Urban Energy Modeling with multiple open data sets and validation at city scale

Kairo SILVEIRA, Benoit DELINCHANT, Yves MARECHAL
Univ. Grenoble Alpes, CNRS, Grenoble INP, G2Elab, 3800 Grenoble, France

ABSTRACT — This work presents a method for modeling electricity demand at multiple spatial and temporal scales using open datasets and data-driven techniques. Despite the lack of accurate data at both levels and the challenges posed by missing and imprecise inputs, data fusion enables reliable modeling across scales. The approach estimates annual consumption at the building level from physical characteristics and then disaggregates this consumption temporally into a detailed time series with a 30-minute time step using national load profile coefficients. The method reconstructs the aggregated time series with a Mean Absolute Percentage Error (MAPE) of less than 10% over the test period, demonstrating its potential for urban energy modeling.

Keywords – urban energy modeling, open datasets, data fusion, top-down, bottom-up, hierarchical model, small-scale modeling.

1. INTRODUCTION

The decentralization of energy systems, driven by increasing reliance on renewable sources, requires localized tools to ensure grid flexibility and stability. These tools are essential at finer spatial scales, such as neighborhoods, where flexibility services can be effectively implemented. Short-term forecasting, covering minutes to hours (e.g., day-ahead), is particularly important for real-time grid operations, enabling proactive management of energy distribution and demand [1].

In cities like Grenoble, the unavailability of precise open data at both spatial and temporal scales highlights the necessity of merging diverse data sources and models to achieve a seamless urban energy modeling framework that addresses these scales.

Several studies have explored data-driven strategies to model building energy consumption. [2] implements and compares different modeling approaches to estimate annual consumption, highlighting the strong performance of machine learning techniques in this context. The use of residential archetypes is also examined in [3], where both black-box and grey-box models are tested, and data-driven methods are found to outperform those grounded in physical assumptions. In [4], time series of building consumption are generated using normalized standard load profiles, an approach conceptually similar to the use of profile coefficients presented in this work.

This study focuses on identifying and merging open data sources while integrating them into a modeling framework for electricity demand at the building level in Grenoble. The goal is to generate a time series of electricity demand with 30-minute intervals, providing granular insights for real-time energy management. By integrating load profiles and machine learning model, this work ensures their suitability for small-scale applications, contributing to the development of precise and practical solutions for urban energy challenges.

2. DATA SOURCES AND INTEGRATION

The city of Grenoble offers an illustrative case for exploring the challenges and opportunities in building-level electricity demand modeling. Some key datasets provide complementary information on electricity consumption and building characteristics. This section presents the datasets used and explains the methods adopted to merge them into a coherent, integrated dataset that supports the modeling pipeline.

2.1. *Distribution System Operator Data*

Electricity consumption measurements were provided by GreenAlp, the distribution system operator (DSO) in Grenoble. The data consist of aggregated electricity demand time series recorded at the level of medium-voltage meshes. Each mesh includes a group of buildings connected downstream of the MV/LV transformers via low-voltage cables. These measurements serve as the main reference for validating the reconstructed demand time series obtained from the bottom-up modeling approach.

2.2. *Base de Données Nationale des Bâtiments*

The *Base de Données Nationale des Bâtiments* (National Building Database, BDNB) is a nationwide dataset that consolidates building-level information from multiple public sources. Managed and developed by the Centre Scientifique et Technique du Bâtiment (CSTB), the BDNB includes detailed attributes such as the *Diagnostic de Performance Énergétique* (DPE, Energy Performance Certificate), surface area, construction year, number of dwellings, main heating energy source, and usage type. Among the available datasets, BDNB stands out for its coverage and richness of descriptive fields.

In this work, BDNB was used as one of the primary sources of building characteristics to support the estimation of annual electricity consumption. Features such as surface area, building age, and declared usage were extracted and linked to other datasets. Although similar information is also available in other databases, BDNB provided the most comprehensive set of descriptors.

2.3. *Observatoire National des Bâtiments*

The *Observatoire National des Bâtiments* (ONB) is another nationwide dataset that provides building-level information, some of which overlaps with the BDNB. In our work, ONB is used in a complementary manner to enrich the dataset when additional or alternative information is needed.

