Understanding Client Heterogeneity and Update Usefulness in Federated Residential Energy Forecasting

Madeeha Laiq[1], Kuldeep Kumar Patel[2], Afreen Fatima[3]

[1]Department of Computer Science and Information Technology,
Sam Higginbottom University of Agriculture, Technology and Sciences, India,
Email: madlaiq259@gmail.com
[2]Department of Electrical Engineering, Rajiv Gandhi Proudyogiki Vishwavidyalaya, India,
Email: kp8858282272@gmail.com
[3]Department of Computer Science and Engineering, Madhu Vachaspati Institute of Engineering and Technology, India, Email: afreensuhail111@gmail.com
Corresponding author: Madeeha Laiq
Email: madlaiq259@gmail.com

Abstract— Federated learning is a distributed learning framework for forecasting residential electricity consumption without centralizing raw client data. Thirty complete households were selected as simulated FL clients from the real-world dataset, after quality controlled preprocessing and hourly aggregation of the data. Next-hour electricity consumption is predicted using a GRU model with a 24-hour lookback and a chronological train–validation–test protocol. Centralized GRU, FedAvg, a drift-aware reliability-weighted prototype, and diagnostic conflict-aware aggregation are evaluated. FedAvg achieves a test MAE of 0.205108 kWh/hour, RMSE of 0.377878, and R^2 of 0.606950, while the drift-aware prototype achieves MAE of 0.208843, RMSE of 0.380713, and R^2 of 0.601030. Client-level analysis shows there is no significant predictive power of the static measure of heterogeneity on the forecasting MAE but update-alignment and conflict measures are associated with post-update validation error. Conflict-aware weighting does not give evidence of improvement and actually causes a 0.000170 increase in MAE compared to FedAvg.

Keywords—federated learning, smart grid, non-IID data, client heterogeneity, adaptive aggregation.

1. INTRODUCTION

The power consumption of households is naturally dispersed, and the electricity readings from smart meters are typically private, geographically distributed, irregularly-timed and heteroscedastic. This problem is studied within the representative real-life context of the London Low Carbon London (LCL) smart-meter collection, which consists of meter readings from 5,567 London households, sampled at half-hourly intervals between November 2011 and February 2014, for both standard and dynamic time-of-use customers [1]. Traditional centralized forecasting pipeline involves combining raw measurements prior to model training. Federated learning instead leaves the training data at the participating clients and exchanges model information with a coordinating server [2], thereby reducing the need for raw-data centralization. This is the canonical baseline setting as it includes several local optimization steps and a sample weighted global model update [3]. But with the statistical heterogeneity of household demand, client updates may go in different directions. In FedProx, a proximal regularizer is added to prevent local divergence [2] and SCAFFOLD uses control variates to counteract client drift [5]. While these methods provide an important methodological baseline, these new methodologies in energy-forecasting literature have advanced towards personalisation, uncertainty quantification, heterogeneity-aware training and robust aggregation [6] - [10].

This development raises a methodological question: When simple client-level signals are employed to adapt aggregation, do such signals reliably identify useful updates? A large drift magnitude need not be bad, as a client might want to see a larger parameter change as an acceptable local consumption pattern. However, even a minor change may be counterproductive if it's in line with the objective of a local narrow perspective. Therefore, the usefulness of updates is not guaranteed by the geometric form of the update, and the relationship between client heterogeneity and the geometric form of the update should be measured, not assumed.

This study does not propose a new state-of-the-art aggregation algorithm. Instead, it provides a controlled empirical investigation of whether easily observable client-level signals can reliably identify useful updates in heterogeneous

residential energy forecasting. However, this study introduces three research questions: (RQ1) What is the strength of the relationship between measured household heterogeneity and the difficulty of forecasting? (RQ2) Does the magnitude of an update provide a reliable indicator of client usefulness? Under a controlled validation protocol, (RQ3) Can simple update-alignment or conflict measures achieve a measurable improvement over FedAvg weighted using samples?

The study makes four contributions. First, it establishes a reproducible 30-client residential federated learning benchmark using real Low Carbon London smart-meter data with strictly chronological training, validation, and testing periods. Second, it characterizes household heterogeneity using distributional and temporal divergence measures. Third, it systematically examines whether local update magnitude, alignment, and conflict are associated with client-level forecasting behaviour. Fourth, it directly evaluates simple adaptive aggregation rules and identifies the conditions under which these single-signal strategies fail to improve the FedAvg baseline. The study therefore provides an empirical basis for developing richer multi-factor and history-aware aggregation strategies.

2. RELATED WORK AND LITERATURE REVIEW

2.1 Federated Learning and Client Heterogeneity

In order to minimize communication, the concept of federated learning was proposed in which decentralized clients conduct local optimization and the server gathers the model information [3]. One of the biggest challenges is that the data that clients provide is not always independent and identically distributed. The authors of the comprehensive review of FL research highlight key open challenges such as statistical heterogeneity, communication limitations, privacy, robustness, and personalization [4]. In order to deal with the statistical and systems heterogeneity, the local objective is augmented by a proximal term that penalizes too large deviations from the current global model [4]. Instead, SCAFFOLD employs control variates to account for drift of local objectives which are not IID [5]. The following methods also serve as important reference points as they illustrate the need for methods other than the global aggregation step for dealing with heterogeneity.

2.2 Federated Load and Energy Forecasting

Energy forecasting is an example of a FL application that is particularly suitable due to the distribution of energy use in households and the possibility to reveal behavioural information. Recently, works have been focused on assessing GRU-based FL and comparing FedAvg, FedProx, and robust aggregation techniques for non-IID load data [6]. Adaptive aggregation and privacy-preserving load forecasting have been introduced in other work [7]. These studies confirm that the design of the aggregation can make a significant difference in performance, but it also reveals that the outcomes can be drawn from partitioning the dataset, architecture of the model, and budget of local training and the horizon of the evaluation. The recent research is more towards personalization. Das et al. have proposed a personalized FL system that considers heterogeneities in its energy forecasting framework [8] with real-world smart-meter data, achieving improvements over a centralized and a federated baseline under their experimental setup. Wang et al. used personalized FL, a mixture-of-experts mechanism and conformal prediction to jointly consider the household-specific behaviour and predictive uncertainty [9]. Knowledge distillation has also emerged as a strong method for cross-regional load forecasting enabling strong heterogeneity awareness personalization [10].

