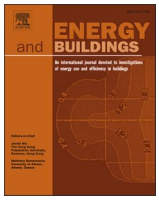
Energy & Buildings 345 (2025) 116033

Contents lists available at ScienceDirect

Energy & Buildings

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



# Temperature-time evaluation as a tool for 'future-proofing' urban building energy modelling (UBEM)

Andrew Marian Zajch [a,*], Yohei Yamaguchi [a,*], Keita Shono [a], Tomoki Shigematsu [a], Hideaki Uchida [a], Tsuyoshi Ueno [b], Yoshiyuki Shimoda [a]

[a] Division of Sustainable Energy and Environmental Engineering, Graduate School of Engineering, The University of Osaka, 2-1 Yamada-oka, Suita, Osaka 565-0871, Japan
[b] Central Research Institute of Electric Power Industry, 2-6-1 Nagasaka, Yokosuka-shi, Kanagawa 240-0196, Japan

ARTICLE INFO

*Keywords:*
Residential electricity demand
Bottom-up type simulation model
Electricity demand validation
Future-proof design

ABSTRACT

Urban Building Energy Modelling (UBEM) application has become a powerful tool for stakeholders to understand the impacts of the building sector and its role in contributing to climate policies through the simulation of future scenarios. Current UBEM validation has not been explicitly tailored to these applications despite the growing demand from model 'consumers', such as policy makers, who may not be aware of UBEM limitations. A model agnostic temperature–time framework was developed and applied to three municipalities in Japan. This identified the influence of temperature-related bias in energy demand estimates by examining the distribution of error across the temperature–time domain as well as comparing the temperature elasticity of modelled energy consumption. All municipalities recorded Coefficient of Variance Root Mean Square Error (CVRMSE) <30 % for both weekday and weekend conditions while only 4 municipal-day type pairs met the Normalized Mean Bias Error (NMBE) criteria of being within ±10 %. Despite meeting these conventional standards, no municipality was able to pass all the temperature–time criteria proposed in the framework highlighting a gap in conventional metrics due to the persistence of temperature–time biases. Notably, no city, and only 1 % of districts, were able to simulate cooling elasticity effectively. The stringent nature of the proposed temperature–time framework suggests it presents an aspirational target for 'future-proof' UBEM and a means for increased model transparency.

## 1. Introduction

Sustainable building design has been relatively preoccupied with reducing Greenhouse gas (GHG) emissions rather than designing buildings resilient against uncertain climate futures [1]. Tradeoffs between climate adaption and mitigation strategies complicates decarbonization policy design typically conducted at the municipal or regional level [2–4]. This can be further complicated by regional variation in energy consumption across as seen in the US (e.g. Wang et al. [5]), Japan (e.g. Honjo & Fujii [6]) and globally (Santamouris et al. [7]) caused by evolving climate, socio-economic and policy conditions. These factors highlight the need for a localizable decarbonization evaluation tool considerate of both building emissions and potential climate impacts on buildings.

Urban Building Energy Modelling (UBEM) is a promising approach to address local, stock-level, carbon reduction strategies [8]. However, Dahlström et al. [9] recognize UBEM has not been widely used to explicitly examine climate change impacts under future scenarios despite the popularization of similar analysis for building stock and building energy models (BEM). Ang et al. [8] emphasize that the UBEM development process, including minimum performance standards, must be cognizant of the desired application of the UBEM. The emerging opportunity for UBEM application to model energy demand for unseen futures scenarios warrants the enhancement of current UBEM approaches. An ideal model for future scenario analysis would perform equally across different scenarios. Extending the idea of a 'future-proof' building resilient to a range of plausible futures, 'future-proof' UBEM's performance would be resilient to a range of projected scenarios [1]. However, conventional UBEM validation methods do not guarantee the development of future-proof models.

This work examined the relationship between estimated energy consumption and temperature-based errors to identify gaps in conventional validation metrics. A new temperature–time framework was proposed for local empirical validation congruent with the UBEM

* Corresponding authors.
*E-mail addresses:* andrew@see.eng.osaka-u.ac.jp (A.M. Zajch), yohei@see.eng.osaka-u.ac.jp (Y. Yamaguchi).

https://doi.org/10.1016/j.enbuild.2025.116033
Received 24 March 2025; Received in revised form 2 June 2025; Accepted 12 June 2025
Available online 15 June 2025

development process for localized decarbonization planning.

### 1.1. Urban building energy modelling (UBEM)

UBEM upscales building energy demand simulations to represent urban level dynamics and interactions often at the cost of individual building model dexterity [10,11]. The UBEM development cycle conducts a process of data aggregation, model compilation, simulation, and validation and calibration tuned to data availability and model purpose [8]. The emergence of smart meter data has granted more spatially and temporally resolved data that can be used as an input (e.g. data-driven modelling leveraging smart meter defined archetypes by [12]) or for providing a baseline to conduct validation (e.g. Kusumoto et al. [13] used it to compare the diurnal patterns of their machine learning approach and observed conditions in Yokohama). Ang et al. [8] subdivided UBEM application into four categories: proactive urban design, energy conservation and decarbonization measure evaluation at both stock and building-levels, and urban energy system analysis considering supply and demand dynamics. The diverse spectrum of applications has unsurprisingly led to a popularization of UBEM use by "consumers", such as policymakers, who are not involved in model development and are unaware the model's inherent limitations to inform decision making [14]. These 'consumers' rely on model validation robustness to confidently use outputs.

While a comprehensive review of UBEM approaches and specific tools is outside the scope of this work it is essential to differentiate UBEM types in the context of their synergy with local decarbonization planning. Fundamentally, urban building energy models can be subdivided into top-down and bottom-up models. Top-down models are inherently incongruent for local decarbonization planning. Top-down models struggle to disaggregate energy behavior and test potential policy solutions as changing socio-economic drivers and technological improvements are often omitted from projections [15,16]. On the other hand, bottom-up models recreate building level dynamics that can be adjusted based on parameterizations and assumptions to test policy scenarios. For example, Papantonis et al. [17] used a bottom-up model to examine decarbonization pathways and the tradeoffs between prescribed policies in Greece to highlight the importance of near-term policy implementation. Ferrando et al. [14] differentiate bottom-up models into data-driven and physics based models, with the latter being further divided into the more popular reduced-order resistor–capacitor (RC) simulation and more detailed dynamic thermal simulation (i.e *Energyplus*). Similar to top-down approaches, data-driven bottom-up models tend to struggle extrapolating energy demand behavior beyond the extent of the historical data making them ill-suited for modelling for uncertain futures [14,18]. Bottom-up physics based urban building energy models are not subject to these limitations, making them the ideal approach for local decarbonization evaluation considering unseen future scenarios.

Bottom-up physics based UBEM, hereafter referred to solely as UBEM unless otherwise specified, provides a high level of detail and granularity at the building scale useful for exploring decarbonization pathways and the tradeoffs between prescribed policies and technology transitions. UBEM has been benefiting recently from the growing availability of data and computing capacity that previously limited their use [14]. Kastner and Dogan [19] showcased the ability of UBEMs to conduct detailed measure evaluation, in their case PV adoption, to support municipal scale decision making in Ithaca, NY, USA. Yamaguchi et al. [20] used UBEM analysis to facilitate a comprehensive feasibility assessment of a net-zero energy system at the municipal level in Japan. However, Ali et al. [21] highlight that while the fine detail of UBEM makes it useful for targeted policy, it's utility can be diminished by the multitude of imbedded assumptions needed to parameterize building dynamics in the absence of building scale data at the municipal or regional level. This emphasizes the need for robust model validation to ensure parameterizations are effective, particularly in areas with poor data coverage.

#### 1.1.1. Challenges with UBEM validation

Validation and calibration, the last step of the UBEM development pipeline, are crucial to ensure UBEM performance is sufficient for its range of users and applications. However, Lefort et al. [22] note there is no standardized form of UBEM validation across the UBEM spectrum with empirical validation heavily relying on historical data availability. The absence of a standardized framework leads to less transparent and comparable models [10]. A review of UBEM by Kong et al. [11] illustrates that UBEM outputs usable for validation tend to focus on energy consumption and energy use intensity at coarse temporal resolutions (i.e. Annual). The Coefficient of Variance Root Mean Square Error (CVRMSE) and Normalized Mean Bias Error (NMBE) co-opted from building energy modelling standards can be used to evaluate these outputs [8]. These provide quantitative assessments of the model's ability to reconstruct energy demand behavior for a validation period for which observed data is available. Model validation relies on an implicit assumption that validating energy consumption empirically will ensure climate-energy dynamics are accurately captured. If this assumption fails, however, a hidden embedded temperature related bias can persist.

Fundamentally, validation metrics have not been designed to explicitly match model applications, such as future scenario analysis. Oraiopoulous and Howard [23] observed a lack novel tailored metrics and validation customized to model purpose. Kong et al [11] noted some progress towards climate focused metrics in the UBEM literature with calibration targets differentiated by heating and cooling regimes representing an intrinsic consideration of climate-energy relationships. For example, Deng et al. [24] demonstrated calibrating their model based on heating intensity rather than aggregate energy demand. Nevertheless, while calibrating models to fit specific circumstances described by the validation data helps improve the model's ability to recreate observed realities, overzealous calibration can lead to over-fitting the model making it less generalizable in other contexts. This is a pitfall for models designed for the purpose of modelling future scenarios in different contexts where validation data provides a limited domain for evaluation.

The range of observable and available electricity demand measurements and corresponding historical outdoor air temperatures constrain model evaluation. This introduces a potential imbedded weakness for bottom-up models relative to top-down approaches. BEM energy demand estimates can be sensitive to the underlying weather data, an issue that extends to UBEM [25]. The constrained extent of weather conditions, particularly the range of outdoor air temperatures, observed during a validation period can present a hurdle to the feedback process of validation and calibration conducted during UBEM development. Top-down UBEM approaches explicitly use outdoor air temperature as an independent variable to model energy demand [26]. This provides an explicit relationship that can be extrapolated using projected 'future' climate conditions in the form of 'future' weather files (e.g. [27]) or by augmenting temperature response functions (TRF) to reflect future scenario conditions (e.g. [28]). Explicit integration of climate dynamics helps circumvent the domain issue by focusing on the climate-energy relationship. The inverse is true for bottom-up physics based UBEM which reconstruct energy demand behavior and, by extension, capture temperature response behavior implicitly. As a result, the ability of bottom-up models to represent energy demand behaviors in response to temperature changes are not as clear as top-down methods.

Insufficient validation resolution further hinders the ability to ensure electricity consumption behaviors are being reflected using UBEM. Hourly building energy demand varies with temperature and time. Local outdoor air temperature dictates heating and cooling demand. Occupancy and activity schedules of residential building occupants cause diurnal variations in energy demand. Model validation needs to assess whether all these underlying patterns are recreated. However, aggregated validation results indicate overall performance of the model but

they do not necessarily pinpoint circumstances where the model performs more poorly [23]. A shift in climate conditions to historically low frequency, extreme, temperatures can exacerbate errors obscured at typical validation aggregation levels such as months or years. Even disaggregation at the monthly or weekly time scales contain intrinsic assumptions about the consistency of weather conditions affecting these months and the historical continuity of heating and cooling periods. The reliability of these assumptions is decreasing as months shift away from their typical conditions due to various scales of climate change. Ultimately, low temporal resolutions introduce a challenge for UBEM validation with a focus on recreating climate-energy relationships.

#### 1.1.2. Towards a 'future-proof' UBEM

A 'future-proof' model should not exhibit heightened errors at certain external air temperatures. It also should be able to capture the building's response to changing external air temperatures. Hekkenberg et al. [15] demonstrated that mischaracterizing the temperature-electricity demand relationship will lead to under or over estimation of future electricity demands. Temperature elasticity describes the underlying relationship between climate and energy. It quantifies the change in electricity demand in response to a unit change in outdoor air temperature. Hu et al. [29] showed temperature elasticity varies nationally or even regionally in response to occupant activity and preferences as well as building and heating/cooling system conditions. Both these factors play a dynamic role in how buildings respond to external air temperatures. A case study of temperature elasticity at the district level in Tokyo revealed temporal (evening-day, weekend-weekday) and spatial differences (urban-suburban) in temperature elasticity [30]. Adequately describing the response of a building stock to climate changes requires validation at hourly temporal resolutions to ensure elasticity is represented. Achieving a 'future-proof' model necessitates a validation resolution capable of representing and investigating variations in temperature–time.