A key feature of ONB used in this study is the column `typobati`, which describes the building usage. This classification was essential for modeling electricity consumption. Additionally, ONB provides annual electricity consumption data by parcel for a subset of buildings across France, which, although not available for buildings in Grenoble, was utilized in the training phase to develop models estimating annual electricity demand based on building characteristics.

2.4. *ENEDIS — Annual Electricity Consumption by Address*

ENEDIS provides datasets containing annual electricity consumption aggregated by address for two main customer categories : residential and non-residential users. The residential dataset includes consumption data for addresses with at least ten residential delivery points, while the non-residential dataset covers enterprises, with consumption values aggregated by activity sector.

Although these datasets offer valuable information for understanding consumption patterns, they come with some limitations. ENEDIS notes that address reliability is not guaranteed, and changes in address normalization procedures since 2021 may introduce inconsistencies across years. Despite these challenges, the address-level consumption data was used in this study as part of the target data for training the annual consumption estimation model.

2.5. *ENEDIS — Dynamic Load Profile Coefficients*

ENEDIS provides dynamic load profile coefficients used to reconstruct electricity consumption curves at a fine temporal resolution. These profiles are unitless time series representing normalized consumption patterns for different usage types. They are derived from measurements collected from a large panel of representative sites and are regularly updated.

For this work, two profiles were selected : `RES1_BASE` for residential use and `PRO1_BASE` for non-residential (professional) use. These dynamic coefficients are designed to reflect typical daily and seasonal variations in consumption and are commonly used in regulatory processes to estimate load curves for sites without direct load measurements.

These empirically derived profiles offer a practical and scalable way to disaggregate annual consumption into sub-daily time series, forming a crucial component of the reconstruction pipeline used in this study.

An example of these dynamic coefficients over the course of one week is shown in Figure 1.

FIG. 1. Example of dynamic load profile coefficients from ENEDIS over one week, showing residential (`RES1_BASE`) and professional (`PRO1_BASE`) usage types. The values represent normalized hourly consumption patterns.

2.6. *Data Linking Methods*

Integrating the diverse data sources used in this study requires different strategies for associating records at the building level. Since each dataset was developed independently and for distinct purposes, they do not share a common identifier. To overcome this limitation, both identifier-based and geospatial-based methods were employed. The main approaches used to associate data across sources and to align them with the physical infrastructure of the distribution network are listed below :

— **ID-Based Linking via BAN-ID and Address Resolution :** One method relies on the *Base Adresse Nationale* (BAN), a national reference system for French addresses. An API provided by the BAN allows querying a structured address to obtain a unique identifier, the BAN-ID. This identifier serves as a bridge between address-based datasets, enabling consistent cross-referencing with building-level information in the BDNB, ONB, and ENEDIS datasets.

— **Geospatial-Based Matching Between Datasets :** When unique identifiers are unavailable or unreliable, a spatial matching strategy is applied. Using GIS data, buildings from different datasets are considered equivalent if their geographic footprints are spatially aligned. This approach is particularly useful for reconciling records that describe the same physical building but originate from different administrative or technical sources.

— **Geospatial-Based Assignment of Buildings to Electrical Meshes :** The GIS data provided by the distribution system operator includes the locations of transformers, low-voltage cableless. By overlaying this infrastructure map with the building locations, it is possible to determine which buildings are electrically connected to which mesh. This spatial assignment allows aggregating estimated building-level consumption to the mesh level for comparison with observed values.

To summarize the data linking strategies and illustrate how the various sources interact, Figure 2 presents an overview of the data integration approach. It highlights the key connections between data sources and the methods employed to establish these links across different spatial and informational layers.