2.3 Placement of the Present Study

Against this background, the present study takes a deliberately different methodological stance. It is not a recent personalized architecture, requiring an additional complex component. Rather, it examines the question of whether the cost of the aggregation can be reduced and the resulting aggregation made interpretable based on signals that the client can provide, such as the heterogeneity score, the magnitude of the update, the alignment, and the conflict of the updates. This is helpful as a negative result can rule out unnecessary complexity: if just one of the scalars is not a good indicator of client utility, then future systems should take into account historical performance, local validation behaviour, temporal context and personalisation, instead of one update statistic.

3. DATASET AND PROBLEM FORMULATION

3.1 Real-World Smart-Meter Dataset

The data sources for the experiments are the "Low Carbon London" smart-meter energy consumption data, made available by UK Power Networks via the London Datastore [1]. The source includes 5567 households in London, half hour readings, unique identifier for each household, times of readings, consumption in kWh per half hour, and tariff data. The original resource is large; when uncompressed, it has around 167 million rows [1]. The calendar year 2013 is used for the FL experiments conducted using computational control. The results of the full household audit were: 5,096 households were fully covered in 2013 and 470 households were partially covered. The 30 complete households were chosen to represent four heterogeneity strata: Low (7), Moderate (7), High (8), and Very High (8). The client IDs that are selected are stored in the experimental records to enable the reproduction of the partition.

3.2 Preprocessing and Temporal Split

Half-hourly observations are aggregated to hourly data. In the selected data, there were no duplicate conflicts, and 360 rows in the data source were duplicates of an existing household-timestamp record in the data base. This gives an hourly 30 client data file, with 262,800 data points, exactly 8,760 hourly data points per client for the total 2013 period. Only 194 hourly values are missing (0.0738%); no imputation is introduced, and invalid sequence windows are discarded. Chronological split means that further information won't be incorporated into the model. The training is from January to September 2013, October is used for validation, and November to December 2013 is the held-out test set. The number of sequences used for training is 195740, those used for validation are 22250, and those used for test are 43672.

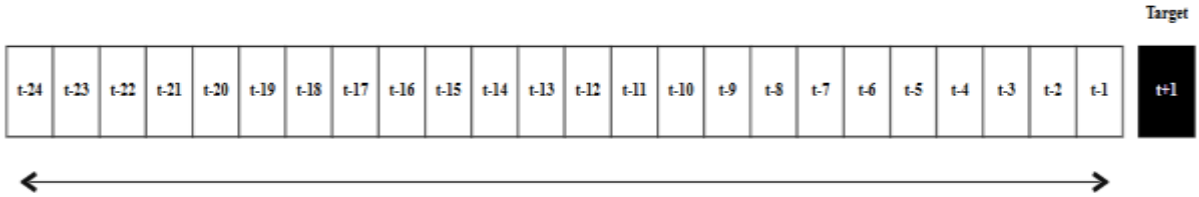


Figure 1. One-step-ahead Forecasting using boundary-aware 24 hour Input Construction.

3.3 Normalization and Forecasting Task

Standardization is only applied to training part. Assume that x represents hourly consumption, μ is the mean of the training set, and σ is the standard deviation of the training set. After transformation this value is:

$$\tilde{X}_t = \frac{x_t - \mu_{train}}{\sigma_{train}} \quad (1)$$

The training mean is 0.4899557 kWh/hour and the training standard deviation is 0.5809915 kWh/hour. A sequence of 24 hours is used to predict the next hour for each household k. The resulting 24-hour input construction is illustrated in Figure 1.

$$X_{k,t} = [\tilde{X}_{k,t} - 23, \ldots\ldots, \tilde{X}_{k,t}], \qquad Y_{k,t+1} = X_{k,t+1} \quad (2)$$

This formulation is able to capture the intraday temporal dependence without losing the household boundary. No random shuffling is done over time.

3.4 Problem Formulation

Let K be the number of clients in the system. D_k be the local training data of the client k and w be the global model parameters. All clients optimise a local forecasting objective, defined over their own data. The server tries to find a global model that is close to the 'empirical risk' function across the entire data set, with also data locality considerations. Here, the question investigated is not just if FL can predict the demand, but whether the statistics collected from the clients can be used to influence the aggregation without affecting the overall generalization.

4. HETEROGENEITY CHARACTERIZATION

4.1 Distributional Divergence

A 20-bin normalized histogram is used to quantify the difference between a household and the global training distribution, to characterize the distribution of the household's hourly consumption. Jensen–Shannon divergence is used:

$$JS(P_k, P_g) = \frac{1}{2} D_{KL}(P_k \parallel M) + \frac{1}{2} D_{KL}(P_g \parallel M), \qquad M = \frac{P_k + P_g}{2} \quad (3)$$

Jensen–Shannon divergence is bounded and symmetric, which is convenient for comparing households rather than a Kullback–Leibler divergence. The mean JS divergence for the full household audit was 0.017926, with a maximum of 0.627447.

4.2 Temporal Heterogeneity

The second measure captures shifts in consumption patterns across the day. An hour-of-day distribution is built and compared for each household to the same one for the world, based on the same JS divergence. The population-level mean temporal JS divergence was 0.012303, suggesting that distributional and temporal differences have different aspects of household behaviour.

4.3 Heterogeneity Score

Due to the complementary nature of the two divergence measures a normalized combined score is used for client stratification:

$$H_k = \frac{1}{2} JS_{dist,k}^{norm} + \frac{1}{2} JS_{temp,k}^{norm} \quad (4)$$

The score is not learned, but used for descriptive grouping. Notably, the analysis does not require that larger values of H_k are correlated with larger values of the forecast error. The resulting group-level relationship between heterogeneity and test MAE is shown in Figure 2.

