Energy and AI 23 (2026) 100680

Contents lists available at ScienceDirect

Energy and AI

journal homepage: www.elsevier.com/locate/egyai

Towards secure federated learning for energy forecasting under adversarial attacks

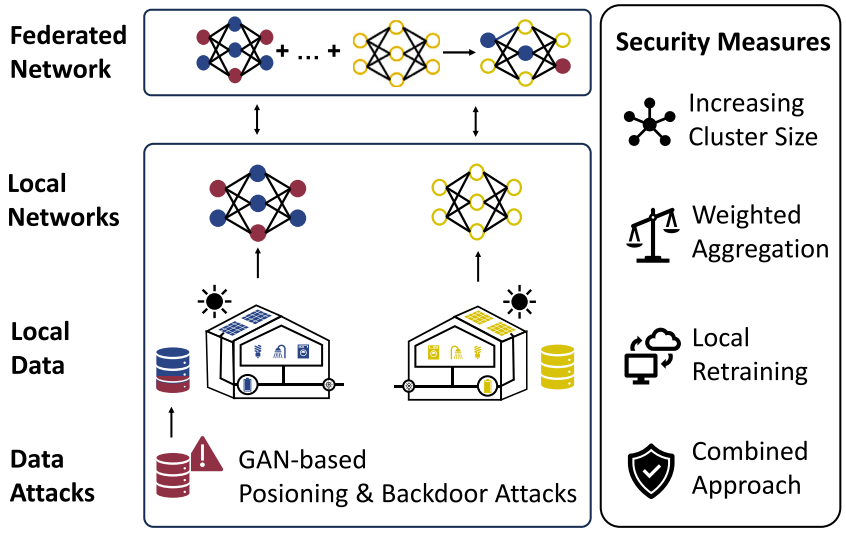
Jonas Sievers [a,*], Krupali Kumbhani [b], Thomas Blank [a], Frank Simon [a], Andreas Mauthe [b]

[a] Karlsruhe Institute of Technology (KIT), Institute for Data Processing and Electronics (IPE), Hermann-von-Helmholtz-Platz 1, Eggenstein-Leopoldshafen, 76344, Germany
[b] University of Koblenz, Institute for Information Systems Research, Universitaetsstrasse 1, Koblenz, 56070, Germany

HIGHLIGHTS

- Comprehensive analysis of attacks and defenses in federated energy forecasting.
- Data poisoning raises global errors by up to 131 %.
- Backdoor attacks increase local errors by up to 48 %.
- Generative manipulations outperform random perturbations.
- Security framework enhances resilience and restores accuracy.

GRAPHICAL ABSTRACT



ARTICLE INFO

Keywords:
Federated learning
Generative adversarial network
Poisoning attack
Backdoor attack

ABSTRACT

Federated learning is increasingly used in energy forecasting, enabling buildings to collaboratively predict load, photovoltaic generation, and prosumption while preserving data privacy. However, this collaborative nature introduces new vulnerabilities, as manipulations by a single participant can propagate across the network. Such attacks can undermine grid balancing, limit flexibility provision, and reduce trust in decentralized energy systems. This work presents a comprehensive study of adversarial threats and defenses in federated energy forecasting. We compare structured manipulations generated with Generative Adversarial Networks against simple random perturbations in two attack scenarios: (i) data poisoning, where corrupted training data degrade global accuracy, and (ii) backdoors, where hidden triggers distort predictions in targeted time windows. Our experiments show that poisoning can increase global forecasting errors by up to 131 %, while backdoors raise local errors by up to 48 %. In both cases, Generative Adversarial Network-based attacks are consistently more effective than random perturbations, with backdoors proving especially challenging to

* Corresponding author.
E-mail address: jonas.sievers@kit.edu (J. Sievers).

https://doi.org/10.1016/j.egyai.2026.100680
Received 15 September 2025; Received in revised form 4 December 2025; Accepted 4 January 2026
Available online 17 January 2026

detect due to their localized effect. To mitigate these threats, we evaluate four defense strategies: weighted aggregation, larger participant clusters, local retraining, and their coordinated integration into a secure framework. Results demonstrate that these defenses substantially reduce the impact of attacks, and in some cases even improve baseline accuracy, thereby enhancing the resilience of federated energy forecasting against adversarial manipulation.

1. Introduction

The increasing integration of renewable energy sources (RES) fundamentally transforms modern power systems. Variability and operational uncertainty increase as intermittent sources such as wind and PV generation become more prevalent. In this context, accurate short-term forecasting of electricity demand, PV generation, and prosumption is essential for efficient and reliable grid operation. Forecasts support a wide range of energy management tasks, including battery scheduling, demand response activation, and the optimization of PV self-consumption. At the residential and community level, precise predictions enable cost minimization, peak load reduction, and coordinated flexibility provision to local energy markets [1]. More generally, high-quality forecasts are essential for maintaining the real-time balance of supply and demand in increasingly decentralized and dynamic grid environments [2].

Achieving high forecasting accuracy typically requires access to high-resolution consumption and generation data. However, the use of such granular data raises substantial privacy concerns. Energy time series may reveal sensitive user information, including occupancy patterns and appliance usage. These concerns have contributed to public opposition against the large-scale deployment of smart metering infrastructure. Traditional privacy-preserving techniques, such as data aggregation, are often unsuitable for forecasting tasks, as they remove key temporal and user-specific patterns necessary for accurate model training [3].

Federated Learning (FL) addresses these challenges by enabling decentralized model training on client devices such as energy management systems (EMS). Only model updates are transmitted to a central server for aggregation, while raw data remains on-site [4]. To further mitigate challenges posed by non independent and identically distributed (non-iid) data, clustering can be employed to group clients with similar consumption or generation characteristics. Such grouping improves model convergence and enhances training stability.

Despite its privacy advantages, FL remains vulnerable to adversarial manipulation. Standard aggregation protocols accept client updates without verifying their integrity, creating opportunities for malicious interference [4]. Adversarial clients can exploit this by performing data poisoning, where local training is deliberately biased to degrade global model performance. Another threat arises from backdoor attacks that embed hidden triggers into local models. These models behave normally for typical inputs, but produce targeted mispredictions when specific patterns, such as temporal markers or atypical load profiles, are present. Because all updates are aggregated jointly, compromised model parameters propagate to every participant (Fig. 1).

While such vulnerabilities have been investigated in computer vision and natural language processing [5], their implications in energy systems are potentially more severe. At the building level, compromised forecasts distort local energy management, leading to misaligned energy storage scheduling, inefficient demand response, and unreliable flexibility provision. Aggregated across participants, these biases accumulate into systematic deviations between expected and actual net load. This undermines local balancing, reduces the reliability of congestion forecasts, complicates the coordination of RES, and reduces confidence in decentralized energy systems [6].

To address this, we provide a comprehensive analysis of adversarial threats in federated energy forecasting and propose a defense framework that mitigates both poisoning and backdoor attacks. Thus, we support the reliable deployment of FL in future energy systems.

1.1. Related work

To provide a comprehensive understanding of the current research and challenges in federated energy systems, we review related work on security measures and adversarial attacks. Additionally, we examine implementations of these attacks in the domains of computer vision and natural language processing. Selected publications are summarized in Table 1.

FL, originally introduced by McMahan et al. [4], enables collaborative model training across distributed clients without direct data sharing. Initially developed for mobile and edge applications, its adoption in the energy sector has expanded to several key domains where privacy preservation is critical.

In energy control, FL has been applied to decentralize the coordination of EMS. Lee et al. [22] propose a privacy-preserving framework for managing shared energy storage across multiple smart buildings, enabling coordinated scheduling of batteries and air conditioning systems through federated reinforcement learning. A consecutive study [23] extends this approach, showing that local model aggregation accelerates convergence and improves appliance-level scheduling under heterogeneous conditions. Rezazadeh et al. [24] further scale this concept to microgrids using a hierarchical architecture, where building-level EMS share hyperparameters with a federated layer. Their results demonstrate improved coordination of prosumption, reduced operating costs, and lower CO_2 emissions, all while preserving data privacy.

FL has also been explored in the context of non-intrusive load monitoring (NILM) and residential load forecasting. Giuseppi et al. [25] propose a decentralized FL variant without a central server, achieving comparable accuracy to federated averaging in appliance disaggregation while eliminating single points of failure. Wang et al. [26] address consumer classification using smart meter data by integrating federated training with privacy-preserving principal component analysis, demonstrating strong performance with three weighted averaging strategies. He et al. [27] combine FL with K-means clustering for short-term load forecasting, showing that intra-cluster federated training improves prediction accuracy and protects household-level privacy.

Despite these advances, systematic investigations of adversarial robustness in federated energy systems remain limited. Qureshi et al. [17] investigate data poisoning in federated load forecasting with Long short-term memory (LSTM) models and show that simple strategies, such as sign flipping or additive noise, can substantially reduce forecasting accuracy. They further demonstrate that a defense based on spectral clustering significantly improves robustness. Building on this, Sievers et al. [21] show that stochastic perturbations are effective for data poisoning but have only limited impact when applied as backdoor attacks.

Most existing research on federated energy systems focuses on defense-oriented techniques rather than on an in-depth analysis of attack mechanisms. Zhao et al. [7] incorporate differential privacy (DP), adding calibrated noise to updates to protect individual consumption profiles. Another direction involves secure aggregation, using cryptographic techniques or similarity-based weighting to protect individual updates and mitigate poisoning. Dong et al. [15] protect gradient exchanges, and Li et al. [11] filters malicious updates based on distance and similarity metrics. Further contributions focus on general robustness to adversarial clients. Manzoor et al. [20] introduce an anomaly-aware aggregation scheme to mitigate backdoor attacks. Other efforts have extended FL to cyber attack detection [12], probabilistic load forecasting [13], and privacy-preserving net energy prediction [9].

Table 1
Review of attack and security research in federated learning for energy forecasting and related domains. We distinguish whether studies investigate adversarial threats, provide security mechanisms, or employ Generative Adversarial Network-based approaches.

Reference	Year	Focus	Domain	Attack	Security	GAN
[7]	2021	Differential privacy for household load forecasting	Energy		✓	
[8]	2022	Byzantine-resilient FL using quantized gradients	Energy		✓	
[9]	2022	Energy forecasting with encrypted aggregation	Energy		✓	
[10]	2023	Personalized FL with differential privacy	Energy		✓	
[11]	2023	Similarity- and distance-based secure aggregation	Energy		✓	
[12]	2023	FL for cyberattack detection in smart grids	Energy		✓	
[13]	2023	Probabilistic individual load forecasting with FL	Energy		✓	
[14]	2023	Scalable architectures for secure FL-based forecasting	Energy		✓	
[15]	2024	Secure aggregation via multiparty computation	Energy		✓	
[16]	2025	Federated reinforcement learning for cost-efficient EMS	Energy		✓	
[17]	2022	Poisoning attacks on FL with clustering defense	Energy	✓		
[18]	2019	Data poisoning attacks on FL in language processing	Language	✓		
[19]	2020	Inference attacks on FL models in vision	Vision	✓		
[20]	2023	Anomaly-aware aggregation to mitigate backdoors	Energy	✓	✓	
[21]	2024	Poisoning and backdoor attacks with stochastic noise	Energy	✓	✓	
This paper	2025	Mitigating GAN-based poisoning and backdoor attacks	Energy	✓	✓	✓

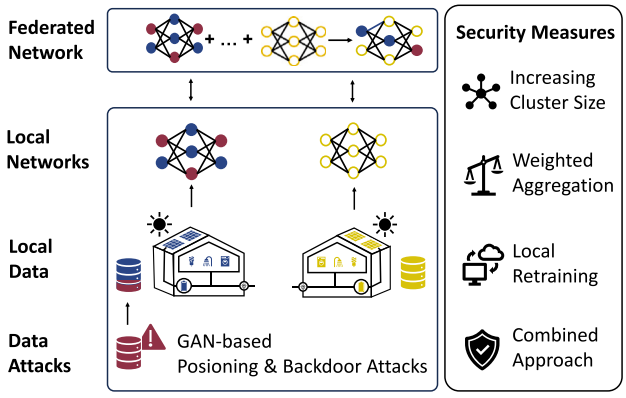


Fig. 1. Clustered federated learning under adversarial attacks. Compromised updates spread through clustered aggregation to the global model, mitigated by larger clusters, weighted aggregation, local retraining, or combined defenses.