#### 1.1.3. Scope and contribution

The current body of literature suggests there is an inherent gap in the validation of bottom-up UBEM due to limited customization and standardization of UBEM validation, temporally constrained historical validation data and the implicit description of climate-energy relationships. Fundamentally, the relationship between conventional validation approaches and temperature-based errors requires better understanding to ensure that temperature-based biases do not persist in conventionally validated models. These biases, while less impactful for historical conditions, can potentially manifest during more extreme future projected climate conditions. Hourly residential electricity consumption simulated using a bottom-up physics model was validated for three case study cities in Japan. This study proposed a novel validation framework for 'future-proof' UBEM. It provided a unique comparison of conventional evaluation methods against the newly proposed tools designed to explicitly examine temperature–time biases and model behavior. This recognized the inability of conventional methods to ensure 'future-proof' models. The benefits of this framework were further demonstrated through a model calibration exercise guided by the new approach.

## 2. Method

Fig. 1 presents the workflow and the novel temperature–time framework presented in this study. The framework (Step 6- Section 2.2) utilizes the same inputs as conventional validation (Step 5- Section 2.1) relying on both simulated (Bottom-up physics-based model −Section 2.3) and measured (smart meter data from case study area, Section 2.4) datasets.

The workflow allowed for a case study application of the temperature–time validation framework to contrast against conventional validation of electricity demand at the local-hourly scale. District scale smart meter data alongside corresponding demographic, temperature and building floor areas were compiled (Step 1). Demographic and temperature data were used in a bottom-up physics-based simulation of the residential electricity demand (Step 2). This simulated data and the measured smart meter data was filtered to include only residential building dominated districts (Step 3) subsequently aggregated at both the municipal and district scales (Step 4). Validation was done at both these scales using the conventional (Step 5) and newly proposed temperature–time (Step 6) validation methods. Juxtaposing, the results of these two evaluations revealed whether conventional approaches obscured model weaknesses identified through temperature–time validation. Finally, a calibration exercise informed by the new temperature–time framework demonstrated the benefits of using the proposed approach.

### 2.1. Conventional model validation

Building energy demand model validation typically relies on calculating the CVRMSE (Eq.1) and the NMBE (Eq. 2) established as industry standards by associations such as ASHRAE [26]. These methods compare the measured (*m*) and simulated (*s*) energy demand for every hour (*i*) in the set of hours within the year or month (*n*). The results were scaled based on the mean measured energy demand ($\overline{m}$). The *p* value was set to 1 [31]. The primary difference between these measures is the range of values. Unlike CVRMSE, NMBE can be both positive and negative reflecting under or overestimation at the cost of being susceptible to cancellation of errors [31].

$$CV(RMSE) = \frac{1}{\overline{m}}\sqrt{\frac{\sum_{i=1}^{n}(m_i - s_i)^2}{n-p}} \tag{1}$$

$$NMBE = \frac{1}{\overline{m}}\frac{\sum_{i=1}^{n}(m_i - s_i)}{n-p} \tag{2}$$

ASHRAE Guideline 14 specifies modelled results should achieve a CVRMSE <30 % and an NMBE < ±10 % for hourly energy data in line roughly with other standards such as the Federal Energy Management Program (FEMP) and the International Performance Measurement and Verification Protocol (IPMVP) [26,31]. These conventional quantitative thresholds assess the validity of the model.

### 2.2. Temperature-time framework

#### 2.2.1. Temperature-time bias

An explicit temperature–time validation framework enabled validation at distinct temperature and times. Fig. 2 illustrates how hourly simulated and measured electricity consumption data was binned by temperature, set at 1 °C interval bins, and time, based on a 24-hour cycle. The CVRMSE and NMBE were calculated using the hourly data assigned to each temperature–time bin. These form the basis of temperature–time (T-T) plots. A single T-T plot describes the timing and temperature bias of errors across the entire year. This provides a streamlined visualization compared to standard approaches using an increasing number of plots to contrast simulated and modelled hourly profiles for a day, week, or month. The occurrence of high CVRMSE or NMBE values identify temperatures and/or times where the model performs poorly. The T-T plots can help identify more complex dynamics that are a combination of temperature and time. For example, Wang and Bielicki [32] showed a lagged response to temperature when occupants are not actively controlling cooling systems. Larger errors for high temperatures overnight would suggest an underrepresentation of such dynamics. Vertical bands of high magnitude CVRMSE or NMBE, indicating poor mode performance, represent time dependent or temperature independent error. Time dependent errors suggest examining occupant activity parameterization due to noticeable diurnal patterns in occupant schedules. Horizontal high magnitude bands indicate a

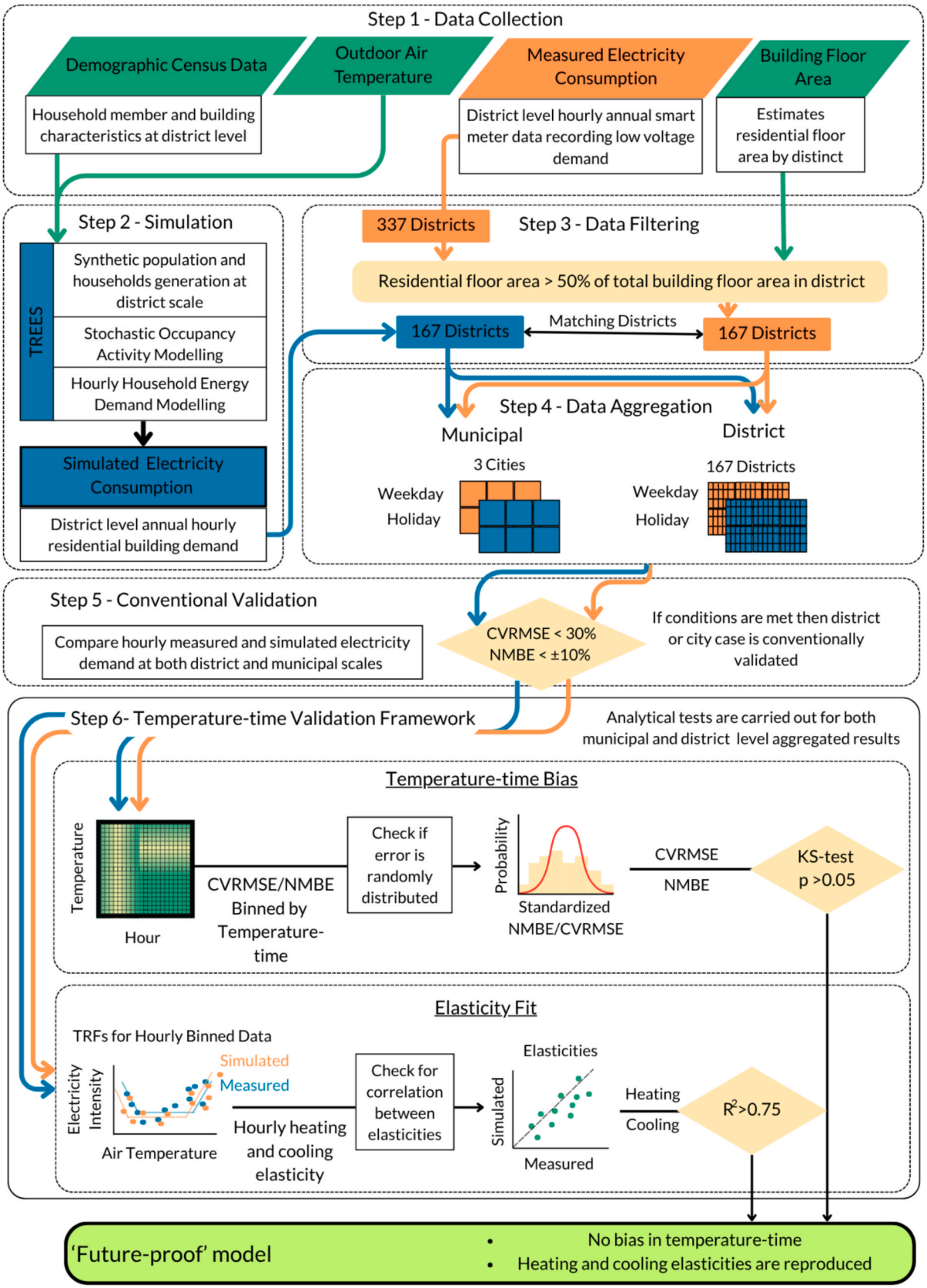



**Fig. 1.** Workflow (Steps 1–6) describing study specific data preparation (Steps 1–4) carried out before context agnostic conventional (Steps 5) and temperature–time validation (Step 6) frameworks. The use of data inputs (simulated-blue, measured-orange) as well as data sources (outdoor air temperature, building floor area, and census household demographic data-green) were delineated by corresponding-colored arrows. Decision points (yellow diamonds) are included to identify criteria for model validation. KS-test refers to the Kolmogorov-Smirnov Test applied to standardized CVRMSE and NMBE while the $R^2$ is the coefficient of determination based on simulated and measured elasticities. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

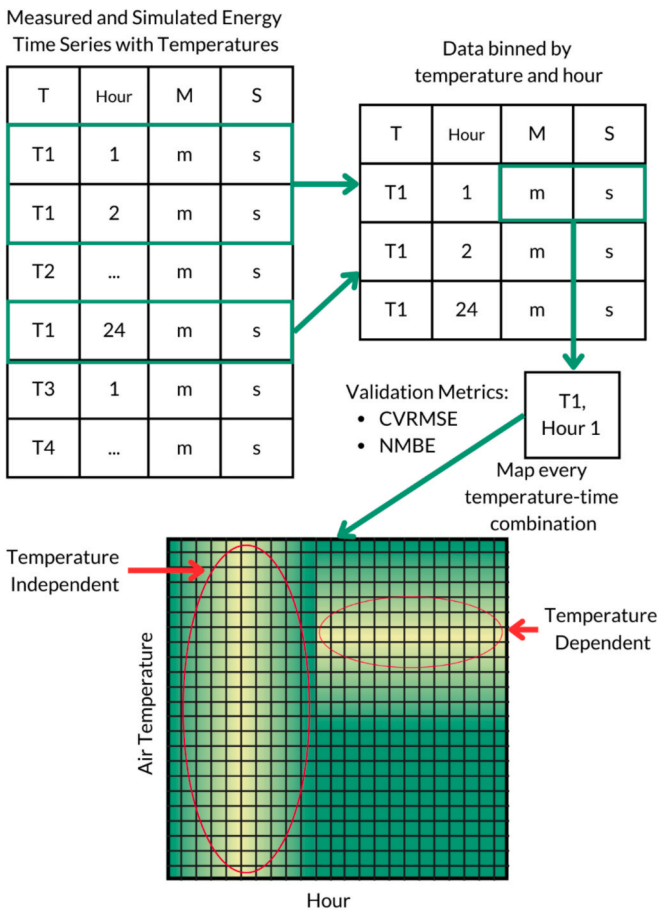



**Fig. 2.** Conversion of hourly annual data by temperature–time bin assignment and aggregation to calculate CVRMSE or NMBE metrics for each temperature–time combination for measured (M) and simulated (S) results. Temperatures (T) are assigned to an interval bin by rounding to the nearest integer value. The results are mapped to a temperature–time grid that can identify temperature independent and temperature dependent error regions in temperature–time.

temperature-bias attributable to systemic issues with the model's heating and cooling parameterization. The temperature–time plot of conventional validation metrics grants a qualitative diagnostic tool for identifying model strengths and weaknesses. The latter can then inform subsequent focused calibration and model improvement.