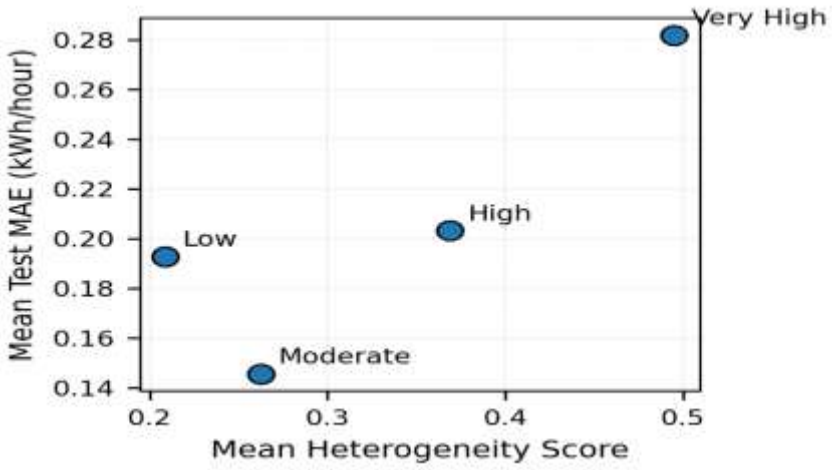


Figure 2. Mean test MAE by the four client strata, using mean heterogeneity.

5. FORECASTING MODEL AND FEDERATED METHODS

5.1 GRU Forecasting Model

The forecasting model is a small gated recurrent unit (GRU) network. It receives a sequence of 24 hourly observation, and has one GRU layer of 64 units, a fully connected layer of 32 ReLU units, and a linear one-unit output layer. This network has a total of 14,977 trainable parameters. Training loss is the mean squared error and Adam is used for optimization with learning rate 0.001.

Layer	Configuration	Parameters
Input	24-time steps × 1 feature	0
GRU	64 units	12,864
Dense	32 units, ReLU	2,080
Output	1 unit, Linear	33
Total		**14,977**

5.2 Mean Squared Error Objective

The local training objective for N forecast samples is:

$$L_{MSE} = \frac{1}{2}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \quad (5)$$

5.3 FedAvg Baseline

The main federated baseline is FedAvg. All 30 clients are involved in each round, every client runs 5 epochs locally with a batch size of 32, and returns the model parameters. The server weights each client update according to the number of local training samples.

$$W^{t+1} = \sum_{K=1}^{K} \frac{n_k}{\sum_{j=1}^{K} n_j} \; W_k^{t+1} \quad (6)$$

This baseline is based on the original idea of iterative model-averaging [3]. After preprocessing, the number of hourly observations in the selected households is the same, so that the effective weight is roughly even across all households. The federated learning system and the evaluated aggregation diagnostics are illustrated in Figure 4.

5.4 Drift-Aware Reliability Prototype

An adaptive aggregator that purely relied on local model drift was developed as an experiment to determine if local model drift is sufficient to identify useful updates. The Euclidean norm of the difference between the local and current global parameter tensors is calculated for a given client k:

$$d_k^t = \left(\sum_{l=1}^{L} \| W_{k,l}^{t+1} - W_l^t \|_2^2\right)^{1/2} \quad (7)$$

The raw drift is divided by the round mean and expressed as a simple reliability score:

$$r_k^t = \frac{1}{1 + d_k^t / \bar{d}^t} \quad (8)$$

This mechanism is treated as a diagnostic prototype rather than as a novel aggregation algorithm. The objective of the experiments is to see if this intuitive single-signal rule beats FedAvg.

5.5 Update Alignment and Conflict Diagnostics

The study does not just explore the size of the local updates, but also the geometric relationships between local updates. Pair-wise cosine similarity is calculated for an update vector Δ_k. Mean positive similarity is interpreted as alignment,

whereas disagreement with other updates contributes to a conflict score. The simple diagnostic conflict-aware weighting is based on the following:

$$C_k = 1 - \frac{1}{K-1}\sum_{j \neq k} \cos(\Delta_k, \Delta_j) \tag{9}$$

And,

$$a_k = \frac{a_k/(1+c_k)}{\sum_{j=1}^{K} n_j/(1+c_j)} \tag{10}$$

This diagnostic is directly validated only with October validation data with the client updates already created. No data from November–December are used for tuning and selecting the weighting rule.

5.6 Experimental Algorithm

The entire controlled procedure is summarized in Algorithm 1. The principle of the design is that the test set does not change until the last evaluation of a frozen model. The end-to-end research workflow is illustrated in Figure 3.

1. Load 2013 Hourly Data (Processed); Load 30-client List (Fixed).
2. Use the data from January to September for fitting μ_{train} and σ_{train}.
3. Build 24-hour input windows and targets for each client.
4. Set up global parameters of the GRU.
5. Send the round's current global parameters to all the clients in each federated round.
6. Adam and MSE are used to perform 5 local epochs for each client.
7. Collect the local models and derive sample statistics and diagnostic measures.
8. Use FedAvg for baseline aggregation and evaluate the drift-aware prototype and conflict-aware diagnostic separately.
9. Use October validation MAE for model selection and for early stopping.
10. Stop model selection, and assess after the November-December test data.
11. Report results on the level of the world, the client, the heterogeneity, and the update characteristic.