Scalability and communication efficiency have also been addressed, for instance by Widmer et al. [14], and Husnoo et al. [8], who propose quantization techniques to resist Byzantine behavior.

In contrast, adversarial attacks, including model and data poisoning, inference-based manipulation, and backdoor insertion, have been extensively studied in domains such as computer vision and natural language processing, where they have demonstrated considerable potential to compromise FL performance. In computer vision, Mayerhofer et al. [28] show that poisoning attacks on image classifiers can degrade integrity. Xiang et al. [29] demonstrate that backdoor patterns reliably induce misclassifications in vision models and that cluster defenses detect such triggers. Further, Zhang et al. [19] highlight that GAN-based attacks can reconstruct private image distributions in federated learning and proposes mitigation strategies. In language processing, Zhai et al. [30] integrate backdoors into large text to image diffusion models at multiple semantic levels while retaining normal generation quality. Wan et al. [31] demonstrate that language models can be poisoned with few crafted examples such that trigger phrases consistently induce failures across tasks. Zhang et al. [18] show that GAN based poisoning in federated learning generates malicious updates that compromise the global model.

An essential development in this context is the GAN, introduced by Goodfellow et al. [32]. By jointly training a generator and a discriminator, classical GANs capture data distributions and produce perturbations that remain statistically consistent with the underlying data. Generator-only approaches, such as those of Mopuri et al. [33] and Wang et al. [34], remove the discriminator and instead focus on directly maximizing the impact of perturbations on a target model.

1.2. Paper contribution and organization

Although FL has recently been adopted in energy systems, its specific vulnerabilities under adversarial conditions remain largely unexplored. Existing studies from other domains involve data and temporal structures that differ fundamentally from those of energy time series. To address this gap, we conduct a comprehensive analysis of adversarial threats and defense strategies in federated energy forecasting. Consequently, our main contributions are:

- We develop GAN-based adversarial manipulations for both data poisoning and backdoor attacks, enabling realistic and structured perturbations of local model updates. For comparison, we also implement stochastic perturbations as a baseline, providing a systematic contrast between random and generative approaches.
- We evaluate the impact of these attacks on forecasting accuracy using two neural network architectures. A basic Multilayer perceptron (MLP) provides a reference baseline, while a more advanced Soft-Gated Dense Neural Network (Soft-Dense) captures the effects of architectural complexity on vulnerability and robustness.
- We propose and analyze four defense strategies to mitigate the risks posed by GAN-based adversaries: secure aggregation, increasing cluster sizes, local model retraining, and an integrated framework of these strategies.

Our results demonstrate that both attack types significantly degrade forecasting accuracy, particularly in small clusters. However, the implementation of suitable defensive measures can mitigate these effects and enhance the reliability of FL-based forecasting against adversarial interference.

The reminder of the paper is organized as follows: Section 2 introduces our methodology, while Section 3 outlines our experimental setup. Building on this, Section 4 presents our results, Section 5 discusses our results, limitations, and future work, and Section 6 provides our conclusion.

2. Methodology

This section outlines the methodological framework, including the federated energy forecasting setup, the design of adversarial attacks, and the implementation of defense strategies.

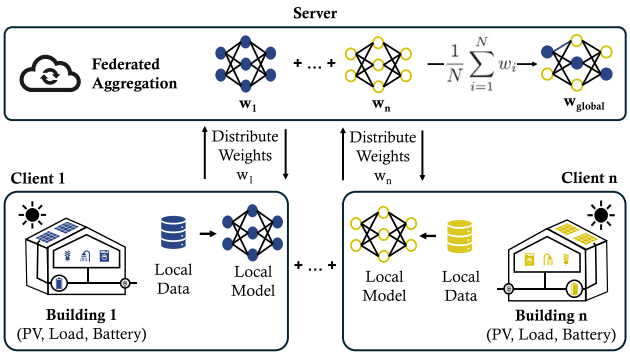


Fig. 2. Federated learning architecture for energy forecasting. Each client trains a local model on private data, and only model weights are sent to the server. The server aggregates updates to form a global model and redistributes the new weights to all participants for the next training round.

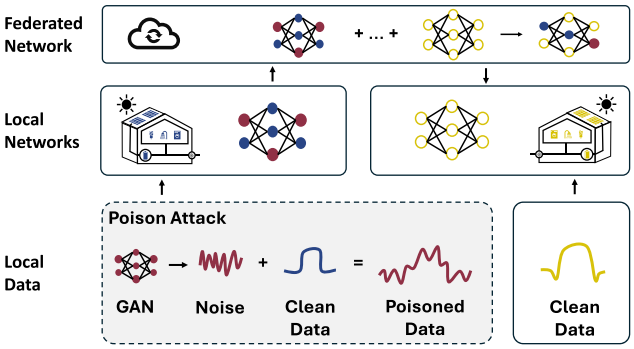


Fig. 3. Generative Adversarial Network-based data poisoning in federated learning. Adversarial noise is added to clean local data, producing poisoned samples used for local training. Aggregated across clients, this leads to a corrupted global model.

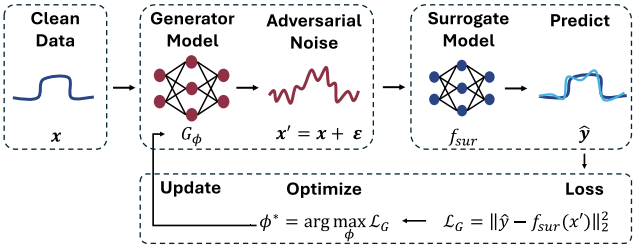


Fig. 4. Generative Adversarial Network training process. The generator perturbs clean data to produce adversarial samples. The GAN parameters are optimized to maximize the surrogate model's prediction loss.

2.1. Federated energy systems

FL enables the decentralized training of forecasting models across N distributed clients, such as buildings, each with access to a private dataset D_i [4]. Instead of exchanging raw data, each client independently trains a local forecasting model f_{w_i} with parameters w_i using only its local data. The training process is coordinated by a central server, which initializes a global model with parameters w_{global} and transmits them to all clients. Each client then updates the received parameters by performing local optimization on its dataset D_i. These local updates w_i are returned to the server, which aggregates them to refine the global model (Fig. 2). A commonly used aggregation method is Federated Averaging, where the updated global parameters are computed as the mean of the client parameters (Eq. (1)):

$$w_{global} = \frac{1}{N}\sum_{i=1}^{N} w_i \tag{1}$$

This iterative process is repeated for T rounds, allowing the global model to progressively adapt to the heterogeneous characteristics of the clients' local data while preserving privacy.

2.2. Data poisoning attack in federated energy systems

While FL supports scalable collaboration, its decentralized structure introduces novel security vulnerabilities. A primary threat is data poisoning, where malicious clients intentionally corrupt their local training data D_i. These manipulations influence the local updates w_i, which, once aggregated, compromise the global model w_{global}, thereby degrading performance across both adversarial and benign clients.

Let (x, y) denote a clean and normalized training sample, where x is the input sequence and y the corresponding target. To assess adversarial scenarios, we implement two poisoning strategies: (i) stochastic noise and (ii) structured perturbations, both applied exclusively to the energy time series while keeping external features unchanged.

In the stochastic case, adversarial inputs are constructed by adding noise ϵ drawn from a distribution D and scaled by a factor γ, as formalized in Eq. (2) [21]:

$$x' = x + \epsilon, \quad \epsilon \sim \mathcal{D}(\gamma), \tag{2}$$

In the structured setting, adversarial perturbations are not randomly sampled but are learned to maximally degrade forecasting accuracy. To this end, a generator network G_ϕ, parameterized by ϕ, is trained to produce input-dependent perturbations. For each clean input sequence x, the generator outputs a perturbation $\epsilon = G_\phi(x)$, which is added to the original input to form the adversarial sample $x' = x + \epsilon$. The generator is optimized to increase the prediction error of a pre-trained surrogate forecasting model f_{sur}, which approximates the behavior of the global model. An illustration of a data poisoning attack using a GAN is shown in Fig. 3.

While classical GANs introduced by Goodfellow et al. [32] train a generator in competition with a discriminator to capture data distributions, our approach departs by omitting the discriminator. Instead, the generator is optimized to maximize the prediction error of a fixed surrogate forecasting model under bounded perturbations. Similar generator-only adversarial frameworks have been proposed by Mopuri et al. [33] and Wang et al. [34].

The resulting training objective is shown in Eq. (3), where the first term maximizes the surrogate's prediction error and the second term penalizes large perturbations via ℓ_2 regularization.

$$\mathcal{L}_{G(\phi)} = -\frac{1}{|D|}\sum_{x\in D}\left\| y - f_{\text{sur}}(x+\epsilon)\right\|_2^2 + \lambda\|\epsilon\|_2^2. \tag{3}$$

To ensure comparability, both stochastic and GAN-based perturbations are restricted to $[-\gamma, \gamma]$, with γ controlling attack intensity. An illustration of the GAN training process is shown in Fig. 4.

Stochastic noise is scaled directly, whereas GAN outputs are passed through a tanh activation and multiplied by γ, as shown in Eq. (4):

$$\epsilon = \gamma \cdot \tanh(G_\phi(x)), \quad \epsilon \in [-\gamma, \gamma]. \tag{4}$$

The perturbations ϵ translate directly into the federated training process: benign clients $k \in \mathcal{B}$ continue to optimize on their clean datasets $D_k = (x_k, y_k)$, whereas adversarial clients $j \in \mathcal{A}$ replace x_j with perturbed inputs $x'_j = x_j + \epsilon$. These manipulated samples enter the standard local optimization step, so that in each round $t \in T$ the parameter update of client i is given by Eq. (5):

$$w_{i,t+1} = w_{i,t} - \eta\nabla_{w_i}\mathcal{L}\left(w_{i,t}, \tilde{x}_i, y_i\right), \tag{5}$$

where $w_{i,t}$ are the model parameters at round t, η is the learning rate, $\mathcal{L}$ the local loss, and ∇_{w_i} its gradient with respect to the parameters. Here, $\tilde{x}_i = x_k$ for benign clients and $\tilde{x}_i = x'_j$ for adversaries. Because the

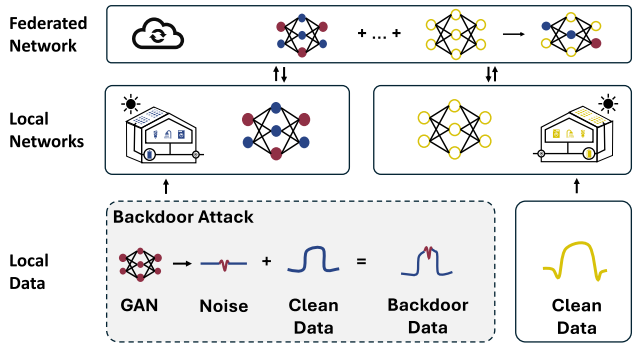


Fig. 5. Generative Adversarial Network-based backdoor injection in federated learning. A generator synthesizes trigger noise that is embedded into clean local data, producing backdoor samples used for local training. During aggregation, these poisoned updates propagate through the federated network, implanting a hidden backdoor in the global model.

resulting updates $w_{j,t+1}$ are incorporated into the federated aggregation step shown in Eq. (1), adversarial perturbations propagate beyond the local model and gradually influence the global forecasting model shared across all participants.