While visual analysis of the temperature–time plot can help identify potential sources of error, users of the approach would benefit from understanding the significance of high error regions of the temperature–time plot. An ideal temperature–time plot would observe randomly distributed CVRMSE and NMBE binned by temperature and time. This would reflect no bias in temperature and time. The Kolmogorov-Smirnov (KS) test (implemented in *Scipy* [33]) provides a statistical test to compare distributions. The KS-test compared the CVRMSE and NMBE values binned by temperature–time against an assumed random distribution of values. Applying the KS-test in this way gauges whether model errors are randomly distributed in temperature–time. Temperature-time bins were omitted from the analysis if they represented less than 2 % of hourly samples. This was done to minimize the impact of low frequency T-T bins while avoiding over-aggressive filtering of T-T bins for subsequent analysis. Standardization was used on the remaining results to center and scale the data to compare against a common normal distribution representative of the ideal, random, distribution of errors. Values lower than the pre-defined p-value (0.05) suggest rejecting the null hypothesis that the distribution of CVRMSE or NMBE follows the normal distribution. Conversely, datasets where this is not the case provide confidence that the CVRMSE or NMBE distribution are effectively random and do not contain any bias. In this case, the dataset passed the temperature–time bias test.

### 2.2.2. *Elasticity fit*

A second layer of temperature-based validation compared simulated and measured temperature electricity demand elasticities. This was done using temperature response functions (TRF). TRFs describe electricity consumption as a function of outdoor air temperature intrinsically describing socio-economic conditions, thermal comfort preferences and system efficiencies [15]. They differentiate electricity consumption between heating and cooling temperature domains which tend to observe negative and positive proportionality with air temperatures, respectively. TRF breakpoints delineate where electricity behavior transitions between heating and cooling regimes. Chen et al. [34] used a single threshold, or stationary point temperature (SPT), for the entire temperature domain due to the dominance of cooling and popularity of natural gas heating in their case study area, California, suggesting a V-shaped TRF. Choi and Yoon [35] explicitly defined the stationarity point temperature as the boundary where the minimum electricity demand, or base load, occurs. Similarly, modelling done for the Tokyo area using a segmented model assigned both heating and cooling thresholds separated by a zone of relatively temperature independent energy use representing a zone of thermal comfort (e.g. [30,36]). More robust Multivariate Adaptative Regression Splines (MARS) have also been used to model TRFs capable of representing non-linear behavior and integrating a range of meteorological information [37]. However, the success and interpretability of a heating, comfort and cooling segment model motivated the use of a three-segment linear model (Fig. 3).

A three-segment TRF model was developed to estimate heating and cooling elasticities for each hour. TRFs were generated for each hour to reduce the impact of differences in activity throughout the diurnal cycle which can manifest as different temperature response functions [37]. Equation 3 presents the TRF fitted assuming two distinct breakpoints for electricity consumption intensity (E, W/m$^2$). The electricity intensity was defined using aggregate electricity consumption scaled by the total floor area estimated at the validation scale, either municipal or district level. The maximum ($t_{max}$) and minimum($t_{min}$) temperatures observed within the annual dataset are set as boundaries of the segmented linear regression. Slopes ($s_H$, $s_F$, and $s_C$) are defined for each linear segment

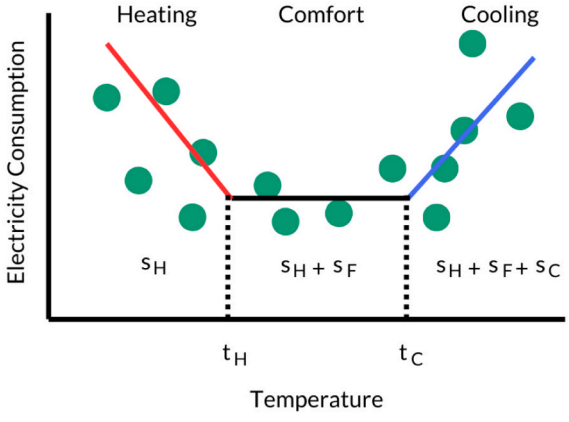



**Fig. 3.** A conceptual schematic of the TRF used in this work based on three-segments (heating, comfort and cooling) defined by heating ($t_H$) and cooling ($t_C$) thresholds. The slope of each segment is based on a superimposition of parameters defined for the slope of previous segments. $S_H$ describes the slope for heating, while $s_F$ is the added parameter for defining the comfort segment. $S_C$ the added parameter for the cooling segment that defines the cooling segment slope described by the sum of all parameters, $s_H$, $s_F$, and $s_C$.

with new terms being added when surpassing a breakpoint to ensure continuity [38]. The slope ($s_H$) of the linear function when temperatures were below the heating threshold temperature ($t_H$) defined heating elasticity. The sum of $s_H$, $s_F$, and $s_C$, or the slope of the function when temperatures exceed the cooling threshold temperature ($t_C$) represented the cooling elasticity. The three-segment linear model simplifies to a similar form typically used; however, this approach does not assume temperature elasticity to be zero in the thermal comfort region ($\geq t_H$ and $< t_C$).

$$E = \begin{cases} E_0 + s_H(T - t_{min}) & t_{min} \leq T < t_H \\ E_0 + s_H(T - t_{min}) + s_F(T - t_H) & t_H \leq T < t_C \\ E_0 + s_H(T - t_{min}) + s_F(T - t_H) + s_C(T - t_C) & t_C \leq T \leq t_{max} \end{cases} \quad (3)$$

The *pwlf* python package was used to determine the breakpoints and slopes of three segment temperature response functions [38]. This package uses a double loop optimization algorithm where break points are optimized based on a least square exploration of segment slopes for the given specified number of segments. The breakpoints were limited to 15–20 °C for heating ($t_H$) and 20–25 °C for cooling ($t_C$) in line with empirical studies and standards identified (e.g. [29,35,36]). Using a smaller range, rather than the entire possible domain where the heating threshold temperature is lower than the cooling threshold temperature ($t_H < t_c$) spurred faster model convergence. It also ensures breakpoints agreed with observed and understood temperature response behavior. Optimizing breakpoint values for a given hour avoided imposing a fixed threshold for the entire diurnal cycle and instead sought to replicate the variability of breakpoint temperatures throughout the diurnal cycle due to acclimatization [32]. Defining TRF slopes on an hourly basis described the diurnal variation of temperature elasticities reflective of temperature-based behavior.

Simulated and measured hourly temperature elasticities for heating and cooling were compared using $R^2$ and CVRMSE. A low $R^2$ would represent a poor representation of the variation in temperature elasticity using the model. An $R^2$ threshold of 0.75 can provide a quantitative test to validate the model's temperature elasticity [31]. This threshold served as the primary determinant of the simulation's ability to capture measured temperature elasticity. This provides additional validation of the model's intrinsic climate-energy relationship.

### 2.3. Bottom-up building energy demand modelling

The bottom-up Total Residential End-use Energy Simulation (TREES) model has been used previously to evaluate residential decarbonization policy by Shimoda et al. [39] and Taniguchi-Matsuoka et al. [40] as well as more comprehensive energy system analysis by Yamaguchi et al. [20]. Fig. 4 outlines how the model uses census district level population information to generate a synthetic population reflective of demographics at the local level. However, prefectural level data was used to inform building construction periods since district level data was not available. The household and building characteristics are then used to predict detailed information on building systems and behavior that is not available at the household level, such as appliances and occupant behavior. A person-based stochastic occupant behavior model is used to generate profiles of household members' occupancy and activity status at 5-min resolution. The model employs a stochastic discrete-event approach to predict the occurrence and sequence of activities characterized by several activity modeling parameters and probability distributions modeled by logistic regression models developed based on Japanese time use survey data considering household member demographics [20,41]. This allows for an estimation of household activities, and resultant appliance use, based on demographic lifestyles. Home appliance operation schedules correspond to occupant activity schedules, for example water heater use is triggered by occupant bathing. These interactions are estimated using probabilistic relationships between appliance operation and occupant activity. System appliance types, for example gas vs electric water heating, are assigned using a logit model relying on household demographic, building geometry and regional characteristics [40]. Aggregating appliance energy use modelled in the building physics model generate temporally and spatially evolved estimates of residential energy demand differentiated by fuel types. The electricity demand can then be extracted to compare against measured electricity demand.

Weather station data informed two model modules impacted by outer air temperatures. A dynamic thermal load simulation relying on a thermal network model estimated heating and cooling energy demand at 5-min intervals [42]. This approach integrated outdoor air temperature alongside building characteristics, and occupant behavior and preset temperatures to determine the operational schedule. Similarly,

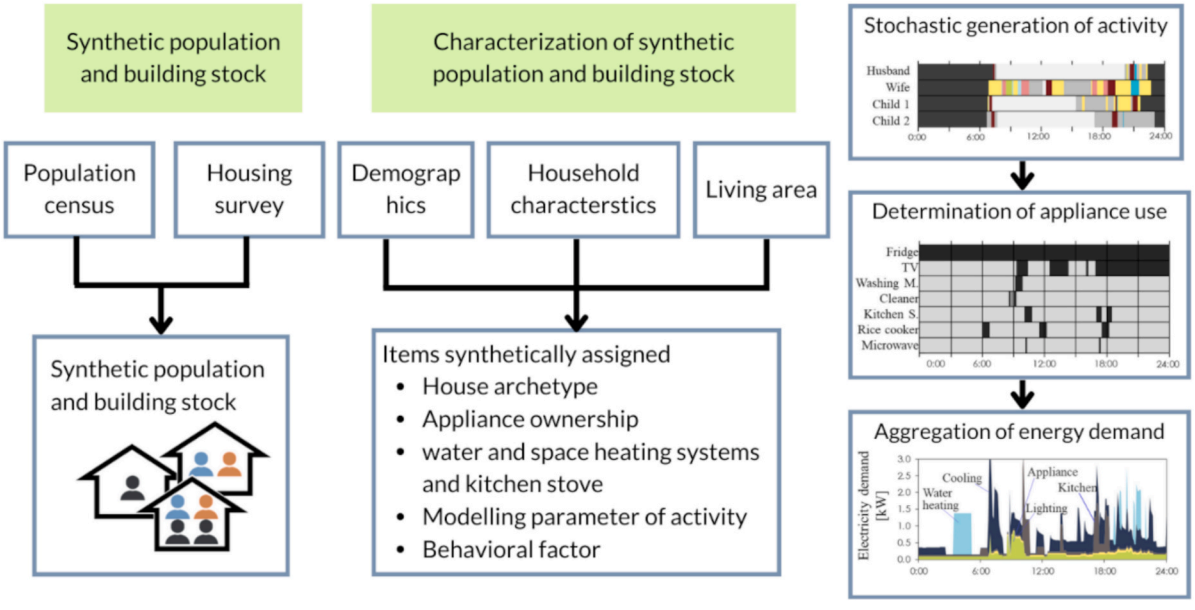



**Fig. 4.** Framework illustrating the data (left) used to generate the synthetic population and building stock and parameters synthetically assigned (middle). Synthetic data was generated for each district based on district-level census data. This information defines occupant activity patterns and subsequently the appliance use schedules (right) used to inform aggregated energy demand at high temporal resolutions.

water heater energy demand was based on occupant bathing schedules and the city water temperature, with the latter being influenced by outdoor air temperature. The intricacies of the imbedded modules within the TREES model are outside the focus of this work intended to be model agnostic. Readers interested in the details of the model's parameterizations are directed to the several case studies conducted using the TREES model such as the work previously highlighted by Taniguchi-Matsuoka et al. [40].

This work presents for the first time validation results for the TREES model at the hourly and district level providing new insights into the localization possible with the bottom-up approach. The model validated in this work was identical to the one presented by Yamaguchi et al. [20] albeit for the addition of a miscellaneous electric load (MELs) parameter used in previous applications of the model by Taniguchi et al. [43] approximated as 70 W/household. Butzbaugh et al. [44] identified that MELs encompass a wide range of technologies and energy consumption behavior, such as entertainment and elective kitchen appliances, that can have a material impact on energy consumption. For example, the 36 of the best understood MELs contribute an estimated 12 % of residential and commercial building consumption in the United States of America [44]. The addition of the MELs parameter to the TREES model output is therefore necessary to account for MELs not explicitly described in the model.

### 2.4. Case study

The temperature–time framework was applied to a case study area in Japan where significant climate goals have been mandated. Japan has set a national emission reduction target of 46 % for the year 2030, relative to its 2013 emissions levels, on the path to achieve net-zero by 2050 [45]. According to Japan's Ministry of Environment ~17 % of national $CO_2$ emissions were attributed to residential buildings in 2013 [46]. Achieving decarbonization in Japan's residential sector would result in significant emissions reductions.

