A Regional Flexibility Estimation Tool Using Open Data, AI-based Timeseries Modelling and Optimization

Robert Ellis
Yosuke Yamaguchi
Yusuke Funaya
Advanced AI Innovation Centre
Hitachi, Ltd.
Tokyo, Japan
robert.ellis.ab@hitachi.com

Masatoshi Kumagai
Environment & Energy
Innovation Centre
Hitachi, Ltd.
Ibaraki, Japan
masatoshi.kumagai.ws@hitachi.com

Takehiro Hagiwara
Financial Business
Strategy Planning Unit
Hitachi, Ltd.
Tokyo, Japan

***Abstract*— A scalable system for estimating regional flexibility in distribution grids is presented, focusing on building-level demand response and thermal demand forecasting. The method integrates open-source mapping, building metadata, and time series foundation models to generate representative historic data and optimize flexibility calculations. Target buildings are identified, their physical characteristics estimated and relevant time series data retrieved for forecasting. Flexibility is quantified by aggregating demand response potential and optimizing power requirements across buildings. Statistical tuning and retrieval-augmented generation address data sparsity and scaling challenges. Results show improved forecast accuracy and flexibility quantification, supporting effective grid planning and increased consumer participation in flexibility schemes. The system enables utilities to assess and activate regional flexibility, aiding renewable energy integration and grid reliability.**

***Index Terms*—Demand response, Foundation models, Power distribution networks, Retrieval augmented generation**

I. Introduction

Increasing renewable energy (RE) curtailment is a global issue. In Japan, curtailment due to supply-demand balance began in 2018 and non-firm grid connections, to address congestion, were introduced in 2022. More regions now face limitations at distribution substations which lack sufficient reverse power flow, restricting further RE integration.

Japan's renewable energy generation relies heavily on PV, which accounts for 10% of annual output in 2024, while wind power contributes only 1.2% due to limited wind-suitable areas. With forests covering 67% of the land, policies promote rooftop PV and use of idle land to balance expansion with forest conservation. As most new PV installations will connect to distribution grids, grid reinforcement or demand-side flexibility is needed. However, the absence of methods to estimate flexibility based on regional user profiles hinders quantitative assessment of grid reinforcement and local consumption potential. This research aims to develop a scalable approach to estimate regional flexibility and its activation costs.

Several advanced tools exist to estimate regional electricity demand and renewable energy deployment. For instance, SP Energy Networks' DFES (Distribution Future Energy Scenarios) uses local demographic and industrial projections to provide detailed insights, but relies on extensive proprietary data, limiting scalability [1].

Distribution grid planning also employs building-level estimates using regional statistics and forecasts for heat pumps (HP) and EV chargers, generating high-resolution time series for PV generation and energy use, however, expanding this approach necessitates updated data processing rules [2].

ETH's EnerPol simulates multi-scale demand by integrating traffic models yet requires significant regional data collection and calibration [3]. Given these limitations, we established the following requirements for regional flexibility estimation tools:

Region: The region used in the analysis is the geographic boundary at the substation level, this varies across Japan but in the relatively rural study area this is maximum of 40 to 60MW.
Flexibility: Estimated by triggering demand response at set times, comparing costs with and without it and summing results across users to assess overall potential.
Targets: Connected consumers with controllable demand response as not all buildings or devices are suitable. For local PV production, users with high demand and DR-controllable devices to optimize flexibility are prioritized, even if fewer qualify. As such the current focus is on those with substantial thermal needs for refrigeration or hot water.

Flexibility estimation can use either a top-down method, which profiles regions, or a bottom-up approach, which optimizes individual DER operations [4]. This study uses the top-down method to analyze buildings and demand patterns, then applies the bottom-up approach to quantify flexibility based on those profiles, estimating regional flexibility as the sum of economically rational DR from EMS-equipped users.

979-8-3315-8138-1

Recent reviews show deep learning (LSTM, GRU, CNN, Transformers), ensemble ML (RF, XGBoost), and hybrid statistical–AI models now dominate energy demand forecasting, with superior accuracy and adaptability [5]. Reference [6] highlights computational and communication challenges in bottom-up flexibility envelope estimation. We developed a flexibility model for distributed computing and hierarchical aggregation to overcome some of these problems.

Section II of this paper presents the overview of data driven energy demand profiling and the optimal operation-based flexibility quantification. Section III presents the evaluation results of the proposed methods and prototype application. We conclude in Section IV with a summary.