2.3. Backdoor attack in federated energy systems

Unlike data poisoning, which generally degrades forecasting accuracy across all time steps, backdoor attacks are designed to remain inconspicuous by preserving overall model performance while embedding targeted vulnerabilities. These vulnerabilities are activated only when a specific trigger condition is met, making detection particularly difficult in safety- and reliability-critical settings such as energy forecasting. Fig. 5 demonstrates the GAN-based backdoor attack.

In our formulation, the trigger is defined by a subset of targeted hours $H \subseteq 0, 1, \ldots, 23$. For each training example (t_j, x_j, y_j), the variable t_j denotes the hour of the day associated with the input x_j. Adversarial clients perturb only those samples whose timestamp falls within the targeted set H. The modified input is shown in Eq. (6)

$$x'_j = x_j + \gamma \cdot \epsilon \cdot \tau(t_j), \tag{6}$$

where ϵ is the perturbation, γ a scaling factor, and $\tau(t_j)$ is an indicator function that equals 1 if t_j lies in the targeted set H and 0 otherwise. More sophisticated triggers can extend $\tau(t_j)$ beyond a simple hour-based switch by incorporating contextual features c_j, for example setting $\tau(t_j, c_j) = 1$ during holidays, under extreme weather conditions, or when frequency-domain patterns indicate congestion. Moreover, $\tau(t_j)$ does not need to be binary: it can be defined as a smooth weighting function $\tau(t_j) \in [0, 1]$ to gradually activate and fade out the perturbation at the beginning and end of the targeted window.

As with data poisoning, ϵ can be drawn from a distribution (e.g. Gaussian, Uniform, Laplace) or learned by a trainable GAN G_ϕ. In the latter case, G_ϕ is optimized with the same adversarial loss as in Eq. (3), with perturbations masked to activate solely under the trigger. By tuning γ and selecting H, attackers can control the severity and evaluate the impact of the backdoor.

2.4. Security measures in federated energy systems

Given the vulnerability of FL to data poisoning and backdoor attacks, robust defense mechanisms are essential for maintaining model integrity. We therefore implement four complementary strategies: clustering, weighted aggregation, local retraining, and their integration within a unified framework.

Clustering reduces the influence of adversarial updates by restricting aggregation to subgroups of clients with similar consumption or generation behavior. Formally, two clients i and j are assigned to the same cluster if their time series representations E_i and E_j satisfy $d(E_i, E_j) \leq \xi$, where $d(\cdot, \cdot)$ is a distance or similarity measure and ξ a threshold controlling cluster granularity.

The choice of $d(\cdot, \cdot)$ and ξ directly affects both robustness and learning efficiency. Small thresholds ξ increase the chance of isolating adversarial clients but may lead to fragmented clusters with limited collaboration, whereas large thresholds enhance knowledge sharing but risk admitting heterogeneous or malicious clients. In practice, common choices for $d(\cdot, \cdot)$ include Euclidean distance for simplicity, correlation-based measures for linear dependencies, and Dynamic Time Warping (DTW) for handling temporal misalignments.

Weighted aggregation mitigates adversarial influence by reducing the contribution of client updates that yield a high loss on a trusted validation set. The rationale is that benign models trained on clean data tend to generalize well, whereas adversarial manipulations typically degrade validation performance. Formally, let $\mathcal{B}$ and $\mathcal{A}$ denote benign and adversarial clients, respectively. The global model update is given by Eq. (7) [21]:

$$w_{\text{global}} = \frac{\sum_{k \in \mathcal{B}} \alpha_k \cdot w_k + \sum_{j \in \mathcal{A}} \alpha'_j \cdot w'_j}{\sum_{k \in \mathcal{B}} \alpha_k + \sum_{j \in \mathcal{A}} \alpha'_j} \tag{7}$$

where α_k and α'_j denote the aggregation weights for benign and compromised clients, respectively, with typically $\alpha'_j \ll \alpha_k$. The defense is effective provided that the validation set is not corrupted, or that the perturbations ϵ used by adversaries are not learnable and therefore still increase validation loss.

Local Retraining enables each client to refine the received global model parameters w_{global} using its private data, thus reducing the impact of adversarial drift introduced during aggregation. After receiving potentially compromised $\mathbf{w}'_{\text{global}}$, benign clients refine their models as in Eq. (5) [21]. This personalized refinement step improves the model's alignment with local data and mitigates the influence of poisoned updates.

In addition to our focus on mitigating adversarial influence during model aggregation and training, another line of research secures the communication layer itself. Methods such as Paillier encryption protect model updates during transmission and restrict unauthorized access or modification. Frameworks like gradient quantization improve the transmission efficiency by decreasing the communication load [8]. However, these methods do not prevent already compromised clients from submitting malicious updates. Therefore, our work focuses on ensuring robustness against adversarial behavior after an attack occurs rather than exclusively protecting data exchange mechanisms.

Together, these strategies form a robust defense framework, enhancing the security and reliability of federated learning in energy forecasting applications.

3. Experimental setup

Building on our methodology, we describe our experimental setup, including data analysis and federated energy forecasting.

3.1. Data analysis

We utilize the Ausgrid dataset [35], which provides half-hourly smart meter measurements of residential electricity consumption and gross PV generation for 300 households in New South Wales, Australia. The dataset spans from July 2010 to June 2013 and distinguishes between general consumption and controlled load. In this study, both load components are aggregated into a single total demand, reflecting the actual electricity requirement of each household. Representative load and PV generation profiles for Building 11 are shown in Fig. 6.

To extend the dataset, we compute prosumption as net demand (load minus PV), which represents the energy that must be supplied by the grid or a storage system at each time step.

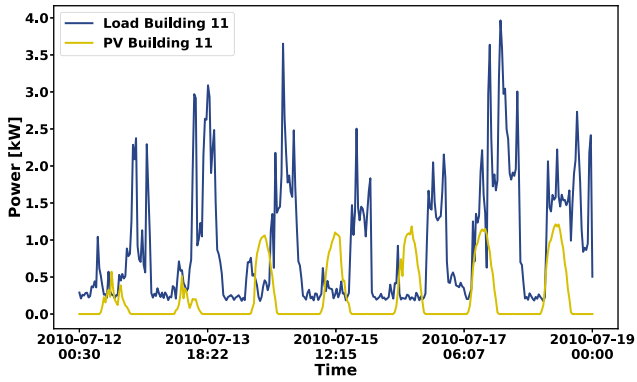


Fig. 6. Measured load and PV patterns of building 11 for a typical week.

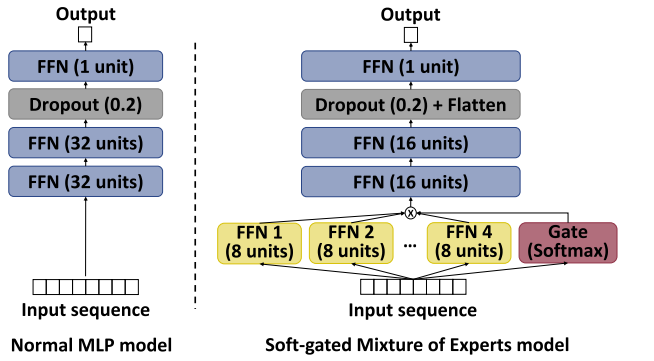


Fig. 7. Overview of the model architectures. The baseline MLP (left) employs stacked dense feed-forward network (FFN) layers with dropout, followed by a single output FFN layer. The soft-gated Mixture-of-Experts model (right) integrates multiple parallel FFN experts and uses a softmax gating network to weight their contributions.

Weather data are sourced from Meteostat [36] for the same region in Australia as the Ausgrid dataset. Weather conditions strongly influence PV output and indirectly affect residential consumption, which makes them valuable predictors for forecasting tasks. Because both the Ausgrid and the weather dataset share a 30-minute resolution, meteorological values are temporally aligned by merging entries on identical timestamps. To focus on the most informative weather features, we calculate the absolute Pearson correlation between each available meteorological variable and the prosumption series and retain the four strongest predictors: temperature, relative humidity, wind speed, and wind direction.

Temporal regularities in residential energy usage are captured through three time-based features: a binary weekday indicator and periodic encodings of the hour of day. Sine–cosine encoding avoids artificial discontinuities at midnight by representing time on a circular domain. For each half-hour step $h \in \{0, \dots, 47\}$ and a period $T = 48$, the encodings are given by Eq. (8):

$$\sin(h) = \sin\left(2\pi \frac{h}{T}\right), \qquad \cos(h) = \cos\left(2\pi \frac{h}{T}\right). \tag{8}$$

For computational efficiency, the analysis of the attacks and mitigation strategies is limited to a randomly selected subset of the first 20 buildings. The data are split into 70% training, 20% validation, and 10% testing.

3.2. Federated energy forecasting

Our FL architecture consists of 3 training rounds, as additional rounds did not yield further improvements. To simulate deployment scenarios and assess robustness, clients are grouped into fixed-size clusters.

Unlike our previous work [21], which used K-Means clustering based on DTW similarity, we adopt random clustering with a fixed seed to ensure reproducibility and controlled evaluation. Unless stated otherwise, cluster size is set to 2. We select this cluster size to expose the system to stronger adversarial influence, thereby enabling a stress-test of federated robustness in realistic heterogeneous environments. Note, that other clustering methods can be chosen.

To address our forecasting task, we evaluate two model architectures: a baseline MLP and a more expressive Mixture of Experts (MoE) model implemented via a soft-gated dense layer. In a previous study [37] the MoE architecture has been extensively benchmarked against state-of-the-art LSTM, CNN, and Transformer models for both local and federated learning. In FL heterogeneous data distributions across buildings often degrade model performance and limit generalization. MoE architectures mitigate this issue by combining multiple specialized sub-models (experts) through a gating mechanism that dynamically selects and weights expert outputs based on the input. This enables the model to adaptively capture complex and nonstationary patterns in energy data. Earlier studies show, that incorporating an MoE layer into standard deep learning models substantially improves forecasting accuracy [37]. The MLP baseline consists of two dense feed-forward layers with 32 units, followed by dropout with a rate of 0.2 and a single output unit. The MoE architecture adopts the soft-gated structure described in [37], employing four expert models with 8 units each, followed by two dense layers with 16 units, dropout (rate 0.2), and a single output unit. Both models are trained using the Adam optimizer (learning rate 10^{-4}), a batch size of 256, 50 training epochs, and early stopping with a patience of 10. A detailed architectural overview is provided in Appendix and Fig. 7.

Stochastic perturbations are sampled independently at each time step from a uniform distribution $U(-\gamma, \gamma)$, with $\gamma \in 0.0, 0.3, 0.6, 0.9$ controlling the perturbation magnitude. The perturbations are then added to the original data of the attacked building, so that we can analyze the impact of increasingly noisy data. Only the energy time series is modified, while all auxiliary features remain unchanged to prevent trivial detection. For data poisoning, the perturbation is applied at every time step of the local training data to maximize degradation in model performance. For the backdoor attack, the perturbation is restricted to four consecutive half-hour intervals between 10:30 and 12:30, thereby establishing a temporal trigger. As illustrated in Fig. 4, learned perturbations are produced using a surrogate–generator setup. The surrogate forecaster f_{sur} approximates the prediction behavior of the target model and provides a differentiable objective for the attack. It is implemented as a shallow MLP with two dense layers of 32 units and a linear prediction layer, trained locally on clean client data using the Adam optimizer with a learning rate of 10^{-3}, batch size 256, mean squared error loss, early stopping with patience 3, and at most 50 training epochs. Once training converges, the surrogate is frozen and solely used to evaluate the effect of perturbations. The generator G_ϕ is an MLP with two dense layers of 32 units and a tanh output, which bounds the generated perturbation to $[-1, 1]$ before scaling. During training, clean inputs are perturbed by G_ϕ, passed through the frozen surrogate, and the generator parameters are updated to increase the surrogate's prediction error while regularizing the perturbation magnitude through an ℓ_2 penalty. Optimization uses Adam with learning rate 10^{-3}, batch size 256, and a regularization weight $\lambda_{\text{reg}} = 10^{-4}$.