The validation of residential energy demand and comparison with temperature-based energy behavior was conducted for three case study areas located in central Japan. Fig. 5 shows the locations of Tama city and Edogawa City in Tokyo prefecture as well as Sosa City in Chiba prefecture. Edogawa City was in the urban core of central Tokyo while Tama City was in Tokyo's suburban fringe. Sosa City represented a rural area located well outside Tokyo along the Pacific Coast boasting larger homes and families. The case study areas represent a diversity of urban form reflected in the underlying synthetic data generated (Appendix A).

Annual hourly electricity consumption smart meter data was compiled for all three cities by census districts from April 2022- March 2023. The dataset recorded low-voltage (50 kW) electricity consumption which typically describes residential buildings and less energy intensive non-residential buildings. Districts with a small number of buildings were omitted due to a lack of data and/or privacy issues. Building point data from the Japanese geospatial data provider Zenrin defined the proportion of non-residential to residential buildings to focus the analysis on residential districts (Appendix B). Tama City, Edogawa City and Sosa City retained districts with at least 50 % residential floor area for analysis resulting in only 36, 76 and 55 districts, respectively. District scale residential floor areas used to filter districts were also used to calculate electricity consumption intensity. Disaggregating district data into weekday and holiday day types yielded 72, 152 and 110 district-day type datasets for subsequent analysis.

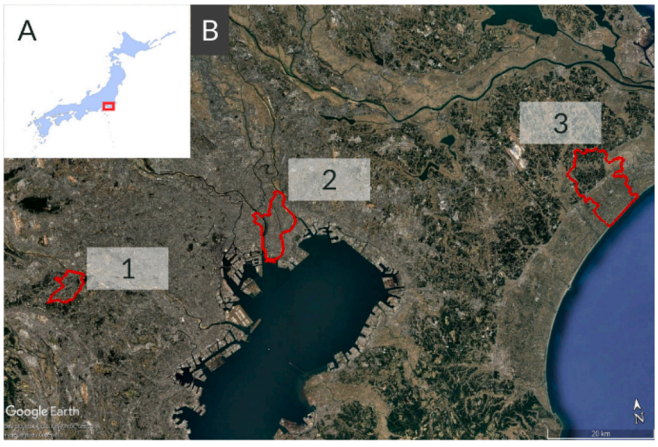



**Fig. 5.** Case study areas (1. Tama City, 2-Edogawa City, 3-Sosa City) located in central Japan (A) spread across the Tokyo and Chiba prefectures(B). Retrieved on 1/9/2025.

## 3. Result

### 3.1. Measured and simulated electricity consumption and elasticity

Examining electricity consumption and temperature related behavior showed a divergence between simulated and measured data. Fig. 6 shows all three cities displayed a good agreement in the diurnal pattern of electricity consumption. However, certain times and months observed poorer agreement. Notably, Sosa City showed consistent underestimation of consumption in May. Meanwhile, Edogawa City overestimated February morning and evening peaks. Fig. 7 revealed Edogawa and Tama cities showed a slight bias towards greater cooling elasticities. This agrees with the higher rate of electrification for cooling systems compared to a mixture of electric and fuel based heating systems which decouples electricity consumption from heating demand [32]. However, Sosa City recorded markedly higher simulated cooling elasticities compared to heating elasticities. Fig. 8 illustrated this poor agreement in Sosa City extended to the number of cooling hours. These were based on cooling and heating temperature thresholds defined from TRFs for simulated and measured datasets.

### 3.2. Model validation at the municipal level

Contrasting the results from conventional metrics against temperature–time approaches in Table 1 revealed a more stringent level of evaluation introduced by the temperature–time framework that was difficult to meet for all cities. Four of the city-day pairs were conventionally validated meeting both CVRMSE and NMBE criteria. Only Edogawa and Tama cities were able to meet multiple temperature–time criteria. None of the city-day pairs met all the temperature–time criteria due to poor cooling elasticity agreement. The Tama-Holiday case came the closest with meeting all but the cooling elasticity criteria. Comparing approaches revealed diverging conclusions despite lower CVRMSE and NMBE values often corresponding to better temperature–time performance. This is exemplified by the ability of all cases to meet the <30 % CVRMSE threshold but not many of the temperature–time criteria. Moreover, comparing CVRMSE calculated for heating and cooling hours and differences in heating and cooling elasticity, respectively, revealed no tangible correlation (Appendix D). This demonstrates that the CVRMSE does not necessarily reflect the ability of the model's simulations to accurately represent the elasticity of the measured data, and temperature related behavior overall.

### 3.3. Temperature-time plots at the municipal level

Temperature-time plots for municipalities revealed temperature dependent and independent errors. Fig. 9 showed regions of high CVRMSE in temperature–time plots across all cases despite passing the CVRMSE threshold tests. Sosa City and Tama City observed a vertical band of high errors from midnight to the early morning (12-4am) extending from ~22 °C to 3 °C. Edogawa City had a more pronounced vertical band of high CVRMSE centered around 5 am. Conversely there is

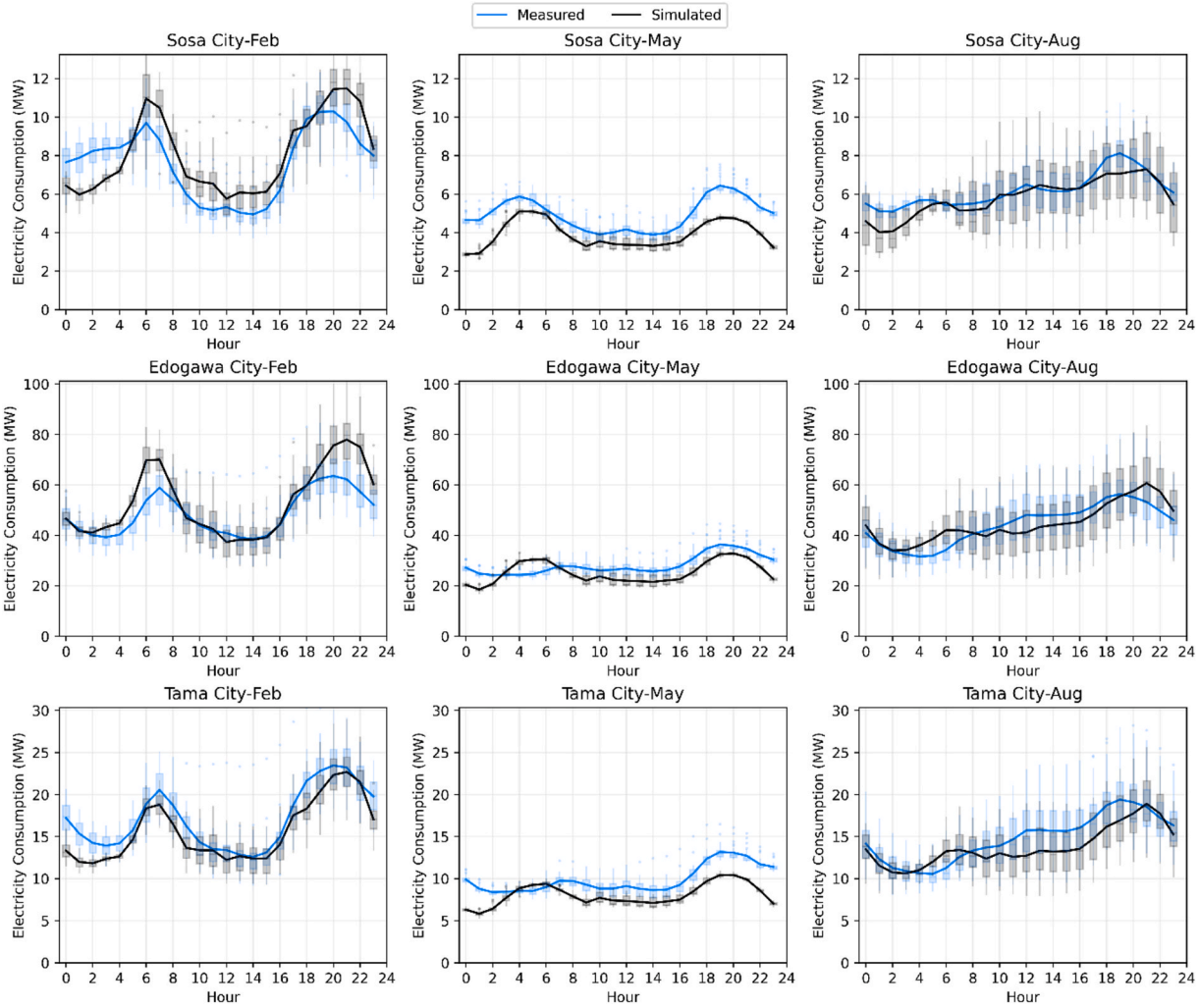



**Fig. 6.** Simulated and measured mean (line) and distribution (boxplots) of hourly electricity consumption for representative months. Note the different scales of electricity consumption for each city (row) were driven by the variable sizes of city building stocks. The boxplot box bounds represent the 1st and 3rd quartiles, with outliers being defined as data points passed the 1st or 3rd quartile by 1.5 times the data's interquartile range.

a band of high errors that extends from ~8 am to 6 pm, coinciding with business hours and commuting times, observed only for temperatures >27 °C for Sosa City. Edogawa City also observed horizontal bands during daytime hours, more obvious in the holiday case, at temperature extremes. Fig. 10 depicted similar patterns in the occurrence of high magnitude NMBE regions in city-day pairs. However, due to the cancellation of errors the aggregate NMBE values can be low despite clear variability in NMBE in temperature–time. Sosa City, which passed the conventional NMBE threshold, observed zones of large magnitude positive and negative NMBEs. Other cities where there was only one mode of NMBE values in temperature–time expectedly observed higher magnitude conventional NMBE values. The added utility of the temperature–time plots is evident through these examples as high error zones, and their potential drivers, appear within a single figure for the entire year.

### 3.4. Heating and cooling elasticities at the municipal level

Comparing estimated elasticities revealed a poorer ability to capture cooling behavior. Fig. 11 revealed high agreement between measured and simulated heating elasticities for Tama and Edogawa cities. Both weekday and holiday conditions for these cities recorded $R^2$ values >0.75. Tama City and Edogawa City- Weekday recorded CVRMSE <30 % for hourly elasticity values. Edogawa-holiday was the only case where interpretation of the fit diverged between the $R^2$ and CVRMSE. Continuously higher simulated heating elasticity led to a larger deviation from the mean measured behavior and higher CVRMSE. This contrasts with the Edogawa-weekday case where simulated elasticities are both under and overestimated across the range of elasticities. Sosa city displayed a low $R^2$ (<0.75) and high CVRMSE (>30 %) driven by an overestimation during daytime hours. Nevertheless, these results were markedly better than those for cooling.

Fig. 12 displayed a lower agreement between measured and simulated cooling elasticities with none of the city-day type pairs observing an $R^2 > 0.75$. Only Tama city cases had a CVRMSE <30 % with Edogawa and Sosa cities being subject to systemic overestimation of cooling elasticities. This was particularly noticeable for Sosa City which displayed daytime and overnight cooling elasticities with more than double the magnitude of measured elasticity. Fig. 13 provided insight into the poorer performance observed for Sosa City and better agreement for Tama and Edogawa cities by showing the estimated heating and cooling threshold temperatures defined by TRFs. The higher simulated cooling thresholds for Sosa City contribute to the overestimation of cooling elasticities since the estimation is based on a smaller and more extreme temperature domain, rather than a larger temperature domain where the change would be anticipated to be more gradual. Edogawa and Tama cities had a good agreement in the heating and cooling thresholds estimated using TRFs.

