

Article

Federated Residual-Element Optimization for Compliance-Aware Environmental Artificial Intelligence in Distributed Effluent Monitoring

Koffka Khan ¹ , Winston Elibox ^{2,*} and Shanta Ramnath ³

¹ Department of Computing and Information Technology, Faculty of Science and Technology, The University of the West Indies, St. Augustine Campus, St Augustine 311126, Trinidad and Tobago; koffka.khan@sta.uwi.edu

² Department of Life Sciences, Faculty of Science and Technology, The University of the West Indies, St. Augustine Campus, St Augustine 311126, Trinidad and Tobago

³ Department of Permit Monitoring and Complaints, Environmental Management Authority, 8 Elizabeth Street, St. Clair, Port of Spain 101101, Trinidad and Tobago; sramnath@ema.co.tt

* Correspondence: winston.elibox@sta.uwi.edu

Abstract

Federated learning (FL) can support environmental monitoring when facilities cannot pool raw operational records, but ordinary FL does not encode permit direction, regulatory margins, censoring, missingness, or facility-specific compliance regimes. This study introduces Federated Residual-Element Optimization (FREO), a compliance-aware environmental artificial intelligence (AI) framework that converts effluent measurements into signed, unit-consistent regulatory residuals and communicates compact residual elements rather than raw samples. Ten de-identified Environmental Management Authority effluent workbooks, treated as workbook-level facility-category clients, produced 4036 assessable observations, 321 facility-month states, and 311 next-period prediction records. A boundary-purged temporal split removed train/test target-input overlap, leaving 204 training and 97 held-out records. On the within-facility next-period high-burden task, FREO achieved AUROC 0.966, AUPRC 0.923, F1 0.852, and Brier score 0.079, with facility-month bootstrap intervals of [0.924, 0.996], [0.810, 0.990], [0.727, 0.941], and [0.056, 0.104], respectively. Matched calibrated baselines indicate that residual features and training-only facility-prior calibration explain much of the gain; FedAvg-Cal slightly exceeded FREO on AUPRC, F1, and Brier. FREO is therefore positioned as an auditable residual-element workflow for raw-data-local regulatory screening, not as a universally superior optimizer.

Keywords: federated learning; environmental artificial intelligence; compliance monitoring; effluent quality; water quality; constrained optimization; residual learning; non-independent and identically distributed data; privacy-preserving analytics; calibration



Academic Editors: Francisco A. Pujol, Amir Javed and Higinio Mora

Received: 11 June 2026

Revised: 6 July 2026

Accepted: 7 July 2026

Published: 10 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

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

1. Introduction

Industrial effluent monitoring is a distributed environmental artificial intelligence (AI) problem with a regulatory structure that differs from ordinary sensor forecasting. Facilities and regulators observe related analytes, but records are operationally local, irregularly sampled, tied to facility-specific permit limits, and governed by reporting conventions such as non-detects, no-flow flags, missing analyte panels and changing monitoring schedules. The same measured concentration may imply compliance, near-exceedance or violation depending on whether its permit condition is upper-bounded, lower-bounded or range-bounded.

Pooling all raw data can be infeasible or undesirable for governance and confidentiality reasons, while a purely local model discards shared structure across facilities. Federated learning (FL) provides a natural middle ground by training a shared model while retaining raw measurements locally [1–3].

This setting creates a representation gap between general-purpose FL and regulatory compliance analytics. Standard FL assumes that clients share a compatible prediction task and that the optimizer mainly needs to manage communication and statistical heterogeneity. Effluent compliance modeling also requires a client-side representation that preserves permit direction, violation margin, missingness and facility history before any model update is sent. The central question addressed in this study is therefore not only how to optimize a federated model but in what compliance-aware state a facility should communicate when raw measurements cannot be pooled and the legally meaningful quantity is a signed margin to a permit limit.

This paper proposes FREO, Federated Residual-Element Optimization, as a response to that representation gap. Each client first maps local measurements into signed compliance residuals, where positive values indicate violation and negative values indicate the compliance margin. The client then forms a residual element, a compact transferable state summarizing residual priors, analyte coverage, missingness, reliability and optional constraint pressure. In the reported experiments, this element is instantiated as residual-summary features, reliability weights and a training-only facility-prior calibration term around a federated logistic regression model; optional compliance-pressure variables are not activated. Federated optimization proceeds over these residual-space objects rather than over raw concentrations alone. In practical terms, FREO is intended to make the unit of federation a bounded, auditable compliance summary rather than a raw effluent record. FREO is evaluated with an auditable generalized linear residual operator because regulatory monitoring requires transparent residual semantics, traceable feature construction and reproducible client-side computation. Deep neural backbones are compatible with the framework but are outside the scope of the present study.

Accordingly, the research contribution is not the invention of another generic server optimizer. It is a compliance-aware representation and client-state layer for federated environmental AI. The contributions are deliberately bounded and explicit: (i) a unit-consistent signed permit-residual coordinate system that handles upper-bound, lower-bound and range effluent constraints with common violation margin semantics; (ii) a compact residual-element client state for compliance-aware FL, combining residual profiles, missingness/coverage, reliability and training-only facility priors without transmitting raw measurements; (iii) an implemented FREO objective that uses those residual elements for compliance-weighted aggregation and auditable next-period risk scoring; and (iv) an empirical evaluation that separates within-regime temporal prediction, cross-category transfer stress testing and non-regulatory auxiliary fixed-feature FL stress tests. The supporting machinery is not claimed as new: Federated Averaging (FedAvg), Federated Proximal (FedProx), Stochastic Controlled Averaging for Federated Learning (SCAFFOLD), federated Adam (FedAdam) and federated Yogi (FedYogi)-style optimization, generalized linear risk modeling, Brier score, area under the receiver operating characteristic curve (AUROC) and area under the precision–recall curve (AUPRC) evaluation, bootstrap uncertainty analysis, and shrinkage calibration are established techniques used here to assess and contextualize the new residual-element formulation. The empirical evaluation includes expanded baselines, ablations, non-independent and identically distributed (non-IID) heterogeneity slices, missingness/noise robustness, rare severe-tail metrics, communication estimates, client dropout, noisy-aggregation sensitivity, facility-month and facility-cluster bootstrap

intervals, paired bootstrap tests, leave-one-facility-category transfer validation and two five-seed auxiliary fixed-feature federated classification stress tests.

2. Related Work and Prior-Art Positioning

2.1. Federated Optimization

Federated Averaging (FedAvg) averages client updates after local optimization and remains the canonical FL reference point [1]. Federated Proximal (FedProx) adds a proximal term to control client drift under heterogeneous local objectives [4]. SCAFFOLD (Stochastic Controlled Averaging for Federated Learning) uses control variates to reduce drift from non-IID clients [5]. Adaptive server optimizers such as federated Adam (FedAdam) and federated Yogi (FedYogi) improve optimization dynamics by changing the server update rule [6]. Federated Normalized Averaging (FedNova) normalizes heterogeneous local progress to reduce objective inconsistency under variable client computation [7]. Personalized FL methods such as Ditto, personalized FedAvg (Per-FedAvg) and personalized federated learning with Moreau envelopes (pFedMe) explicitly adapt global models to non-IID client distributions [8–10].

These methods address communication efficiency, client drift, optimizer stability and personalization after the local learning problem has already been defined. They generally do not specify how a regulated measurement should be transformed when clients differ in permit limits, allowable ranges, analyte panels and reporting flags. FREO is therefore positioned as a representation layer and client-state construction that can operate inside or alongside established optimizers. The comparison against FedAvg, FedProx, SCAFFOLD-style, FedAdam and FedYogi baselines is meant to test whether the residual-element state adds value beyond these known optimization mechanisms, not to claim that FREO replaces the broader FL optimizer literature.

2.2. Environmental and Water Quality Federated Learning

Environmental FL has been applied to decentralized air quality and water quality sensing, and recent reviews identify environmental monitoring as a growing application domain for FL [11]. Water quality FL studies include privacy-preserving forecasting across distributed water-monitoring nodes [12] and federated deep learning for sensor-based water quality management [13]. These studies motivate decentralized environmental intelligence and show that local environmental data can contribute to shared models without routine raw-data centralization.

The present setting is narrower and more regulatory. Effluent-monitoring records are not only environmental time-series measurements; they are evidence against permit-specific compliance conditions. Existing environmental FL work usually treats the client as a sensor, station or monitoring node and models physical values, class labels or quality indices. FREO instead treats the client as a regulated facility whose measurements must first be interpreted relative to upper-bound, lower-bound or range permit rules. The literature therefore motivates distributed environmental modeling, while the residual-element formulation addresses the additional compliance semantics needed for effluent screening use.

2.3. Compliance-Aware Learning, Residualization and Calibration

The proposed residualization step is related to residual learning and statistical residualization in the broad sense of transforming observations into deviations from a reference structure. In this application, the reference structure is the permit limit. This differs from generic residual connections in deep networks [14]: FREO does not add residual blocks to a neural architecture; it constructs a regulatory residual coordinate system before federated

optimization. The transformation is also different from simply normalizing raw concentrations because the sign of the residual encodes the regulatory direction and the magnitude represents a stabilized compliance margin.

Calibration is important because compliance screening is a risk-scoring task rather than a pure ranking problem. Proper scoring rules and calibration studies motivate Brier score and negative-log-likelihood-style diagnostics, while receiver operating characteristic (ROC) and precision–recall analyses measure ranking behavior [15–17]. FREO uses calibration as a known supporting tool, not as a new calibration theory. The distinctive step is to estimate facility-prior calibration from training-period labels and embed it in the residual element so that probability adjustment remains tied to the same leakage-controlled client state used for federated compliance modeling.

2.4. Imbalanced Event Prediction, Privacy and Personalization

Compliance exceedances are often imbalanced, making precision–recall analysis more informative than receiver operating characteristic (ROC) curves alone in rare-event settings [18,19]. This motivates this paper’s emphasis on AUPRC, F1-score operating points and recall at a fixed false-positive-rate target in addition to AUROC. Because rare severe events are especially sparse in the available held-out set, the severe-tail analysis is treated as a stress test rather than as the primary evidence.

FL also raises privacy and communication questions. Raw-data locality is not equivalent to formal privacy; secure aggregation and differential privacy provide additional guarantees when required [20–22]. FREO’s compact residual elements reduce what must be exchanged, but they do not by themselves provide a mathematical privacy guarantee. Facility priors connect the method to personalized FL because client-level calibration can improve local relevance, but this paper enforces a training-only construction so that priors cannot leak held-out compliance labels into evaluation.

2.5. Novelty Boundary

The closest intellectual neighbors are FL optimization, federated environmental monitoring, residualization, calibrated risk prediction and federated personalization. Taken together, the literature provides necessary pieces but not the specific object needed here: an auditable federated state whose coordinates already encode regulatory direction, the compliance margin, analyte coverage and facility history. To make the boundary explicit, FREO does not claim novelty for FedAvg-style federated averaging, proximal/control-variate/adaptive server updates, residualization as a statistical idea, shrinkage calibration, Brier scoring, bootstrap uncertainty analysis or generalized linear risk modeling. These are known components used as baselines, diagnostics or implementation scaffolding. This paper’s novel contribution is the following compliance-specific composition: (i) converting heterogeneous upper-bound, lower-bound and range permit rules into signed residuals with shared violation margin semantics; (ii) packaging residual profiles, positive residual burden, observation coverage, reliability and training-only facility priors into a compact residual element; and (iii) using that element as the auditable client state in a federated effluent-screening objective. Facility priors are therefore not presented as a new calibration method; they are a leakage-controlled, training-only client calibration term embedded in the residual element. This bounded positioning clarifies the contribution relative to the existing literature while avoiding unsupported claims of having no intellectual neighbors.

2.6. Synthesis Across the Four Strands

The four related work strands play distinct but connected roles in FREO. Federated optimization supplies the communication and aggregation substrate; environmental FL supplies the decentralized monitoring setting; compliance-aware residualization supplies the

permit-relative coordinate system; and personalization/calibration supplies the training-only facility adjustment that keeps risk scores locally meaningful. FREO bridges these areas through the residual element, which acts as the shared interface among them: it is the object transported by the FL procedure, the environmental client summary exchanged across regulated facilities, the carrier of signed compliance margins and missingness information, and the location where facility-specific calibration is stored without using held-out labels. Table 1 summarizes this mapping. In this sense, FREO is not a simple application of a single literature strand to effluent data; it is a synthesis that makes the regulated facility, rather than an abstract data silo, the auditable unit of federation.