To defend against these attacks, we implement four strategies. First, increasing the cluster size from 2 to 5 to reduce the relative influence of any single adversarial client. Second, weighted aggregation to scale client contributions based on local validation loss. Third, local retraining to allow benign clients to fine-tune the global model on their private data for up to 50 epochs with early stopping. Fourth, a combination of the previous three defense mechanisms. Performance is evaluated using Root Mean Squared Error (RMSE).

Table 2
Performance of unmodified buildings under different poisoning intensities.

Setup	RMSE Uniform/GAN	STD Uniform/GAN	Diff Uniform/GAN
Load			
FL	0.1154/0.1157	0.03/0.03	0.00%/0.00%
N0.3	0.1205/0.1383	0.03/0.02	4.38%/19.56%
N0.6	0.1312/0.1456	0.03/0.03	13.66%/25.85%
N0.9	0.1386/0.1685	0.03/0.03	20.13%/45.70%
PV			
FL	0.0673/0.0679	0.01/0.01	0.00%/0.00%
N0.3	0.0743/0.0958	0.01/0.01	10.37%/41.15%
N0.6	0.0845/0.1309	0.01/0.02	25.58%/92.79%
N0.9	0.0952/0.1571	0.01/0.01	41.45%/131.38%
Prosumption			
FL	0.1030/0.1034	0.03/0.03	0.00%/0.00%
N0.3	0.1101/0.1297	0.03/0.02	6.93%/25.39%
N0.6	0.1261/0.1543	0.04/0.04	22.43%/49.16%
N0.9	0.1308/0.1312	0.04/0.04	27.06%/26.83%

Note: N0.3 indicates a noise scale of 0.3; Diff shows the change from the FL baseline (noise 0.0); noise is sampled from a uniform distribution or a GAN.

4. Results

This section presents experimental results for data poisoning, backdoor attacks, and the effectiveness of defense strategies. For each attack scenario, we distinguish between the *adversarial* models (trained on manipulated data) and the *unmodified* models (affected indirectly via federated aggregation). Unless otherwise noted, all metrics are computed on the test set and averaged over all buildings and clusters. The model architectures, attack scenarios and defense strategies are identical for different energy forecasting types (PV, load, prosumption).

4.1. Data poisoning attack

We start by analyzing the vulnerability of federated energy forecasting to data poisoning. The evaluation examines both overall and per-building performance, comparing the robustness of MLP and MoE models under uniform and GAN-based perturbations.

Table 2 shows the RMSE and standard deviation (STD) for load, PV, and prosumption forecasts across all unmodified clients under different perturbation scales. For each energy type, we compare the impact of uniform and GAN-generated noise, while the FL baseline corresponds to the clean federated model without any attack.

The results demonstrate a consistent increase in forecasting error as the perturbation scale rises, across both attack types and all forecasting tasks. GAN-based perturbations lead to significantly greater degradation than uniform noise, with the strongest effects observed in PV forecasting.

For the FL baseline, the performance of models under both noise types is nearly identical, confirming a comparable starting point for fair evaluation. For instance, the RMSE for load is 0.1154 (uniform) versus 0.1157 (GAN). Among the three forecasting tasks, PV prediction is the most vulnerable to adversarial manipulation. At a perturbation scale of 0.6, the RMSE increases by 25.58% with uniform noise and by 92.79% with GAN-generated perturbations. At scale 0.9, these increases rise to 41.45% and 131.38%, respectively. In load forecasting, uniform noise with scale 0.9 raises the RMSE from 0.1154 to 0.1386 (+20.13%), whereas GAN-based perturbations result in a +45.70% increase. Interestingly, for prosumption forecasting the performance degradation converges at the highest perturbation scale to an RMSE increase of 27.06% (uniform) and 26.83% (GAN).

Fig. 8 complements the previous analysis by comparing the performance of the two model architectures, MLP and MoE, under data poisoning attacks. For each energy type, the figure reports the RMSE of unmodified clients and annotates the relative increase compared to the clean baseline. The results highlight that the MoE architecture is consistently more robust than the MLP, particularly under GAN-based perturbations.

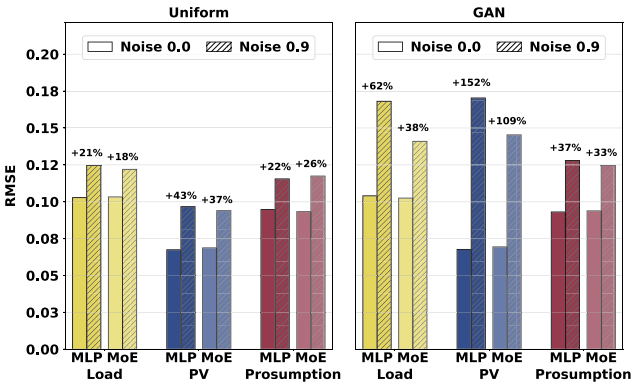


Fig. 8. Performance of MLP vs. MoE on unmodified buildings under poisoning attacks.

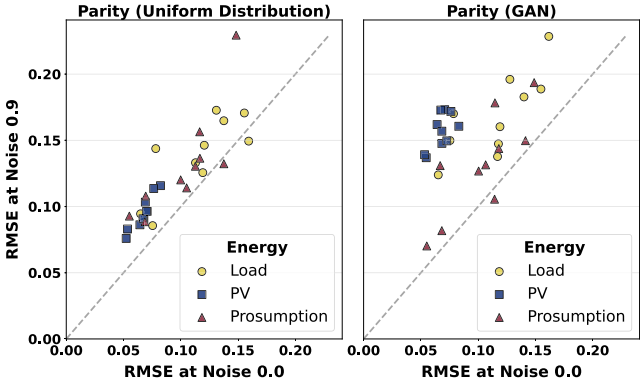


Fig. 9. Performance parity of unmodified buildings under poisoning at noise scales 0.0 and 0.9.

Specifically, the MoE demonstrates stronger resilience for both load and PV forecasting when facing adversarially generated noise. At a perturbation scale of 0.9, the RMSE for load increases by only +38% with the MoE, compared to +62% with the MLP. Similarly, for PV, the MoE exhibits an increase of +109%, while the MLP reaches +152%, indicating a substantially higher vulnerability. Under uniform noise, the performance differences between the two models are less pronounced. At a noise scale of 0.9, the MoE achieves slightly lower RMSE values for load (0.122 vs. 0.125) and PV (0.094 vs. 0.097), while the MLP performs marginally better for prosumption. Overall, the MoE architecture offers increased robustness and more reliable forecasting performance in the presence of data poisoning, making it the preferred choice in adversarial settings.

While the previous results considered mean performance across all unmodified buildings, we now evaluate the forecasting accuracy at the building-level. Fig. 9 presents parity plots comparing the RMSE of unmodified buildings under attack (perturbation scale 0.9) to the clean baseline (scale 0.0) for both uniform and GAN-based perturbations. Each point corresponds to a single benign building, with most lying above the diagonal, indicating consistent performance degradation. Notably, GAN-based attacks induce stronger and more variable impacts compared to uniform noise.

The proportion of affected clients is high across both attack types. Under uniform noise, RMSE increases in 93% of buildings, while GAN-based perturbations lead to degradation in 97% of cases. Beyond this high prevalence, GAN-induced errors exhibit greater dispersion across buildings, particularly for load and PV forecasts. For instance, the

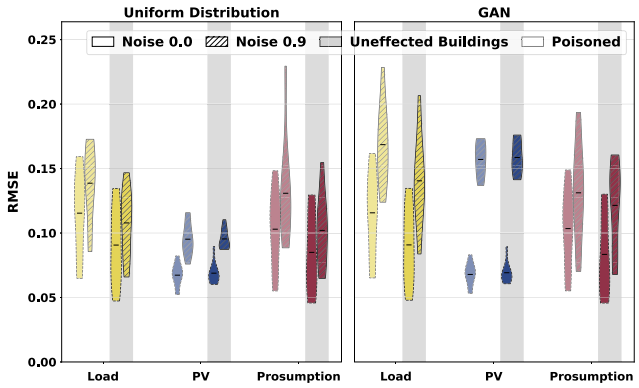


Fig. 10. Performance comparison between poisoned and unmodified buildings at noise scales 0.0 and 0.9 for the poisoning attack.

building-wise RMSE range for load expands from $[0.0856, 0.1727]$ under uniform noise to $[0.1239, 0.2285]$ under GAN perturbations. For PV, the range shifts from $[0.0759, 0.1158]$ to $[0.1369, 0.1732]$, indicating a marked increase in both error magnitude and variability. In the case of prosumption, the distribution slightly narrows, shifting from $[0.0885, 0.2294]$ (uniform) to $[0.0702, 0.1935]$ (GAN), although the overall trend still suggests increased model instability.

To complement the building-level perspective, Fig. 10 analyzes the distribution of forecasting errors using violin plots. The figure contrasts adversarial clients, which introduce manipulated updates, with benign clients, which contribute only clean data. Comparisons are shown across energy types, attack settings, and perturbation scales (0.0 and 0.9). This perspective highlights not only the direct degradation of adversarial clients but also how their influence propagates through federated aggregation to impair the performance of otherwise unmodified buildings.

The results show that under uniform noise at scale 0.9, RMSE increases are similar for both adversarial and unmodified clients, suggesting that stochastic perturbations diffuse evenly across the network. For example, load forecasting shows increases of +23.81% (adversarial) and +20.13% (unmodified); PV increases are +32.77% and +41.45%, respectively; and prosumption increases are +21.33% and +27.06%. In contrast, GAN-based perturbations produce more asymmetric effects, with adversarial clients exhibiting notably higher errors. For load, the RMSE increases by +48.37% (adversarial) versus +45.68% (unmodified), and for prosumption by +59.09% versus +26.83%. In PV forecasting, both groups are heavily affected, with increases of +127.86% (adversarial) and +131.48% (unmodified).

Thus, our findings underscore the vulnerability of federated energy forecasting to poisoning attacks. Across all forecasting tasks, increasing perturbation intensity consistently degrades performance, with GAN-based manipulations causing more severe effects than uniform noise. While both adversarial and benign clients are negatively impacted through the aggregation process, the propagation of GAN-based perturbations is particularly harmful, with PV forecasting being the most affected. Between model architectures, the MoE demonstrates consistently greater robustness than the MLP, offering more reliable performance under adversarial conditions.

4.2. Backdoor attack

While data poisoning attacks degrade model performance globally, backdoor attacks represent a more covert threat by targeting specific conditions while largely maintaining overall accuracy. In the following section, we assess the effectiveness of backdoor attacks by injecting perturbations into a single adversarial client per cluster, with activation limited to a fixed temporal window (10:30–12:30 am). The objective is to impair forecast accuracy for benign clients during this interval, without affecting performance at other times.

Table 3
Performance of unmodified buildings under different backdoor intensities.

Setup	RMSE Noise/GAN	STD Noise/GAN	Diff Noise/GAN
Load			
FL	0.0762/0.0770	0.0304/0.0307	0.00%/0.00%
N0.3	0.0830/0.0824	0.0319/0.0307	8.92%/7.01%
N0.6	0.0863/0.0833	0.0339/0.0315	13.25%/8.18%
N0.9	0.0867/0.0840	0.0336/0.0363	13.78%/9.09%
PV			
FL	0.1174/0.1170	0.0181/0.0186	0.00%/0.00%
N0.3	0.1245/0.1392	0.0193/0.0201	6.05%/18.97%
N0.6	0.1354/0.1612	0.0239/0.0282	15.33%/37.78%
N0.9	0.1380/0.1731	0.0216/0.0300	17.55%/47.95%
Prosumption			
FL	0.0714/0.0726	0.0261/0.0268	0.00%/0.00%
N0.3	0.0838/0.0820	0.0312/0.0228	17.37%/12.95%
N0.6	0.0864/0.0849	0.0299/0.0227	21.01%/16.94%
N0.9	0.0884/0.0918	0.0288/0.0260	23.81%/26.45%

Note: N0.3 indicates a noise scale of 0.3; Diff shows the change from the FL baseline (noise 0.0); noise is sampled from a uniform distribution or a GAN.