II. Implementation

A. System overview

This section provides an overview of the developed system for flexibility estimation, a schematic is shown in Fig. 1. The approach has several steps: 1) a calendar range of 1 – 14 days ahead is chosen. 2) A distribution area is selected. 3) Target buildings in the area are extracted based on predetermined tags available in mapping APIs e.g. supermarket, nursing home, hospital, family restaurant, and hotel. API query response is then cross-referenced to offline Japanese building metadata and returns building lists, class tags, locations, and footprint polygons. 4) For each building, the polygon is used to calculate floor area, building height is estimated from this and other metadata, then total floor space is found by dividing height by 3m (average Japanese floor height) and multiplying by the single floor area. 5) For each building, metadata, seasonal data, and total area are used to select short representative time series from a knowledge base. These are combined to form a single recent historic input, usually covering the week before the chosen calendar date and input into the forecasting model. 6) Thermal demand forecasts for each building are passed to refrigeration or heat pump models. Numerical optimization determines the minimum cost and demand response power needed at each time step and flexibility is calculated by summing all building power requirements in the region for each interval. 7) results are visualized via a UI.

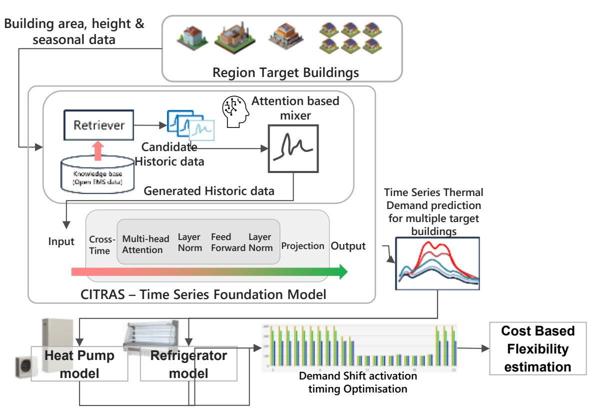

Fig. 1. Schematic overview of flexibility estimation tool.

B. Modelling challenges

Time series analysis (TSA) is crucial for making decisions based on trends and patterns of data in many business domains and is a core requirement of flexibility estimation. The prevailing approach often suffers from two problems.

1) Lack of standardization for modelling approach

Time series analysis tasks, such as forecasting, classification, and anomaly detection, often require specialized models for each building type or device. This approach is inefficient and limits scalability and generalization for regional applications.

2) Data sparsity

Specialized models depend on closed data, which is often unavailable or limited. Even with access, models may not adapt well to new cases, resulting in poor performance across different tasks.

C. Time Series Foundation Model

To mitigate these issues a foundation model approach was used.

1) Hitachi CITRAS

Recent TSFMs, such as Timer and TimesBert, have enabled zero-shot forecasting on first-seen time series based on temporal dynamics. However, they cannot utilize covariates that represent external factors affecting future values, such as weather or calendar information. In contrast, CITRAS [9] flexibly incorporates these covariates into a decoder-only transformer, thereby enableing forecasts based on temporal dyamics and external infleunce. We utilized a pretrained version of CITRAS as our TSFM, which can generalize across domains without requiring task- or building-specific model training.

2) TS-RAG application

In most zero-shot TSFM applications, recent real historic data is assumed to be available as input at inference time. In the practical setting considered here, however, closed behind-the-meter data from devices is difficult to obtain for individual consumers and infeasible to collect at scale for regional applications. As a result, the system operates in a cold-start deployment setting, where data sparsity arises at the forecast input stage rather than during model training. To address this, we propose the use of Time Series Retrieval-Augmented Generation (TS-RAG) to construct a representative historical input sequence for TSFM inference. In contrast to typical TS-RAG applications that augment real historical inputs, retrieval here substitutes for missing history to enable forecasting under these conditions. Retrieval-Augmented Generation (RAG) uses relevant information from an external knowledge base for use in downstream model tasks. In this system, the knowledge base consists of homogeneous, semi-anonymized open EMS data [10], enabling a straightforward retrieval process as illustrated in Fig. 2. Building metadata from the mapping API query (e.g., building type, floor area, and target date range) is used to retrieve and filter candidate time series sequences, which are normalized, dissimilar samples are then excluded based on RMSE.

