	0.192	0.067	0.077	0.090
GeoMed	0.169	0.068	0.079	0.091
FedRASR	0.142	0.066	0.078	0.089
FoolsGold	0.181	0.072	0.081	0.095
WAFFLE-FT	0.166	0.076	0.086	0.100
IF-Detector	0.203	0.068	0.078	0.091

Table S9: Aggregation-mechanism statistics for the ablation study. Hist. and Prof. denote Hist-Residual and Profile-LIE. Mal. share is the combined final normalized aggregation contribution of all selected malicious clients after sample-count weighting and renormalization. Mal. wt. and Benign wt. are the average client-level defense factors $a_i^{(r)}$ assigned to selected malicious and benign clients before normalization. For each seed, all three statistics are evaluated at the final communication round $R = 50$. Reported values are the mean and sample standard deviation over five seeds. Lower Mal. share and Mal. wt. are better, while a Benign wt. near one indicates that benign clients remain largely unsuppressed.

Attack	Variant	Mal. share	Mal. wt.	Benign wt.
Hist.	Full	**0.1037±0.0346**	**0.2573±0.0401**	0.9592±0.0048
Hist.	Hard iso.	0.1426±0.0414	0.3862±0.0514	0.9917±0.0067
Hist.	Time-dom.	0.1590±0.0305	0.3842±0.0236	0.8517±0.0263
Hist.	Basic frame	0.1896±0.0359	0.4865±0.0308	0.8748±0.0315
Prof.	Full	**0.1035±0.0370**	**0.2565±0.0448**	0.9608±0.0042
Prof.	Hard iso.	0.1410±0.0389	0.3811±0.0439	0.9905±0.0067
Prof.	Time-dom.	0.1629±0.0244	0.3949±0.0158	0.8542±0.0250
Prof.	Basic frame	0.1931±0.0309	0.5028±0.0273	0.8819±0.0283

I.2 Ablation Mechanism Statistics

The forecasting-error ablation table is reported in the main paper. The ablation isolates the two central design choices of frequency-domain versus time-domain trajectory scoring and soft reweighting versus hard isolation while holding regional calibration, trajectory construction, and the remaining settings fixed. The time-domain variant replaces frequency-domain scoring with a time-domain trajectory score while keeping cross-round trajectory memory and soft reweighting. The hard-isolation variant retains frequency-domain scoring but replaces soft reweighting with binary filtering. The basic-frame variant combines the two reduced choices.

For completeness, the time-domain variant uses the same residual trajectory $X_i^{(r)} \in \mathbb{R}^{L_r \times p}$ as FedRASR. It first computes the adjacent-round difference

$$D_{i,\ell}^{(r)} = X_{i,\ell}^{(r)} - X_{i,\ell-1}^{(r)}, \qquad \ell = 1, \dots, L_r - 1. \tag{64}$$

The normalized change magnitude and trajectory level are

$$d_{i,\ell}^{(r)} = \frac{\left\| X_{i,\ell}^{(r)} - X_{i,\ell-1}^{(r)} \right\|_2}{\sqrt{p}}, \tag{65}$$

$$v_{i,\ell}^{(r)} = \frac{\left\| X_{i,\ell}^{(r)} \right\|_2}{\sqrt{p}}. \tag{66}$$

The resulting time-domain score is

$$S_{i,\mathrm{time}}^{(r)} = \frac{\dfrac{1}{L_r - 1} \displaystyle\sum_{\ell=1}^{L_r-1} d_{i,\ell}^{(r)}}{\dfrac{1}{L_r} \displaystyle\sum_{\ell=0}^{L_r-1} v_{i,\ell}^{(r)} + \epsilon}. \tag{67}$$

This score measures the average adjacent-round change relative to the trajectory's average magnitude. The remaining trajectory construction, regional risk calibration, directional evidence, and soft aggregation steps are unchanged.

The full method produces the lowest malicious contribution share and average malicious defense factor under both attacks. Its benign factors remain close to one, showing that the reduction does not arise from uniformly suppressing all clients. Hard isolation retains larger malicious contribution shares, while the time-domain and basic-frame variants provide weaker separation between malicious and benign participants. These diagnostics support the roles of frequency-domain scoring and continuous soft reweighting.