Table 3 reports the RMSE and STD for load, PV, and prosumption forecasts during the targeted window for the unmodified buildings. For each energy type, we compare the effects of uniform and GAN-based perturbations across increasing intensity levels. Results show a consistent degradation in forecasting performance with higher perturbation scales, with GAN-based attacks causing notably larger errors, particularly in PV forecasting.

Backdoor effectiveness is most pronounced in PV forecasting. At a perturbation scale of 0.9, the RMSE for unmodified clients rises from 0.1174 to 0.1380 under uniform noise (+17.55%), and to 0.1731 under GAN-based perturbations (+47.95%), the largest absolute degradation observed across all forecasting tasks. A similar vulnerability is evident in prosumption, where the RMSE increases from 0.0726 to 0.0918 (+26.45%) under GAN-based noise. In contrast, load forecasting exhibits greater resilience and an inverse trend: uniform noise results in higher degradation (+13.78%) than GAN-based perturbations (+9.09%).

These patterns are already apparent at lower intensities. At a noise scale of 0.3, PV forecasting shows a modest increase of +6.05% under uniform noise, but a substantially larger rise of +18.97% under GAN-generated triggers. Prosumption follows a similar trajectory, with increases of +17.37% and +12.95% for uniform and GAN-based noise, respectively. In load forecasting, however, the difference remains marginal, with uniform noise yielding +8.92% and GAN-based perturbations +7.01%.

Overall, these results demonstrate that localized backdoor triggers, introduced by a single adversarial client, can propagate through federated aggregation and significantly impair unmodified clients during the targeted window.

To assess whether backdoor attacks degrade forecasting accuracy globally or primarily within the targeted window, we compare model performance across all hours of the day. While Table 3 focuses on the impact during the triggered time slots, the following analysis evaluates whether these localized perturbations also influence non-targeted periods. Figs. 11 to 13 present the hourly RMSE across the full 24-hour cycle for unmodified clients, under both uniform and GAN-based perturbations at increasing noise scales.

Focusing first on load forecasting (Fig. 11), the impact of backdoor perturbations remains limited outside the targeted hours. For example, at a noise scale of 0.9, the RMSE during clean hours increases only marginally—from 0.0993 to 0.1038 under uniform noise (+4.53%) and from 0.1006 to 0.1025 under GAN-based noise (+1.89%). In contrast, during the backdoor window (10:30–12:30 am), the effect is significantly more pronounced: uniform noise increases the RMSE from

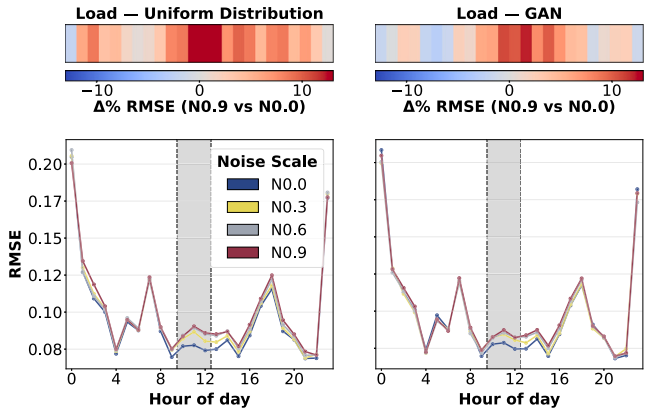


Fig. 11. Performance of unmodified buildings for each hour of the day under backdoor attacks on Load.

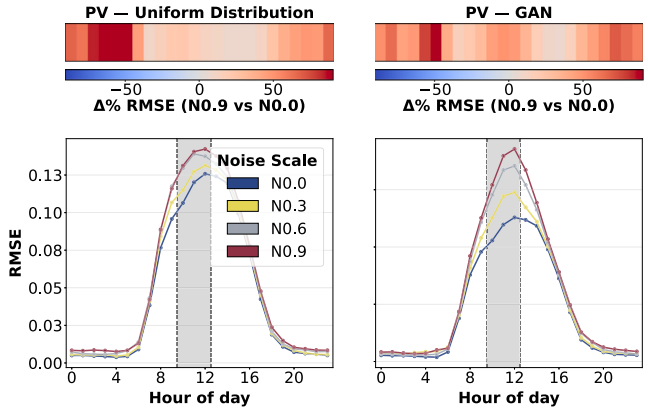


Fig. 12. Performance of unmodified buildings for each hour of the day under poisoning attacks on PV.

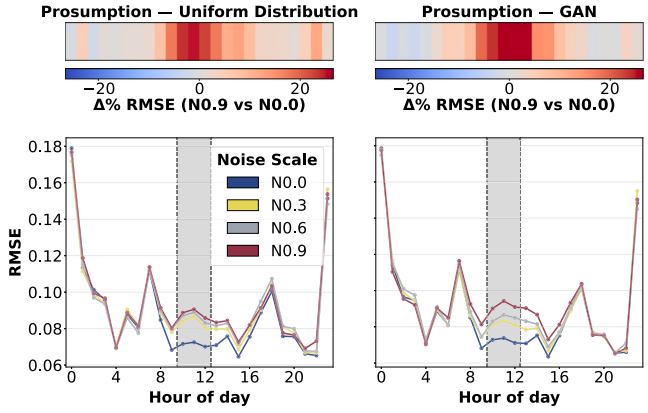


Fig. 13. Performance of unmodified buildings for each hour of the day under poisoning attacks on prosumption.

0.0762 to 0.0867 (+13.78%), while GAN-based perturbations yield a rise from 0.0770 to 0.0840 (+9.09%).

For PV (Fig. 12), degradation is strong both outside and inside the backdoor window. During clean hours, RMSE increases from 0.0361 to 0.0422 under uniform noise (+16.90%) and from 0.0361 to 0.0450 with the GAN-based perturbations (+24.65%). Within the backdoor window, the impact intensifies, with RMSE rising from 0.1174 to 0.1380 (+17.55%) under uniform noise, and to 0.1731 (+47.95%) under GAN-based perturbations

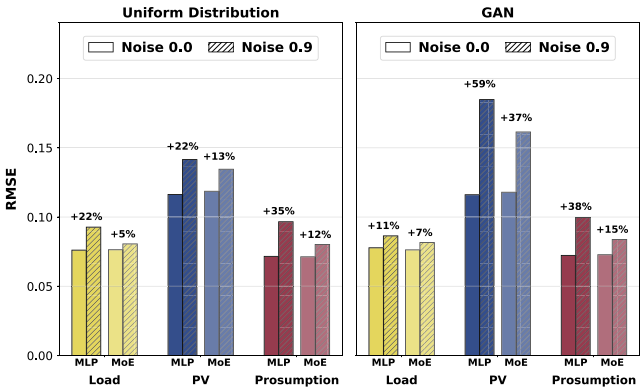


Fig. 14. Performance of MLP vs. MoE on unmodified buildings under backdoor attacks.

Prosumption forecasting (Fig. 13) exhibits intermediate sensitivity. For non-targeted hours, RMSE rises from 0.0906 to 0.0945 (+4.30%) under uniform noise and from 0.0902 to 0.0940 (+4.21%) with GAN-based perturbations. Within the backdoor window, however, the impact is more pronounced: uniform noise results in an increase from 0.0714 to 0.0884 (+23.81%), and GAN-based perturbations from 0.0726 to 0.0918 (+26.45%).

In summary, backdoor effects vary across forecasting tasks. In the context of load and prosumption, degradation remains temporally confined, thereby preserving performance outside the targeted window and enhancing stealth. Conversely, PV forecasting demonstrates the most significant increase in error, with substantial impact on non-targeted hours, resulting in reduced temporal precision and enhanced detectability.

Following the temporal analysis of backdoor attacks, we evaluate the architectural robustness of the MLP and MoE models, focusing on unmodified clients during the targeted window. This complements the prior poisoning results (Fig. 8) and examines whether the MoE's resilience extends to localized, trigger-based attacks.

Fig. 14 presents the RMSE averaged over the backdoor interval under uniform and GAN-based perturbations at noise scale 0.9. Across all energy types and attack settings, the MoE consistently outperforms the MLP, confirming its enhanced robustness in adversarial environments.

For load forecasting, the MLP exhibits a 21.81% increase in RMSE under uniform noise (from 0.0761 to 0.0927), whereas the MoE limits this to 5.50% (from 0.0764 to 0.0806). Under GAN-based perturbations, the respective increases are 11.05% for the MLP and 6.95% for the MoE.

In PV forecasting, where vulnerability is highest, the MLP degrades by 21.67% under uniform noise and 59.17% under GAN-based noise, while the MoE shows markedly lower increases of 13.41% and 36.90%, respectively.

Prosumption forecasting exhibits a similar pattern. The MLP RMSE increases by 34.92% (uniform) and 37.90% (GAN), while the MoE maintains lower increases of 12.48% and 15.11%.

Beyond lower relative degradation, the MoE consistently achieves lower absolute RMSE across all tasks and perturbation types, reinforcing its suitability for federated forecasting under adversarial conditions.

While the previous analysis focused on average performance across unmodified clients, we now examine the client-level consistency of backdoor effects during the targeted hours. This complements the model-level results by assessing how reliably performance degradation manifests across individual benign buildings. Fig. 15 shows parity plots comparing the RMSE of each unmodified client under attack (perturbation scale 0.9) to the clean FL baseline, averaged over the backdoor interval. Points above the diagonal indicate increased error due to the attack.

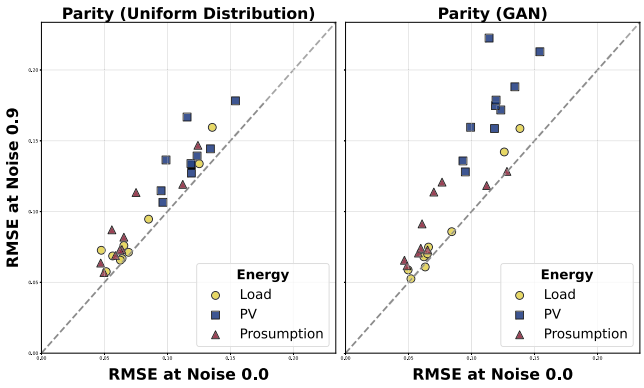


Fig. 15. Performance parity of unmodified buildings under backdoor attacks at noise scales 0.0 and 0.9.

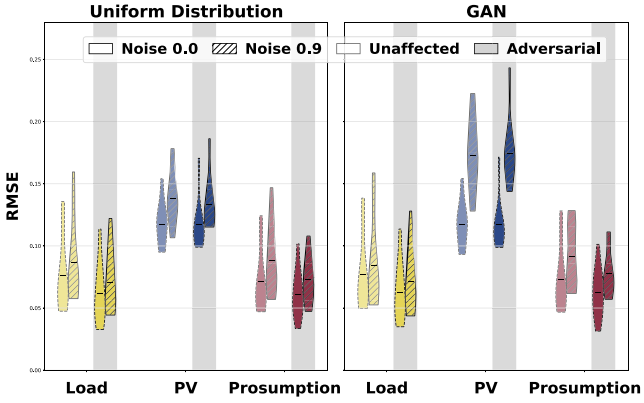


Fig. 16. Performance comparison between poisoned and unmodified buildings at noise scales 0.0 and 0.9 for the backdoor attack.