Table 1. Conceptual map of how FREO connects the four literature strands reviewed in Section 3.

Literature Strand	What the Strand Contributes	How FREO Connects It to Effluent Compliance
Federated optimization	Communication-efficient training, client-update aggregation and mechanisms for heterogeneous local objectives.	FREO uses these optimizers as the transport layer while changing the local object being transported from ordinary feature updates to residual-element summaries plus model updates.
Environmental FL	A distributed monitoring setting in which facilities, stations or sensors cannot always pool raw observations.	FREO specializes this setting to regulated effluent facilities and keeps raw measurements local while sharing bounded compliance summaries.
Compliance-aware residualization	A way to express measurements relative to permit limits, directions and margins.	FREO converts upper-bound, lower-bound and range rules into signed residuals so that heterogeneous analytes become comparable violation margin coordinates.
Personalization and calibration	Client-specific adjustment for non-IID data and probability scores that remain meaningful locally.	FREO embeds training-only facility priors, reliability and coverage information in the residual element, improving local relevance without using held-out labels.

3. Materials and Methods

3.1. Data Source and Construction

The empirical study used ten Environmental Management Authority (EMA) spreadsheet workbooks, each representing one de-identified facility-category client rather than a named facility. These workbook-level clients corresponded to the following industrial or utility categories: biscuit manufacturing, bread/cake manufacturing, cigarette manufacturing, commercial catering, concrete products, drinks manufacturing, poultry processing, quarrying, wastewater treatment and water treatment. The extraction pipeline identified parameter names, stations, sample dates, numeric values, regulatory limits, limit direction, censoring/no-flow flags and compliance status. Rows that could not be assessed against a usable limit were excluded from modeling.

The extracted table contained 6072 observation rows, of which 4036 current-date observations were assessable against a limit. Aggregating by workbook-level client and sample month yielded 321 facility-months (workbook client months). Unless stated otherwise, the facility and client therefore denote this workbook-level facility-category client. The independent units are the ten workbook-level clients, not the 4036 assessable measurement rows; the latter are source measurements used to construct within-client residual summaries. For prediction, each facility-month was paired with the next observed month from the same workbook-level client, producing 311 next-period records. Because the interval is determined by the next observed workbook-client month rather than a guaranteed fixed calendar interval, the manuscript uses next-period terminology. The gap was one calendar

month for 309 of the 311 pairs (99.4%); two pairs involved longer gaps of 8 and 9 calendar months (243 and 275 days), giving a median gap of 31 days, an interquartile range of 30–31 days and a range of 28–275 days. The temporal split was then revised to avoid train/test boundary overlap. Within each workbook-level client, the first held-out input month defines the boundary; any candidate training row whose target month is on or after that boundary is removed from training. This removed ten boundary training pairs, one per workbook-level client, and produced 204 training rows and 97 held-out rows. Under this non-overlap split, no month used as a training target later appears as a held-out input. Because multiple facility-month records came from the same workbook-level clients, the 311 prediction records should be interpreted as repeated temporal observations clustered within ten workbook-level clients rather than as 311 independent regulatory sites. Table 2 summarizes the resulting data construction and temporal split. The primary target was next-period high-burden compliance, defined as more than 50% of assessable parameters exceeding their limit in the next observed period. Training and held-out prevalences were 27.9% and 26.8%, respectively. A secondary severe-tail task used the 80th percentile of the training distribution of next-period maximum stabilized residuals as the threshold.

Table 2. Data construction and temporal split summary. IQR denotes interquartile range.

Quantity	Value
Uploaded facility-category workbooks	10
Extracted observation rows	6072
Assessable current-date compliance observations	4036
Modeled parameter categories	12
Facility-month aggregates	321
Next-period prediction records	311
One-calendar-month prediction gaps	309/311
Longer irregular prediction gaps	2 (8 and 9 months; 243 and 275 days)
Prediction gap days	median 31; IQR 30–31; range 28–275
Independent workbook-level client clusters	10
Training/held-out records under non-overlap split	204/97
Boundary training pairs removed	10
Train-target/held-out-input overlaps after revision	0
Primary training/held-out positive prevalence	27.9%/26.8%
Severe-tail held-out positive prevalence	7.2%

3.2. Signed Compliance Residuals

For a measurement y and constraint descriptor c belonging to parameter category p , define a signed residual $g(y; c)$ such that $g > 0$ denotes violation and $g \leq 0$ denotes compliance. Before residualization, measurement and permit values are expressed in the same canonical reporting unit for parameter category p . To avoid unit-dependent constants, the residual is a dimensionless ratio whose denominator is the corresponding permit-bound scale. Let $s_p > 0$ be a prespecified parameter-specific reporting scale, expressed in the same canonical unit, that is used only if a permit bound is zero or otherwise unusable as a denominator, and define

$$D(a; p) = \begin{cases} |a|, & |a| > 0, \\ s_p, & |a| = 0. \end{cases} \quad (1)$$

For an upper-bound limit u ,

$$g(y; c) = \frac{y - u}{D(u; p)}. \quad (2)$$

For a lower-bound limit ℓ ,

$$g(y; c) = \frac{\ell - y}{D(\ell; p)}. \quad (3)$$

For a range constraint $[\ell, u]$,

$$g(y; c) = \max \left\{ \frac{\ell - y}{D(\ell; p)}, \frac{y - u}{D(u; p)} \right\}. \quad (4)$$

Multiplying y , ℓ and u by any positive unit-conversion factor multiplies the numerator and denominator by the same factor and leaves g unchanged. The definition therefore avoids a hard-coded $\max(|u|, 1)$ -style operation on arbitrary source units. In the processed effluent dataset, every assessable permit bound used as a denominator was positive, so no fallback reporting scale was invoked and the reported numerical results are unchanged. The stabilized residual is

$$z = \operatorname{asinh}(g), \quad (5)$$

which preserves the sign while reducing leverage from extreme concentration excursions.

3.3. Residual Elements and Federated Objective

Let client k denote a workbook-level facility-category client and t denote a sampling month. The residual element $E_{k,t}$ is a compact state built only from the training portion of that client's records, as follows:

$$E_{k,t} = (\mu_{k,t}, \sigma_{k,t}, b_{k,t}, m_{k,t}, q_{k,t}, a_{k,t}, \pi_k, \lambda_k), \quad (6)$$

where μ and σ are residual location and scale summaries, b is the current violation burden, m is missingness/coverage, q is the parameter-wise residual profile, a is a reliability weight derived from sample count and residual heterogeneity, π_k is a shrinkage facility prior, and λ_k is an optional dual variable for compliance pressure. Raw measurements and sample-level permit records are not transmitted. Equation (6) defines the framework-level residual element, while Equation (7) gives the objective; the experiments below use the concrete logistic implementation in Section 3.4. In that implementation, compliance pressure is represented through burden/severity sample weights, and no separate dual-variable optimizer is run.

For clients $k = 1, \dots, K$, the FREO objective is

$$\min_{\theta, \{E_k\}} \sum_{k=1}^K \alpha_k(E_k) \mathbb{E}_{(x,y,c) \sim \mathcal{D}_k} [w(g(y; c), E_k) \ell(f_\theta(\phi(x, c, E_k)), Y)] + R(\theta) + \Gamma(E_1, \dots, E_K), \quad (7)$$

where ϕ is the residualized feature map, Y is a next-period compliance target, w is a compliance weight, α_k is a transport/reliability weight, R regularizes the shared residual operator and Γ controls residual-element complexity. The terms residual-element dictionary and residual transport are used operationally here to denote the collection of compact client residual-summary vectors and their reliability-weighted use in server aggregation; they are not separate learned neural modules in the reported experiments.

3.4. Implementation-Level Residual-Element Specification and Exact Experimental Method

The reproducible implementation uses $P = 12$ modeled parameter categories and represents each client state by $2P + 6$ transmitted floating-point summaries. For client k at month t , let $\mathcal{I}_{k,t,p}$ be the local observations for parameter category p , and let $z_i = \operatorname{asinh}(g(y_i; c_i))$. The transmitted residual-element vector is

$$\tilde{E}_{k,t} = \left[\bar{z}_{k,t,p}, \bar{z}_{k,t,p}^+ \right]_{p=1}^P \oplus \left(\bar{z}_{k,t}, \bar{z}_{k,t}^+, z_{k,t}^{\max}, \rho_{k,t}^{\text{obs}}, \rho_{k,t}^{\text{miss}}, n_{k,t} \right), \quad (8)$$

where $\bar{z}_{k,t,p} = |\mathcal{I}_{k,t,p}|^{-1} \sum_{i \in \mathcal{I}_{k,t,p}} z_i$ when the parameter is observed and zero otherwise, $\bar{z}_{k,t,p}^+$ is the analogous mean of $\max(z_i, 0)$, $\rho_{k,t}^{\text{obs}}$ is the fraction of the modeled parameter categories observed, $\rho_{k,t}^{\text{miss}} = 1 - \rho_{k,t}^{\text{obs}}$, and $n_{k,t}$ is the local observation count. Observation indicators remain in the local feature map and prevent zero-filled residual means from being interpreted as true zero margins.

The facility-prior calibration term is estimated only from training-period labels. Let $\hat{p}_k$ be the empirical training prevalence for facility k , $\bar{p}$ be the global training prevalence, $\bar{q}_k$ be the mean uncalibrated model probability on training rows from facility k , and n_k be the number of training rows for that facility. The shrinkage prior added to the logit score is

$$\pi_k = \text{clip}\{\text{logit}(s_k \hat{p}_k + (1 - s_k) \bar{p}) - \text{logit}(\bar{q}_k), -1.1, 1.1\}, \quad s_k = \frac{n_k}{n_k + 7}. \quad (9)$$

If a held-out facility or facility category has no labeled training history, π_k is set to zero. Intuitively, this prior is a conservative facility-specific baseline adjustment. A facility with a higher-training-period, higher-burden rate receives a modest upward shift in its risk score, while a lower-risk facility receives a downward shift; when few training records are available, the shrinkage factor pulls the adjustment toward the global average. This helps because facilities with similar current residuals can still operate under different permit regimes, process histories and sampling patterns. The compliance sample weights are

$$w_i = \frac{\omega_i^{\text{class}} (1 + 1.05b_i + 0.30\tilde{s}_i)}{\frac{1}{n_k} \sum_{j \in k} \omega_j^{\text{class}} (1 + 1.05b_j + 0.30\tilde{s}_j)}, \quad (10)$$

where b_i is the current violation burden and $\tilde{s}_i$ is the current severity, scaled after capping at the 90th percentile of training severity values. Residual-information aggregation uses normalized weights

$$\alpha_k = \frac{v_k}{\sum_j v_j}, \quad v_k = \frac{\sqrt{n_k} (0.25 + 0.75H(\hat{p}_k))}{1 + 0.55\|\bar{x}_k - \bar{x}\|_2 / \sqrt{d}}, \quad (11)$$

where $H(\hat{p}_k)$ is binary entropy, $\bar{x}_k$ is the client feature mean, $\bar{x}$ is the global training feature mean and d is feature dimension. The implementation numerically clips logits to $[-40, 40]$ before the sigmoid transformation. No differential-privacy gradient-clipping norm is claimed for the reported experiments; the noisy-aggregation experiment below is therefore a robustness sensitivity analysis rather than a privacy guarantee.

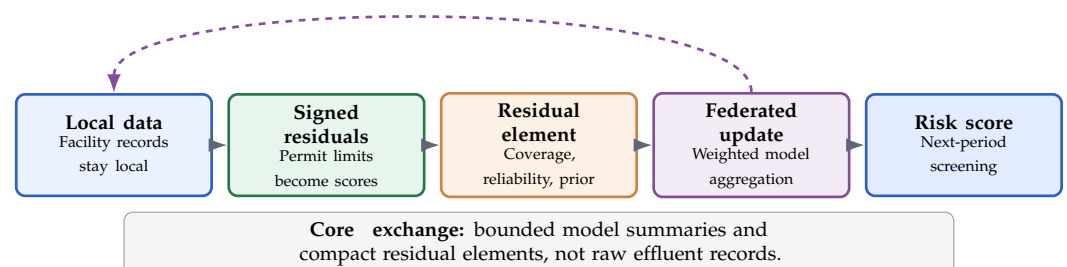
Table 3 gives the exact method used to generate all effluent-monitoring results. This table should be read as the operational definition of FREO in the experiments, separating the implemented logistic residual-summary method from optional framework-compatible extensions.