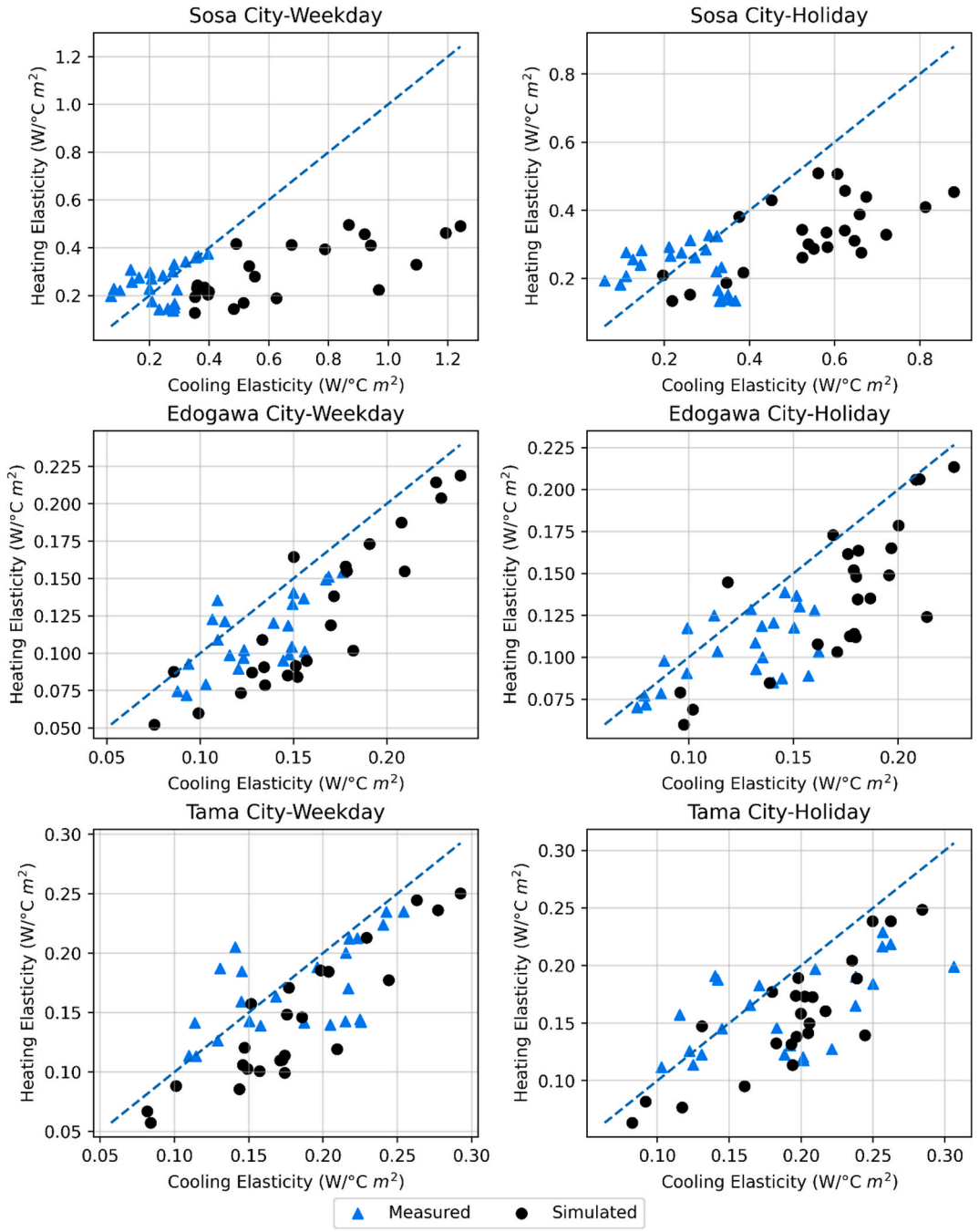



**Fig. 7.** Magnitude of hourly heating and cooling elasticities (W/°C $m^2$) compared for all six city-day type pairs highlighting the tendency for higher magnitude cooling elasticities. The dotted line represents a one-to-one fit between heating and cooling elasticity magnitudes.

### 3.5. Model validation by district

Validating model results for each district-day pair highlighted the varied ability of the model across cities and relatively poor performance when evaluated using the temperature–time approaches (Table 2). Notably, 212 (63 %) of district-day pairs met the ASHRAE guideline of CVRMSE (%) < 30 %. Only 39 % of district-day pairs meet the prescribed level of NMBE (%), a slight increase compared to the 38 % of district-day pairs which met both NMBE and CVRMSE threshold guidelines. Fig. 14 indicated that both underestimation and overestimation were occurring across districts within the cities due to the U-shape of the scatter plots of CVRMSE and NMBE. Districts in Sosa City underestimated demand indicated by positive NMBE. The inverse situation was observed for Edogawa and Tama cities which predominantly overestimated demand. The higher incidence of meeting the < 30 % CVRMSE criteria agrees with the higher success rate for CVRMSE (7 %) KS tests compared to the NMBE (1 %) KS tests.

Heating and cooling elasticities showed a tangible contrast in test results. Heating elasticity demonstrated a higher agreement with 59 % of district-day pairs observing an $R^2$ over 0.75. Only three district-day pairs surpassed this threshold for cooling elasticities. The lower performance of cooling elasticity estimates is evident by comparing the $R^2$ values for heating and cooling elasticities in Fig. 14. Unsurprisingly, no district-day pair passed all six tests. Only one district-day pair was able to pass four tests, failing the KS-tests. These results suggest that the combination of all six tests presents a significantly higher validation threshold when compared to standard CVRMSE and NMBE guidelines. This also underscores that energy estimates may not be representing

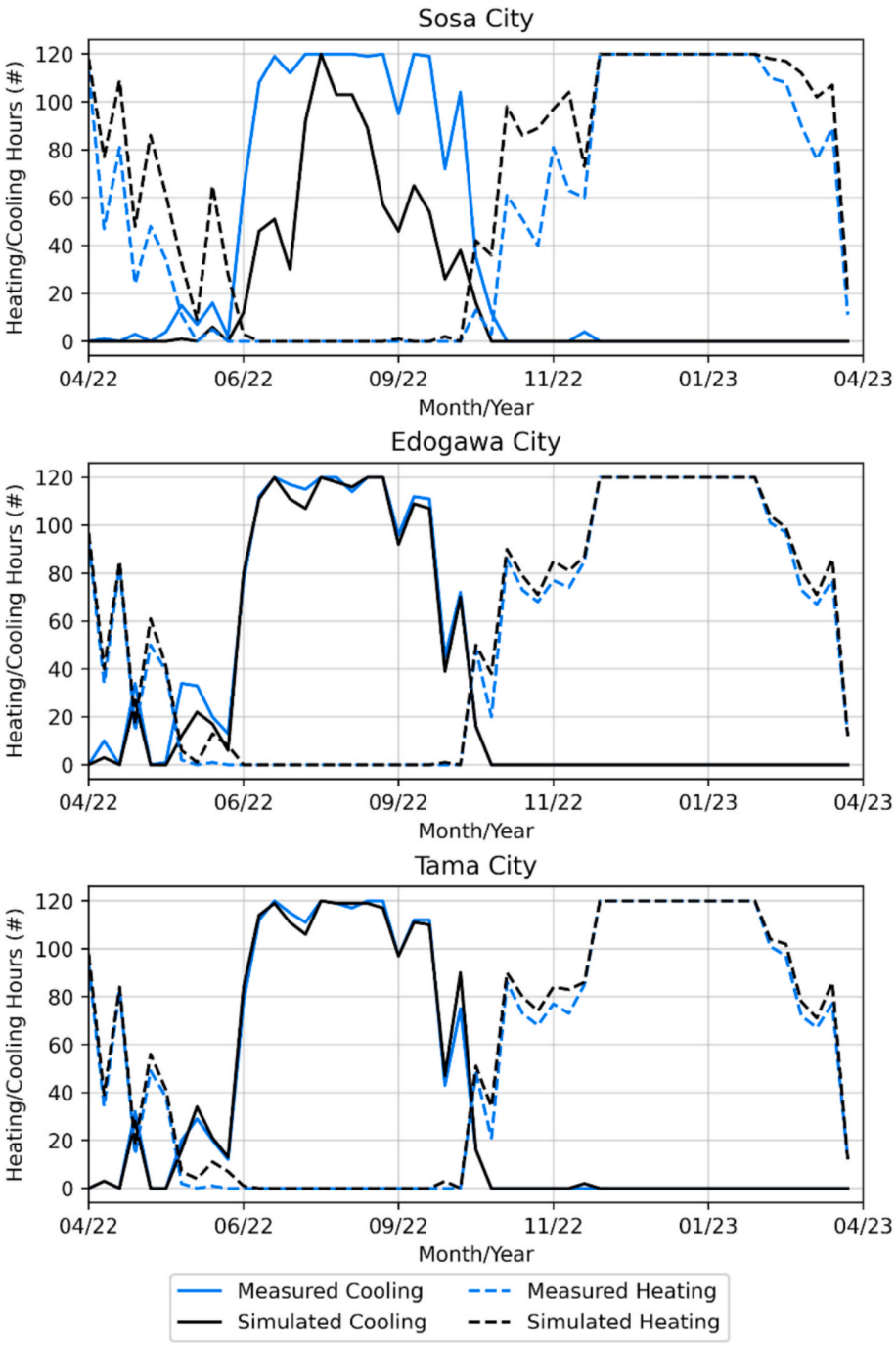



**Fig. 8.** The number of heating and cooling hours estimated per week for weekdays based on the temperature thresholds determined from hourly TRFs for both measured and simulated conditions. Note the last week saw a marked decrease in hours due to it being an incomplete week.

**Table 1**
Results for six validation tests used to evaluate each city-day type pair. Values annotated with a * represent that the criteria were met for the validation metric. Temperature-time bins with less than 5 and 2 incidences in the annual hourly dataset for weekday and holiday type days, respectively, were omitted for KS-tests.

| City | Day type | CVRMSE (%) | NMBE (%) | CVRMSE T-T KS | NMBE T-T KS | Heating elasticity $R^2$ | Cooling elasticity $R^2$ |
|---|---|---|---|---|---|---|---|
| Sosa | Weekday | 20.88* | 9.88* | | * | 62.2 | 21.9 |
| | Holiday | 21.85* | 5.71* | | * | 35.5 | 68.5 |
| Edogawa | Weekday | 20.63* | −7.28* | | * | 92.8* | 59.3 |
| | Holiday | 24.62* | −14.74 | | * | 91.0* | 62.1 |
| Tama | Weekday | 19.35* | 13.21 | | * | 95.3* | 60.2 |
| | Holiday | 17.56* | 8.71* | * | * | 80.8* | 53.9 |

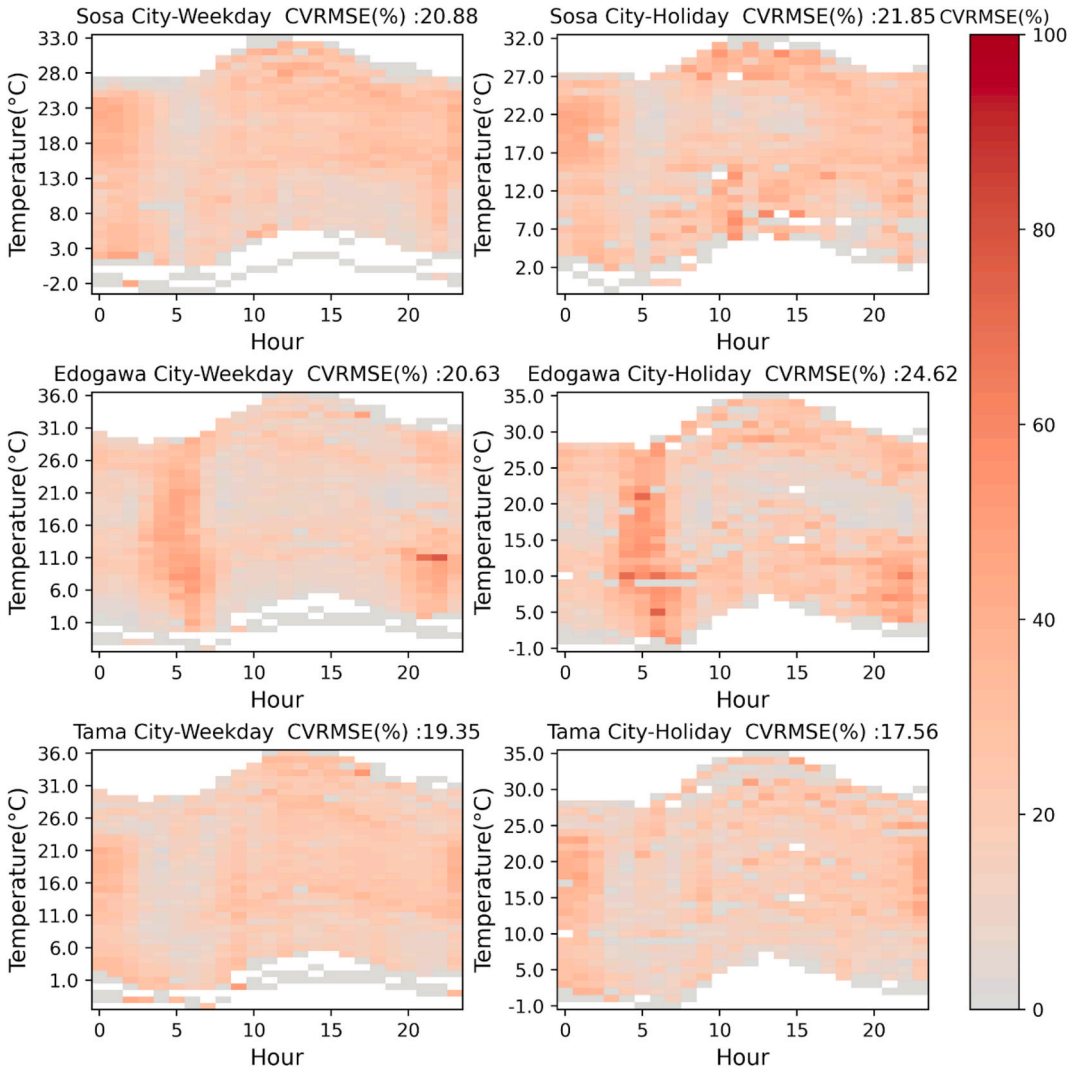



**Fig. 9.** CVRMSE (%) calculated for hourly data assigned to each temperature–time bin. Aggregate CVRMSE is shown in the Fig. titles for comparison.

temperature related behavior effectively on a local scale despite the relatively high rate of success based on conventional metrics.