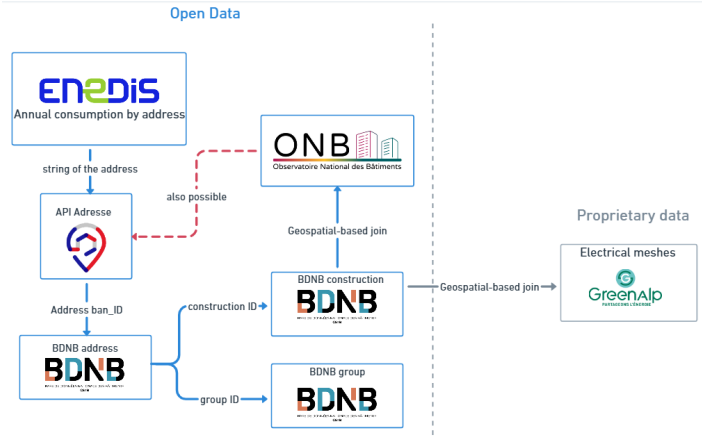

FIG. 2. Overview of the data linking strategy between the various datasets used in this work.

2.7. *Uncertainties and Assumptions*

Despite the richness of the datasets used, several uncertainties and assumptions must be acknowledged. These come

from the limitations of each source as well as the procedures adopted to integrate them. The main points are summarized below :

— **DSO Measurements :** The time series of electricity demand at the mesh level contains missing values for certain periods and meshes, and includes outliers that were removed during preprocessing. Additionally, the assignment of buildings to meshes is not always accurate due to ambiguities in the cable and transformer mapping.

— **Building Features (BDNB and ONB) :** Although both databases often contain similar descriptors for the same buildings, inconsistencies exist between them. Furthermore, the geospatial matching used to link buildings across datasets introduces uncertainty and is not guaranteed to be perfectly precise.

— **ENEDIS — Annual Electricity Consumption :** The data is provided at the address level rather than per building, and its reliability is not guaranteed. Some addresses correspond to multiple buildings, while some buildings are associated with more than one address. Complex cases were excluded from the analysis. Additionally, the dataset only includes addresses with at least 10 dwellings, which may limit representativeness for smaller buildings.

— **ENEDIS — Dynamic Load Profile Coefficients :** These profiles are derived from national-level behavior and may not reflect the specific consumption patterns of Grenoble, particularly with respect to temperature sensitivity. The RES1_BASE and PRO1_BASE profiles were chosen for their simplicity, but they may not be the most accurate representations of local demand dynamics.

3. MODELING APPROACH

This section describes the methodology used to estimate and reconstruct the electricity demand of buildings in Grenoble. The process involves several steps, from estimating annual consumption based on building characteristics to reconstructing detailed load profiles. Figure 3 provides a high-level overview of the main stages of the approach.

Each step is explained in detail in the following sections.

3.1. Estimation of Annual Consumption at the Building Level

This step aims to estimate the annual electricity consumption of each building based on its physical and usage characteristics. This information is not systematically available for all buildings in the study area, particularly within the city of Grenoble. However, annual consumption data can be obtained for a subset of buildings through the ENEDIS dataset (consumption by address) and the ONB dataset, as previously described. These records serve as the target for training a predictive model.

To construct the training dataset, features from multiple sources were merged, including those from the BDNB and ONB, with the corresponding annual consumption values as targets. The training was restricted to buildings located in departments surrounding Grenoble (Isère — 38, Ain — 01, and Rhône — 69). In total, more than 20,000 buildings were used for model development.

An XGBoost regression model was selected due to its strong performance on structured data and its ability to automatically handle heterogeneous feature sets and non-linear relationships. It is also particularly well suited for feature importance evaluation and robustness in the presence of multicollinearity.