Workflow 1: Exact FREO implementation for the effluent experiments.

Phase	Action
Initialization	Initialize a global logistic residual operator θ^0 and training-only client residual summaries. At round r , broadcast θ^r ; each participating facility computes signed residual features, updates residual summaries, performs four full-batch local weighted-logistic steps with the proximal and L_2 terms in Table 3, applies the numerical bounds and returns only the bounded coefficient delta plus residual-element summaries.
Client update	
Server aggregation and prediction	Aggregate coefficient deltas using α_k , update the residual-summary table for weighting and diagnostics, broadcast θ^{r+1} and score held-out records as $\sigma(\text{clip}(\theta^\top \phi_{k,t} + \pi_k, -40, 40))$. Figure 1 summarizes the same workflow visually.

Table 3. Exact FREO implementation used in the effluent experiments.

Step	Operational Definition in the Reported Experiments
Local calculation	Each facility keeps raw measurements, permit limits and sample-level records locally. For every assessable observation, it computes the signed stabilized residual $z_i = \text{asinh}(g(y_i; c_i))$. For each facility-month, it then computes per-parameter mean residuals, positive-residual means, observation indicators, current violation burden, capped severity, missingness/coverage fractions, aggregate residual summaries and the training-only empirical facility prevalence used for shrinkage calibration.
Transmitted or shared state	Participating clients return bounded logistic-model coefficient deltas, local count/reliability information and the $2P + 6$ residual-element summaries in Equation (8). The facility prior π_k is a bounded training-only calibration scalar associated with the facility. No raw concentrations, individual samples, sample-level residuals, permit rows or held-out labels are transmitted.
Optimized model	The shared predictor is a full-batch logistic regression residual operator on the residualized facility-month feature vector. FREO uses class/compliance-weighted logistic loss, $L_2 = 0.0025$, four local gradient steps per participating client round, the proximal coefficient 0.018 and the server aggregation weights α_k from Equation (11).
Final prediction	At evaluation, the client constructs the same current-month residual feature vector using training-fitted scaling. The global logistic operator produces an uncalibrated logit $\eta_{k,t} = \theta^\top \phi_{k,t}$. The training-only facility-prior offset π_k is added when available, the resulting logit is clipped to $[-40, 40]$, and the sigmoid value is the next-period high-burden risk score. Binary decisions use thresholds selected on training predictions only.
Not separate active modules	No deep residual block, learned neural residual-transport network, learned token dictionary, separate dual-variable constraint update or formal differential-privacy optimizer is used in the reported effluent experiments. These are framework-compatible extensions or sensitivity considerations, not active components of the empirical FREO results.

**Figure 1.** FREO workflow. Facility-held records remain local; the shared model is trained through signed compliance residualization, compact residual elements, reliability-weighted federated aggregation and next-period risk scoring. The dashed arrow denotes the broadcast of the updated global model and residual-element summary state back to clients.

3.5. Leakage Prevention

All feature scaling, threshold selection, facility priors, residual-element summaries, calibration parameters and class weights were estimated using training-period data only. Held-out months were not used in parameter tuning, threshold selection, missingness imputation or residual-element construction, and were only used for final evaluation. Because the primary effluent prediction table contains only 311 next-period records clustered within ten facilities, no separate environmental validation split was created for hyperparameter search; doing so would have further reduced the already small temporal training and held-out sets. The numerical constants in Table 4 were therefore locked before final held-out evaluation and selected from training-period information, standard federated-learning baseline practice, and numerical stability considerations only. The leave-one-facility-category

validation further removed all labeled records from the held-out category during training and evaluated a transfer variant without held-out facility priors. The same rule was applied in the transfer and cluster-bootstrap sensitivity analyses so that facility priors would not be able to indirectly carry test-period information.

3.6. Experimental Protocol and Reproducibility

The implementation used deterministic Python/Numpy code and a generalized linear residual operator. Features included per-parameter stabilized residual means, positive residual means, observation indicators, month seasonality, observation fraction, missingness fraction and aggregate current residual summaries. The residual columns were generated after unit harmonization using the unit-consistent denominator in Section 3.2; no fallback reporting scale was invoked because every assessable permit-bound denominator was positive. To make the raw-value comparison fairer, the raw baseline was augmented with the same permit metadata available to the residual construction: per-parameter raw means, permit low and high limit summaries, max/min/range limit direction indicators, observation indicators, observation/missingness fractions, seasonality and sample count. Thus, the representation comparison tests the effect of signed, stabilized permit residualization rather than merely giving FREO access to permit limits or limit direction.

The primary effluent constants should therefore be interpreted as fixed implementation choices rather than validation-optimized hyperparameters. The sample is too small and too clustered to support a reliable additional environmental validation split, so the held-out test set is used only once for final evaluation. The only operating quantities selected from the model outputs in the primary effluent task are the F1 and recall@FPR thresholds, and those thresholds are selected on training predictions only. In contrast, the auxiliary CIFAR-10 and FEMNIST-10 experiments below use explicit train/validation/test partitions; calibration strength and the FREO-Adapter backbone are selected on validation data and then evaluated once on the held-out test set.

The baseline suite comprised local-only logistic residual models, centralized pooled logistic residual models, FedAvg, FedProx, a SCAFFOLD-style correction, FedAdam and FedYogi. All primary effluent results used the non-overlap split described above; the one-month training target used is therefore never later reused as a held-out input month. To isolate the effect of residual features and facility-prior calibration from the rest of FREO, we also added matched FedAvg-Cal and FedAdam-Cal references. These matched references use the same compliance-residual feature matrix and the same training-only facility-prior logit adjustment as FREO, but omit FREO's compliance weighting, residual-information aggregation and residual-transport state. The raw-representation diagnostic was also updated to include permit lower/upper limits and max/min/range limit direction indicators, so that the raw baselines were not disadvantaged simply because the residual features encoded permit information. The ablations removed residual elements, compliance weights, residual transport and facility priors. Metrics included AUROC, AUPRC, F1 score, recall at a training-selected false-positive-rate (FPR) target of 0.10 and Brier score. Uncertainty was summarized at two levels. Facility-month bootstrap CIs, computed with 3000 resamples of held-out facility-month rows, describe variability conditional on the observed facility set. Because rows from the same facility can be correlated, we gave interpretive priority to the facility-cluster bootstrap for evidence across independent facilities. That bootstrap resampled the ten held-out workbook-level clients with replacement and retained all held-out rows from the sampled facilities. Paired bootstrap tests resampled the same held-out units for FREO and each comparator; positive deltas favor FREO, with Brier deltas defined as comparator Brier minus FREO Brier. Resamples containing only one class were discarded

for AUROC/AUPRC. Leave-one-facility-category transfer uncertainty was reported with a category-cluster bootstrap over the seven evaluable held-out categories.

Table 4. Reproducibility table.

Item	Specification
Python environment	Python 3.13.5; NumPy 2.3.5; scikit-learn 1.8.0; pandas 2.2.3 available but not required for the final experiment script.
Random seeds	Split is deterministic by date order. Seeds: 42 for main FL training, 1000–1007 for missingness/noise perturbations, 200–203 for client dropout, 4000–4007 for noisy-update sensitivity, and 20260610 for bootstrap resampling.
Optimizer	Full-batch logistic residual operator; 48 global rounds. Hyperparameters were fixed, not searched: learning rate/local steps/ L_2 were FedAvg 0.045/4/0.002, FedProx 0.043/4/0.002, SCAFFOLD-style 0.035/3/0.002, FedAdam/FedYogi 0.035/4/0.002, and FREO 0.043/4/0.0025. Local-step counts are fixed per participating client per round, not maxima chosen adaptively within a round.
Hyperparameter-selection policy	The primary effluent experiment used no separate validation set. Values in this table were prespecified before final held-out evaluation from training-only convergence checks, conservative numerical bounds and standard FL baseline settings; the held-out set was not used. The learning-rate and L_2 entries are method-specific fixed values rather than endpoints of a searched range.
FedProx	Proximal coefficient $\mu = 0.025$ for FedProx baseline and $\mu = 0.018$ for FREO's proximal local update.
SCAFFOLD-style method	Client and server control-variate correction applied to local gradients; three local steps and learning rate 0.035.
FedAdam/FedYogi	Adaptive server updates with first-moment coefficient 0.9; second-moment coefficient 0.99; server gains 0.30 for FedAdam and 0.22 for FedYogi; numerical stabilizers 10^{-5} and 10^{-3} .
FREO residual element	Per-client element contained $2P + 6$ floating-point summaries for $P = 12$ parameter categories, plus training-only facility-prior calibration. Shrinkage used $s_k = n_k / (n_k + 7)$, priors were clipped to $[-1.1, 1.1]$ on the logit scale, residual-information weights used entropy- and distance-adjusted $\sqrt{n_k}$ factors, and logits were numerically clipped to $[-40, 40]$ before sigmoid evaluation.
Threshold selection	Thresholds for F1 score and recall at a false-positive-rate (FPR) target of 0.10 were selected on training predictions only and then applied unchanged to held-out records.
Scaling and imputation	Feature means/standard deviations were fitted on training rows only. Missing analyte channels were encoded through zero-filled residual features plus explicit observation indicators and missingness fractions.
Communication accounting	A 64-bit floating-point model and residual-element payloads counted in both directions over 48 rounds. Protocol, authentication, serialization and secure-aggregation overheads were excluded.

3.7. Auxiliary Non-Regulatory Fixed-Feature Stress Test Protocol

Two attached image classification datasets were used as auxiliary FL stress tests: the Canadian Institute for Advanced Research 10-class image dataset (CIFAR-10) [23] and Federated Extended MNIST (FEMNIST) [24]. These experiments do not provide environmental validation, do not use permit residuals and are not presented as standard image-classification benchmarks. Their purpose is narrower: to test whether residual-element personalization behaves sensibly in conventional non-IID FL settings where client identity, label imbalance and training-only client-prior calibration are known to matter. All auxiliary methods use the same handcrafted deterministic feature matrices and generalized

linear classification heads; no convolutional, transformer or learned visual feature extractor is trained. Method names therefore refer to the federated optimizer or personalization logic implemented on this shared fixed-feature head, not to canonical convolutional or transformer leaderboard implementations.

The CIFAR-10 fixed-feature stress test used 10,000 training images, 2000 validation images and 2500 held-out test images, balanced at 1000/200/250 images per class. Images were represented by pooled red–green–blue (RGB) intensities, pooled horizontal/vertical gradient magnitudes, channel mean/standard deviation summaries and channel-contrast summaries. The FEMNIST fixed-feature stress test used digit classes 0–9 parsed from the attached FEMNIST tensors and stratified into 10,000 training images, 2000 validation images and 2500 held-out test images, balanced at 1000/200/250 images per class. FEMNIST features used inverted pooled pixel intensities, row/column ink projections, total ink mass and pooled stroke-gradient summaries. For both datasets, feature scaling used the training data only.

For each dataset, 24 clients were generated using Dirichlet non-IID class partitions. CIFAR-10 used concentration parameter 0.20 and FEMNIST-10 used concentration parameter 0.25. Five independent seeds were used: 11, 22, 33, 44 and 55. To make the auxiliary comparison easier to follow, we interpreted the methods in four groups: ordinary federated optimizers (FedAvg, FedProx, SCAFFOLD, FedAvgM, FedAdam, FedYogi and FedNova), personalized uncalibrated references (FedNova+Ditto and FedNova+pFedMe-style), matched calibrated references (especially FedAvgM-Cal) and FREO variants. The “-Cal” variants use the same training-only client logit-prior residual and the same validation-selected calibration grid. FREO-full-Cal is the direct non-regulatory analogue of the full FREO variant, whereas FREO-Adapter is the main auxiliary method: it applies the residual-element client-prior adapter over candidate federated backbones and selects the backbone and calibration strength using the validation data only. Thus, the intended auxiliary comparison is FREO-Adapter against the strongest matched calibrated reference, which is not a claim that every FREO variant must dominate every baseline on every row.