### *3.6. Temperature-time framework application to calibration at the municipal scale*

Edogawa City temperature–time plots (Figs. 9 and 10) displayed a noticeable vertical high error band around 5 am that was targeted for temperature–time guided calibration. The error bands represented consistent overestimation across the entire annual temperature domain during the morning. The proportion of electric water heaters was calibrated for Edogawa City since water heating tends to maximize output in the early morning. The proportion of electric water heating was decreased by 50 % for the simulation based on hourly electric water heating estimates. In addition, the miscellaneous electric load added in this study was also reduced by 50 % to address the consistent overestimation in T-T exhibited by the NMBE T-T plot. The temperature–time framework was re-applied to gauge the impact of the calibration exercise.

Fig. 15 revealed both CVRMSE and NMBE values improved with calibration. CVRMSE decreased by ~1 % and ~4 % for weekday and holiday cases, respectively. NMBE showed a more noticeable improvement, particularly for the holiday case which observed a reduction of ~10 %. This resulted in the calibrated Edogawa City – Holiday case meeting the conventional NMBE threshold. These improvements were also reflected in the KS-test results. The Edogawa City – Weekday no longer exhibited T-T bias based on the CVRMSE KS-test result. Finally, heating elasticity $R^2$ values improved while cooling elasticity $R^2$ values were slightly lower following calibration. This did not change the interpretation of the model's ability to reflect elasticity in Edogawa City. However, this exercise was not expected to significantly influence elasticity since the vertical, temperature-independent, band of error was targeted. Further calibration is required to reduce the under and over estimations of electricity consumption at high temperatures (>26 °C) in overnight and daytime periods, respectively, still evident in NMBE T-T plots. Nevertheless, this example highlights how temperature–time guided calibration efforts helped Edogawa City achieve conventional validation thresholds and improve T-T metrics.

## 4. Discussion

Comparing conventional and temperature–time validation results in Section 3.2 revealed a variable performance across cities and heating/cooling regimes. Tama City- Holiday saw the best performance recording both favorable CVRMSE and NMBE values and a lower degree of temperature–time biases. Section 3.3 showed high error regions, notably a vertical band overnight and into the morning, punctuating T-T plots despite meeting conventional NMBE and CVRMSE thresholds. The worst performing case was Sosa City which was only able to meet the T-T NMBE KS-test. The poor representation of both heating and cooling

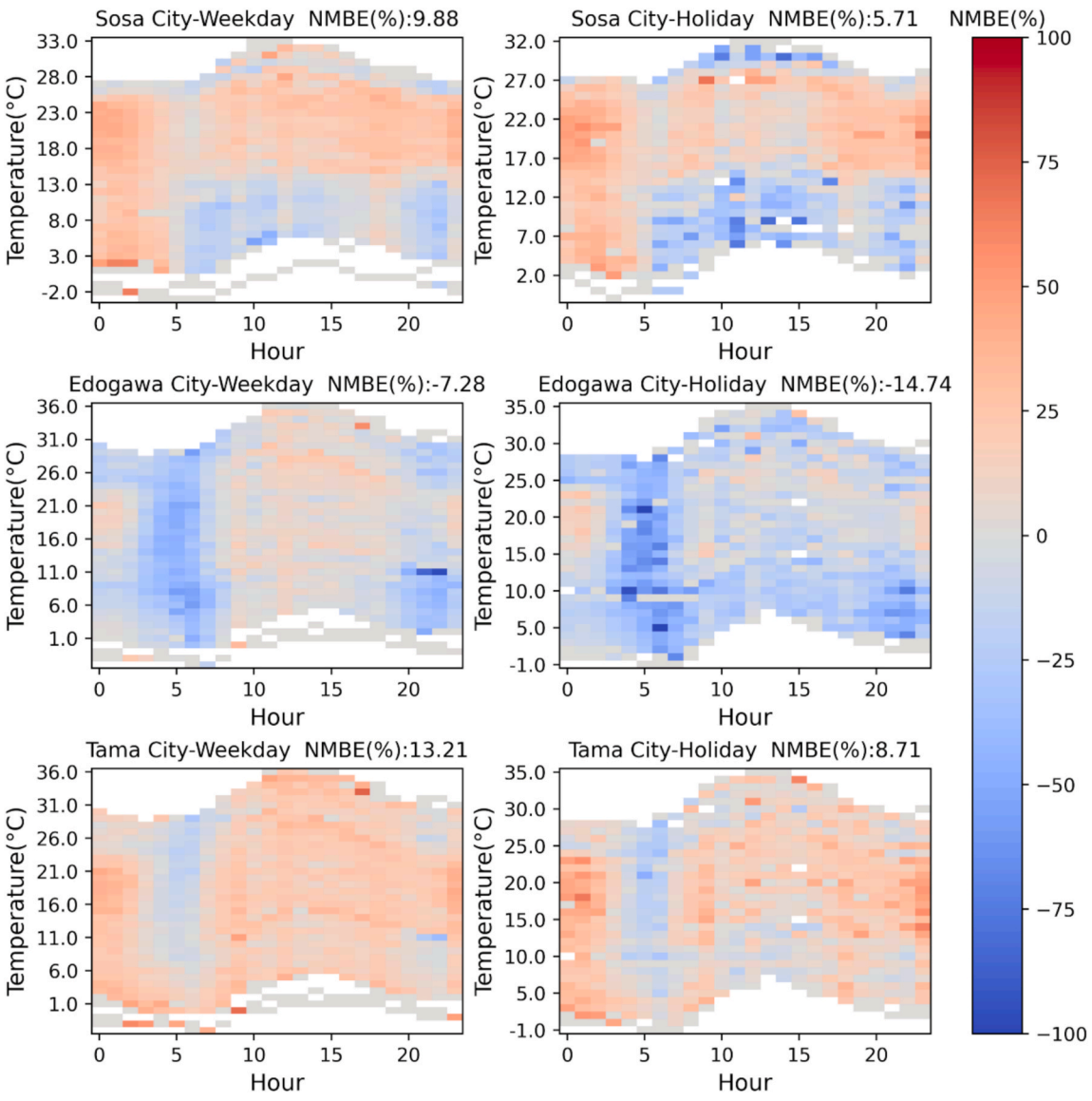



**Fig. 10.** NMBE (%) calculated for hourly data assigned to each temperature–time bin. Aggregate NMBE is shown in the Fig. titles for comparison.

elasticities, shown in Section 3.4, agreed with higher errors at temperature extremes, more noticeable in the holiday case. Other cities also observed the consistent divergence between measured and simulated cooling elasticity. The model does not consider passive cooling which may be more likely in more open sub-urban and rural areas leading to deviations in simulated and actual cooling behavior, particularly for Sosa City.

District level validation, described in Section 3.5, mirrored conventional results at the municipal level but differed in their T-T framework results. For example, hardly any districts were able to pass the T-T NMBE KS-test unlike in the city cases where all cities met this criterion. This is due to the absence of cancellation effects at the district scale which manifests at the municipal scale. Moreover, since the construction period was estimated based on prefectural estimates it did not capture district level variations. This introduces errors at the district level that are muted at the aggregate municipal level. Comparing cities, Edogawa City performed well at the district scale. The lower variability in demographic conditions across districts in Edogawa City may have countered this behavior as there was less variability required by parameterizations of the population and building stock. Nevertheless, the results show that localization of the bottom-up model may still require improvement if stakeholders are seeking district level results.

Despite general agreement between the implied performance of the model using both conventional and temperature–time approaches, the suite of temperature–time analysis emphasized that there are biases in temperature–time that persist in conventionally validated simulations. As a result, intrinsic temperature-based bias or errors can persist despite conventional validation. This has implications particularly for analysis which are concerned with energy demand at higher temporal resolutions and temperature extremes such as peak energy demand timings. For example, while Tama City – Holiday was validated at the municipal level based on conventional NMBE and CVRMSE conditions it failed to capture cooling elasticity. While this may appear irrelevant since the model captured electricity consumption directly, it is important to recognize that cooling elasticity represents the change in consumption with increasing temperatures. As a result, an overestimation of elasticity will result in a growing divergence between actual and simulated electricity consumption as temperatures grow. Conventional approaches failed to recognize the model's limitation in terms of cooling elasticity. This can influence the ability of the model to inform analysis pertaining to hourly cooling demand, such as resilience to extreme heat, where the cooling reactivity and consumption at extreme temperatures needs to be reflected accurately. Accounting for these effects and accurately capturing demand at higher temporal resolutions is expected to grow as