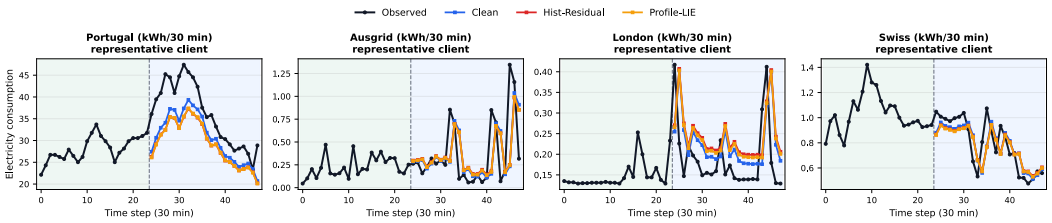

Figure S2: Representative FedRASR forecasting traces under clean training, Hist-Residual poisoning, and Profile-LIE poisoning. The black curve is the observed load, colored curves are seed-averaged forecasts, and the dashed line separates observed history from the forecast horizon.

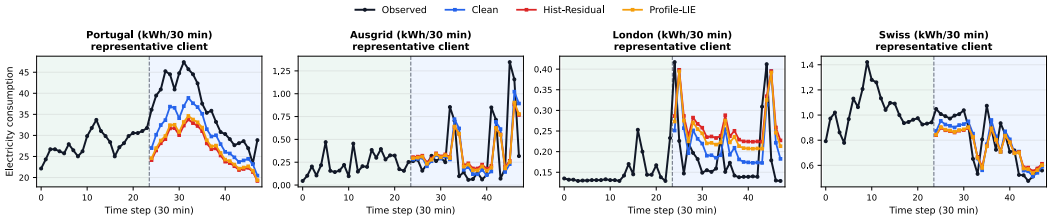

Figure S3: Representative GeoMed forecasting traces under clean training, Hist-Residual poisoning, and Profile-LIE poisoning. The black curve is the observed load, colored curves are seed-averaged forecasts, and the dashed line separates observed history from the forecast horizon. The same trace-selection protocol is used for the FedRASR and GeoMed panels.

I.3 Representative Forecast Trace Selection

GeoMed is selected for the representative baseline panels because it has the best non-FedRASR RMSE under both poisoning scenarios in the main results table, with a tie under Hist-Residual. This does not imply that it is the best baseline for every metric; Krum, for example, has the strongest non-FedRASR MAE under the two attacks.

The forecast panels are intended as qualitative illustrations rather than as the primary performance metric. To avoid method-specific cherry-picking, the same representative client and time window are used across clean training, Hist-Residual poisoning, and Profile-LIE poisoning for a given source and method. The plotted forecast region is a rolling first-horizon trace: at each successive evaluation window, the first value of the model's two-step output is plotted, and the next input window is formed from the observed evaluation sequence rather than from earlier predictions. The selected traces are then averaged over the five random seeds before plotting.

For the proposed method, the displayed clients are fixed as Ausgrid client 1, London client MAC000018, Portugal client MT_028, and Swiss client 45bc26f3b84499091c5baf481a0d3203. For the GeoMed comparison panel, the same region-level display rule and seed averaging are used. The panels therefore compare scenario-level forecast behavior under a fixed visualization protocol; the quantitative claims in the main paper are based on the fixed per-seed 10-client-per-region evaluation subset rather than these representative traces alone.

Table S10: Forecast-region WAPE for representative forecasting traces. Lower is better.

Method	Scenario	Portugal	Ausgrid	London	Swiss
FedRASR	Clean	15.12%	56.19%	39.46%	12.89%
FedRASR	Hist-Residual	18.86%	56.57%	44.93%	13.31%
FedRASR	Profile-LIE	18.85%	55.99%	43.33%	13.30%
GeoMed	Clean	16.14%	55.89%	38.60%	12.99%
GeoMed	Hist-Residual	25.42%	58.92%	51.16%	14.03%
GeoMed	Profile-LIE	23.75%	57.07%	46.30%	13.69%

I.4 FORECAST-REGION WAPE

Table S10 reports WAPE for representative forecasting traces. The values are computed over the displayed forecast region using seed-averaged forecasts and therefore summarize the plotted ensemble traces rather than per-seed run-level WAPE.