Workflow 2: Auxiliary fixed-feature stress test evaluation.

Phase	Action
Split and partition	For each dataset and seed, construct class-balanced training, validation and held-out test partitions; draw a Dirichlet client label partition and apply it separately to training, validation and test indices.
Train and calibrate	Train each federated backbone with the same standardized features, client partition, round budget, local-step budget and L_2 penalty; for calibrated methods, estimate client logit-prior residuals from training labels only and select calibration strength on validation macro-F1 with NLL as a tie-breaker.
Select and evaluate	For FREO-Adapter, select the backbone and calibration strength on validation data only, evaluate once on the held-out test partition and report accuracy, macro-F1 and NLL as mean $\pm$ standard deviation over five seeds.

For interpretation, the primary auxiliary contrast is ordinary FedAvg versus the strongest matched calibrated reference (FedAvgM-Cal) versus FREO-Adapter. Table 5 lists the shared fixed-feature protocol and hyperparameter grid, and the remaining methods document the breadth of the comparison set rather than forming separate headline claims.

Table 5. Auxiliary fixed-feature stress test implementation and hyperparameter grid. All entries share the same handcrafted features and generalized linear head.

Component	Implementation	Grid or Setting
Data split	Class-balanced training/validation/test split, followed by Dirichlet client partition.	CIFAR-10: $\alpha = 0.20$; FEMNIST-10: $\alpha = 0.25$; 24 clients; seeds 11, 22, 33, 44, 55.
Common training	Full-batch local logistic updates with standardized fixed features.	18 global rounds; one local step; $L_2 = 10^{-4}$; CIFAR-10 learning rate 0.32; FEMNIST-10 learning rate 0.42.
FedProx	FedAvg plus proximal term to the current global model.	$\mu = 0.025$.
SCAFFOLD-style	Client/server control-variate correction applied to local gradients.	Same round and step budget as FedAvg.
FedAvgM	Server momentum applied to aggregated model delta.	Momentum 0.9.
FedAdam/FedYogi	Adaptive server updates using first and second moments.	$\beta_1 = 0.9$, $\beta_2 = 0.99$; FedAdam server gain 0.25, $\epsilon = 10^{-5}$; FedYogi server gain 0.20, $\epsilon = 10^{-3}$.
FedNova	Local-step-normalized aggregation for variable client progress.	Local steps $1 + ((k + r) \bmod 3)$, normalized by effective steps.
Ditto/Per-FedAvg/pFedMe-style	Personalized fine-tuning from FedAvgM, FedYogi and FedNova global references.	Ditto: 8 steps, $\mu = 0.08$; Per-FedAvg-style: 4 steps; pFedMe-style: 10 steps, $\mu = 0.14$; fine-tuning rate 0.35 times the dataset rate.
Calibrated variants	Client logit-prior residual estimated from training labels only.	$\gamma \in \{0, 0.25, 0.50, 0.75, 1.00, 1.25, 1.50\}$ selected on validation macro-F1, with validation negative log-likelihood (NLL) as tie-breaker.
FREO-full-Cal	FREO residual-information aggregation plus bounded local updates and client-prior calibration.	Residual weights $\sqrt{n_k}(0.30 + 0.70H(\hat{p}_k))$; update norm bound 8.0; same γ grid.
FREO-Adapter	Validation-selected residual-element adapter over the federated backbone set.	Backbone and γ selected on validation data only; all individual calibrated backbones are also reported.

4. Optimization Statement

Proposition 1 (Descent with bounded residual-element drift). *Assume that each client residual objective is L -smooth in θ , stochastic local gradients have bounded variance σ^2 , client residual heterogeneity is bounded by Δ_{het} , and residual-element compression or drift has a mean square error of at most ϵ_E^2 . For a constant step size $\eta \leq 1/(4L)$ and bounded local steps, FREO iterates satisfy*

$$\frac{1}{T} \sum_{r=0}^{T-1} \mathbb{E} \|\nabla F(\theta^r)\|^2 \leq \frac{2(F(\theta^0) - F^*)}{\eta T} + C_1 \eta \sigma^2 + C_2 \eta \Delta_{\text{het}} + C_3 \epsilon_E^2, \quad (12)$$

for constants C_1, C_2, C_3 , independent of T .

This proposition is a standard smooth nonconvex FL stationarity statement adapted to residual-element compression. It is not intended to prove universal superiority over other FL algorithms; it identifies where residual-element approximation enters the usual descent bound. A full proof is provided in Appendix C.

5. Results

5.1. Residual Representation and Main Held-Out Performance

Compliance residualization improved both ordinary FedAvg and FREO relative to a permit-metadata-matched raw-value baseline (Table 6). The raw baseline includes raw concentration summaries, permit low/high limit summaries and limit direction indicators, so it has access to the same permit information used to construct the residuals. With this fairer raw baseline, FedAvg achieved AUPRC 0.652 and F1 0.492; with compliance-residual features, it achieved AUPRC 0.850 and F1 0.733. FREO also improved from the raw-plus-permit-metadata AUPRC 0.848 to the residual-feature AUPRC 0.923, with F1 increasing from 0.800 to 0.852. This supports the premise that the signed, stabilized residual coordinate system is more useful than raw concentrations with explicit permit metadata, rather than merely showing that permit information matters.

Table 6. Effect of residual representation on held-out next-period high-burden detection under the non-overlap temporal split. The raw-value baseline includes permit lower/upper limits and max/min/range limit direction indicators. Lower Brier is better; higher other metrics are better.

Representation	Method	AUROC	AUPRC	F1	Brier
Raw + permit metadata features	FedAvg	0.749	0.652	0.492	0.200
Raw + permit metadata features	FREO	0.930	0.848	0.800	0.109
Compliance-residual features	FedAvg	0.920	0.850	0.733	0.125
Compliance-residual features	FREO	0.966	0.923	0.852	0.079

The 4036 assessable observation rows are inputs for the residual-summary construction and non-independent prediction cases. The main prediction evidence comes from 97 held-out facility-month records nested within ten facilities after ten boundary training pairs were removed. On the revised non-overlap split, FREO achieved AUROC 0.966, AUPRC 0.923, F1 0.852 and Brier score 0.079 (Table 7); Figure 2 visualizes the same main held-out comparison. These values remain above ordinary FedAvg and FedProx on ranking, operating-point and calibration metrics, but the matched FedAvg-Cal baseline is slightly higher on AUPRC and F1 and lower on Brier. The detailed uncertainty and paired comparisons are introduced after the prediction-gap audit. FREO should therefore be interpreted as a residual-element framework that shows an improvement over ordinary uncalibrated federated baselines while being practically comparable to the strongest matched calibrated FedAvg reference. Centralized pooled regression remains a non-deployment-equivalent upper-reference when raw-data pooling is infeasible.

The prediction gap and split audit address the reviewer concern directly (Table 8). Most transitions are of one calendar month, but the task is labeled next-period because two prediction pairs have longer gaps. The revised split removes target/input boundary reuse by deleting boundary rows from the training set rather than deleting held-out evaluation records.

Table 7. Main temporal held-out results for next-period high-burden detection under the non-overlap split. KB denotes estimated two-way communication in kilobytes over 48 rounds, and FPR denotes the false positive rate. FedAvg-Cal and FedAdam-Cal are matched residual-feature baselines with the same training-only facility-prior calibration as FREO but without FREO compliance weighting, residual-information aggregation or residual-transport state. Lower Brier is better; higher other metrics are better.

Method	AUROC	AUPRC	F1	Recall@FPR0.10	Brier	KB
Local-only	0.943	0.898	0.800	0.885	0.099	0.0
Centralized pooled	0.965	0.914	0.873	0.923	0.070	0.0
FedAvg	0.920	0.850	0.733	0.731	0.125	337.5
FedAvg-Cal	0.968	0.928	0.868	0.885	0.071	337.5
FedProx	0.920	0.851	0.733	0.731	0.125	337.5
SCAFFOLD-style	0.920	0.849	0.667	0.654	0.125	337.5
FedAdam	0.907	0.804	0.730	0.846	0.116	337.5
FedAdam-Cal	0.926	0.840	0.721	0.846	0.087	337.5
FedYogi	0.911	0.823	0.667	0.846	0.121	337.5
FREO	0.966	0.923	0.852	0.923	0.079	562.5

Table 8. Prediction gap and non-overlap split audit. The non-overlap split removes candidate training pairs whose target month is on or after the first held-out input month for that facility.

Audit item	Value
Prediction records	311
One-calendar-month input-to-target gaps	309
Irregular gaps greater than one month	2
Irregular gap lengths	8 and 9 months (243 and 275 days)
Prediction gap days	median 31; IQR 30–31; range 28–275
Boundary training pairs removed	10
Training/held-out records after revision	204/97
Train-target/held-out-input overlaps after revision	0

The facility-cluster bootstrap gives a more conservative and more relevant view of uncertainty for this small regulatory dataset because it resamples complete held-out facilities rather than individual facility-month rows (Table 9). The point estimates are unchanged, but intervals widen materially, reflecting the fact that only ten facilities contribute independent facility-level evidence. The facility-month bootstrap is useful for describing temporal prediction variability within the observed facility set, whereas the facility-cluster bootstrap governs how cautiously the findings should be generalized beyond those facilities. The paired comparisons with the matched calibrated rows further narrow the claim: FREO is not higher than FedAvg-Cal on AUPRC, F1 or Brier under the revised split. The evidence therefore supports FREO as a raw-data-local residual-element framework that remains strong relative to ordinary federated baselines, not as a universally superior optimizer or calibration method.

The facility-month bootstrap intervals in Table 10 and paired comparisons in Table 11 further sharpen the attribution: much of the improvement over ordinary federated logistic regression comes from permit-residual representation and training-only facility-prior calibration.

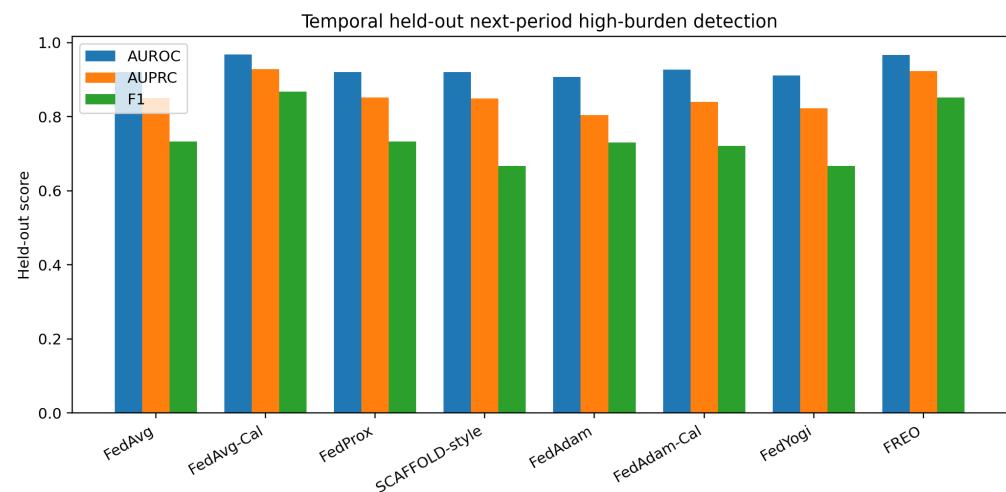


Figure 2. Main held-out comparison among federated baselines and matched calibrated references under the non-overlap split. FREO improves over ordinary federated baselines, while the calibrated FedAvg row shows that facility-prior calibration explains much of the improvement.

Table 9. Facility-cluster bootstrap sensitivity for the primary held-out task under the non-overlap split. Intervals resample the ten held-out workbook-level clients with replacement and are treated as the conservative uncertainty summary for independent facility-level evidence; lower Brier is better. CI denotes confidence interval.