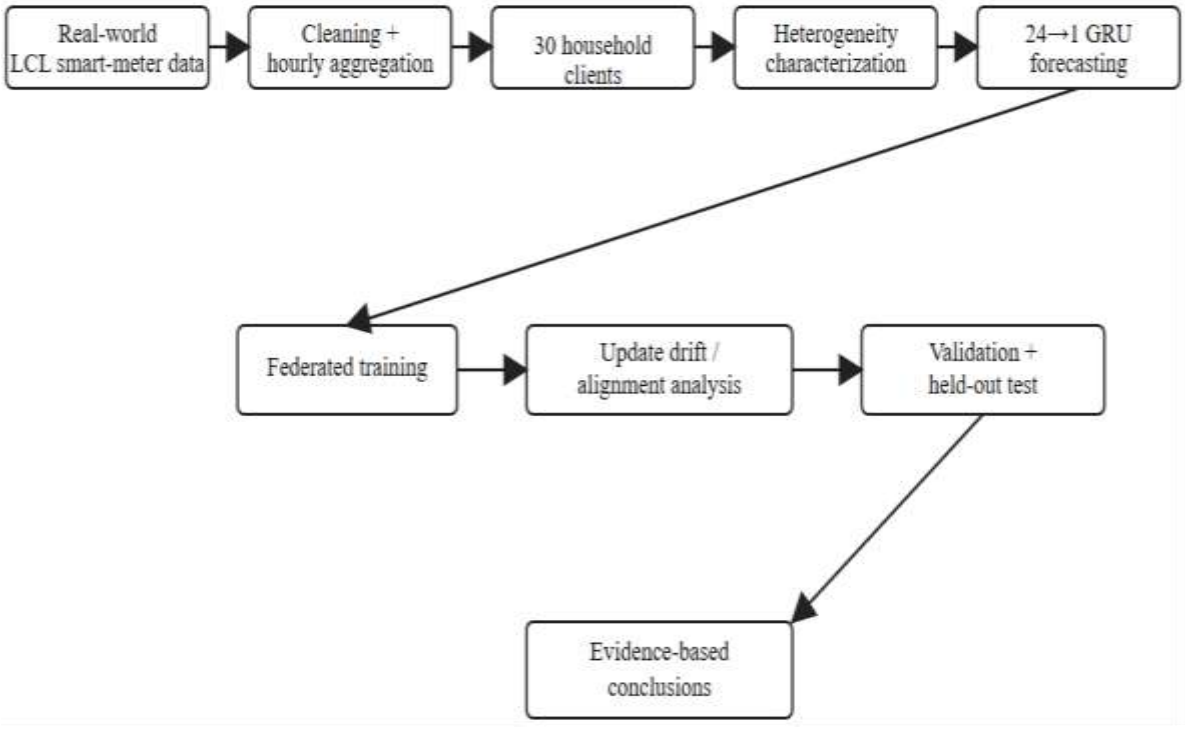


Figure 3. Research workflow for the study: end-to-end research workflow.

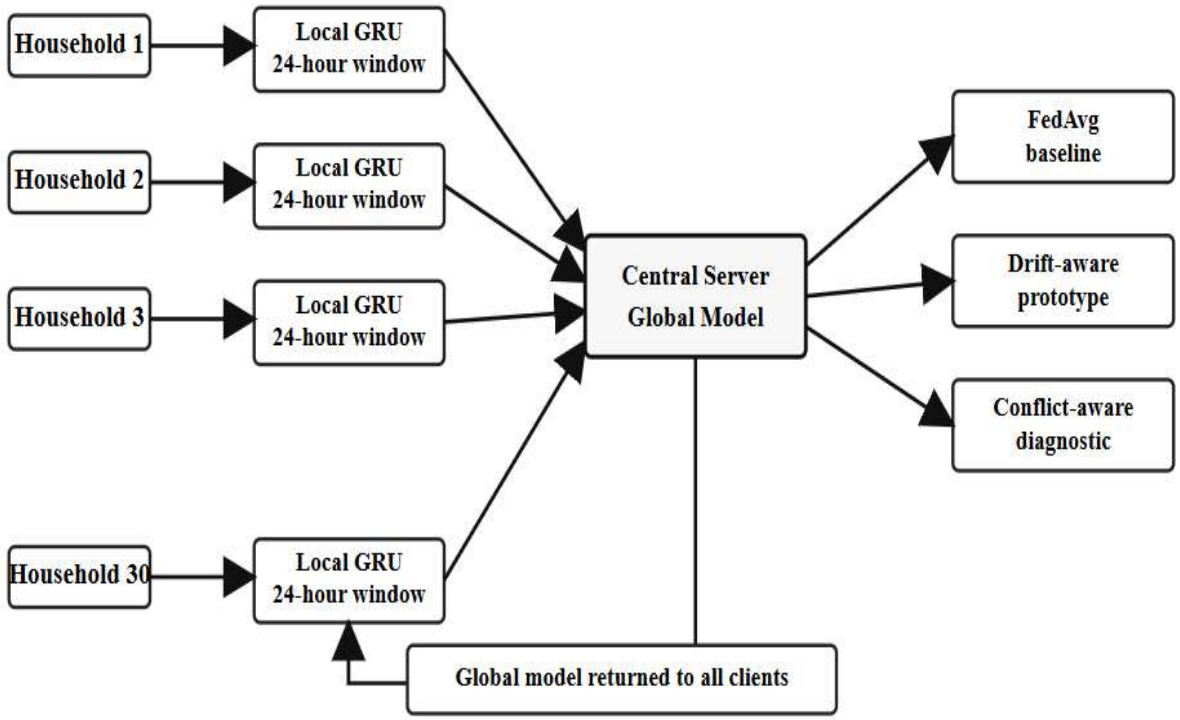


Figure 4. Federated learning systems, and aggregation diagnostics.

6. EXPERIMENTAL PROTOCOL AND METRICS

Setting	Configuration
Clients	30 household clients
Federated rounds	20
Local epochs	5
Batch size	32
Optimizer	Adam
Learning rate	0.001
Sequence length	24 hours
Target	Next-hour consumption
Validation	October 2013
Test	November–December 2013
Framework	TensorFlow 2.20.0
Execution	Google Colab, CPU

6.1 Evaluation Metrics

The primary error metric is mean absolute error:

$$\text{MAE}=\frac{1}{N}\sum_{i=1}^{N}|y_i-\hat{y}_i| \quad (11)$$

And root mean squared error is used to emphasize larger forecasting deviations:

$$\text{RMSE}=\sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i-\hat{y}_i)^2} \quad (12)$$

And the coefficient of determination is reported as:

$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y}_i)^2}$

(13)

The test values reported are all calculated from the held-out set of November and December data. There are no test observations for client selection, for grouping clients into heterogeneous groups, to use for aggregation tuning, or to use for checkpoint selection.