The WAPE values are consistent with the visual forecast panels: on Portugal and London, the proposed method has lower WAPE than GeoMed under both poisoning scenarios, while on Ausgrid and Swiss the displayed forecasts are closer.

I.5 NOTATION SUMMARY

Table S11: Notation for the federated task, considered attacks, and residual trajectory construction.

Symbol	Meaning	Symbol	Meaning
Federated task and indexing			
N	Total number of clients.	i, j, k	Focal-client, aggregation-sum, and peer-client indices, respectively.
r	Global communication-round index.	$\mathcal{S}^{(r)}$	Clients selected in round r.
n_i	Observations in client i's chronological training segment.	M_i	Number of local forecasting windows of client i.
m, t	Local-window index and scalar time index.	$x_{i,m}, y_{i,m}$	Historical input window and two-step forecast target.
$F_\theta, \mathcal{L}_i$	Forecasting model and client i's local regression objective.	$\theta^{(r)}, \theta_i^{(r)}$	Global model and locally trained client model at round r.
$\Delta_i^{(r)}$	Complete client update $\theta_i^{(r)} - \theta^{(r-1)}$.	$w_i^{(r)}, a_i^{(r)}$	FedAvg coefficient and FedRASR defense factor.
$\mathcal{R}, \zeta_i$	Set of represented regions and the region associated with client i.	$\mathcal{S}_\zeta^{(r)}$	Selected round-r clients associated with region ζ.
$\mathcal{M}, \mathcal{M}^{(r)}$	Fixed malicious pool and its selected members in round r.	id_i	Stable identifier of client i's time series.
Attack notation			
$\mu_{i,y}, \sigma_{i,y}$	Historical target mean and standard deviation for malicious client i.	$\psi_i(t)$	Bounded profile-dependent target perturbation pattern.
$\alpha, \lambda_{\mathrm{hist}}$	Hist-Residual amplitude and historical deviation scale.	χ_r	Alternating perturbation sign used at round r.
$\widetilde{y}_i^{(r)}(t)$	Poisoned scalar target component.	κ	Full-model coordinate index.
$\Delta_i^{\mathrm{temp},(r)}$	Update obtained after local training on Hist-Residual targets.	$\mu_{\Delta,\kappa}^{(r)}, \sigma_{\Delta,\kappa}^{(r)}$	Benign round-level mean and standard deviation for coordinate κ.
$d_{i,\kappa}^{(r)}, z_{\mathrm{lie}}$	LIE displacement direction and displacement scale.	$\Delta_i^{\mathrm{LIE},(r)}$	Complete coordinate-wise LIE-shaped update.
$\Delta_i^{\mathrm{adv},(r)}$	Final adversarial update received by the server.	λ_{mix}	Mixing coefficient between temporary and LIE-shaped updates.
Compact features and trajectories			
U	One non-head update tensor.	$f_i^{(r)}, p$	Compact update feature and its dimension.
$\widetilde{c}_\zeta^{(r)}, c_\zeta^{(r)}$	Instantaneous regional median and stored smoothed regional center.	β_{center}	EMA coefficient for the regional center.
$e_i^{(r)}$	Region-calibrated residual compact feature.	T, L_r	Maximum and currently available trajectory lengths.
$T_{\min}, L_{\mathrm{stale}}$	Minimum observed residuals for frequency scoring and maximum directional-evidence age.	$X_i^{(r)}$	Aligned residual trajectory in $\mathbb{R}^{L_r \times p}$.
ℓ, q	Trajectory-position and compact-feature-coordinate indices.	$Z_{i,\ell,q}^{(r)}$	Temporally centered trajectory coordinate.

Table S12: Notation for frequency-domain scoring, risk estimation, aggregation, optimization, and evaluation. Calligraphic symbols denote sets, while ordinary uppercase and lowercase symbols denote values, vectors, or indices according to context.