Method	AUPRC 95% CI	F1 95% CI	Brier 95% CI
FedAvg	[0.284, 0.991]	[0.323, 0.918]	[0.051, 0.205]
FedAvg-Cal	[0.325, 1.000]	[0.000, 0.976]	[0.028, 0.126]
FedProx	[0.310, 0.991]	[0.323, 0.917]	[0.052, 0.205]
FedAdam	[0.209, 0.977]	[0.261, 0.939]	[0.045, 0.196]
FedAdam-Cal	[0.124, 0.989]	[0.000, 0.944]	[0.029, 0.161]
Centralized pooled	[0.502, 1.000]	[0.000, 1.000]	[0.022, 0.136]
FREO	[0.511, 1.000]	[0.457, 0.976]	[0.032, 0.134]

Table 10. FREO 95% bootstrap confidence intervals (CIs) under the non-overlap split. Bootstrap resampling unit: held-out facility-month row; CI denotes confidence interval.

Metric	Point Estimate	95% Bootstrap Confidence Interval (CI)
AUROC	0.966	[0.924, 0.996]
AUPRC	0.923	[0.810, 0.990]
F1	0.852	[0.727, 0.941]
Recall@FPR0.10	0.923	[0.808, 1.000]
Brier	0.079	[0.056, 0.104]

5.2. Ablation Study

The ablation study identifies residual elements and facility priors as the dominant contributors (Table 12 and Figure 3). Removing residual elements reduced AUPRC from 0.923 to 0.851 and F1 from 0.852 to 0.733. Removing facility priors reduced F1 to 0.667 and worsened the Brier score to 0.143. The matched baselines in Table 7 confirm the same mechanism from the opposite direction: adding the same training-only facility-prior calibration to residual-feature FedAvg recovers most of the original FedAvg-to-FREO gap and slightly exceeds FREO on AUPRC, F1 and Brier under the non-overlap split. Compliance weights and residual transport primarily affected ranking and operating-point behavior rather than uniformly improving all calibration metrics. The active mechanism supported most strongly by the evidence is residual representation with facility-specific calibration.

Table 11. Paired facility-month bootstrap comparisons under non-overlap split. Positive deltas favor FREO; for Brier, positive means the comparator has higher Brier than FREO. CI denotes confidence interval.

Comparator	Metric	Δ	95% CI	Two-Sided p
FedAvg	AUPRC	0.073	[−0.001, 0.175]	0.0527
FedAvg	F1	0.119	[0.024, 0.227]	0.0147
FedAvg	Brier improvement	0.046	[0.029, 0.065]	0.0003
FedAvg-Cal	AUPRC	−0.005	[−0.048, 0.031]	0.8040
FedAvg-Cal	F1	−0.016	[−0.055, 0.000]	0.7120
FedAvg-Cal	Brier improvement	−0.008	[−0.015, −0.000]	0.0473
FedProx	AUPRC	0.072	[−0.001, 0.166]	0.0540
FedProx	F1	0.119	[0.023, 0.228]	0.0140
FedProx	Brier improvement	0.046	[0.028, 0.065]	0.0003
FedAdam	AUPRC	0.119	[0.028, 0.238]	0.0040
FedAdam	F1	0.122	[0.020, 0.243]	0.0227
FedAdam	Brier improvement	0.037	[0.007, 0.071]	0.0153
FedAdam-Cal	AUPRC	0.083	[0.008, 0.186]	0.0260
FedAdam-Cal	F1	0.131	[0.041, 0.230]	0.0040
FedAdam-Cal	Brier improvement	0.008	[−0.019, 0.039]	0.5620
Centralized pooled	AUPRC	0.009	[−0.042, 0.069]	0.7367
Centralized pooled	F1	−0.021	[−0.088, 0.036]	0.6227
Centralized pooled	Brier improvement	−0.009	[−0.025, 0.009]	0.2880

Table 12. FREO ablation study on the primary held-out task under the non-overlap split.

Variant	AUROC	AUPRC	F1	Recall@FPR0.10	Brier
FREO full	0.966	0.923	0.852	0.923	0.079
Without residual elements	0.920	0.851	0.733	0.731	0.125
No compliance weights	0.970	0.930	0.852	0.923	0.076
No residual transport	0.965	0.922	0.868	0.885	0.073
No facility priors	0.920	0.843	0.667	0.731	0.143

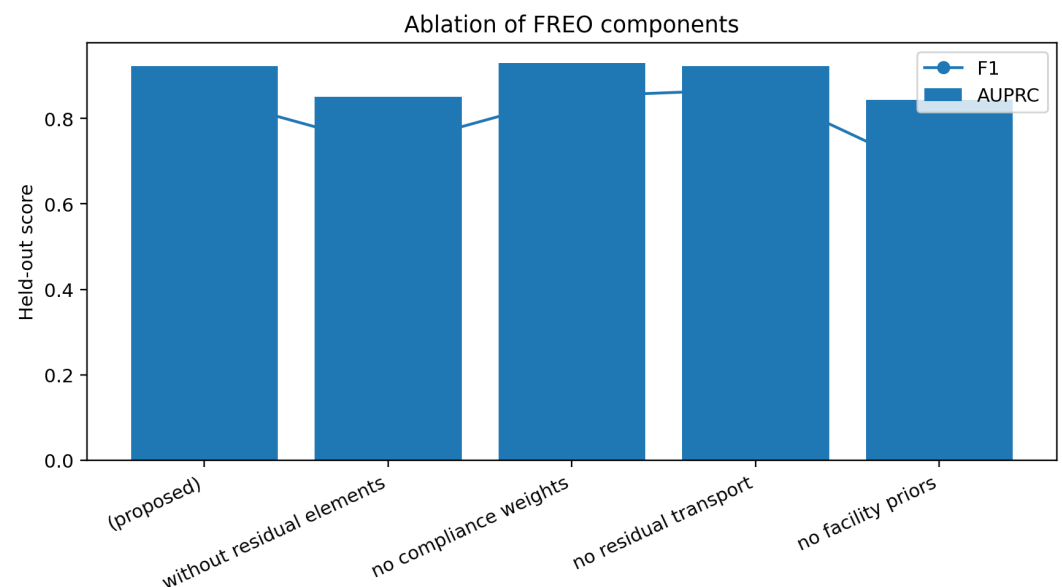


Figure 3. Ablation of FREO components. Residual elements and facility priors explain the largest changes in held-out F1 and AUPRC.

5.3. Client Heterogeneity and Pseudo-External Transfer

Federated environmental data are non-IID by construction: facilities differ by process type, sample count, permitted parameters and violation prevalence. Table 13 reports representative slices. FREO remained competitive in large-client and high non-IID slices

and improved the operating-point F1 in the small-client slice, although its small-client AUPRC was lower than FedAvg. Slices with very few positives were retained in the supplementary comma-separated-value (CSV) files but should not be overinterpreted.

Table 13. Client-heterogeneity stress slices under the non-overlap split.

Slice	Method	<i>n</i>	Positives	AUPRC	F1
Small clients	FedAvg	36	14	0.819	0.733
Small clients	FREO	36	14	0.782	0.828
Large clients	FedAvg	61	12	0.954	0.733
Large clients	FREO	61	12	0.957	0.880
High non-IID clients	FedAvg	64	25	0.947	0.846
High non-IID clients	FREO	64	25	0.950	0.902
Food and beverage manufacturing	FedAvg	47	19	0.917	0.780
Food and beverage manufacturing	FREO	47	19	0.931	0.850

To assess whether the method was merely memorizing the ten supplied facility categories, we also performed leave-one-facility-category transfer validation (Table 14). Each evaluable category was held out entirely, the model was trained on all other categories, and the held-out category was evaluated without labeled held-out priors. This is a harder transfer setting than the main temporal split. FREO-transfer was comparable to FedAvg in AUPRC and slightly higher in F1, but it did not dominate in all metrics and had a worse Brier score. The result should therefore be read as a general stress test rather than a superiority claim.

A category-cluster bootstrap over the seven evaluable held-out categories confirms that transfer uncertainty is large (Table 15). FREO-transfer and FedAvg have nearly identical AUPRC weighted means, FREO-transfer has a modestly higher F1 weighted mean, and FedAvg has better AUROC and Brier. The transfer experiment therefore supports exploratory cross-category transportability, not broad new-sector superiority.

Table 14. Leave-one-facility-category transfer validation. Values are test-size-weighted means over evaluable held-out categories. FREO-transfer excludes held-out facility priors; Pos. denotes positive cases.

Method	Categories	<i>n</i>	Pos.	AUROC	AUPRC	F1	Brier
FedAvg	7	210	85	0.682	0.575	0.416	0.228
FedAdam	7	210	85	0.581	0.483	0.193	0.290
FREO-transfer	7	210	85	0.660	0.576	0.432	0.265

Table 15. Leave-one-facility-category transfer with category-cluster bootstrap intervals over seven evaluable categories. Values are test-size-weighted means with 95% bootstrap intervals.

Method	AUROC	AUPRC	F1	Brier
FedAvg	0.682 [0.538, 0.842]	0.575 [0.306, 0.887]	0.416 [0.164, 0.677]	0.228 [0.153, 0.323]
FedAdam	0.581 [0.486, 0.666]	0.483 [0.223, 0.754]	0.193 [0.060, 0.331]	0.290 [0.153, 0.453]
FREO-transfer	0.660 [0.509, 0.823]	0.576 [0.308, 0.889]	0.432 [0.205, 0.649]	0.265 [0.174, 0.393]

5.4. Missingness, Noise, Communication and Dropout

Under simultaneous 35% feature masking and 0.20 measurement-feature noise, FREO retained AUPRC 0.897, F1 0.819 and Brier score 0.092. FedAvg under the same stress retained AUPRC 0.760, F1 0.626 and Brier score 0.138 (Table 16). Figure 4 shows the degradation trend across masking and noise levels. These results indicate that residual elements reduce sensitivity to missing analyte channels and noisy current measurements.

Table 16. Robustness to missing analyte channels and measurement-feature noise under the non-overlap split. Perturbation denotes 35% feature masking and 0.20 measurement-feature noise.

Method	Clean AUPRC	Perturbed AUPRC	Perturbed F1	Perturbed Brier
FedAvg	0.850	0.760 ± 0.044	0.626 ± 0.055	0.138 ± 0.009
FedProx	0.851	0.761 ± 0.044	0.626 ± 0.055	0.138 ± 0.009
FedAdam	0.804	0.508 ± 0.089	0.517 ± 0.086	0.210 ± 0.030
FREO	0.923	0.897 ± 0.042	0.819 ± 0.056	0.092 ± 0.010

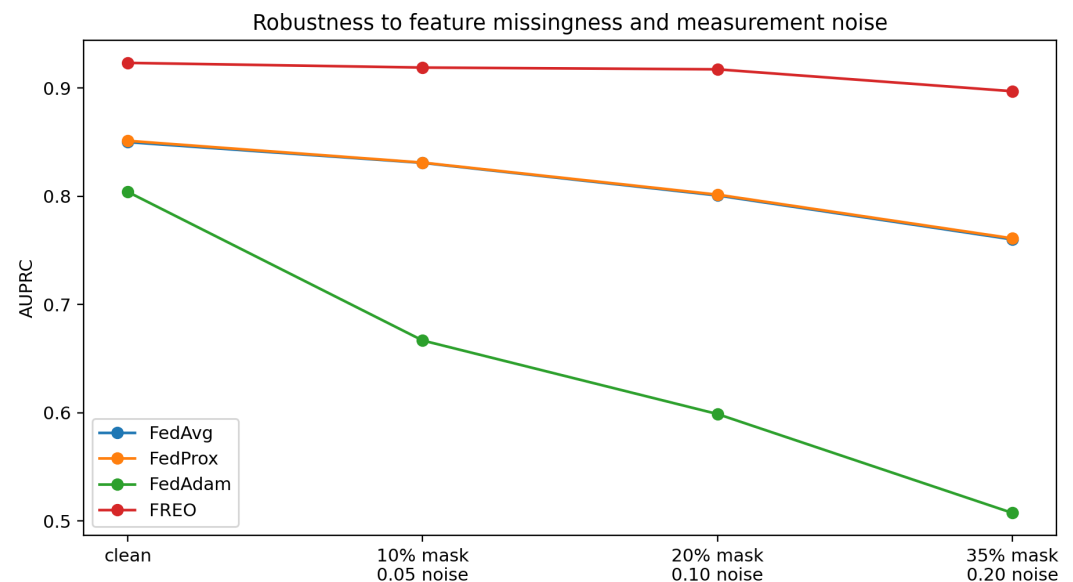


Figure 4. AUPRC under increasing feature missingness and measurement-feature noise. FREO degrades more slowly than the generic federated baselines.