7. RESULTS

7.1 A centralized reference model

The upper reference scenario is a centralized GRU, where all of the household training observations are shared. After five epochs, the model obtains a test MAE of 0.191413 kWh/hour, RMSE of 0.354992, and R^2 of 0.653117. The results provide quantitative estimates of what could be lost in a distributed training scenario with the current partition of clients and the local optimization budget.

7.2 FedAvg Performance

Among the evaluated federated models, FedAvg achieves the lowest final test MAE. The best October validation MAE is 0.339732 at round 14 and the last round is 0.340839. The final frozen test performance is MAE 0.205108 kWh/hour, RMSE 0.377878, and R^2 0.606950.

7.3 Drift-Aware Prototype

The drift-aware α=1 prototype reaches its best October validation MAE of 0.344789 at round 16. Its final test results are MAE 0.208843, RMSE 0.380713, and R^2 0.601030. Relative to FedAvg, MAE increases by 1.82%, RMSE by 0.75%, and R^2 decreases by 0.98%. Thus, the experiment failed to support a hypothesis that the single-signal drift reliability rule provides an improvement in the baseline. The global test performance is summarized in Table 1. The corresponding test MAE comparison is shown in Figure 5, the RMSE comparison in Figure 6, and the R^2 comparison in Figure 7.

Table 1: Global test performance of centralized reference, FedAvg and drift-aware prototype.

Model	**MAE**	**RMSE**	**R^2**
Centralized GRU	0.191413	0.354992	0.653117
FedAvg	0.205108	0.377878	0.606950
Drift-aware α=1	0.208843	0.380713	0.601030

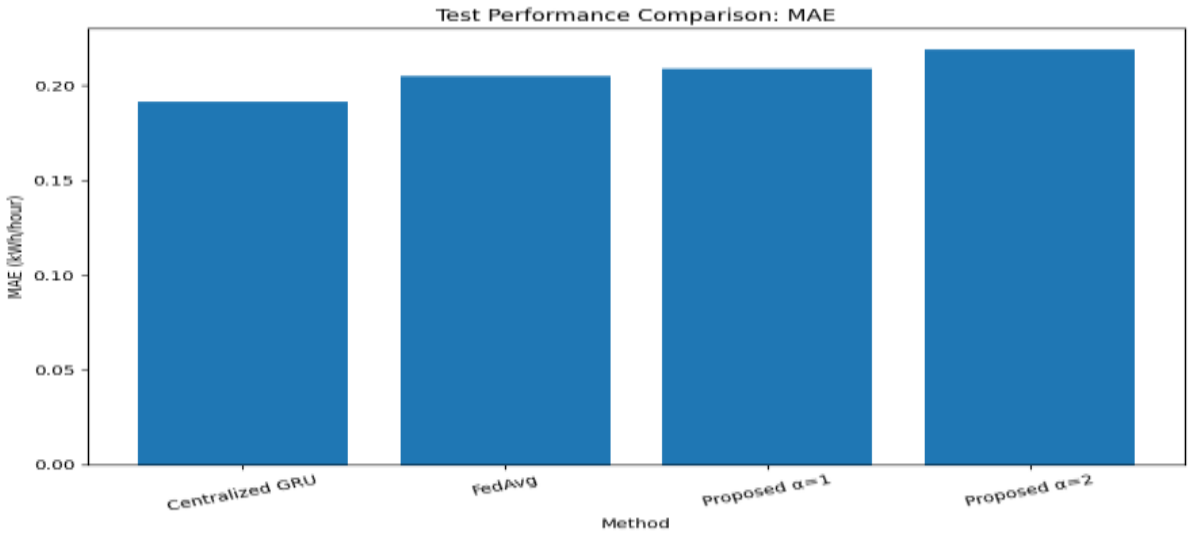


Figure 5. Test MAE comparison of the centralized GRU, FedAvg, and drift-aware aggregation variants.

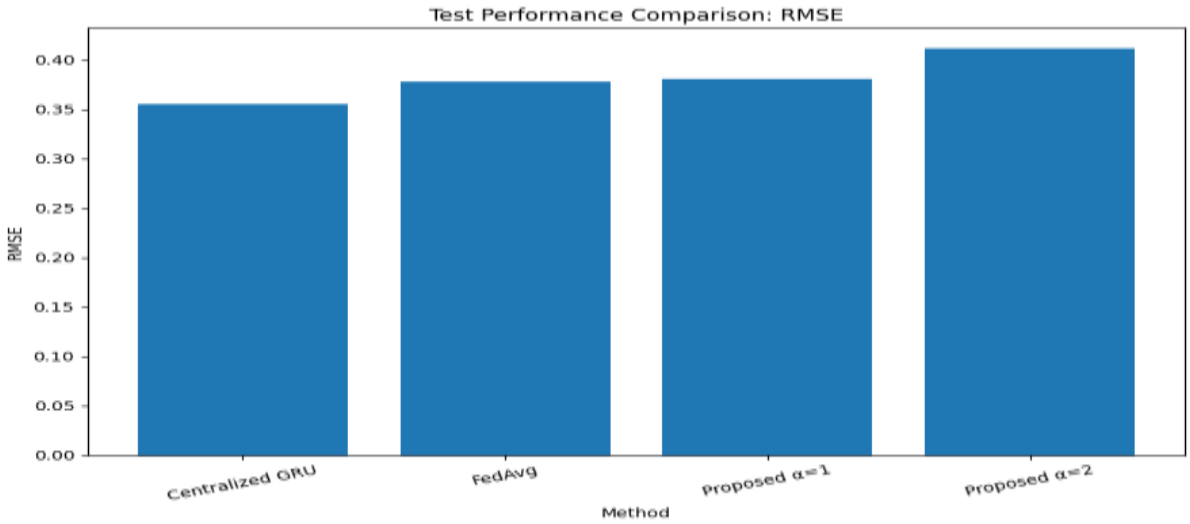


Figure 6. Test RMSE comparison of the centralized GRU, FedAvg, and drift-aware aggregation variants.