979-8-3315-8138-1

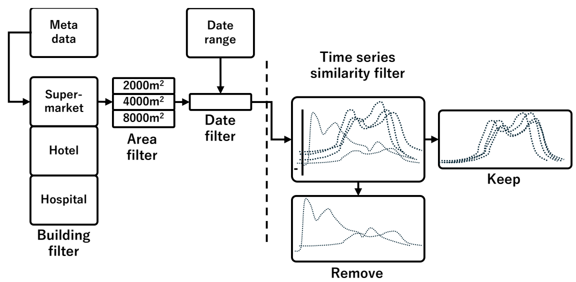

Fig. 2. Overview of straightforward data retrieval method implementation.

The retrieved time series are then aggregated using a lightweight attention-based encoder–decoder mixer. Given 1–10 retrieved EMS time series, the encoder computes similarity-weighted attention scores and produces a single latent sequence that lies in the same space and units as the input data (kWh at hourly resolution). The mixer is trained offline using a standard supervised learning loop, where the mixed sequence is decoded back to the retrieved series and optimized using a reconstruction loss. The TSFM (CITRAS) is not required during mixer training and remains fixed and pretrained, it is not fine-tuned or adapted. At inference, only the encoder component of the mixer is used to generate a single representative historical time series, which is passed directly to the TSFM for forecasting the selected date range. The resulting thermal demand forecasts for each building are subsequently used in flexibility estimation and optimization.

D. Flexibility optimization and calculation

Flexibility modeling aims to simply express power and energy with demand response (DR) as a "virtual storage battery." This raises the following challenges:

1. Unlike a physical battery, a DER such as a heat pump is primarily operated to deliver a service. As a result, its flexibility is conditional on its operating state, and the equivalent virtual battery parameters, including state of charge, power, energy, and cost, are not fixed.
2. Increasing parameters to represent time-varying power or energy complicates aggregation and allocation.
3. Methods requiring revisions to existing DER operation models make it difficult for DER users or aggregators to adapt easily, preventing societal implementation.

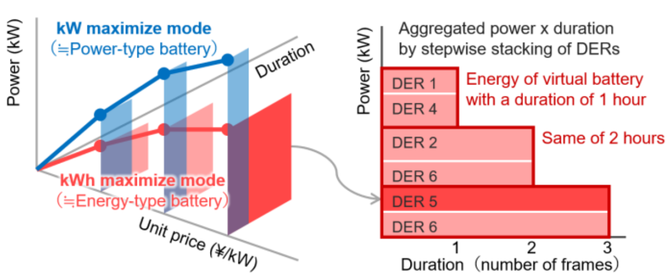

Fig. 3. The flexibility supply curve and stepwise aggregation.

Regarding challenge 1, we defined a "flexibility supply curve" using three metrics: unit price, power, and duration. This is a supply curve, common in electricity trading, augmented with a duration axis. In the flexibility supply curve on the left of Fig. 3, for a given incentive (flexibility unit price), the height of the rectangle corresponds to the power (kW) as a virtual battery and the area corresponds to the energy (kW·h). The kW maximization mode and the kW·h maximization mode correspond to power-type batteries and energy-type batteries, respectively. The advantage of the flexibility supply curve is that when aggregating multiple DERs focused on a specific flexibility unit price, the corresponding rectangles can be extracted and stacked in a stepwise manner, as shown on the right of Fig. 3, enabling easy aggregation as virtual batteries with discrete durations.

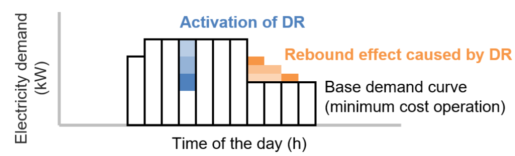

Fig. 4. The idea for DR cost definition based on DER operation planning.

Regarding challenge 2, calculating such a flexibility supply curve requires identifying acceptable demand response for a given incentive while satisfying the DER's intended usage. Fig. 4 illustrates this concept. While a downward DR is used as an example here, the same applies to upward DR.