FREO communicated approximately 562.5 KB over 48 rounds because each client sent residual-element components in addition to model deltas. The cost was larger than FedAvg/FedProx/FedAdam but small in absolute terms, as shown in Figure 5. Under 30% client dropout, FREO retained AUPRC 0.927, F1 0.877 and Brier score 0.079, averaged over dropout replicates (Table 17).

Table 17. Communication and representative 30% client-dropout results under the non-overlap split. Communication is reported in kilobytes (KB), and rounds denote rounds to within 1% of final training loss.

Method	KB	Rounds	Clean F1	Dropout AUPRC	Dropout F1	Dropout Brier
FedAvg	337.5	26	0.733	0.848 ± 0.020	0.707 ± 0.020	0.126 ± 0.005
FedProx	337.5	26	0.733	0.848 ± 0.021	0.716 ± 0.020	0.126 ± 0.005
FedAdam	337.5	26	0.730	0.675 ± 0.155	0.612 ± 0.100	0.187 ± 0.044
FREO	562.5	41	0.852	0.927 ± 0.008	0.877 ± 0.011	0.079 ± 0.003

5.5. Rare Severe-Tail Detection and Noisy-Aggregation Sensitivity

Severe-tail detection is reported as a secondary stress test rather than the primary claim because the held-out severe-tail set contains only seven positive events (Table 18). FedAvg and FedProx achieved the highest F1 at the training-selected threshold, while FREO retained high AUROC and recall at the fixed-FPR operating point. More severe-tail data are needed before making strong claims about rare exceedance superiority.

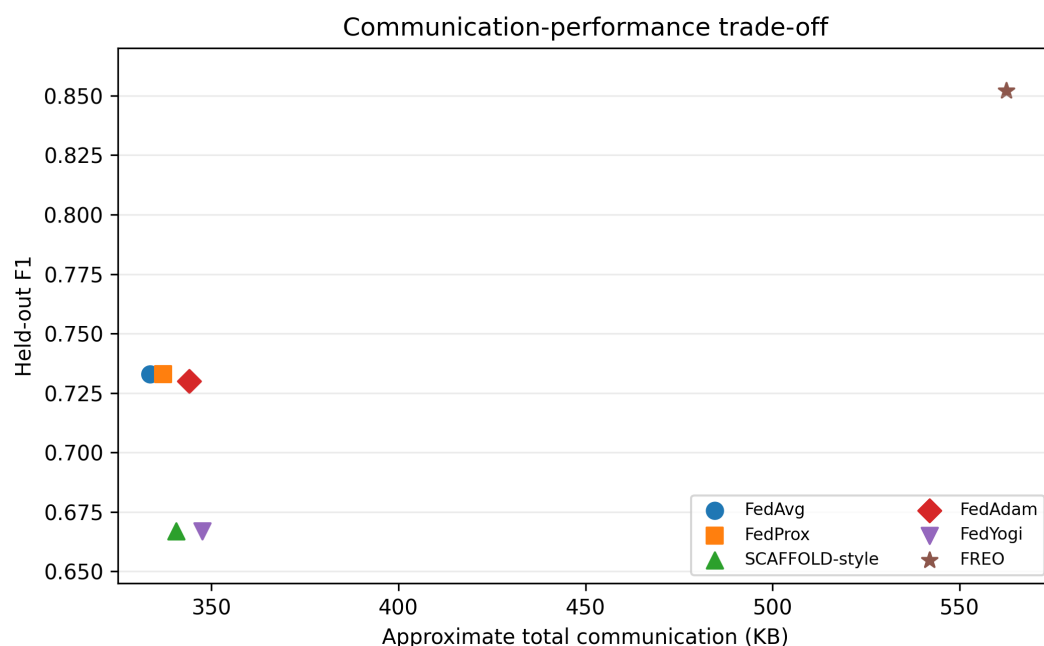


Figure 5. Communication-estimated performance. FREO uses additional residual-element summaries but remains lightweight in absolute communication volume; small horizontal offsets separate methods with identical estimated communication.

Table 18. Secondary rare severe-tail task under the non-overlap split. The held-out set contains seven positive events.

Method	Positives	AUROC	AUPRC	F1	Recall@FPR0.10	Brier
FedAvg	7	0.970	0.788	0.667	0.714	0.096
FedProx	7	0.970	0.788	0.667	0.714	0.097
FedAdam	7	0.970	0.647	0.625	0.571	0.048
FREO	7	0.957	0.713	0.500	0.714	0.057

The final experiment adds Gaussian noise to aggregated update summaries. This is a sensitivity analysis only; it is not a formal differential privacy guarantee and no ϵ value is claimed. FREO remained stable under moderate update perturbation (Table 19).

Table 19. Noisy-aggregation sensitivity for FREO under the non-overlap split. SD denotes standard deviation.

Noise SD	Replicates	AUROC	AUPRC	F1	Brier
0.000	1	0.966	0.923	0.852	0.079
0.005	8	0.967	0.921 $\pm$ 0.003	0.864 $\pm$ 0.014	0.079 $\pm$ 0.001
0.010	8	0.967	0.923 $\pm$ 0.005	0.860 $\pm$ 0.015	0.079 $\pm$ 0.002
0.020	8	0.966	0.921 $\pm$ 0.016	0.869 $\pm$ 0.014	0.080 $\pm$ 0.004
0.040	8	0.962	0.904 $\pm$ 0.042	0.848 $\pm$ 0.024	0.083 $\pm$ 0.007

5.6. Auxiliary Fixed-Feature, Linear-Head Image Stress Tests

The auxiliary fixed-feature, linear-head image stress tests provide non-regulatory checks of the residual-element adapter against a wider FL comparison set. They should not be read as standard image-classification benchmarks or as a vision-leaderboard claim because every method operates on the same handcrafted fixed features with a generalized linear head. Their role is limited to testing whether the same client-state mechanism remains useful in controlled non-IID federated classification settings.

The auxiliary comparison is intentionally read in two layers. Table 20 gives the take-home result, while Table 21 retains the full method inventory for reproducibility. The key finding is that FREO-Adapter is practically tied with the matched calibrated FedAvgM-Cal reference: CIFAR-10 macro-F1 differs by -0.002 and FEMNIST-10 macro-F1 differs by -0.001 . The larger gains are relative to ordinary FedAvg, not to FedAvgM-Cal. FREO-full-Cal is therefore reported only as a diagnostic full-FREO variant; because CIFAR-10 and FEMNIST-10 have no permit limits or compliance weights, the cross-domain residual-element result is the adapter comparison.

Table 20. Take-home auxiliary macro-F1 comparison. Values are mean $\pm$ standard deviation over five seeds. The main comparison is FREO-Adapter versus the matched calibrated FedAvgM-Cal reference; the ordinary FedAvg column shows the gain relative to an uncalibrated non-IID FL baseline.

Dataset	FedAvg	Best Uncal./Pers.	FedAvgM-Cal	FREO-Adapter	Δ vs. FedAvgM-Cal
CIFAR-10	0.478 ± 0.006	0.684 ± 0.022	0.741 ± 0.029	0.739 ± 0.029	-0.002
FEMNIST-10	0.908 ± 0.006	0.946 ± 0.004	0.961 ± 0.005	0.960 ± 0.004	-0.001

To make the full table easier to scan, Table 21 should be read in the following groups: ordinary FL optimizers (FedAvg through FedNova), uncalibrated personalization references (FedNova+Ditto and FedNova+pFedMe-style), the strongest matched calibrated reference (FedAvgM-Cal), and the FREO variants. This grouping separates the main adapter result from the diagnostic FREO-full-Cal row and avoids interpreting the table as a single undifferentiated leaderboard.

The headline table is the intended reading of the larger auxiliary comparison. On CIFAR-10, FREO-Adapter achieved macro-F1 0.739 ± 0.029 , compared with 0.478 ± 0.006 for ordinary FedAvg and 0.741 ± 0.029 for the strongest matched calibrated federated reference, FedAvgM-Cal. On FEMNIST-10, FREO-Adapter achieved macro-F1 0.960 ± 0.004 , compared with 0.908 ± 0.006 for ordinary FedAvg and 0.961 ± 0.005 for FedAvgM-Cal. Figure 6 visualizes these macro-F1 patterns, and Table 22 gives the paired seed-level comparisons. Thus, the auxiliary stress tests show a large improvement over ordinary non-IID FL baselines and practical comparability with the strongest matched calibrated reference, rather than superiority over that reference. This is why the main auxiliary conclusion is a match to FedAvgM-Cal under the same client-prior calibration protocol.

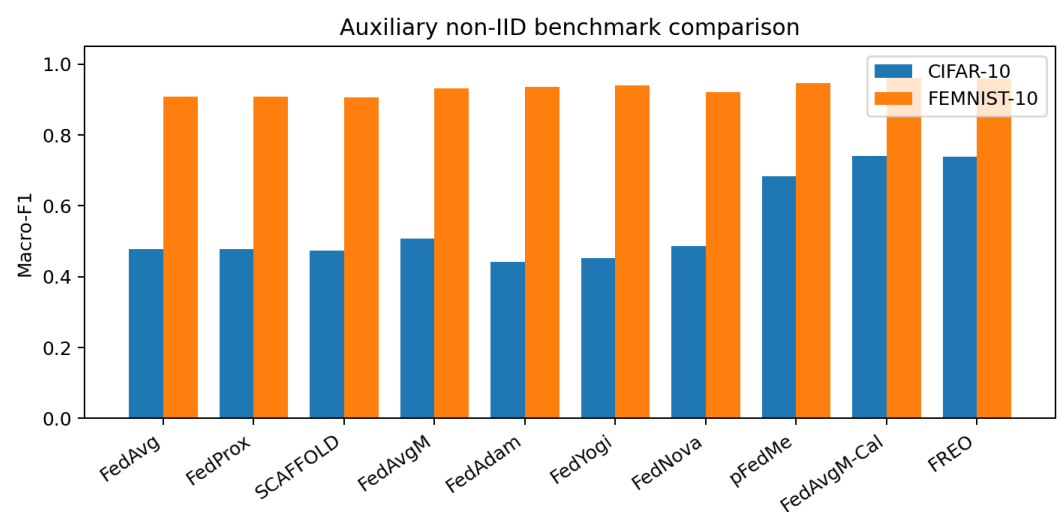


Figure 6. Macro-F1 on auxiliary non-IID CIFAR-10 and FEMNIST-10 fixed-feature stress tests. Error bars denote standard deviation over five seeds; the FREO bar denotes FREO-Adapter.

Table 21. Detailed auxiliary non-IID CIFAR-10 and FEMNIST-10 fixed-feature stress test results over five seeds. Values are mean $\pm$ standard deviation. Calibrated variants use client-prior residuals estimated from training data and select calibration strength on validation data only. Lower negative log-likelihood (NLL) is better; higher other metrics are better.