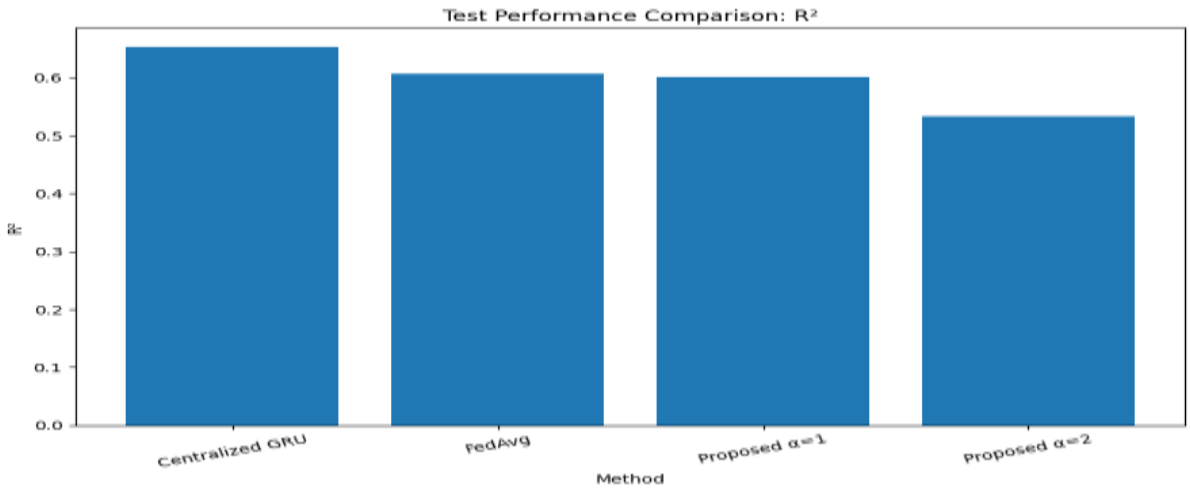


Figure 7. Test R^2 comparison between the centralized GRU, FedAvg and drift-aware aggregation variants.

7.4 Sensitivity to the Drift Exponent

A subtler sensitivity experiment was performed on α=2, for five rounds, where the state from the first round was used again and only two local epochs were used in the other rounds. The best validation round was round 3. On the held-out test set, the α=2 prototype produced MAE 0.219109, RMSE 0.411538, and R^2 0.533808. This is 6.83% more MAE,

8.91% more RMSE, and 12.05% less R^2 compared to FedAvg. As this experiment uses a reduced training budget, it is not treated as a direct comparison with the full-budget experiments, but rather as a sensitivity check. The detailed drift-weight sensitivity results are presented in Table 2.

Table 2: Results of Drift-weight sensitivity

Method	Budget	Selected Round	MAE	RMSE	R^2
FedAvg	20 × 5 local epochs	20	0.205108	0.377878	0.606950
Drift α=1	20 × 5 local epochs	16	0.208843	0.380713	0.601030
Drift α=2	5 rounds; reduced local budget	3	0.219109	0.411538	0.533808

8. CLIENT-LEVEL ANALYSIS

8.1 Variation Across Households

The global mean does not accurately reflect the heterogeneity in client-level performance. The mean and standard deviation of the MAE for the FedAvg model are 0.204632 and 0.107078, respectively, while the minimum and maximum are 0.075328 and 0.505490, respectively. The mean R^2 for the clients is 0.359052. The mean client MAE and standard deviation for the drift-aware model are 0.208365 and 0.107957, respectively, while the mean client R^2 is 0.350143.

8.2 Heterogeneity Does Not Directly Explain MAE

The relationship between the four heterogeneity groups and test error is non-monotonic. Under the drift-aware evaluation, Low, Moderate, High, and Very High groups have mean heterogeneity scores of 0.208689, 0.262514, 0.369034, and 0.495025, while their corresponding mean MAEs are 0.192889, 0.145588, 0.203253, and 0.281948. Pearson correlation of heterogeneity score with MAE at client level is −0.090847 (p=0.633050) and the Spearman correlation at the client level is −0.178198 (p=0.346127). Therefore, higher observed heterogeneity should not be interpreted as a measure to predict the level of difficulty. The group-level relationship is shown in Figure 8.

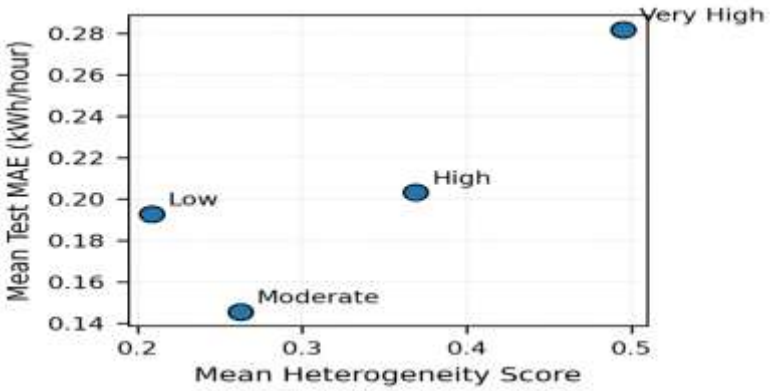


Figure 8. Group level relationship between measured heterogeneity and test MAE is not monotonic.