Backdoor perturbations impact nearly all unmodified clients during the targeted window, though the severity of degradation varies. Under uniform noise, all clients exhibit some level of deviation from the clean baseline, while GAN-based perturbations affect 97% of clients. However, not all deviations result in substantial performance loss. The magnitude and variability of the impact are significantly higher under GAN-based attacks, particularly in load and PV forecasting. For load, the client-wise RMSE range broadens from [0.0856, 0.1727] (uniform) to [0.1239, 0.2285] (GAN), and for PV from [0.0759, 0.1158] to [0.1369, 0.1732], indicating both higher error levels and increased dispersion. In contrast, prosumption forecasting shows a slightly narrower distribution under GAN-based noise ([0.0702, 0.1935]) compared to uniform noise ([0.0885, 0.2294]), though the overall RMSE remains elevated. These findings demonstrate that, despite being temporally localized, backdoor perturbations reliably propagate to benign clients. Moreover, GAN-based attacks introduce greater heterogeneity and model instability, making them especially challenging to detect and mitigate.

To complement the previous analysis of unmodified buildings, we now examine how backdoor attacks impact individual clients during the targeted time window, distinguishing between the attacker and unaffected participants. This enables a direct assessment of whether adversarial updates from a single client degrade the performance of others through the federated aggregation process.

Fig. 16 displays violin plots of the RMSE at noise scales 0.0 and 0.9, contrasting adversarial and benign clients. The results confirm that performance degradation is not confined to the attacking client, but extends to unmodified buildings, with the severity depending on both the perturbation type and the forecasting task.

Under uniform noise, forecasting errors increase similarly across both groups. For load, the RMSE rises by +14.94% for adversarial clients and +13.78% for the unmodified buildings. Prosumption follows a comparable trend, with increases of +18.79% and +23.81%, respectively. For PV, the effect is slightly stronger on unmodified buildings, with a +17.55% increase compared to +13.87% in the adversarial case.

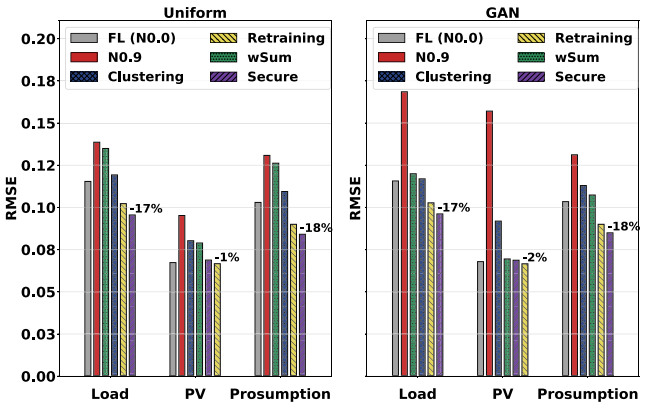


Fig. 17. Performance of the unmodified buildings under poisoning attack and selected security measures.

In contrast, GAN-based attacks produce significantly stronger effects. For PV, the RMSE increases by +49.27% in compromised clients and +47.95% in the unmodified buildings, indicating substantial transfer of the backdoor effect. Similar propagation is observed for prosumption (+24.56% vs. +26.45%). For load, the increase is more moderate, with +13.06% in adversarial clients and +9.09% in the benign buildings.

In summary, backdoor attacks reliably impair forecasting performance across other clients. The effect is strongest in PV forecasting and most severe under GAN-based perturbations, highlighting the ability of structured adversarial updates to propagate through the federated model. For load and prosumption, the degradation remains largely confined to the backdoor window, preserving overall model accuracy. In contrast, PV forecasting experiences broader degradation that extends beyond the targeted hours. These results highlight the need for defensive measures to assure a reliable deployment of federated energy forecasting.

4.3. Security in federated energy forecasting

While previous results show that a single adversarial client per cluster can significantly degrade performance for unaffected participants, this subsection evaluates the effectiveness of various mitigation strategies. Specifically, we assess the impact of (i) increasing cluster size to dilute adversarial influence, (ii) retraining local models to adapt to clean data, (iii) applying weighted aggregation to down-weight anomalous updates, and (iv) integrating all measures into a unified secure framework.

We begin by evaluating the effectiveness of the proposed defense strategies against data poisoning. Fig. 17 presents the resulting RMSE for unmodified buildings across all energy types and both perturbation types. Detailed RMSE and STD values are provided in Table 4.

For load forecasting, unprotected models exhibit substantial degradation, with RMSE increasing from 0.1154 to 0.1386 under uniform noise (+20.13%) and to 0.1685 under GAN-based perturbations (+45.70%). Clustering significantly mitigates this effect, reducing the RMSE to 0.1193 (+3.34%) for uniform noise and 0.1169 (+1.09%) for GAN-learned noise, recovering 83% and 97% of the added error, respectively. Weighted aggregation is less effective for uniform noise (+16.93%) but performs markedly better under GAN-based attacks, limiting the increase to +3.70% and eliminating 92% of the degradation. Local retraining further improves robustness, decreasing RMSE by −11.39% (uniform) and −11.20% (GAN), thereby outperforming the clean baseline. The Secure approach achieves the best overall results,

Table 4
Performance of unmodified buildings under poisoning attacks and selected security measures.

Setup	RMSE Uniform/GAN	STD Uniform/GAN	Diff vs. FL (N0.0) Uniform/GAN
Load			
FL	0.1154/0.1157	0.03/0.03	0.00%/0.00%
N0.9	0.1386/0.1685	0.03/0.03	20.13%/45.70%
Cluster	0.1193/0.1169	0.03/0.03	3.34%/1.09%
Retrain	0.1023/0.1027	0.03/0.03	−11.39%/−11.20%
wAgg	0.1349/0.1200	0.03/0.04	16.93%/3.70%
Secure	0.0955/0.0961	0.03/0.03	−17.23%/−16.88%
PV			
FL	0.0673/0.0679	0.01/0.01	0.00%/0.00%
N0.9	0.0952/0.1571	0.01/0.01	41.45%/131.38%
Cluster	0.0803/0.0919	0.01/0.01	19.24%/35.35%
Retrain	0.0667/0.0666	0.01/0.01	−0.97%/−1.89%
wAgg	0.0789/0.0695	0.01/0.01	17.24%/2.32%
Secure	0.0688/0.0687	0.01/0.01	2.21%/1.21%
Prosumption			
FL	0.1030/0.1034	0.03/0.03	0.00%/0.00%
N0.9	0.1308/0.1312	0.04/0.04	27.06%/26.83%
Cluster	0.1094/0.1130	0.03/0.03	6.29%/9.22%
Retrain	0.0900/0.0900	0.02/0.02	−12.60%/−12.99%
wAgg	0.1262/0.1073	0.04/0.02	22.57%/3.77%
Secure	0.0841/0.0849	0.02/0.02	−18.34%/−17.88%

Note: N0.9 denotes a noise scale of 0.9; Cluster increases cluster size, Retrain applies local retraining, wAgg uses weighted aggregation, and Secure combines all security measures. Diff shows the change from the FL baseline (noise 0.0). Noise is sampled from a uniform distribution or a GAN.

Table 5
Performance of unmodified buildings under backdoor attacks and selected security measures (attacked hours only).

Setup	RMSE Uniform/GAN	STD Uniform/GAN	Diff vs. N0.9 Uniform/GAN
Load			
FL	0.0759/0.0764	0.03/0.03	0.00%/0.00%
N0.9	0.0863/0.0840	0.03/0.03	13.61%/9.85%
Cluster	0.0751/0.0738	0.03/0.03	−1.04%/−3.52%
Retrain	0.0730/0.0732	0.02/0.03	−3.93%/−4.19%
wAgg	0.0853/0.0862	0.03/0.03	12.32%/12.82%
Secure	0.0671/0.0670	0.03/0.03	−11.60%/−12.30%
PV			
FL	0.1191/0.1187	0.02/0.02	0.00%/0.00%
N0.9	0.1379/0.1716	0.02/0.03	15.76%/44.61%
Cluster	0.1319/0.1358	0.02/0.02	10.73%/14.45%
Retrain	0.1188/0.1182	0.02/0.02	−0.28%/−0.44%
wAgg	0.1340/0.1310	0.02/0.02	12.48%/10.37%
Secure	0.1201/0.1197	0.02/0.02	0.84%/0.87%
Prosumption			
FL	0.0713/0.0721	0.02/0.02	0.00%/0.00%
N0.9	0.0871/0.0914	0.03/0.02	22.20%/26.68%
Cluster	0.0747/0.0778	0.02/0.02	4.74%/7.88%
Retrain	0.0700/0.0695	0.02/0.02	−1.79%/−3.65%
wAgg	0.0912/0.0872	0.03/0.02	27.85%/20.91%
Secure	0.0645/0.0643	0.02/0.02	−9.54%/−10.83%

Note: N0.9 denotes a noise scale of 0.9; Cluster increases cluster size, Retrain applies local retraining, wAgg uses weighted aggregation, and Secure combines all security measures. Diff shows the change from the FL baseline (noise 0.0). Noise is sampled from a uniform distribution or a GAN.

lowering RMSE up to −17.23%. Low STD (0.03) across clients confirm the consistency of all mitigation strategies.

For PV forecasting, the attack causes more severe degradation. Under uniform noise, the RMSE increases from 0.0673 to 0.0952 (+41.45%), while under the GAN attack it reaches 0.1571 (+131.38%). Clustering reduces these errors to 0.0803 (+19.24%) and 0.0919 (+35.35%), mitigating 54% and 73% of the degradation. Weighted aggregation performs especially well under the GAN-based noise, achieving an RMSE of 0.0695, which represents only a +2.32% increase over the clean baseline and recovers nearly 98% of the added error. Local retraining fully restores the model performance, yielding an RMSE of 0.0667 (−0.97%) and 0.0666 (−1.89%) for uniform and GAN, respectively. The Secure defense is similarly effective, achieving an RMSE of 0.0688 (+2.21%) and 0.0687 (+1.21%), corresponding to mitigation rates of 95% and 99%.

For prosumption, both attack types yield comparable increases in RMSE: from 0.1030 to 0.1308 (+27.06%) under uniform noise and to 0.1312 (+26.83%) under learned GAN perturbations. Clustering lowers the RMSE to 0.1094 (+6.29%) and 0.1130 (+9.22%), recovering 77% and 66% of the error, respectively. Weighted aggregation shows clear asymmetry: under GAN, it reduces RMSE to 0.1073 (+3.77%), reversing 86% of the attack impact, while under uniform noise it only reaches 0.1262 (+22.57%), achieving just 17% mitigation. Local retraining again offers full protection, lowering RMSE to 0.0900 in both cases (−12.60% and −12.99%), effectively eliminating the attack's effect and improving model accuracy. The Secure strategy achieves the strongest overall performance, with an RMSE of 0.0841 and 0.0849, and outperforming even the clean baseline.

Summarizing the results for data poisoning attacks, local retraining and the Secure combination consistently achieve the lowest RMSE, closely followed by clustering. Weighted aggregation alone proves insufficient in the presence of uniform noise, but shows moderate success under GAN.

Building on the poisoning results, Fig. 18 shows the RMSE achieved by each defense strategy under backdoor attacks, evaluated across all energy types and both perturbation methods. The results are averaged over the targeted interval (10:30–12:30) and computed solely for unaffected clients, isolating the indirect impact and its mitigation. Complementary to this, Table 5 provides exact RMSE values, STD across clients, and the relative improvement compared to the unprotected setting at scale 0.9, enabling a detailed assessment of all countermeasures.