Dataset	Method	Accuracy	Macro-F1	NLL
<i>CIFAR-10: ordinary federated optimizers</i>				
CIFAR-10	FedAvg	0.488 $\pm$ 0.007	0.478 $\pm$ 0.006	1.482 $\pm$ 0.004
CIFAR-10	FedProx	0.488 $\pm$ 0.007	0.478 $\pm$ 0.006	1.482 $\pm$ 0.004
CIFAR-10	SCAFFOLD	0.484 $\pm$ 0.012	0.474 $\pm$ 0.014	1.494 $\pm$ 0.018
CIFAR-10	FedAvgM	0.514 $\pm$ 0.008	0.508 $\pm$ 0.008	1.456 $\pm$ 0.021
CIFAR-10	FedAdam	0.445 $\pm$ 0.012	0.442 $\pm$ 0.011	5.395 $\pm$ 0.116
CIFAR-10	FedYogi	0.458 $\pm$ 0.005	0.453 $\pm$ 0.005	3.538 $\pm$ 0.113
CIFAR-10	FedNova	0.495 $\pm$ 0.011	0.486 $\pm$ 0.015	1.440 $\pm$ 0.014
<i>CIFAR-10: personalization, calibration and FREO variants</i>				
CIFAR-10	FedNova+Ditto	0.682 $\pm$ 0.021	0.680 $\pm$ 0.022	1.058 $\pm$ 0.048
CIFAR-10	FedNova+pFedMe-style	0.686 $\pm$ 0.022	0.684 $\pm$ 0.022	1.046 $\pm$ 0.049
CIFAR-10	FedAvgM-Cal	0.743 $\pm$ 0.029	0.741 $\pm$ 0.029	0.753 $\pm$ 0.075
CIFAR-10	FREO-full-Cal	0.721 $\pm$ 0.033	0.719 $\pm$ 0.033	0.793 $\pm$ 0.076
CIFAR-10	FREO-Adapter	0.741 $\pm$ 0.028	0.739 $\pm$ 0.029	0.756 $\pm$ 0.074
<i>FEMNIST-10: ordinary federated optimizers</i>				
FEMNIST-10	FedAvg	0.909 $\pm$ 0.006	0.908 $\pm$ 0.006	0.427 $\pm$ 0.007
FEMNIST-10	FedProx	0.909 $\pm$ 0.006	0.908 $\pm$ 0.006	0.427 $\pm$ 0.007
FEMNIST-10	SCAFFOLD	0.907 $\pm$ 0.007	0.907 $\pm$ 0.007	0.430 $\pm$ 0.010
FEMNIST-10	FedAvgM	0.932 $\pm$ 0.006	0.932 $\pm$ 0.006	0.298 $\pm$ 0.031
FEMNIST-10	FedAdam	0.935 $\pm$ 0.005	0.935 $\pm$ 0.005	0.708 $\pm$ 0.061
FEMNIST-10	FedYogi	0.940 $\pm$ 0.006	0.940 $\pm$ 0.006	0.500 $\pm$ 0.051
FEMNIST-10	FedNova	0.921 $\pm$ 0.004	0.921 $\pm$ 0.004	0.371 $\pm$ 0.007
<i>FEMNIST-10: personalization, calibration and FREO variants</i>				
FEMNIST-10	FedNova+Ditto	0.944 $\pm$ 0.004	0.944 $\pm$ 0.004	0.272 $\pm$ 0.010
FEMNIST-10	FedNova+pFedMe-style	0.946 $\pm$ 0.004	0.946 $\pm$ 0.004	0.267 $\pm$ 0.010
FEMNIST-10	FedAvgM-Cal	0.961 $\pm$ 0.005	0.961 $\pm$ 0.005	0.179 $\pm$ 0.021
FEMNIST-10	FREO-full-Cal	0.947 $\pm$ 0.003	0.947 $\pm$ 0.003	0.247 $\pm$ 0.017
FEMNIST-10	FREO-Adapter	0.960 $\pm$ 0.004	0.960 $\pm$ 0.004	0.208 $\pm$ 0.065

Table 22. Paired seed-level auxiliary comparisons for FREO-Adapter. Delta is FREO-Adapter minus its comparator for macro-F1; positive values favor FREO-Adapter. The interval is a paired bootstrap over the five seeds; CI denotes confidence interval.

Dataset	Comparator	Mean Δ F1	95% CI	Sign-Flip p
CIFAR-10	FedAvg	0.261	[0.246, 0.280]	0.063
CIFAR-10	FedNova+pFedMe-style	0.055	[0.048, 0.062]	0.063
CIFAR-10	FedNova+Ditto	0.059	[0.051, 0.068]	0.063
CIFAR-10	FedAvgM-Cal	−0.002	[−0.006, 0.000]	1.000
FEMNIST-10	FedAvg	0.052	[0.050, 0.054]	0.063
FEMNIST-10	FedNova+pFedMe-style	0.014	[0.012, 0.017]	0.063
FEMNIST-10	FedNova+Ditto	0.016	[0.014, 0.018]	0.063
FEMNIST-10	FedAvgM-Cal	−0.001	[−0.003, 0.000]	1.000

The paired seed-level analysis further clarifies the interpretation. FREO-Adapter is consistently above ordinary FedAvg and the strongest uncalibrated or personalized FedNova references, but it is essentially tied with FedAvgM-Cal. The direct FREO-full-Cal variant was weaker than FREO-Adapter on these non-regulatory fixed-feature stress tests, indicating that when permit residuals and compliance weights are absent, the transferable mechanism is mainly the training-only residual-element adapter rather than the full

compliance-weighted environmental objective. We therefore treat FREO-Adapter versus FedAvgM-Cal as the focal auxiliary comparison, not as evidence that FREO-Adapter beats the matched calibrated reference.

6. Discussion

FREO is best interpreted as a compliance-aware federated environmental AI framework, but its empirical implementation is intentionally simpler than the framework terminology may suggest. In the reported experiments, FREO denotes residual-summary feature construction, a full-batch federated logistic residual operator, compliance/sample weighting, reliability-weighted aggregation and training-only facility-prior calibration; optional deep residual, learned-transport and dual-constraint modules were not used. The updated raw baseline includes permit lower/upper limits and max/min/range limit direction indicators, so the representation result now isolates the additional value of signed residualization beyond access to permit metadata. The empirical evidence supports two claims. First, signed compliance residuals provide a useful common coordinate system for heterogeneous effluent records. Second, residual elements and training-only facility priors provide a compact way to transport compliance-relevant state without centralizing raw measurements. The ablations and the new matched calibrated FedAvg/FedAdam references sharpen this interpretation: residual features and facility-prior calibration were the components most responsible for the main gains, while compliance weights and residual transport were modular refinements that did not uniformly improve every calibration metric in this dataset. FedAvg-Cal nearly matched FREO in AUPRC and F1 and had a slightly lower Brier score, so the full FREO stack should not be described as the sole source of improvement. Accordingly, the observed gains should be attributed primarily to the auditable residual representation plus training-only facility calibration, with FREO contributing the integrated residual-element state, reliability/transport weighting and raw-data-local compliance-screening workflow, not to a claim that conventional FL optimization or shrinkage calibration is new.

The paired bootstrap tests strengthen the statistical interpretation of the main result, but the facility-cluster bootstrap also shows why the result should be interpreted cautiously. Row-level paired resampling gave positive intervals for FREO over ordinary FedAvg, FedProx and FedAdam on several screening metrics, but those intervals treat facility-period records as the resampling unit and are best read as conditional on the observed facilities. Facility-cluster resampling widened the intervals because only ten held-out facilities were available. We therefore treat the cluster-level analysis as the stronger guide to generalization across facilities. After adding the matched calibrated references and revising the split to remove target/input boundary overlap, the strongest statement is more specific: FREO improves the AUPRC/F1 screening operating point relative to ordinary uncalibrated federated baselines, but FedAvg-Cal is slightly stronger on AUPRC, F1 and Brier in the main non-overlap split. Centralized pooled regression retained a strong Brier score and F1, it but pools raw training records and is not a deployment-equivalent comparator when such pooling is infeasible. We therefore treat it as an upper-reference and diagnostic calibration baseline, while interpreting FREO as a raw-data-local federated alternative, rather than as a method that must dominate either centralized learning or matched calibrated federated references on every metric.

The transfer validation is intentionally conservative. When an entire facility category is withheld, held-out facility priors are unavailable. FREO-transfer performed comparably to FedAvg in AUPRC and slightly better in F1, but did not dominate AUROC or Brier, and the category-cluster intervals were wide. This suggests that facility-prior calibration

is valuable when training history exists, and that additional cross-regime data would be needed for stronger claims about new-sector transfer.

The CIFAR-10 and FEMNIST-10 auxiliary fixed-feature stress tests cannot be used for regulatory external validation, but they show that residual-element personalization is not merely a quirk of the EMA spreadsheets. The compact headline comparison is the intended reading: FREO-Adapter matches FedAvgM-Cal while substantially outperforming ordinary FedAvg and uncalibrated personalization references. This is useful because the auxiliary experiments no longer rely on comparison against uncalibrated FL alone, and all calibrated methods use the same training-only client-prior construction and validation-selection rule.

Regulatory Deployment Pathway

In a real regulatory setting, FREO would be deployed as a decision-support layer rather than an autonomous enforcement mechanism. A regulator would first freeze a data dictionary for analyte names, units, limit directions, range constraints, non-detect/no-flow flags and reporting periods. Each facility would then run a local FREO client behind its own data-governance boundary to map submitted measurements into signed residuals, compute residual elements and train on local records; only bounded model summaries and residual-element summaries would be sent to the federated aggregation service.

Before operational use, the regulator would prespecify the prediction target, validation period, risk threshold and escalation rule. FREO risk scores would prioritize inspections, requests for clarification, targeted sampling or technical assistance, while legal compliance determinations would still be made from submitted measurements, permit conditions and human regulatory review. Routine use would require audit logs for the model version, residual-element schema, participating facilities, feature dictionary, threshold and calibration state, together with monitoring for calibration drift, missingness shifts, new facilities without historical priors and periodic retraining. If stronger confidentiality guarantees are required, the same deployment can add secure aggregation or differential privacy mechanisms around the model-update channel.

There are limitations. First, the effective independent regulatory sample is small: it comprises ten workbook-level facility-category clients. The 4036 assessable observations are source measurements used to compute residual features, and the 311 prediction records are repeated facility-period observations within those workbook-level clients. Although 99.4% of the paired periods are one calendar month apart, two irregular gaps occur, so the predictive target is more accurately described as next-period rather than strictly calendar-month. Accordingly, the present evidence supports within-facility temporal prediction and exploratory cross-category transfer in this environmental context rather than universal environmental compliance generalization. The fixed-feature, linear-head CIFAR-10 and FEMNIST-10 results add non-regulatory FL stress tests but do not replace external environmental replication with additional facilities or serve as standard image-classification benchmark evidence. As such, the primary effluent hyperparameters were prespecified or selected from training-only convergence checks rather than being tuned on a separate environmental validation set; larger multi-facility studies should use nested facility-level validation for hyperparameter selection. Second, the current prototype uses auditable fixed-feature generalized linear residual operators; nonlinear backbones may improve accuracy but would change the comparison set and require separate tuning. Third, the auxiliary stress test labels denote optimizer and personalization logic implemented on a shared handcrafted fixed-feature head; they should not be interpreted as canonical deep vision leaderboard implementations or standard image-classification benchmark results. Fourth, raw-data locality is not a formal privacy guarantee. Secure aggregation and calibrated differential privacy should be added for deployments that require mathematical

privacy bounds. Fifth, operational regulatory use would require implementation of the deployment controls described above, including domain validation of permit semantics, unit harmonization, parameter-specific residual-scale floors, sampling flags, non-detect conventions, facility-specific reporting rules, audit logs, monitoring and human review.

7. Conclusions

This paper presents FREO, a federated residual-element framework for compliance-aware environmental AI. The methodological novelty is the residual-element compliance state and its permit-residual training protocol, not the standard FL optimizers or calibration routines used as comparators and components. FREO maps heterogeneous effluent records into signed regulatory residuals, communicates with compact residual elements rather than raw samples, and optimizes a shared next-period compliance risk operator intended for human-in-the-loop regulatory screening rather than automatic enforcement. In deployment, FREO should be paired with a prespecified data dictionary, unit-conversion policy, parameter-specific residual-scale floors, validation period, operating thresholds, audit logs, calibration/drift monitoring and, where required, secure aggregation or differential privacy. On this small ten-workbook Environmental Management Authority dataset, FREO was stronger than ordinary uncalibrated federated baselines on a primary non-overlap within-facility temporal high-burden task, but the matched calibrated FedAvg-Cal baseline was slightly stronger on AUPRC, F1 and Brier. This reinforces the main attribution: residual features and facility-prior calibration account for much of the improvement over ordinary FedAvg, while FREO contributes the integrated residual-element state, reliability-weighted aggregation and raw-data-local compliance-screening workflow. FREO also remained robust under missingness, noise, client dropout and update perturbation. The added CIFAR-10 and FEMNIST-10 fixed-feature, linear-head stress tests, implemented with handcrafted features and generalized linear heads, show that residual-element personalization can improve non-IID FL classification relative to ordinary federated baselines; the key auxiliary result is a practical tie between FREO-Adapter and FedAvgM-Cal, not superiority over that calibrated reference. Transfer validation was encouraging but mixed, and the cluster-level intervals remained wide, so the evidence supports FREO as a reproducible within-facility temporal environmental-AI and federated-learning framework with auxiliary non-regulatory FL support, rather than as a universally superior federated optimizer. Future work will examine concrete nonlinear and deep-model extensions rather than treating them as an unspecified placeholder. Promising directions include monotone or shape-constrained neural residual operators that can learn nonlinear analyte interactions while preserving auditable directionality; temporal sequence models over facility-month residual histories for persistent or delayed compliance stress; graph or message-passing variants that share information among facilities with similar sectors, permit and analyte profiles that do not expose raw records, and multimodal or deep encoder variants for high-frequency sensor streams and textual inspection metadata. Each extension would require external environmental validation, matched calibrated deep federated baselines and the same deployment controls emphasized here, including secure aggregation where required, drift monitoring and human regulatory review.