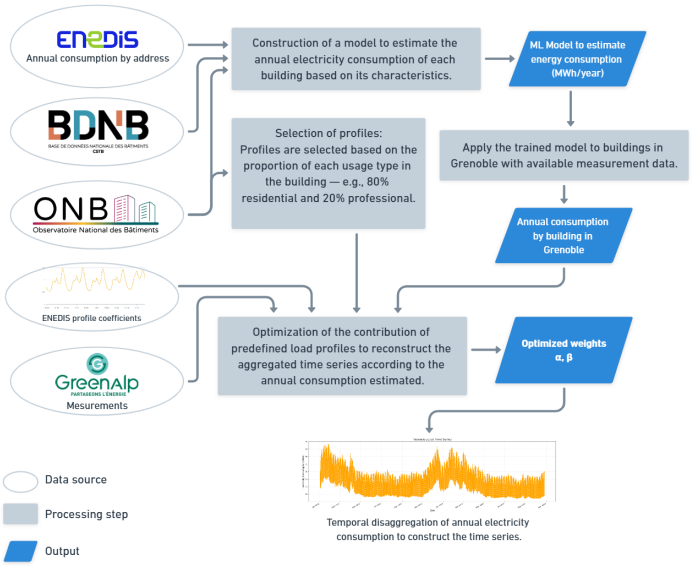

FIG. 3. Overview of the modeling approach used to estimate and reconstruct electricity demand. The diagram illustrates data sources, model training, load profile optimization, and the reconstruction of the time series.

Several variables extracted from the various data sources were found to be particularly influential in predicting annual electricity consumption. As shown in Figure 4, many of the most important features originate from the DPE tables, which provide detailed information on building characteristics, energy systems, and efficiency indicators. These top-ranking features reflect structural, temporal, and usage-related aspects of the buildings and exhibited strong predictive power in preliminary model evaluations.

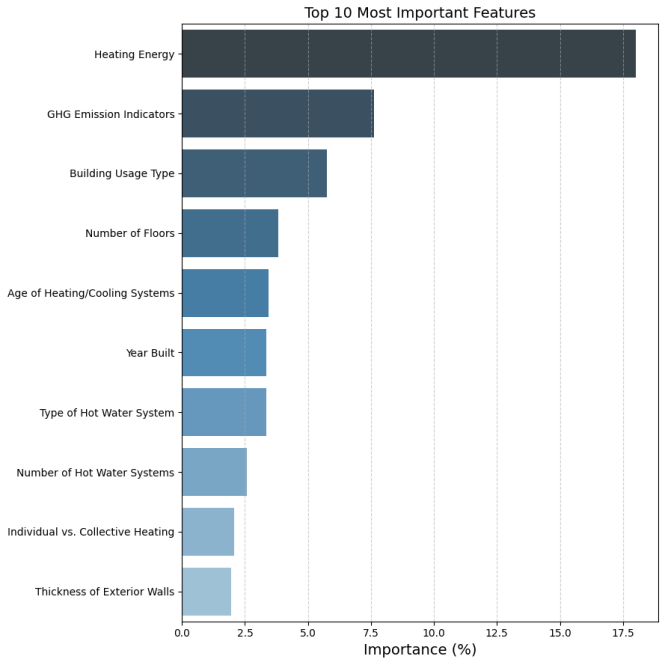

FIG. 4. Top 10 most important features for predicting annual electricity consumption, based on the trained model. Feature importance was aggregated across related variables from different data sources.

The trained model is later applied to buildings in Grenoble where direct consumption data is missing, enabling a complete estimation of annual energy demand across the urban area.

3.2. *Selection of Buildings for Simulation in Grenoble*

To apply the proposed approach within the city of Grenoble, a subset of buildings was selected based on the availability of corresponding measurement data from the distribution network. In the selection process, industrial buildings were avoided, as their consumption behaviors and profiles differ significantly from residential and tertiary buildings.

Figure 5 displays the buildings retained for simulation. These were chosen based on the availability and reliability of aggregated electrical measurements at the local scale, ensuring sufficient data coverage for model evaluation.

The trained model described in the previous section was applied to this selected set in order to estimate the annual electricity consumption for each building.