Symbol	Meaning	Symbol	Meaning
Frequency-domain scoring			
$\widehat{X}^{(r)}_{i,\nu,q}$	DFT coefficient at frequency bin ν.	ν, ν'	Discrete frequency-bin indices.
$\mathcal{F}^+, \mathcal{F}^h$	Retained positive-frequency bins and their high-frequency subset.	$\sqrt{-1}, \pi$	Imaginary unit and circle constant in the DFT.
$\varpi^{(r)}_{i,\nu,q}$	Normalized power assigned to frequency bin ν.	$H^{(r)}_{i,q}, P^{(r)}_{i,q}, Q^{(r)}_{i,q}$	Coordinate-level high-frequency energy, peak concentration, and entropy-based periodicity.
$\overline{H}^{(r)}_i, \overline{P}^{(r)}_i, \overline{Q}^{(r)}_i$	Frequency statistics averaged over compact-feature coordinates.	$\omega_H, \omega_P, \omega_Q$	Weights of the three frequency statistics.
$S^{(r)}_i$	Combined frequency-domain score.	$S^{(r)}_{i,\mathrm{time}}$	Time-domain trajectory-variation score used in the ablation.
Risk estimation and aggregation			
$\mathcal{I}^{(r)}_\zeta$	Regional score-reference set.	$\mathrm{med}^{(r)}_\zeta, \mathrm{MAD}^{(r)}_\zeta$	Regional score median and median absolute deviation.
$\widetilde{\tau}^{(r)}_\zeta, \tau^{(r)}_\zeta$	Instantaneous and stored regional thresholds.	$\lambda_{\mathrm{mad}}, \beta_\tau$	MAD threshold scale and threshold EMA coefficient.
$\xi^{(r)}_i, \rho^{(r)}_i$	Positive score margin and normalized within-region rank.	$\rho^{\mathrm{lo}}_b, \rho^{\mathrm{hi}}_b$	Lower and upper endpoints of the frequency-rank gate.
$\rho^{\mathrm{lo}}_u, \rho^{\mathrm{hi}}_u$	Lower and upper endpoints of the directional-rank gate.	$b^{(r)}_i, u^{(r)}_i$	Frequency-score and directional-risk gates.
s_{scale}	Scale used to normalize the positive score margin.	$r^-_i, C^{(r)}_i$	Latest earlier observed round and residual cosine similarity.
τ_{cos}	Directional cosine threshold.	$R^{(r)}_{i,\mathrm{freq}}, R^{(r)}_{i,\mathrm{dir}}$	Frequency and directional risk components.
$\lambda_{\mathrm{dir}}, R_{\max}$	Directional-risk coefficient and total-risk cap.	$R^{(r)}_i$	Capped total client risk.
$\gamma, a_{\min}$	Exponential decay coefficient and defense-factor floor.	$\Delta^{(r)}_{\mathrm{def}}$	Soft-reweighted aggregate of complete client updates.
Optimization, metrics, and safeguards			
E, h	Number of local Adam steps and local-step index.	$\mathcal{B}_{i,h}, \eta_c$	Local mini-batch and client learning rate.
$N_{\mathrm{eval}}, \widehat{y}_m$	Number of evaluated scalar targets and predicted target value.	$\mathcal{T}_{\mathrm{pred}}$	Displayed forecast region used for WAPE.
$\mathrm{Share}^{(r)}_{\mathrm{mal}}$	Combined normalized aggregation contribution of selected malicious clients.	ϵ	Operation-specific positive numerical safeguard used to avoid unstable division or logarithms.
ρ_{mal}, f_B	Malicious-client ratio and assumed Byzantine count used by applicable baselines.	$\mathrm{clip}(x, l, u)$	Bounds x to the interval $[l, u]$.
$\overline{a}^{(r)}_{\mathrm{mal}}, \overline{a}^{(r)}_{\mathrm{ben}}$	Round-level mean defense factors assigned to selected malicious and benign clients.		

J CODE AND DATA SUPPLEMENT

An anonymous Code and Data Supplement accompanies this submission. It contains the implementation of federated training, the two attacks, aggregation baselines, FedRASR, and the ablation variants, together with machine-readable experiment configurations, preprocessing, evaluation, and plotting scripts, software requirements, and execution instructions.

The archive does not contain the complete raw or processed datasets. Instead, it records the four official data sources, expected filenames, and access notes, and provides the preprocessing scripts needed to reconstruct the processed input format after the original data have been obtained under their respective terms.

K LIMITATIONS

The present evaluation focuses on two structured persistent-poisoning settings at fixed malicious-client and participation ratios. Future work will extend the study to additional temporal attack patterns, participation regimes, and broader client coverage. The current global-round trajectory representation uses zero filling for non-participation; mask-aware or event-indexed alternatives, together with formal robustness analysis, are promising extensions.