For load forecasting, the unprotected backdoor attack moderately increases RMSE, rising from 0.0759 to 0.0863 (+13.61%) under uniform noise and from 0.0764 to 0.0840 (+9.85%) with GAN-based perturbations. Clustering effectively mitigates the impact, reducing RMSE to 0.0751 (−1.04%) and 0.0738 (−3.52%), slightly improving over the clean baseline. Local retraining achieves even better results with 0.0730 (−3.93%) and 0.0732 (−4.19%), indicating full mitigation and performance gains. In contrast, weighted aggregation offers limited protection, with RMSE values of 0.0853 (+12.32%) and 0.0862 (+12.82%), close to the unprotected case. The Secure approach yields the best results, reducing RMSE to 0.0671 (uniform) and 0.0670 (GAN), corresponding to improvements of −11.60% and −12.30% relative to the clean baseline. Standard deviations remain low (0.03), confirming consistent effectiveness across clients.

For PV, the unprotected backdoor introduces substantial error, raising RMSE from 0.1191 to 0.1379 (+15.76%) under uniform noise and from 0.1187 to 0.1716 (+44.61%) under GAN-based perturbations. Clustering reduces these values to 0.1319 (+10.73%) and 0.1358 (+14.45%), corresponding to 32% and 68% recovery, respectively. Weighted aggregation performs slightly worse, achieving RMSE values of 0.1340 (+12.48%) and 0.1310 (+10.37%), reducing the impact by only 21% and 77%. Local retraining effectively neutralizes the attack, lowering RMSE to 0.1188 (−0.28%) and 0.1182 (−0.44%), corresponding to a full reversal of the induced degradation. The Secure strategy performs similarly, reaching 0.1201 (+0.84%) and 0.1197 (+0.87%), implying 95–98% mitigation of the original RMSE increase.

For Prosumption, the unprotected backdoor leads to the most pronounced relative degradation across all three energy types, with RMSE rising from 0.0713 to 0.0871 (+22.20%) under uniform noise and from 0.0721 to 0.0914 (+26.68%) under GAN perturbations. Clustering provides partial mitigation, reducing the RMSE to 0.0747 (+4.74%) and 0.0778 (+7.88%), corresponding to 79% and 70% recovery, respectively. Weighted aggregation again performs poorly under uniform

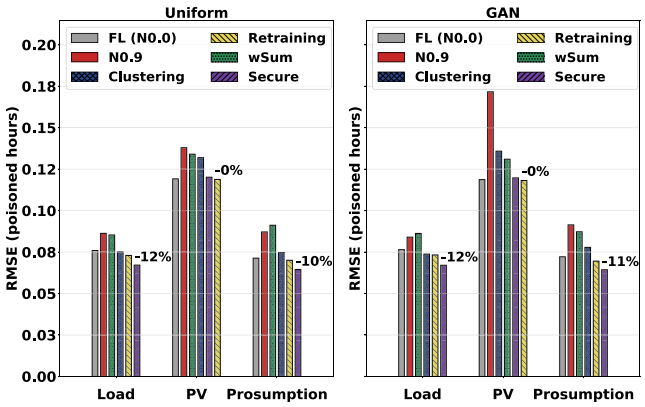


Fig. 18. Performance of the unmodified buildings under backdoor attack and selected security measures.

noise, with RMSE rising further to 0.0912 (+27.85%), indicating no effective mitigation. Under GAN, it performs slightly better at 0.0872 (+20.91%), still far from neutralizing the attack. Local retraining, by contrast, is consistently effective, yielding RMSE of 0.0700 (−1.79%) and 0.0695 (−3.65%), corresponding to full mitigation and a slight performance gain over the clean FL baseline. The secure combination achieves the strongest performance, reducing RMSE to 0.0645 and 0.0643, which translates to a complete mitigation of the backdoor effect and further improvement beyond baseline (−9.54% and −10.83%).

In summary, the Secure defense and local retraining again emerge as the most effective strategies across all energy types and perturbation methods. Clustering provides moderate protection, especially for Load and Prosumption, but fails to fully restore clean performance. Weighted aggregation proves largely ineffective against the backdoor attack, particularly under uniform noise, where RMSE remains comparable to or worse than the unprotected attack level. These findings reinforce the importance of combining multiple mitigation mechanisms, as done in the secure framework, to achieve reliable and generalizable backdoor robustness across domains and attack types.

5. Discussion and limitations

In this section, we discuss our results on data poisoning, backdoor attacks, and the evaluated mitigation strategies for federated energy forecasting. We first interpret how attack type, perturbation structure, energy type, and model architecture jointly shape vulnerability, and then assess how the proposed mitigation strategies restore or even improve forecasting performance. Finally, we relate these technical insights to their implications for operating federated forecasting in practical energy systems.

Our results presented in Section 4.2 demonstrate that federated energy forecasting is vulnerable to data poisoning attacks, even when only a single client is compromised. We observe consistent performance degradation across all unaffected clients, with GAN-generated perturbations causing significantly more harm than uniform noise. This effect increases with perturbation scale and varies across energy types and model architectures. The pronounced difference between GAN and uniform noise could be attributed to the structured nature of the adversarial signals. While uniform noise introduces random, uncoordinated deviations that may partially cancel out during aggregation, GAN perturbations are explicitly optimized to mislead a surrogate forecaster. As a result, they may learn transferable patterns that propagate more effectively through the global model and degrade performance across the federated buildings. The degree of degradation also differs by energy type. Load forecasting appears more resilient, which could be due to the inherent variability and stochasticity of load profiles across clients, limiting the influence of a single poisoned input. In contrast, PV time series are typically more homogeneous and predictable, which may enable structured perturbations to generalize more easily, thereby amplifying their impact. Model architecture further influences robustness. MoE models consistently exhibit lower error increases than standard MLP. This improvement may stem from its higher representational capacity and the expert gating mechanism, which could reduce sensitivity to localized corruptions.

Building on the poisoning results, Section 4.2 shows that even a temporally restricted backdoor from a single client can impair the performance of unaffected participants. Unlike poisoning, which impacts the full day, backdoor effects are confined to the trigger window yet still propagate effectively through the cluster. While our trigger targets a fixed two-hour slot, such attacks could be aligned with critical periods (e.g., grid congestion), amplifying their operational impact. Degradation is strongest under GAN-based perturbations, highlighting that learned signals can manipulate the global model even when temporally constrained. Uniform noise also induces errors but is less effective, particularly for PV and prosumption. Load forecasting remains comparatively robust, likely due to its higher intrinsic variability, while PV is the most susceptible. Consistent with earlier findings, the MoE architecture provides increased robustness over the MLP, especially under GAN-based perturbations.

The results in Section 4.3 confirm that federated energy forecasting can safeguarded against both poisoning and backdoor attacks through targeted defenses. Most notably, local retraining and our integrated security framework consistently restore performance, often surpassing the clean FL baseline. This suggests that retraining not only mitigates malicious influence but also corrects residual errors in the global model. Clustering provides moderate protection by diluting poisoned updates across more clients. Its effectiveness varies, offering stronger defense for load and prosumption, but limited for PV, where similarity across clients amplifies attack transferability. Weighted aggregation performs inconsistently: it partially mitigates structured GAN-based attacks but fails against unstructured uniform noise, likely due to difficulty in identifying random perturbations. Across all settings, PV remains the most vulnerable, reflecting its predictable structure, while load is more resilient due to its inherent uncertainty.

Overall, these results show that FL-based forecasting in residential energy systems is not secure by default. Even a single compromised client applying moderate perturbations can propagate errors to the aggregated model, distort cluster-wide predictions, and thereby influence downstream processes such as grid balancing, flexibility activation, and market clearing. At the same time, the observed benefits of MoE architectures, together with the effectiveness of local retraining and larger cluster sizes, demonstrate that robust and privacy-preserving forecasting is feasible when adversarial behavior is explicitly considered in system design. For operators and regulators, this underscores the necessity of complementing accuracy benchmarks with systematic robustness assessments and integrating security mechanisms, continuous monitoring, and adversarial testing into the deployment and certification of federated forecasting services. A combination of weighted aggregation, increased cluster sizes, local retraining, advanced forecasting models, and anomaly detection for noisy or manipulated data emerges as a practical approach to achieving secure and reliable FL in energy systems.

5.1. Limitations and future work

While our findings highlight critical vulnerabilities and effective defenses, several limitations remain, motivating future research directions. First, the GAN-based attacks rely on access to a white-box surrogate model trained on similar data. While this enables structured, worst-case perturbations, real-world attackers may face restricted access or operate under black-box conditions. Future work could explore more these constrained threat models, including transferability from

unrelated domains. Second, our evaluation assumes a single compromised client per cluster. Although this already causes substantial performance degradation, coordinated attacks involving multiple malicious participants could further stress the system. Investigating such collusion scenarios and their implications on defense robustness is a promising next step. Third, the backdoor trigger is fixed to a static time window. While this design facilitates analysis, adaptive triggers, e.g., those aligned with high congestion hours or calendar-based events, could lead to more effective and harder-to-detect attacks. Future work should explore context-aware backdoors and methods to detect them. Fourth, our experiments are limited to two neural architectures (MLP and MoE). While these are commonly used, evaluating more complex architectures such as LSTM, convolutional neural network, or transformers may offer deeper insights into architectural resilience and defense compatibility. Fifth, the study is based on a single dataset with structurally similar clients. Real-world federated systems typically exhibit greater heterogeneity in data quality, scale, and behavior. Future evaluations should account for such non-iid settings, client dropout, and variable participation rates to assess generalizability. Beyond addressing current limitations, future work should aim to develop more sophisticated attack strategies, particularly by advancing the training of the GAN. For example, by including a discriminator network or diffusion-based generative models could enable the creation of smoother, more temporally consistent perturbations that better mimic natural energy consumption patterns, making the attack harder to detect.

6. Conclusion

This paper presented a comprehensive analysis of security vulnerabilities and corresponding mitigation strategies in federated energy forecasting. We demonstrated that both data poisoning and backdoor attacks can substantially impair global model performance. Specifically, perturbations generated using a Generative Adversarial Network increased the RMSE by up to 131% in poisoning scenarios, while backdoor attacks reduced prediction accuracy by up to 48% during the targeted intervals. Among the forecasting tasks, photovoltaic predictions were most susceptible to adversarial manipulation, whereas load forecasting exhibited greater robustness. To address these vulnerabilities, we systematically evaluated four mitigation strategies. Among them, local model retraining and the proposed integrated security framework proved most effective, consistently mitigating both attack types and, in several cases, restoring or surpassing baseline performance. By contrast, clustering-based defenses and weighted aggregation achieved only limited mitigation, particularly in the presence of unstructured or adaptive perturbations. In all evaluated scenarios, the Mixture of Experts architecture demonstrated superior robustness to adversarial interference compared to the standard Multilayer Perceptron, emphasizing the role of architectural choices in enhancing system resilience. Overall, our results highlight the critical importance of integrating robust security measures into federated energy forecasting frameworks to ensure their reliable deployment in real-world energy systems.

CRediT authorship contribution statement

Jonas Sievers: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Krupali Kumbhani:** Methodology, Formal analysis. **Thomas Blank:** Writing – review & editing, Supervision, Funding acquisition. **Frank Simon:** Writing – review & editing, Supervision, Funding acquisition. **Andreas Mauthe:** Writing – review & editing, Supervision, Project administration.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work the author(s) used Grammarly, ChatGPT, and DeepL in order to improve the readability and language of the work. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Jonas Sievers reports financial support was provided by German Research Foundation. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We acknowledge support by the KIT-Publication Fund of the Karlsruhe Institute of Technology, Germany. The presented work was funded by the German Research Foundation (DFG) as part of the Research Training Group 2153: Energy Status Data — Informatics Methods for its Collection, Analysis, and Exploitation.