9. CLIENT UPDATE CHARACTERISTICS

9.1 One-Epoch Controlled Update Study

To separate the update behaviour from long-term training effects, all clients were reset using the same round-20 global model and trained for a single local epoch. New validation performance was evaluated before and after the update based on data from October only. The analysis resulted in 30 client level records and a 30 x 14,977 update matrix. Drift magnitude has only a weak and non-significant association with post-update validation MAE: Pearson r=0.299215 (p=0.108213) and Spearman ρ=0.261846 (p=0.162180). The immediate MAE improvement is also barely related to drift

(Pearson r=0.084861, p=0.655701). These results directly weaken the assumption that a larger local parameter change is inherently less reliable.

9.2 Alignment and Conflict Signals

The rank-based correlation for mean update alignment is 0.471858, p=0.008476; and for update coherence, it is 0.385095, p=0.035602. The conflict score is inversely associated with post-update MAE (Pearson r=−0.411694, p=0.023795; Spearman ρ=−0.500779, p=0.004823). These are not enough to prove cause-and-effect and they do not mean that if the signal was reversed it would be a better aggregator. The relationships between update geometry and post-update validation MAE are shown in Figure 9, and the corresponding correlation statistics are summarized in Table 3.

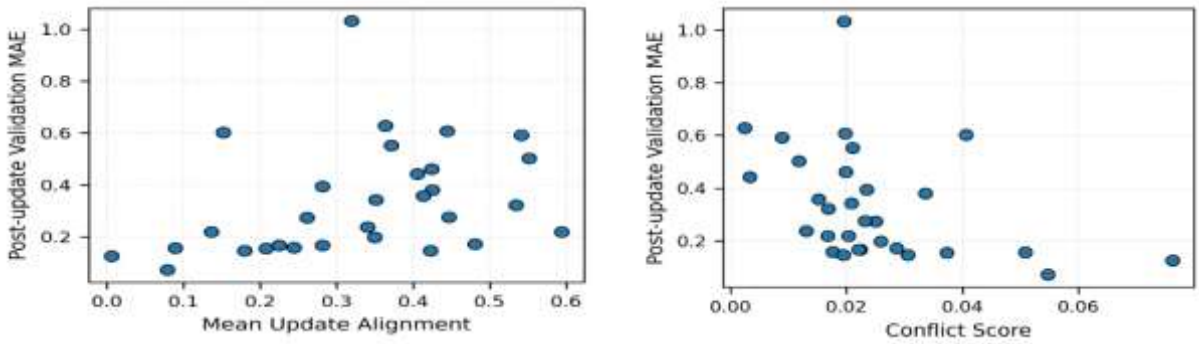


Figure 9. Relationships between update geometry and post-update validation MAE for diagnostics.

Table 3: Correlation analysis of client update characteristics with after-update MAE and MAE improvement.

Signal	Pearson r	Pearson p	Spearman ρ	Spearman p
Drift vs. after-update MAE	0.2992	0.1082	0.2618	0.1622
Alignment vs. after-update MAE	0.3422	0.0642	0.4719	0.0085
Coherence vs. after-update MAE	0.2928	0.1163	0.3851	0.0356
Conflict vs. after-update MAE	−0.4117	0.0238	−0.5008	0.0048
Drift vs. MAE improvement	0.0849	0.6557	0.0590	0.7570

10. DIRECT VALIDATION OF CONFLICT-AWARE AGGREGATION

The conflict-aware weighting rule was directly tested with the already generated 30 client-updates. The server used ordinary sample-weighted FedAvg, and weights were proportional to n_k/(1+c_k) (October validation only). For FedAvg, the validation values were: validation MAE: 0.367296, validation loss: 0.453161. Conflict-aware weighting resulted in MAE 0.367465 and loss 0.452658. Therefore, the conflict-aware weighting slightly decreases the validation loss while increasing the MAE by 0.000170. FedAvg is better for 23 of 30 clients, the conflict-aware model is better for 7.

The direct validation, however, does not support replacing FedAvg with the conflict-aware weighting rule. Most important, the example shows that correlation is insufficient on its own – a statistic can be correlated with the outcome, but a monotonic weighting transformation of the statistic might not substantially increase the forecasting metric. The validation comparison is summarized in Table 4.

Table 4: Validation comparison between FedAvg and conflict-aware aggregation

Aggregation	Validation MAE	Validation Loss	Clients Better
FedAvg	0.367296	0.453161	23
Conflict-aware	0.367465	0.452658	7

11. DISCUSSION

The most significant result is methodological. The experiments do not support the common intuition that any one of the simple properties of a client that is easily measured can directly be translated into a reliable aggregation weight. While the drift-aware prototype is slightly worse than FedAvg on final test set, the conflict-aware diagnostic does not improve validation MAE.

One possible explanation is that update geometry is context-dependent. A large update can be a beneficial adaptation to a pattern that occurs in a household instead of a negative drift. Likewise, an update that differs from other updates may be informative if it includes a different but valid temporal regime. This interpretation is in line with the overall FL literature, which considers the issue of statistical heterogeneity as an important optimization problem and not merely an error ranking [4], [5].

The results also reveal that there is a difference between ‘difficulty’ and ‘heterogeneity'. The highest MAE is found in the highest heterogeneity group, and the average MAE of the Moderate group is lower than that of the Low group, but the client-level correlation tests do not reveal any significant monotonic relationship. Therefore, an adaptive server should not assume that the most heterogeneous client is automatically the least trustworthy.

Recent energy-forecasting studies have addressed these limitations through richer mechanisms. Heterogeneity-aware personalization, mixture-of-experts architectures, uncertainty quantification, and knowledge distillation provide more expressive ways to represent household-specific behaviour [8]- [10]. The present results thus suggest a direction for the future that is grounded on principles: Usefulness should be measured through several factors and past performance rather than based on a single instantaneous statistic.