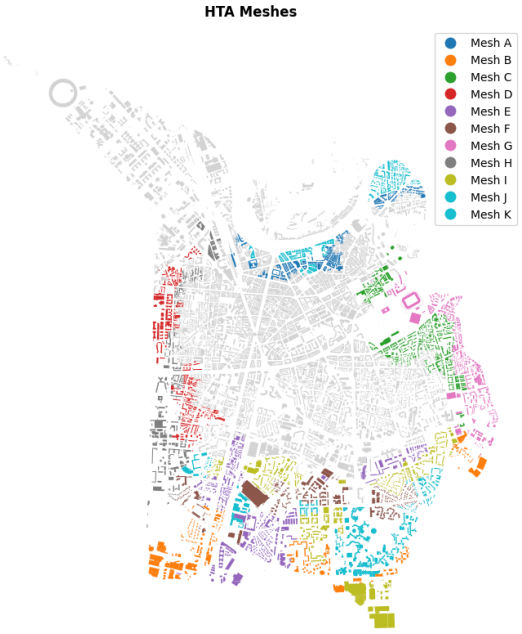

FIG. 5. Map of Grenoble showing the buildings selected for simulation based on measurement availability and data quality constraints.

3.3. *Assigning Usage Type Based on Area Distribution*

To better characterize the electricity consumption profiles of each building, proportions of residential and professional usage were assigned based on the building's usage classification. This classification is derived from the ONB dataset's *typobati* attribute, which indicates whether a building is used for *habitation*, *activité*, or a combination of both (*habitation et activité*).

Figure 6 presents examples illustrating these usage types.

The attribution rules for consumption based on the *typobati* value are as follows :

Habitation : 100% of the building's consumption is assigned to the residential profile.

Activité : 100% of the building's consumption is assigned to the professional profile.

Habitation et activité : A proportion of the area equivalent to one floor (representing the ground floor) is assumed to be professional profile, and the remaining area is assigned to the residential profile.

3.4. *Optimization of Load Profile Contributions*

To reconstruct the aggregated electricity consumption time series, the contributions of predefined dynamic load profiles

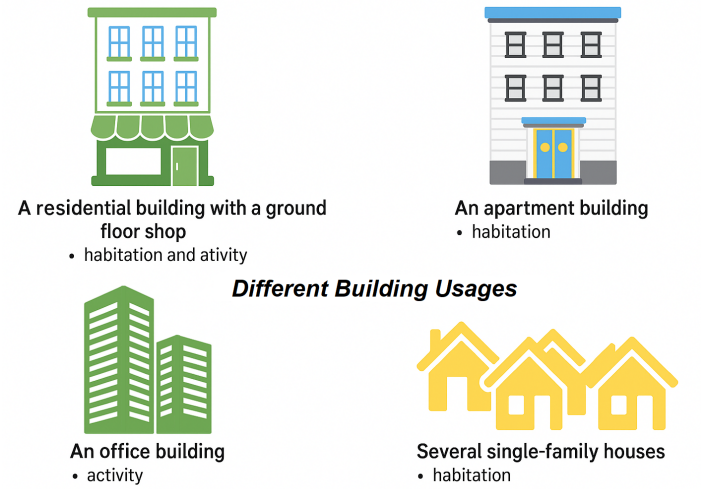

FIG. 6. Examples of building usage types according to the *typobati* classification.

are optimized based on the estimated annual consumption of each building. The modeled consumption at time t, denoted $\hat{C}(t)$, is given by the equation :

$$\hat{C}(t) = \sum_{i=1}^{N} \hat{E}_i \cdot (\alpha \cdot A_i \cdot D_{\text{pro}}(t) + \beta \cdot B_i \cdot D_{\text{res}}(t))$$

where α and β are scaling factors applied to the professional and residential dynamic load profiles, respectively. These coefficients are optimized to minimize the mean absolute percentage error (MAPE) between the measured consumption $C(t)$ and the estimated consumption $\hat{C}(t, \alpha, \beta)$:

$$\min_{\alpha,\beta} \text{MAPE}\left(C(t), \hat{C}(t, \alpha, \beta)\right)$$

The optimization was performed using the Tree-structured Parzen Estimator (TPE), a Bayesian optimization method. The calibration used data from 2020 and 2021, while evaluation was done on data from 2022 and 2023. Table 1 summarizes the notation used in this formulation.

4. RESULTS

This section presents the outcomes from both modeling stages : the prediction of annual electricity consumption at the building level and the reconstruction of the aggregated time series.