Appendix. Hyperparameters of the Multilayer Perceptron and Mixture of experts

In this section we briefly describe our MLP and the MoE gating mechanism. For more details we refer to [37]. The MLP serves as a compact reference architecture. The input is a sequence of length T with F features per time step, representing one observation window of the time series. By flattening the $T \times F$ input to a single vector, all temporal samples become jointly accessible to the dense layers. The MLP provides a baseline against which adaptive specialization of the MoE can be evaluated. The MoE extends this setup through a learnable routing mechanism. For each input x_t, the gating network generates a vector of scores, which are normalized into expert weights via a softmax transformation

$$G(x_t) = \mathrm{softmax}(x_t W_g),$$

where W_g is a trainable gating matrix. The output of the MoE layer is a convex combination of expert responses,

$$M(x_t) = \sum_{i=1}^{n} p_i \, E_i(x_t),$$

with $p_i \geq 0$ and $\sum_i p_i = 1$. This formulation preserves differentiability and enables the model to interpolate across experts, allowing each expert to specialize on different sub-tasks. During training, the gating network may repeatedly assign high probability to the same experts, which limits specialization and results in the well-known dying-experts phenomenon. To mitigate this, we regularize based on expert similarity rather than solely on routing imbalance. By discouraging convergence of expert parameters towards identical solutions, the architecture maintains complementary expert behaviors, which improves robustness in heterogeneous energy settings. This design allows the MoE to capture distinct consumption dynamics, device usage patterns, or context-specific temporal signals, while the gating mechanism adaptively selects the most informative experts for each input segment [37].

Data availability

Data will be made available on request.

References

[1] Plaum F, Ahmadiahangar R, Rosin A, Kilter J. Aggregated demand-side energy flexibility: A comprehensive review on characterization, forecasting and market prospects. Energy Rep 2022;8:9344–62. http://dx.doi.org/10.1016/j.egyr.2022.07.038.

[2] Das U, Tey K, Seyedmahmoudian M, Mekhilef S, Idris MYI, Deventer WV, Horan B, Stojcevski A. Forecasting of photovoltaic power generation and model optimization: A review. Renew Sustain Energy Rev 2018;81:912–28. http://dx.doi.org/10.1016/J.RSER.2017.08.017.

[3] Sievers J, Blank T. Secure short-term load forecasting for smart grids with transformer-based federated learning. In: 2023 international conference on clean electrical power. ICCEP, 2023, p. 229–36. http://dx.doi.org/10.1109/ICCEP57914.2023.10247363.

[4] McMahan HB, Moore E, Ramage D, Hampson S, y Arcas BA. Communication-efficient learning of deep networks from decentralized data. In: International conference on artificial intelligence and statistics. 2016, http://dx.doi.org/10.48550/arXiv.1602.05629.

[5] Zhang J, Zhu H, Wang F, Zhao J, Xu Q, Li H, Wang Z. Security and privacy threats to federated learning: Issues, methods, and challenges. Secur Commun Netw 2022;2022. http://dx.doi.org/10.1155/2022/2886795.

[6] Jasiūnas J, Lund PD, Mikkola J. Energy system resilience – A review. Renew Sustain Energy Rev 2021;150:111476. http://dx.doi.org/10.1016/j.rser.2021.111476.

[7] Zhao Y, Xiao W, Shuai L, Luo J, Yao S, Zhang M. A differential privacy-enhanced federated learning method for short-term household load forecasting in smart grid. In: 2021 7th international conference on computer and communications. ICCC 2021, Institute of Electrical and Electronics Engineers Inc.; 2021, p. 1399–404. http://dx.doi.org/10.1109/ICCC54389.2021.9674514.

[8] Husnoo MA, Anwar A, Hosseinzadeh N, Islam SN, Mahmood AN, Doss R. A secure federated learning framework for residential short term load forecasting. 2022, http://dx.doi.org/10.48550/arXiv.2209.14547.

[9] Badr MM, Ibrahem MI, Mahmoud M, Alasmary W, Fouda MM, Almotairi KH, Fadlullah ZM. Privacy-preserving federated-learning-based net-energy forecasting. vol. 2022-March, Institute of Electrical and Electronics Engineers Inc.; 2022, p. 133–9. http://dx.doi.org/10.1109/SoutheastCon48659.2022.9764093.

[10] Qu X, Guan C, Xie G, Tian Z, Sood K, Sun C, Cui L. Personalized federated learning for heterogeneous residential load forecasting. Big Data Min Anal 2023;6:421–32. http://dx.doi.org/10.26599/BDMA.2022.9020043.

[11] Li J, Li H, Wang R, Guo Y, Wu S. Fed-SAD:A secure aggregation federated learning method for distributed load forecasting. 2023, http://dx.doi.org/10.22541/au.169028986.64063960/v1.

[12] Husnoo MA, Anwar A, Reda HT, Hosseinzadeh N, Islam SN, Mahmood AN, Doss R. FedDiSC: A computation-efficient federated learning framework for power systems disturbance and cyber attack discrimination. Energy AI 2023;14. http://dx.doi.org/10.1016/j.egyai.2023.100271.

[13] Liu Y, Dong Z, Liu B, Xu Y, Ding Z. FedForecast: A federated learning framework for short-term probabilistic individual load forecasting in smart grid. Int J Electr Power Energy Syst 2023;152. http://dx.doi.org/10.1016/j.ijepes.2023.109172.

[14] Widmer F, Nowak S, Bowler B, Huber P, Papaemmanouil A. Data-driven comparison of federated learning and model personalization for electric load forecasting. Energy AI 2023;14. http://dx.doi.org/10.1016/j.egyai.2023.100253.

[15] Dong Y, Wang Y, Gama M, Mustafa MA, Deconinck G, Huang X. Privacy-preserving distributed learning for residential short-term load forecasting. IEEE Internet Things J 2024. http://dx.doi.org/10.1109/JIOT.2024.3362587.

[16] Sievers J, Henrich P, Beichter M, Mikut R, Hagenmeyer V, Blank T, Simon F. Federated reinforcement learning for sustainable and cost-efficient energy management. Energy AI 2025;21:100521. http://dx.doi.org/10.1016/j.egyai.2025.100521.

[17] Qureshi NBS, Kim DH, Lee J, Lee EK. Poisoning attacks against federated learning in load forecasting of smart energy. Institute of Electrical and Electronics Engineers Inc.; 2022, http://dx.doi.org/10.1109/NOMS54207.2022.9789884.

[18] Zhang J, Chen J, Wu D, Chen B, Yu S. Poisoning attack in federated learning using generative adversarial nets. Institute of Electrical and Electronics Engineers Inc.; 2019, p. 374–80. http://dx.doi.org/10.1109/TrustCom/BigDataSE.2019.00057.

[19] Luo X, Zhu X. Exploiting defenses against GAN-based feature inference attacks in federated learning. 2020, http://dx.doi.org/10.1145/3719350, CoRR abs/2004.12571.

[20] Manzoor HU, Khan AR, Sher T, Ahmad W, Zoha A. Defending federated learning from backdoor attacks: Anomaly-aware FedAVG with layer-based aggregation. Institute of Electrical and Electronics Engineers Inc.; 2023, http://dx.doi.org/10.1109/PIMRC56721.2023.10293950.

[21] Sievers J, Kumbhani K, Blank T, Simon F, Mauthe A. Security and attacks on federated energy forecasting. In: ENERGY 2025 : The fifteenth international conference on smart grids, green communications and IT energy-aware technologies. International Academy Research and Industry Association (IARIA); 2024, p. 23–8. http://dx.doi.org/10.5445/IR/1000181787.

[22] Lee S, Xie L, Choi D-H. Privacy-preserving energy management of a shared energy storage system for smart buildings: A federated deep reinforcement learning approach. Sens 2021;21(14). http://dx.doi.org/10.3390/s21144898.

[23] Lee S, Choi D-H. Federated reinforcement learning for energy management of multiple smart homes with distributed energy resources. IEEE Trans Ind Inform 2022;18(1):488–97. http://dx.doi.org/10.1109/TII.2020.3035451.

[24] Rezazadeh F, Bartzoudis N. A federated DRL approach for smart micro-grid energy control with distributed energy resources. 2022, http://dx.doi.org/10.1109/CAMAD55695.2022.9966919.

[25] Giuseppi A, Manfredi S, Menegatti D, Pietrabissa A, Poli C. Decentralized federated learning for nonintrusive load monitoring in smart energy communities. In: 2022 30th mediterranean conference on control and automation. MED, 2022, p. 312–7. http://dx.doi.org/10.1109/MED54222.2022.9837291.

[26] Wang Y, Bennani IL, Liu X, Sun M, Zhou Y. Electricity consumer characteristics identification: A federated learning approach. IEEE Trans Smart Grid 2021;12(4):3637–47. http://dx.doi.org/10.1109/TSG.2021.3066577.

[27] He Y, Luo F, Ranzi G, Kong W. Short-term residential load forecasting based on federated learning and load clustering. In: 2021 IEEE international conference on communications, control, and computing technologies for smart grids. SmartGridComm, 2021, p. 77–82. http://dx.doi.org/10.1109/SmartGridComm51999.2021.9632314.

[28] Mayerhofer R, Mayer R. Poisoning attacks against feature-based image classification. In: Proceedings of the twelfth ACM conference on data and application security and privacy. CODASPY '22, New York, NY, USA: Association for Computing Machinery; 2022, p. 358–60. http://dx.doi.org/10.1145/3508398.3519363.

[29] Xiang Z, Miller DJ, Kesidis G. A benchmark study of backdoor data poisoning defenses for deep neural network classifiers and a novel defense. In: 2019 IEEE 29th international workshop on machine learning for signal processing. MLSP, 2019, p. 1–6. http://dx.doi.org/10.1109/MLSP.2019.8918908.

[30] Zhai S, Dong Y, Shen Q, Pu S, Fang Y, Su H. Text-to-image diffusion models can be easily backdoored through multimodal data poisoning. In: Proceedings of the 31st ACM international conference on multimedia. MM '23, New York, NY, USA: Association for Computing Machinery; 2023, p. 1577–87. http://dx.doi.org/10.1145/3581783.3612108.

[31] Wan A, Wallace E, Shen S, Klein D. Poisoning language models during instruction tuning. In: Krause A, Brunskill E, Cho K, Engelhardt B, Sabato S, Scarlett J, editors. Proceedings of the 40th international conference on machine learning. Proceedings of machine learning research, vol. 202, PMLR; 2023, p. 35413–25. http://dx.doi.org/10.48550/arXiv.2305.00944.

[32] Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y. Generative adversarial nets. In: Advances in neural information processing systems. vol. 27, Curran Associates, Inc.; 2014, p. 2672–80. http://dx.doi.org/10.48550/arXiv.1406.2661.

[33] Mopuri KR, Ganeshan A, Babu RV. NAG: Network for adversary generation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. CVPR, IEEE; 2018, p. 742–51. http://dx.doi.org/10.1109/CVPR.2018.00084.

[34] Wang X, He K, Song C, Wang L, Hopcroft JE. AT-GAN: An adversarial generative model for non-constrained adversarial examples. Int Conf Learn Represent 2021. http://dx.doi.org/10.48550/arXiv.1904.07793.

[35] Ratnam EL, Weller SR, Kellett CM, Murray AT. Residential load and rooftop PV generation: An Australian distribution network dataset. Int J Sustain Energy 2017;36(8):787–806. http://dx.doi.org/10.1080/14786451.2015.1100196.

[36] Meteostat. Sydney airport wetterrückblick & klimadaten. 2025, URL https://meteostat.net/de/station/94767?t=2025-02-26/2025-03-05.

[37] Sievers J, Blank T, Simon F. Advancing accuracy in energy forecasting using mixture-of-experts and federated learning. In: Proceedings of the 15th ACM international conference on future and sustainable energy systems. E-energy '24, New York, NY, USA: Association for Computing Machinery; 2024, p. 65–83. http://dx.doi.org/10.1145/3632775.3661945.