12. LIMITATIONS

- The federated environment is simulated by using a single real-world dataset; physical communication delays, device failures or asynchronous participation are not modelled.
- The controlled FL benchmark is based on just 30 households, but the analysis of household heterogeneity was based on a much larger household audit. FedProx and SCAFFOLD were not fully trained due to the computational cost of the selected configuration of the GRU/local-epoch. Only a smoke test was performed with SCAFFOLD to assess computational feasibility.
- The drift-aware and conflict-aware mechanisms are diagnostic prototypes and should not be interpreted as state-of-the-art algorithms.
- In the final forecasting trials, communication privacy is not assessed, data locality is conceptually maintained, and neither differential privacy noise nor secure aggregation or formal privacy accounting is used.
- The study only aims to quantify the performance of one-step-ahead hourly forecasting, and does not claim performance for longer forecasting horizons.

13. FUTURE SCOPE

Moving from single signal weighting to multi factor client utility estimation should be done in future work. Candidate signals include client-specific uncertainty, long-term contribution to global generalization, update alignment over several rounds, temporal regime stability, and historical validation improvement. Another possible direction is personalized FL since the findings indicate that a single global model might not be able to represent the behaviour of the individual households.

Another way is communication-aware personalization. With only 14,977 parameters in the GRU in this study, future experiments can measure the impact of the local epochs, number of communication rounds, compression of parameters, and the predictive performance. A third direction is the privacy-preserving deployment by secure aggregation and formal differential-privacy accounting, instead of an informal privacy policy.

Finally, the benchmark should be evaluated on a wider range of real-world energy datasets. Validation of cross-datasets would help separate out general FL behaviour and LCL-specific effects and would give a more solid foundation for claims published at the level of the dataset.

14. CONCLUSION

This study presents a controlled empirical evaluation of federated residential energy forecasting using heterogeneous real-world smart-meter data. The results show that FL can achieve competitive results for one-step ahead forecasting without data pooling in raw data, but the evaluated simple adaptive aggregation mechanisms do not improve on FedAvg. FedAvg's test MAE is 0.205108, test RMSE is 0.377878, and test R^2 is 0.606950, which beats the tested drift-aware $\alpha=1$ prototype.

Measured heterogeneity does not significantly predict MAE at the client level and drift magnitude is not a good predictor of the usefulness of the updates. Alignment and conflict statistics are found to be more related to post-update validation behaviour, while the direct conflict-aware weighting does not benefit the MAE for validation. The central conclusion is therefore not that adaptive aggregation is impossible, but that single-signal adaptation is insufficiently reliable for this setting. The negative result gives a concrete methodological boundary, and inspires the future development of aggregation mechanisms that include historical, temporal, personalized and validation-based evidence.

Acknowledgement

I would like to sincerely thank **Mr. Animesh Kushwaha**, my colleague, for his valuable support and helpful suggestions during this research. I also express my sincere gratitude to **Mr. Rohit Mishra** for his support and encouragement throughout this work.

Data availability statement

The dataset used in this study is publicly available from the London Datastore as part of the Low Carbon London project. The data can be accessed at: https://data.london.gov.uk/dataset/smartmeter-energy-consumption-data-in-london-households-vqm0d

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

No external funding was received for this research.

REFERENCES

[1] UK Power Networks, “SmartMeter Energy Consumption Data in London Households,” London Datastore, Low Carbon London project, Nov. 2011–Feb. 2014. Available: https://data.london.gov.uk/dataset/smartmeter-energy-consumption-data-in-london-households-vqm0d

[2] P. Kairouz et al., “Advances and open problems in federated learning,” Foundations and Trends in Machine Learning, vol. 14, nos. 1–2, pp. 1–210, 2021, doi: 10.1561/2200000083.

[3] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, “Communication-efficient learning of deep networks from decentralized data,” in Proc. 20th Int. Conf. Artificial Intelligence and Statistics (AISTATS), PMLR, vol. 54, pp. 1273–1282, 2017.

[4] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "Federated optimization in heterogeneous networks," Proc. Machine Learning and Systems, vol. 2, 2020.

[5] S. P. Karimireddy, S. Kale, M. Mohri, S. Reddi, S. U. Stich, and A. T. Suresh, "SCAFFOLD: Stochastic controlled averaging for federated learning," in Proc. 37th Int. Conf. Machine Learning (ICML), PMLR, vol. 119, pp. 5132–5143, 2020.

[6] A. Tibermacine, I. Naidji, I. E. Tibermacine, S. Sahraoui, M. Zouai, A. Rabehi, and M. H. Habib, "Privacy-preserving load forecasting in smart grids using federated learning: A comparative analysis of aggregation strategies," Frontiers in Energy Research, vol. 14, 2026, Art. no. 1854654, doi: 10.3389/fenrg.2026.1854654.

[7] "Privacy-Preserving Electrical Load Forecasting Through Federated Learning," in Proc. 2025 IEEE Int. Conf. on Smart Power, Energy, Renewables, and Transportation (SPERT), 2025, doi: 10.1109/SPERT67079.2025.11469461.

[8] P. Das, S. Pandey, C. Pandey, and D. Kumar et al., "Heterogeneity-aware personalised federated learning for household energy forecasting on multi-source real-world smart meter data," Scientific Reports, vol. 16, Art. no. 24957, 2026, doi: 10.1038/s41598-026-53020-6.

[9] J. Wang, D. Xu, L. Xing, T. Chen, Y. Liu, M. Yan, and T. Yang, "Personalized federated learning with mixture of experts and conformal prediction for household energy forecasting," Expert Systems with Applications, vol. 300, Art. no. 130417, 2026, doi: 10.1016/j.eswa.2025.130417.

[10] Z. Zuo, H. Ye, J. Li, and Y. Ge, "A robust and heterogeneity-aware federated learning framework with knowledge distillation for cross-regional load forecasting," IEEE Transactions on Smart Grid, early access, 2026, doi: 10.1109/TSG.2026.3660096.

[11] S. H. Tindemans, "Low Carbon London smart meter data (refactored)," 4TU.ResearchData, dataset, 2023, doi: 10.4121/FBBE775B-48D8-469F-A39B-B64488BFD6FD.