4.1. *Annual Electricity Consumption Model*

The performance of the annual consumption prediction model was evaluated on a held-out test set comprising 20% of the available buildings. Table 2 summarizes the main performance metrics, including MAE, RMSE, MAPE, and the coefficient of determination R^2.

To further explore model performance across different building categories, Figure 7 presents the RMSE broken down by building type as defined in the ONB_typobati classification. The figure highlights that prediction errors are highest for buildings categorized primarily as "activite," suggesting a greater heterogeneity or complexity in their energy usage patterns.

Figure 8 shows the distribution of prediction errors across main building usage types. The error is calculated as the difference between the predicted and actual annual consumption

TABLE 1. Notation used in the load profile optimization

Symbol	Description
$\hat{C}(t)$	Estimated total electricity consumption at time t, aggregated over all buildings
$C(t)$	Measured consumption of the aggregated buildings
N	Total number of buildings in the dataset
$\hat{E}_i$	Annual predicted energy consumption of building i
A_i	Share of building i's area attributed to professional (activité)
B_i	Share of building i's area attributed to residential
$D_{\text{pro}}(t)$	Enedis dynamic consumption profile for professional buildings at time t
$D_{\text{res}}(t)$	Enedis dynamic consumption profile for residential buildings at time t
α	Scaling factor for professional profile contribution (optimized)
β	Scaling factor for residential profile contribution (optimized)

TABLE 2. Test set performance of the annual consumption model. MAE = Mean Absolute Error, RMSE = Root Mean Squared Error, MAPE = Mean Absolute Percentage Error, R^2 = Coefficient of Determination.

Metric	Value
MAE (MWh)	13.39
RMSE (MWh)	22.90
MAPE	8.06%
R^2	0.69

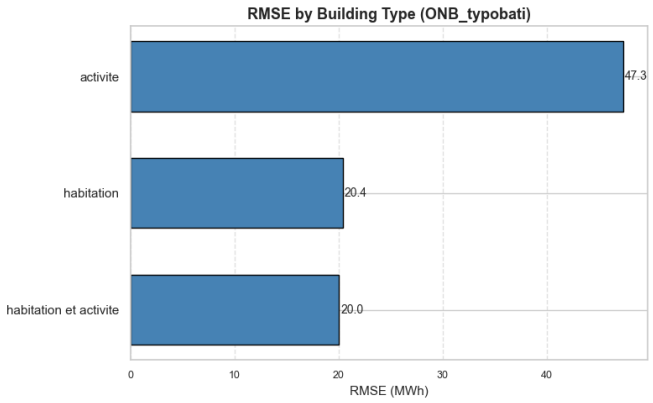

FIG. 7. Root Mean Squared Error (RMSE) of annual electricity consumption predictions across building types (ONB_typobati classification).

(predicted – real), expressed in MWh. The results are grouped into four categories : Residential, Secondary, Tertiary, and Outbuilding.

Residential buildings display a relatively narrow error range, indicating more consistent predictions, which may be attributed to their more uniform consumption patterns. On the other hand, tertiary and secondary buildings show wider and more skewed error distributions, reflecting greater variability in usage or the challenges in capturing their operational complexity with the current feature set. Outbuildings present a moderate error spread, likely due to their diverse functions and generally lower consumption levels. These differences

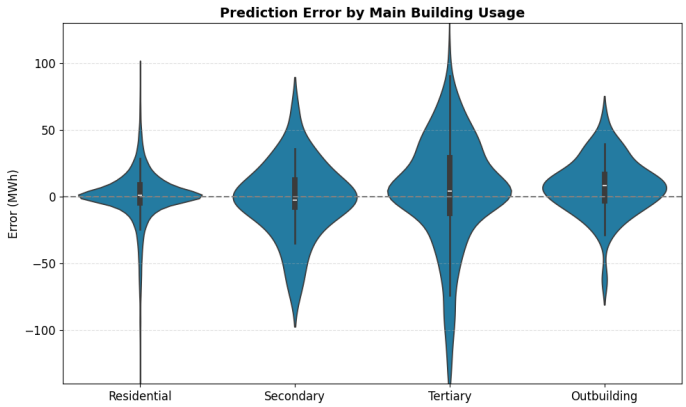

FIG. 8. Distribution of annual electricity consumption prediction errors (in MWh) across main building usage types : Residential, Secondary, Tertiary, and Outbuilding.

suggest that considering building usage in the modeling process is important, and that including more detailed or specific features could help improve prediction accuracy for some building types.