DERs equipped with EMS are generally operated to minimize operating costs under given constraints, which is termed the base operation plan. DR operation plans differing from the base operation plan increase operating costs. The additional cost of the DR operation plan relative to the base operation plan represents the activation cost of the flexibility. If the activation cost is less than or equal to the flexibility incentive, the user can accept those DR operation plans. This idea is expressed as the following DR constraints.

$$K_{\text{direction}} C_{\text{unit}} \left(W_{\text{base}\,F_{\text{activate}}} - w_{F_{\text{activate}}} \right) n_{\text{duration}} \geq \sum_f Y_{\text{usage}\,f} w_f - \sum_f Y_{\text{usage}\,f} W_{\text{base}\,f} \;, \tag{1}$$

$$W_{\text{base}\,f} - w_f = W_{\text{base}\,F_{\text{activate}}} - w_{F_{\text{activate}}}, \quad f = \{F_{\text{activate}} + 1, \cdots, F_{\text{activate}} + (n_{\text{duration}} - 1)\} \,. \tag{2}$$

where the constant $W_{\text{base}\,f}$ and the variable w_f is electricity demand at time frame f in the base operation plan and DR operation plan, respectively. The variable n_{duration} is the number of sustained frames for DR duration. The constant $K_{\text{direction}}$ represents a downward or upward DR when set to 1 or -1, respectively. C_{unit} is the given flexibility unit price. $Y_{\text{usage}\,f}$ is the given electricity usage charge at frame f. F_{activate} is set to indicate the frame of DR activation. The left-hand side of (1) represents the incentive income for DER users, while the right-hand side represents the DR activation cost mentioned above. In addition, the DR power is constrained to

979-8-3315-8138-1

remain the same throughout its duration by (2). The rebound effect can be managed by additional inequality constraints between $W_{\text{base}\,f}$ and w_f in the specific time window.

Furthermore, under these constraints, the DR objective function that maximizes flexibility becomes:

$$\max.\ K_{\text{direction}}\left(W_{\text{base}\,F_{\text{activate}}} - w_{F_{\text{activate}}}\right)n_{\text{duration}} \\ -K_{\text{base}}Obj_{\text{base}}. \tag{3}$$

where Obj_{base} is the objective function of the base operation plan and K_{base} is the parameter for the linear weighted sum. The first term of (3) represents maximizing DR energy while the second term represents the cost-minimizing DR operation plan against redundancy under sufficient incentive. Equation (3) is defined for kW·h maximization mode, while kW maximization mode can be defined by excluding n_{duration}. Though these equations are described using quadratic programming for the variables w_f and n_{duration}, they can also be described using the mixed integer linear programming.

Regarding challenge 3, as seen in equations (1), (2) and (3), the original base operation plan with device/comfort constraints does not require modification in the calculation of the DR operation plan. That is, the DR operation plan can be implemented as an additional module that calls the operation planning algorithms for various DERs or combination of DERs, making use of existing software assets. Furthermore, this approach is expected to contribute to reducing iterative calculations by using the solution of the base operation plan as the initial solution for the DR operation plan.

III. Results

Results are presented as two sections, A) covering forecast accuracy of the TS-RAG implementation and B) Flexibility calculation.

A. Demand Forecasting

Thermal demand forecasts are generated using a recursive approach, where each day's forecast uses a seven-day input window, as forecasting progresses, previous predictions replace historical data. While longer horizons are possible, practical deployment is currently limited to one week, with a maximum horizon of two weeks. It is important to note that when recent historical data is available, direct TSFM (as well as more typical TS-RAG implementations) inference would be expected to outperform the current approach. In the target deployment setting considered here, however, such behind-the-meter data is unavailable prior to deployment. TS-RAG therefore enables the use of a pretrained TSFM by constructing a representative historical input under cold-start conditions. A key objective of this work is to achieve a forecast error of ≤35%, corresponding to the maximum allowable error for PPA subsidy eligibility and consumer participation in flexibility schemes. Fig. 5 shows unscaled and normalized demand forecasts for four supermarket size categories. Direct comparison of unscaled forecasts reveals large variation in raw energy error across sizes, with an average error of 72%, driven primarily by overestimation for the largest building.

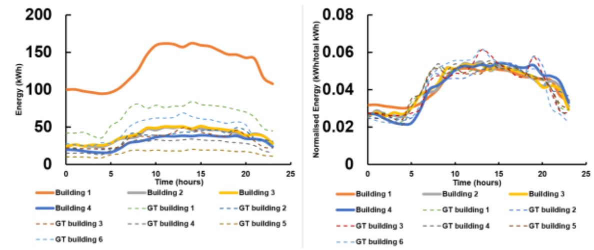

Fig. 5. Unscaled raw (left) and normalized (right) demand forecasts.