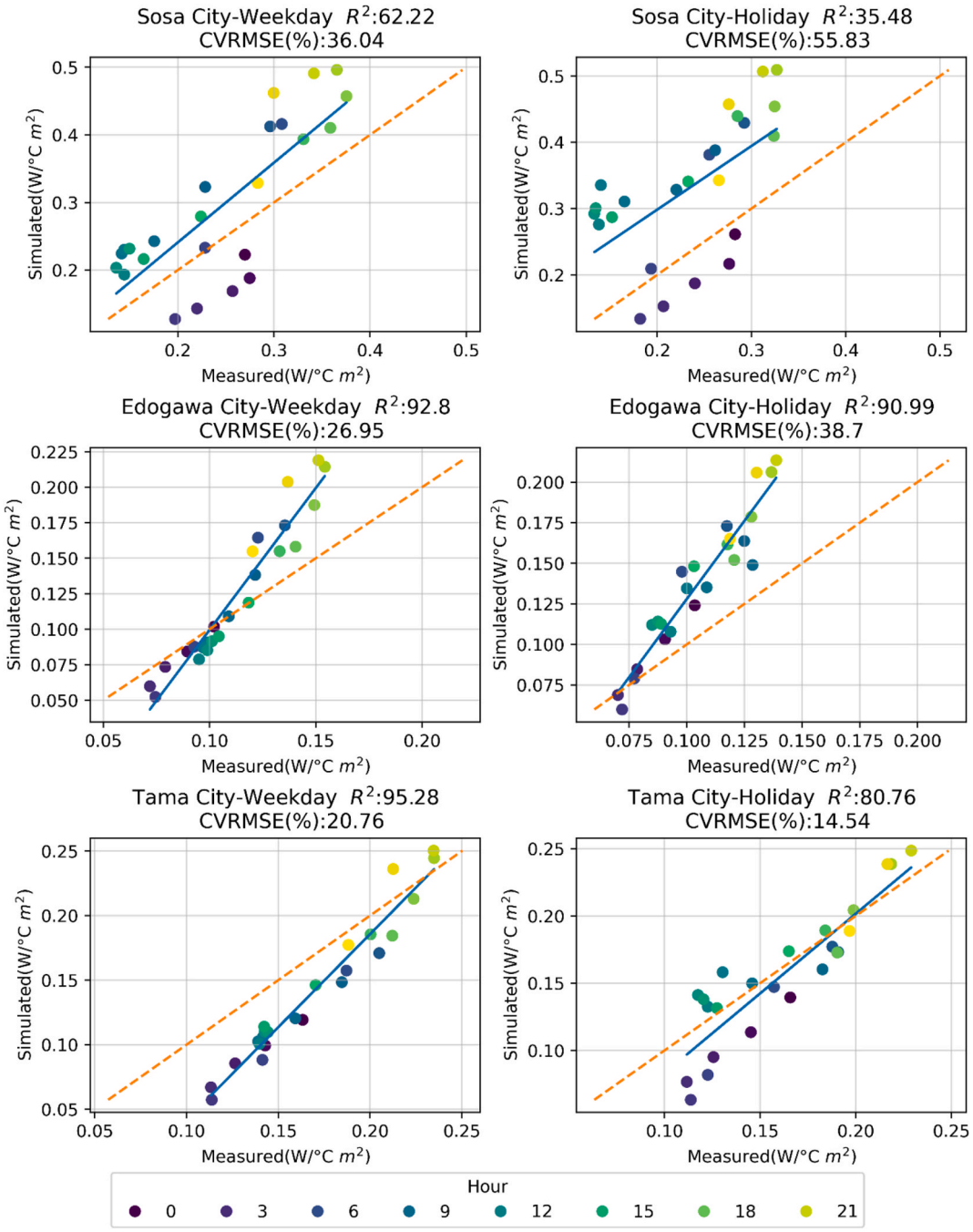



**Fig. 11.** Comparison of measured and simulated heating elasticities. CVRMSE (%) was calculated based on hourly plotted values. The blue solid line represents the $R^2$ fit. The red dotted line represents an agreement between measured and simulated results. Note that each plot's axis is scaled to their minimum and maximum values. Note that $R^2$ values have been multiplied by a factor of 100. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

demand response programs and smart design become more widespread.

Temperature-time validation grant modelers a sober second look at the model performance to inform specific calibration targets towards building a 'future-proof' model as demonstrated in Section 3.6. The ability of the temperature-based validation framework to inform more targeted calibration can streamline the validation-calibration process. Poor fits between measured and simulated elasticities can reflect potential issues with the parameterization of system efficiencies and geometric properties of the household. Weak representation of the heating or cooling elasticity can be indicative of inaccurate parameterization of heating system electrification and the prevalence of passive cooling, respectively. On the other hand, T-T plots differentiate between temperature dependent and independent errors caused by heating and cooling parameterization and occupant behaviors, respectively. The framework therefore supports more targeted calibration that can directly improve the temperature-based validation of the model by targeting underlying model weaknesses.

The intention of the framework is to provide an additional layer of analysis for modelers, and by extension policymakers and stakeholders, to understand the limitations of their model in the context of anticipated future climate changes. The low success rate of model validation using the temperature–time framework, with no city or district meeting all six criteria, suggests that the 'future-proof' model may present an aspirational target rather than a goalpost to discredit models deemed sufficient

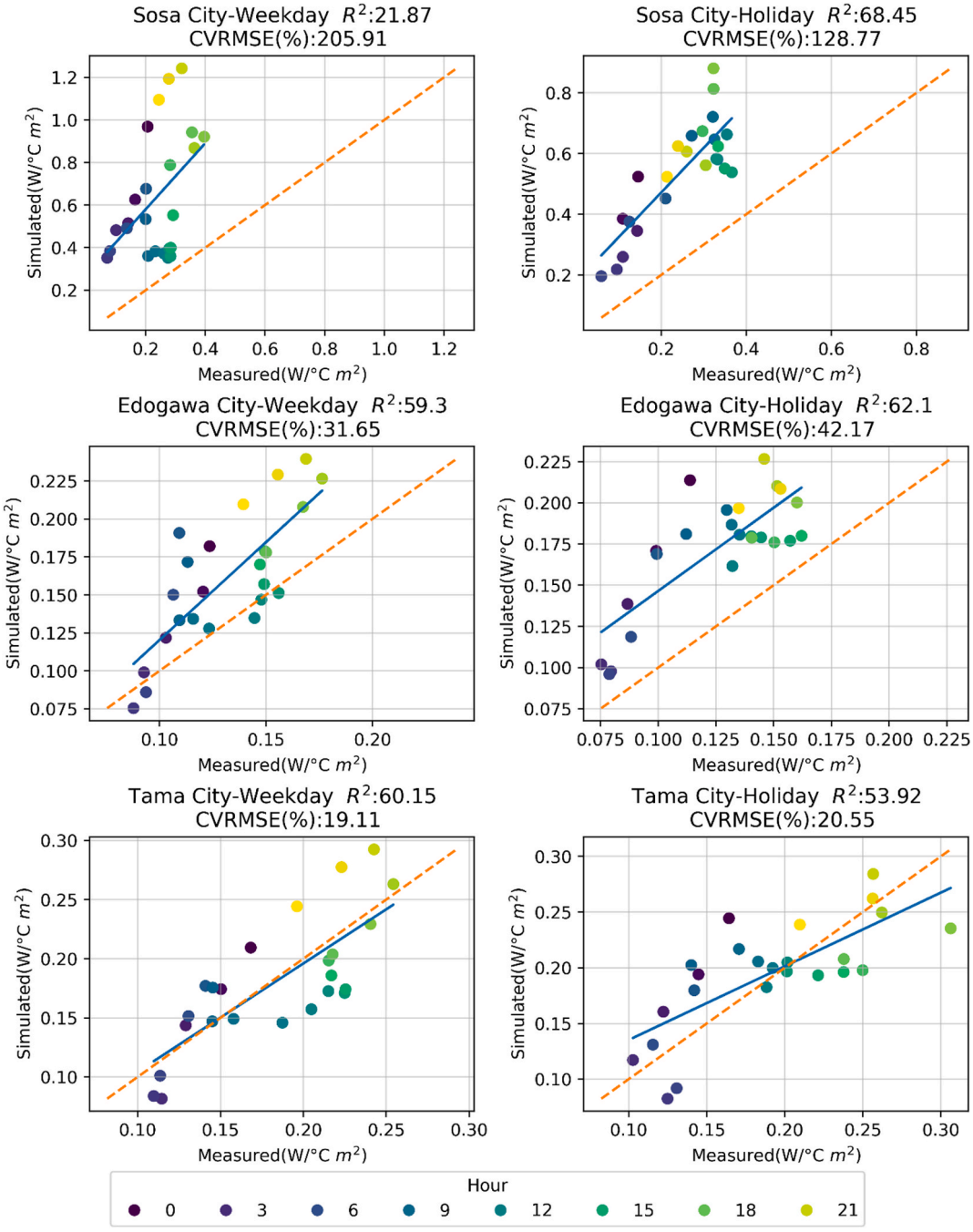



**Fig. 12.** Comparison of measured and simulated cooling elasticities. CVRMSE (%) was calculated based on hourly plotted values. $R^2$ fit is represented by the blue solid line. The red dotted line represents an agreement between measured and simulated results. Note that $R^2$ values have been multiplied by a factor of 100. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

using conventional metrics.

The increased resolution and focus of the analysis treads new ground that will ideally start a discussion about more tailored validation frameworks as data resolution and availability improve. The dissemination of Artificial Intelligence of Things (AIoT) technology and Home Energy Management Systems (HEMS) will undoubtedly spur the demand for more sophisticated validation approaches as stakeholders seek to leverage the deluge of data from these technologies. On the other hand, low data demands promote the adoption of the framework for a wider array of contexts. The framework's design deliberately considers both these opportunities. The framework leverages low data demands, relying solely on electricity demand and weather data, to provide a more informative validation process whose resolution can be improved in response to available data. Modelers can gauge their model's 'future-proof' status using the proposed framework intended to be model agnostic. While the framework is presented for the physics-based bottom-up TREES model, it can be valuable for other bottom-up approaches as well. Data-driven bottom-up UBEM stands to benefit the most since building dynamics are more implicit than for physics-based models. Other physics-based models relying on more sophisticated building model engines (i.e *Energyplus*) can use the framework to validate and calibrate parameterizations rather than internal model dynamics. Ultimately, modelers can flag models failing to meet 'future-proof' status for stakeholders at the interface of science and policy. Hong et al. [47]

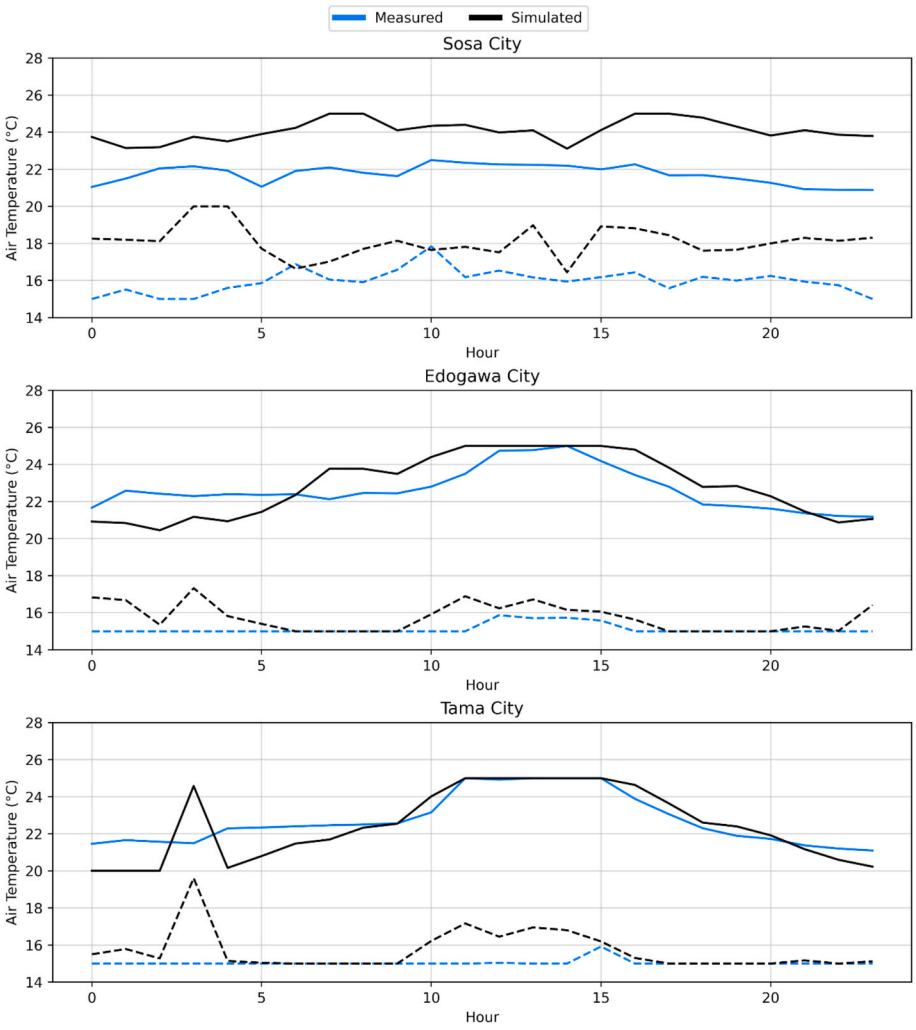



**Fig. 13.** Temperature threshold for heating (dashed lines) and cooling (solid lines) defined by hourly TRFs. Temperatures above and below the ranges identified by vertical lines represent cooling and heating conditions, respectively.

**Table 2**
Number of district-day type pairs which passed the validation evaluation for different validation metrics.

| City | Districts | CVRMSE (%) | NMBE (%) | CVRMSE T-T KS | NMBE T-T KS | Heating elasticity $R^2$ | Cooling elasticity $R^2$ |
|---|---|---|---|---|---|---|---|
| Sosa | 110 | 60 | 41 | 4 | 1 | 3 | 1 |
| Edogawa | 152 | 114 | 71 | 10 | 2 | 146 | 2 |
| Tama | 72 | 38 | 19 | 8 | 2 | 47 | 0 |
| Total | 334 | 212 | 131 | 15 | 5 | 196 | 3 |
| % of total | | 63 | 39 | 7 | 1 | 59 | 1 |

advocate for such improved transparency through the identification of potential limitations, particularly for policy evaluation considering uncertain futures.