4.2. *Reconstruction of Aggregated Load Curve*

For the reconstruction of the aggregated electricity demand time series during 2022–2023, the performance of the optimized load profile contributions is summarized in Table 3. The model achieved an R^2 of 0.7341, with optimized scaling factors for the residential and professional profiles.

TABLE 3. Performance of the aggregated load curve reconstruction (2022–2023).

Metric	Value
MAE (MWh)	2.31
RMSE (MWh)	8.25
MAPE	9.85%
R^2	0.7341

Parameter	Optimized Value
α (professional)	4.79×10^{-4}
β (residential)	3.61×10^{-5}

4.3. *Qualitative Comparison and Discussion*

To further assess the model's performance, Figure 9 compares the average daily load profiles between measured and simulated values across three seasons : winter, summer, and intermediate. The simulated signal captures the general shape and timing of daily variations across all seasons.

Figure 10 shows the full simulated and measured aggregated time series for 2022 and 2023. The model successfully tracks the seasonal dynamics, although some deviations appear, particularly during winter peaks.

The seasonal comparison confirms a generally good fit, though winter presents the most notable discrepancies. A likely explanation is that the national consumption profiles used may not fully reflect local temperature-dependent consumption behavior in Grenoble. Introducing alternative or localized profiles could help address this issue.

The annual consumption model tended to overestimate the energy use of buildings in Grenoble. Although this bias was partially mitigated through the optimization of α and β, further refinement of the annual model is necessary. One factor that might contribute to this overestimation is the presence of individual residences, which represent around 18% of the se-

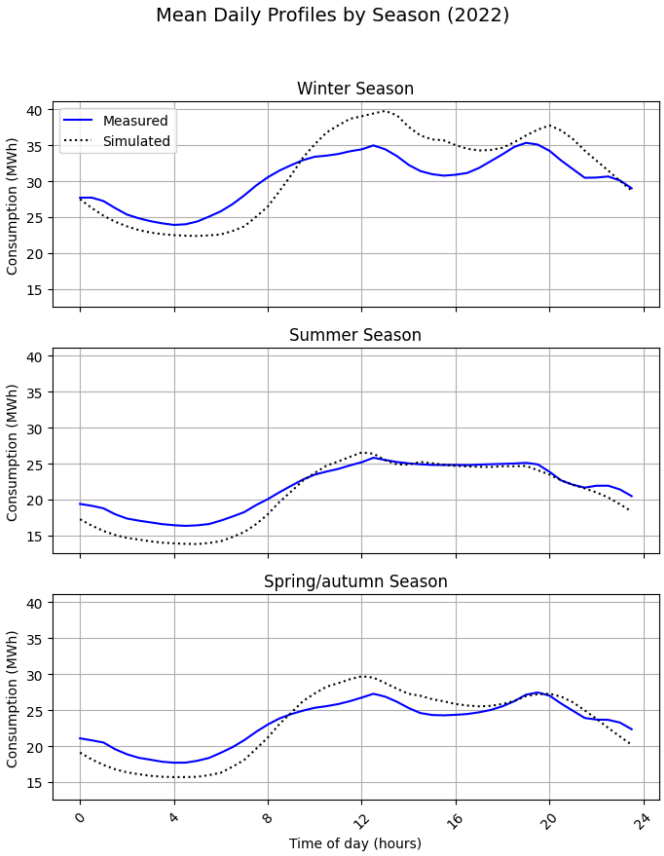

FIG. 9. Average daily consumption by season : comparison between simulated and measured values.