Normalizing by total daily energy substantially improves shape accuracy, reducing average error to 10%, indicating that occupancy-driven temporal patterns are well captured, while absolute power scale remains sensitive to unobserved factors such as efficiency, building age, and device characteristics. Applying statistical scaling using limited closed datasets further reduces raw energy error to 23% and normalized error to 7% (Fig. 6), meeting the threshold required for business feasibility.

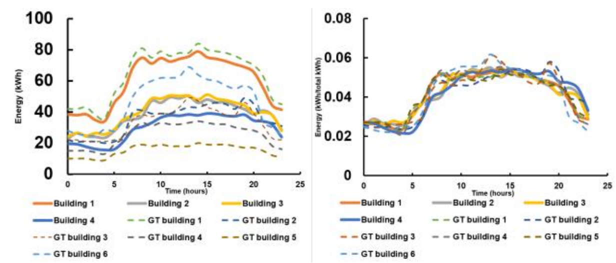

Fig. 6. Scaled raw (left) and normalized (right) demand forecasts.

Additionally, we compare the proposed TSFM-based approach against a conventional boosting ensemble trained separately for weekdays, Saturdays, and Sundays, with multiple regressors aggregated per day type. While effective for single-building forecasting, this baseline requires maintaining multiple models and does not scale naturally to regional deployment.

Under the same evaluation and scaling procedure, TSFM (with RAG) consistently outperforms the boosting ensemble across size categories, reducing median MAPE from 31.05% to 26.38% (−17%) and RMSE-based error from 34.19% to 30.41% (−11.0%) for statistically scaled data.

B. Flexibility optimization and calculation

Flexibility modeling is evaluated by single DER simulation. We examined the DR behaviors and price signal requirement, focusing on a HP. The rated power is 17 kW, and the base operation plan is optimized for the market-linked charge shown in Fig. 7 (a). DR operation plans for flexibility unit prices of 2 and 4 JPY/kW·h are shown in Fig. 7 (c1) and (c2), respectively, while (d1) and (d2) show the corresponding thermal balances. Thermal demand follows the case study for a nursing care home in the heat pump guidebook [11].

Graphs (c1) and (c2) both represent DR operation plans with an activation time of 14:00. For each flexibility unit price, upward DR of 3.7 kW × 2 hours and 5.1 kW × 5 hours was achieved. The decrease in operational efficiency due to DR and rebound, along with differences in electricity charges, leads to

979-8-3315-8138-1

increased operating costs. Because these factors are nonlinear as shown in the graphs (a) and (b), the DR quantity does not increase linearly with the flexibility unit price.

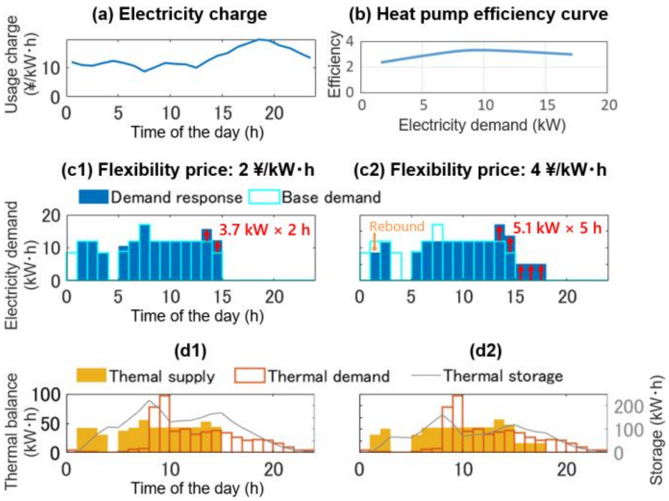

Fig. 7. Examples of DR operation planning for a hot water HP.

The daily operation costs of the base operation, DR operation (c1), and (c2) are 17,499 JPY, 17,598 JPY, and 18,469 JPY respectively. For DR operation (c1) and (c2), the cost increase relative to base operation is 99 JPY and 970 JPY respectively, equivalent to the flexibility incentives. This result indicates that the DER user can perform economically rational DR with a small signal price. Such quantification provides crucial insights for grid operators seeking to utilize DERs as an alternative to grid reinforcement.

C. Prototype application

We have prototyped a web application combining the technologies described above. This features a UI for selecting distribution areas, dates, and building types, and performs map rendering of identified buildings, demand forecasting, and quantification of flexibility as shown in Fig. 8.