Supplementary Materials: The following supporting information can be downloaded at <https://www.mdpi.com/article/10.3390/ai7070253/s1>.

Author Contributions: Conceptualization, K.K., W.E. and S.R.; methodology, K.K. and W.E.; software, K.K.; validation, W.E. and S.R.; formal analysis, K.K. and W.E.; investigation, W.E. and S.R.; resources, S.R.; data curation, S.R. and K.K.; writing—original draft preparation, K.K.; writing—review and

editing, K.K., W.E. and S.R.; visualization, K.K.; supervision, W.E.; project administration, S.R. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable. This study analyzes facility-level effluent-monitoring records and does not involve human participants or animals.

Informed Consent Statement: Not applicable.

Data Availability Statement: Due to regulatory and facility confidentiality constraints, raw facility-identifiable workbooks are available only subject to permission from the data owner and, where applicable, Environmental Management Authority approval. De-identified processed tables, feature matrices, experiment scripts and result summaries are provided as Supplementary Material.

Acknowledgments: The authors acknowledge the Environmental Management Authority data providers for the effluent quality workbooks used in this study.

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

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial intelligence
AUPRC	Area under the precision–recall curve
AUROC	Area under the receiver operating characteristic curve
CSV	Comma-separated value
EMA	Environmental Management Authority
FEMNIST	Federated Extended Modified National Institute of Standards and Technology
FEMNIST-10	10-class FEMNIST subset used in this study
FL	Federated learning
FREO	Federated residual-element optimization
IID	Independent and identically distributed
KB	Kilobyte
ROC	Receiver operating characteristic

Appendix A. Reproducibility Details

The script provided in the Supplementary Materials `reproduce_mdpi_ai_freo_final.py` regenerates the processed result tables from the de-identified observation comma-separated-value (CSV) file. Running the script creates a results directory containing the main baseline, ablation, robustness, dropout, update-noise, rare-tail, bootstrap, cluster-bootstrap and transfer-validation CSV files. The default bootstrap count is 3000; a smaller exploratory count can be set with the environment variable `FREO_BOOTSTRAP`. The auxiliary scripts are `reproduce_sota_aux_benchmarks.py` and `run_one_sota_aux_seed.py`. These scripts parse the attached CIFAR-10 and FEMNIST archives, form the fixed-feature stress test subsets, and run the five-seed non-IID experiments against the expanded federated baseline suite. The script `compute_sota_aux_paired_tests.py` regenerates the paired seed-level comparison table.

Appendix B. Supplementary Comma-Separated-Value (CSV) Files

The Supplementary Materials also include CSV files that contain de-identified observations, the next-period facility-month matrix, the facility-code/sector/workbook-client map, baseline metrics, permit-metadata-matched raw-representation checks, ablations, bootstrap confidence intervals, paired bootstrap tests, facility-cluster bootstrap sensitivity, temporal gap/split boundary audit, heterogeneity slices, leave-one-facility-category

transfer validation with category-cluster intervals, missingness/noise robustness, communication/dropout results, update-noise sensitivity results, rare severe-tail results and auxiliary CIFAR-10/FEMNIST-10 fixed-feature stress test run, summary and paired seed-level comparison tables, and residual unit-scale metadata.

Appendix C. Proof of Proposition

Proof. Let $F(\theta) = \sum_k p_k F_k(\theta, E_k)$ denote the residual-element objective at the current round, with aggregation weights p_k . Let $\hat{g}_r$ be the aggregated stochastic update returned to the server at round r . Decompose the update as

$$\hat{g}_r = \nabla F(\theta^r) + \xi_r + d_r + e_r, \quad (\text{A1})$$

where ξ_r is the zero-mean stochastic gradient noise, d_r is local client drift caused by multiple local steps and non-IID objectives, and e_r is residual-element compression or drift error. By assumption, $\mathbb{E}\|\xi_r\|^2 \leq \sigma^2$, $\mathbb{E}\|d_r\|^2 \leq C_d \Delta_{\text{het}}$, and $\mathbb{E}\|e_r\|^2 \leq C_e \epsilon_E^2$. The server update is $\theta^{r+1} = \theta^r - \eta \hat{g}_r$.

By L -smoothness,

$$F(\theta^{r+1}) \leq F(\theta^r) - \eta \langle \nabla F(\theta^r), \hat{g}_r \rangle + \frac{L\eta^2}{2} \|\hat{g}_r\|^2. \quad (\text{A2})$$

Taking conditional expectation and using the decomposition above gives a descent term in $\|\nabla F(\theta^r)\|^2$ plus variance, heterogeneity and residual-element approximation terms. Applying Young's inequality to the inner products involving d_r and e_r , and using $\eta \leq 1/(4L)$, yields

$$\mathbb{E}F(\theta^{r+1}) \leq \mathbb{E}F(\theta^r) - \frac{\eta}{2} \mathbb{E}\|\nabla F(\theta^r)\|^2 + C_1 \eta^2 \sigma^2 + C_2 \eta^2 \Delta_{\text{het}} + C_3 \eta \epsilon_E^2. \quad (\text{A3})$$

Summing over $r = 0, \dots, T-1$, telescoping the objective values and using $F(\theta^T) \geq F^*$ gives

$$\frac{1}{T} \sum_{r=0}^{T-1} \mathbb{E}\|\nabla F(\theta^r)\|^2 \leq \frac{2(F(\theta^0) - F^*)}{\eta T} + C'_1 \eta \sigma^2 + C'_2 \eta \Delta_{\text{het}} + C'_3 \epsilon_E^2, \quad (\text{A4})$$

which is the stated bound after absorbing constants. The proof follows the usual nonconvex FL descent template; the only additional term is the residual-element approximation term. $\square$

References

1. McMahan, B.; Moore, E.; Ramage, D.; Hampson, S.; y Arcas, B.A. Communication-efficient learning of deep networks from decentralized data. In Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, Fort Lauderdale, FL, USA, 20–22 April 2017; pp. 1273–1282.
2. Yang, Q.; Liu, Y.; Chen, T.; Tong, Y. Federated machine learning: Concept and applications. *ACM Trans. Intell. Syst. Technol.* **2019**, *10*, 12. [CrossRef]
3. Kairouz, P.; McMahan, H.B.; Avent, B.; Bellet, A.; Bennis, M.; Bhagoji, A.N.; Bonawitz, K.; Charles, Z.; Cormode, G.; Cummings, R.; et al. Advances and open problems in federated learning. *Found. Trends Mach. Learn.* **2021**, *14*, 1–210. [CrossRef]
4. Li, T.; Sahu, A.K.; Zaheer, M.; Sanjabi, M.; Talwalkar, A.; Smith, V. Federated optimization in heterogeneous networks. In Proceedings of the 3rd Conference on Machine Learning and Systems, Austin, TX, USA, 2–4 March 2020; pp. 429–450.
5. Karimireddy, S.P.; Kale, S.; Mohri, M.; Reddi, S.J.; Stich, S.U.; Suresh, A.T. SCAFFOLD: Stochastic controlled averaging for federated learning. In Proceedings of the 37th International Conference on Machine Learning, Virtual Event, 13–18 July 2020; pp. 5132–5143.
6. Reddi, S.J.; Charles, Z.; Zaheer, M.; Garrett, Z.; Rush, K.; Konečný, J.; Kumar, S.; McMahan, H.B. Adaptive federated optimization. In Proceedings of the International Conference on Learning Representations, Virtual Event, 3–7 May 2021.

7. Wang, J.; Liu, Q.; Liang, H.; Joshi, G.; Poor, H.V. Tackling the objective inconsistency problem in heterogeneous federated optimization. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*; Curran Associates Inc.: Red Hook, NY, USA, 2020; Volume 33, pp. 7611–7623.
8. Li, T.; Hu, S.; Beirami, A.; Smith, V. Ditto: Fair and robust federated learning through personalization. In *Proceedings of the 38th International Conference on Machine Learning*, Virtual, 18–24 July 2021; pp. 6357–6368.
9. Fallah, A.; Mokhtari, A.; Ozdaglar, A. Personalized federated learning with theoretical guarantees: A model-agnostic meta-learning approach. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*; Curran Associates Inc.: Red Hook, NY, USA, 2020; Volume 33, pp. 3557–3568.
10. Dinh, C.T.; Tran, N.H.; Nguyen, T.D. Personalized federated learning with Moreau envelopes. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*; Curran Associates Inc.: Red Hook, NY, USA, 2020; Volume 33, pp. 21394–21405.
11. Miller, T.; Durlik, I.; Kostecka, E.; Puzskarek, A. Federated learning for environmental monitoring: A review of applications, challenges, and future directions. *Appl. Sci.* **2025**, *15*, 12685. [[CrossRef](#)]
12. Jenifel, M.G.; Linda, M.M. Privacy-preserving water quality forecasting using federated learning across distributed water monitoring nodes and optimized RPART modelling. *Environ. Sci. Water Res. Technol.* **2025**, *11*, 3143. [[CrossRef](#)]
13. Das, B.; Adel, A.; Jan, T.; Wahiduzzaman, M.D. Water quality management using federated deep learning in developing Southeastern Asian country. *Water Resour. Manag.* **2025**, *39*, 1893–1909. [[CrossRef](#)]
14. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778. [[CrossRef](#)]
15. Brier, G.W. Verification of forecasts expressed in terms of probability. *Mon. Weather Rev.* **1950**, *78*, 1–3. [[CrossRef](#)]
16. Gneiting, T.; Raftery, A.E. Strictly proper scoring rules, prediction, and estimation. *J. Am. Stat. Assoc.* **2007**, *102*, 359–378. [[CrossRef](#)]
17. Guo, C.; Pleiss, G.; Sun, Y.; Weinberger, K.Q. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, Sydney, Australia, 6–11 August 2017; pp. 1321–1330.
18. Davis, J.; Goadrich, M. The relationship between precision-recall and ROC curves. In *Proceedings of the 23rd International Conference on Machine Learning*, Pittsburgh, PA, USA, 25–29 June 2006; pp. 233–240.
19. Fawcett, T. An introduction to ROC analysis. *Pattern Recognit. Lett.* **2006**, *27*, 861–874. [[CrossRef](#)]
20. Bonawitz, K.; Ivanov, V.; Kreuter, B.; Marcedone, A.; McMahan, H.B.; Patel, S.; Ramage, D.; Segal, A.; Seth, K. Practical secure aggregation for privacy-preserving machine learning. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, Dallas, TX, USA, 30 October–3 November 2017; pp. 1175–1191. [[CrossRef](#)]
21. Dwork, C.; McSherry, F.; Nissim, K.; Smith, A. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography*; Springer: Berlin/Heidelberg, Germany, 2006; pp. 265–284. [[CrossRef](#)]
22. Abadi, M.; Chu, A.; Goodfellow, I.; McMahan, H.B.; Mironov, I.; Talwar, K.; Zhang, L. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, Vienna, Austria, 24–28 October 2016; pp. 308–318. [[CrossRef](#)]
23. Krizhevsky, A. *Learning Multiple Layers of Features from Tiny Images*; Technical Report; University of Toronto: Toronto, ON, Canada, 2009.
24. Caldas, S.; Duddu, S.M.K.; Wu, P.; Li, T.; Konečný, J.; McMahan, H.B.; Smith, V.; Talwalkar, A. LEAF: A benchmark for federated settings. *arXiv* **2018**, arXiv:1812.01097.

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