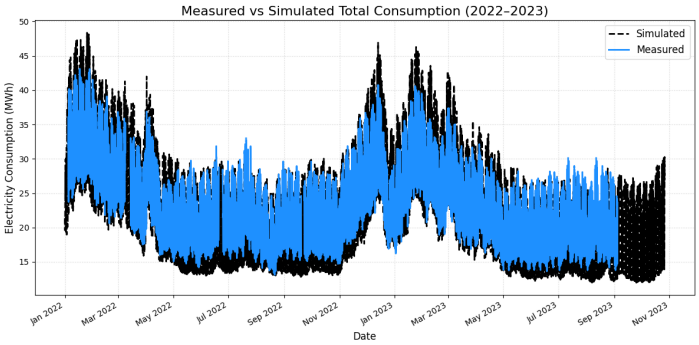

FIG. 10. Measured vs. simulated aggregated consumption during the test period (2022–2023).

lected buildings but are underrepresented in the training dataset. These buildings often have distinct consumption patterns and physical characteristics that differ from larger collective buildings. Enhancing the treatment of data uncertainty, particularly related to building type classification and representation, could significantly improve model performance.

Finally, introducing additional parameters or seasonal segmentation—e.g., distinct α and β values per season—might better capture seasonal variations in behavior and improve temporal disaggregation accuracy.

5. CONCLUSION

This study developed a method to estimate and simulate electricity consumption at the building level by combining annual consumption predictions with dynamic load profiles. Despite relying on partial and heterogeneous data sources, the approach successfully reconstructed aggregated demand with reasonable accuracy, achieving a test period MAPE of 9.85% and an R^2 of 0.73.

Seasonal analysis revealed that the model generally captures consumption patterns well but exhibits notable discrepancies during winter peaks. These differences likely stem from the use of national consumption profiles that may not fully reflect Grenoble's local, temperature-dependent usage. Introducing alternative or localized dynamic profiles, along with seasonal segmentation of model parameters—such as distinct scaling coefficients per season—could better represent temporal variations and improve accuracy in disaggregation.

Moreover, the underrepresentation of individual residences in the training dataset, which account for approximately 18% of the building sample, appears to contribute to overestimation. These buildings often have unique consumption behaviors and physical characteristics that differ significantly from larger collective buildings. Enhancing the treatment of data uncertainty, particularly in building type classification and representation, is therefore critical to improving model robustness.

Additionally, exploring more refined estimation techniques for usage shares and better handling of data heterogeneity will help reduce prediction errors and increase model reliability. These efforts, combined with the integration of local climatic and behavioral factors, will facilitate a more accurate and context-sensitive simulation of urban electricity demand.

Overall, the proposed methodology presents a scalable framework for simulating electricity consumption at a fine spatial scale using open data sources. Its continued development promises to provide valuable support for local energy planning and management, enabling better anticipation of demand patterns and more effective integration of renewable energy resources.

6. ACKNOWLEDGMENTS

This work is part of the Fine4Cast project, funded by France 2030 (Grant No : ANR22-PETA-0008). The authors would like to thank GreenAlp, particularly Damien Fresier and Vincent Martin, for providing the energy consumption data essential to this study.

7. REFERENCES

[1] M. C. Ruiz-Abellón, L. A. Fernández-Jiménez, A. Guillamón, A. Falces, A. García-Garre, and A. Gabaldón, "Integration of Demand Response and Short-Term Forecasting for the Management of Prosumers' Demand and Generation," Energies, vol. 13, no. 1, p. 11, 2019.

[2] Kontokosta, C. E., & Tull, C. (2017). A data-driven predictive model of city-scale energy use in buildings. Applied energy, 197, 303-317.

[3] González-Vidal, A., Ramallo-González, A. P., Terroso-Saenz, F., & Skarmeta, A. (2017, December). Data driven modeling for energy consumption prediction in smart buildings. In 2017 IEEE International Conference on Big Data (Big Data) (pp. 4562-4569). IEEE.

[4] A. Alhamwi, W. Medjroubi, T. Vogt, and C. Agert, "Modelling urban energy requirements using open source data and models," Applied Energy, vol. 231, pp. 1100–1108, Dec. 2018. [Online]. Available : https ://doi.org/10.1016/j.apenergy.2018.09.170.