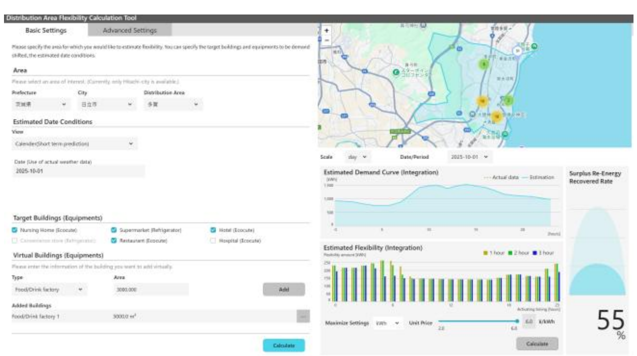

Fig. 8. Overview of the regional flexibility estimation tool.

Fig. 9 illustrates the stepwise aggregation of flexibility within the region view of the graph in the lower right corner of the tool. The horizontal axis represents the activation time of flexibility, and the vertical axis represents the corresponding power. The differently colored bar graphs represent durations of 1, 2, and 3 hours, respectively. Comparing graph (a) and graph (b), it was confirmed that in kW maximization mode, the power tends to be higher while the duration is shorter, whereas in kWh maximization mode, the energy tends to be increased by extending the duration.

Fig. 9. Comparison of kW and kW·h maximization modes.

IV. CONCLUSION

The paper introduces a system for estimating building-level grid flexibility, featuring data collection, profiling, and time series demand forecasting. It employs statistical tuning and retrieval-augmented generation to manage data sparsity and scaling, resulting in better forecast accuracy and flexibility assessment. The approach supports renewable integration and is scalable. Future work aims to enhance models, broaden the knowledge base, improve retrieval methods, and develop new flexibility calculations for industrial use.

References

[1] "Distribution Future Energy Scenarios," SP Energy Networks, [Online]. Available: https://www.spenergynetworks.co.uk/pages/distribution_future_energy_scenarios.aspx.

[2] C. Y. Evrenosoglu, T. Demiray, P. Buchecker, A. Nazaré, D. Incesu and R. Tessier, "Distribution grid planning in Switzerland considering local flexibility," in *Proc. CIRED 2025 Conf.*, 2025.

[3] P. Plagowski, A. Saprykin, N. Chokani and R. S. Abhari, "Impact of electric vehicle charging – An agent-based approach," *IET Gener. Transm. Distrib.,* vol. 15, no. 18, pp. 2605-2617, 2021.

[4] M. S. Bravo, M. N. Pincetic and A. Kiprakis, "Demand-side energy flexibility estimation for day-ahead models," *Applied Energy,* vol. 347, 2023.

[5] A. Mystakidis, P. Koukaras, N. Tsalikidis, D. Ioannidis and C. Tjortjis, "Energy Forecasting: A Comprehensive Review of Techniques and Technologies," *Energies,* vol. 17, no. 7, 2024.

[6] N. Hekmat, H. Cai, T. Zufferey, G. Hug and P. Heer, "Data-Driven Demand-Side Flexibility Quantification: Prediction and Approximation of Flexibility Envelopes," in *Proc. 2023 IEEE Belgrade PowerTech.*, 2023.

[7] Y. Liu, H. Zhang, C. Li, X. Huang, J. Wang and M. Long, "Timer: Generative Pre-trained Transformers Are Large Time Series Models," in *Proc. of the 41 st Int. Conf. on Machine Learning*, Vienna, 2024.

[8] H. Zhang, Y. Liu, Y. Qiu, H. Liu, Z. Pei, J. Wang and M. Long, "TimesBERT: A BERT-Style Foundation Model for Time Series Understanding," *arXiv,* no. arXiv:2502.21245, https://doi.org/10.48550/arXiv.2502.21245, 2025.

[9] Y. Yamagichi, I. Suemitsu and W. Wei, "CITRAS: Covariate-Informed Transformer for Time Series Forecasting, https://doi.org/10.48550/arXiv.2503.24007, 2025.

[10] "EMS open data," Sustainable open Innovation Initiative, [Online]. Available: https://www.ems-opendata.jp/.

[11] *Design Guidebook for Hot Water Heat Pump Systems*, the Heat Pump and Themal Storage Technology Center of Japan, 2018.

979-8-3315-8138-1