## 5. Conclusions

The residential bottom-up TREES model was validated at the district and municipal levels for three municipalities in Japan. This study observed the persistence of temperature-based errors despite sufficient scores based on conventional industry standards. Notably, elasticity was identified as a blind spot of conventional metrics. The growing popularization of Urban Building Energy Modelling (UBEM) application to evaluate future energy demand, emission and policy scenarios should be cognizant of these inherent limitations, particularly as the desired level of detail increases to hourly temporal and local spatial scales. Temperature-time plots provide a more robust qualitative assessment of model errors and can ideally help address this gap and inform future model improvements.

Framework improvement is needed to address potential shortcomings of the approach. Examining cooling elasticity using outdoor air temperatures introduced a potential weakness to the approach. Analysis by Cao et al. [48] of cooling demand and a wet-bulb temperature-based cooling degree day index revealed the latter's effectiveness at describing cooling energy demand since it encompassed both sensible and latent energy demands. Future work will look to assess the applicability of a wet-bulb temperature-based analysis expected to have added utility in hot and humid climate conditions such as Japan's megacities and other rapidly developing nations. A parsimonious approach was sought in this

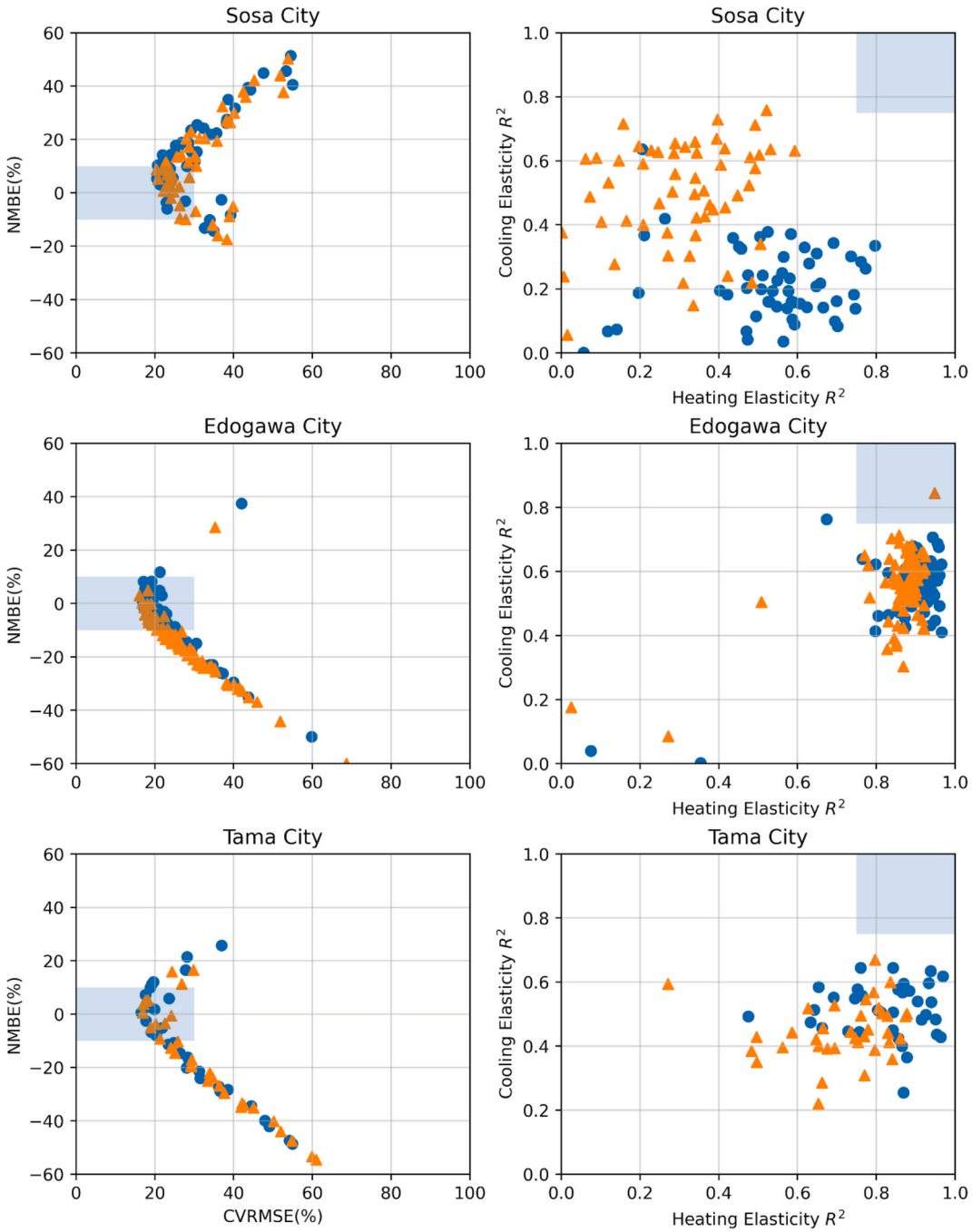



**Fig. 14.** CVRMSE (%) and NMBE (%) metrics (left) and $R^2$ scores for heating and cooling elasticities (right) for each district differentiated by weekday (blue circle) and holiday (orange triangle) conditions. Shaded regions represent desired validation metrics. Elasticity $R^2$ values were based on the linear relationship between measured and simulated hourly elasticity values in W/C°m$^2$. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

work to reduce the barrier to entry for framework implementation. Future framework iterations can use more robust temperature response functions if modelling capacity and data availability are sufficient. Furthermore, focusing solely on-air temperature fails to consider the impact from covariates, such as daylight hours, which can also explain the error during these periods. A tiered approach should be applied to investigate the most likely issues first. If there is minimal change by exploring that domain within logical bounds, the covariates can then be explored. Ultimately, the temperature–time approaches presented in this work describe a starting point for developing a more robust suite of analysis to inform calibration pathways that modelers can develop holistically to ensure model resilience against climate change and shifting socio-economic conditions.

The new framework presents a pathway forward for UBEM application for policy analysis. It is an inherent goal of this work to propel discourse past the science-policy divide about the limitations and critical assumptions of modelling future projected scenarios. This study provides a more transparent and rigorous approach to improve the interpretation of model outputs for future climate scenarios. This is essential to avoid the propagation of erroneous conclusions by model consumers as model complexity and scenario uncertainty expands. Future work will apply this framework to improve the development of more robust UBEM

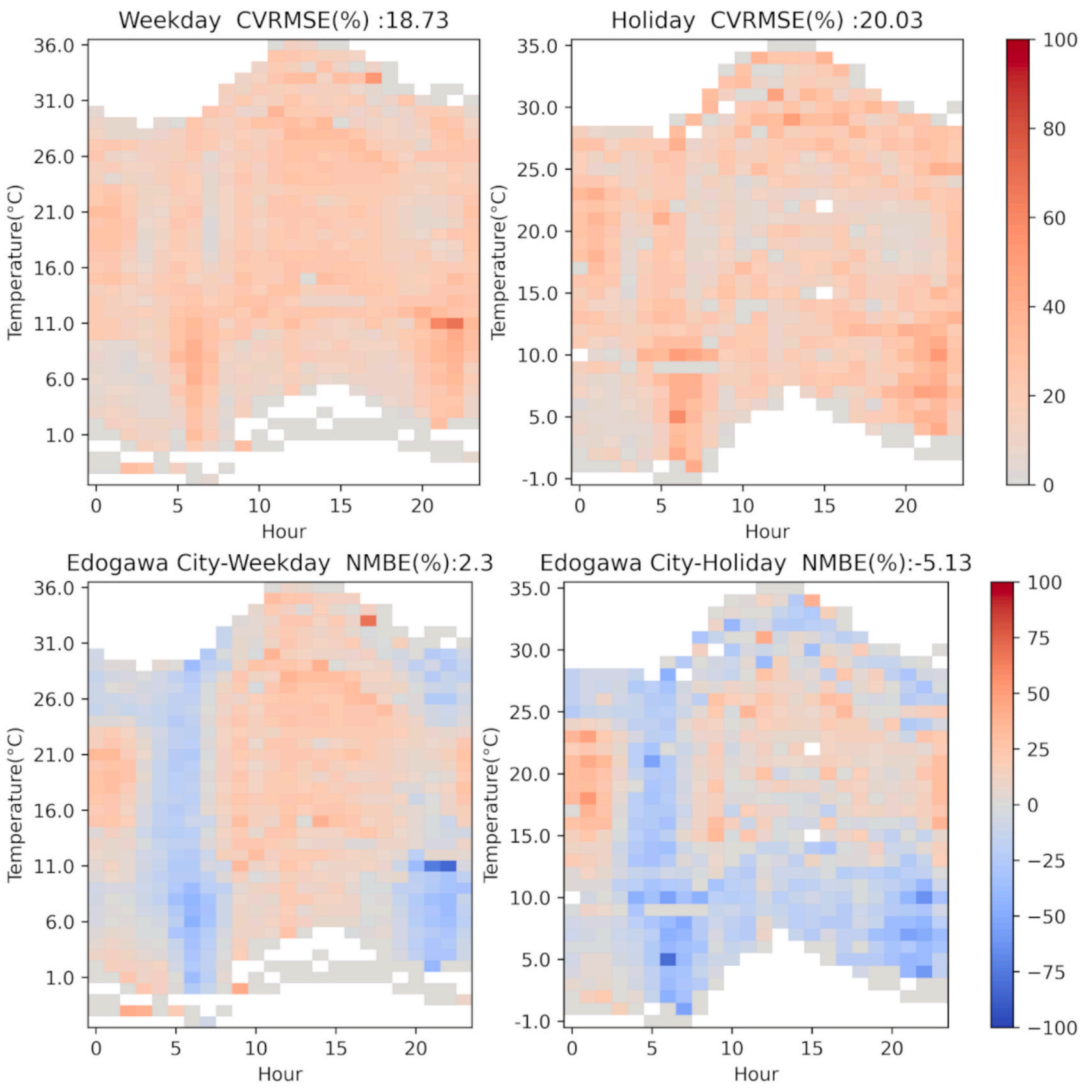



**Fig. 15.** Calibration results in T-T for Edogawa City following the reduction of electric water heating and miscellaneous electric loads by 50% displayed subdued errors in T-T.

and explore purpose-driven calibration towards the creation of a value-driven model for future scenario modelling and policy assessments.

## CRediT authorship contribution statement

**Andrew Marian Zajch:** Writing – original draft, Visualization, Methodology, Formal analysis, Conceptualization. **Yohei Yamaguchi:** Writing – review & editing, Methodology, Funding acquisition, Conceptualization. **Keita Shono:** Writing – review & editing. **Tomoki Shigematsu:** Writing – review & editing, Formal analysis. **Hideaki Uchida:** Writing – review & editing. **Tsuyoshi Ueno:** Data Curation, Writing-Review and Editing. **Yoshiyuki Shimoda:** Writing – review & editing, Supervision.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgements

This work was supported by the Council for Science, Technology and Innovation, Japan (CSTI), Cross-ministerial Strategic Innovation Promotion Program (SIP), the 3rd period of SIP “Smart energy management system” Grant Number JPJ012207 (Funding agency: JST).

## Appendix A. Case study conditions

**Table A1**
Population and household numbers are based on the 2020 national census[2].

| | Tama City | Edogawa City | Sosa City |
|---|---|---|---|
| Population (#) | 146,951 | 697,932 | 35,040 |
| Households (#) | 68,354 | 332,895 | 12,848 |
| Area ($km^2$) | 20.8 | 49.3 | 96.9 |
| Population Density (persons/ $km^2$) | 7065 | 14,157 | 362 |
| Districts (#) | 89 | 200 | 63 |

[2][Population, Households, Sex, Age and Marital status] Number of households and Household members by Type of household – Japan, Prefectures, Municipalities (including Municipalities as of 2000). Accessed on January 24th, 2025. https://www.e-stat.go.jp/en/stat-search/database?statdisp_id = 0003445098.

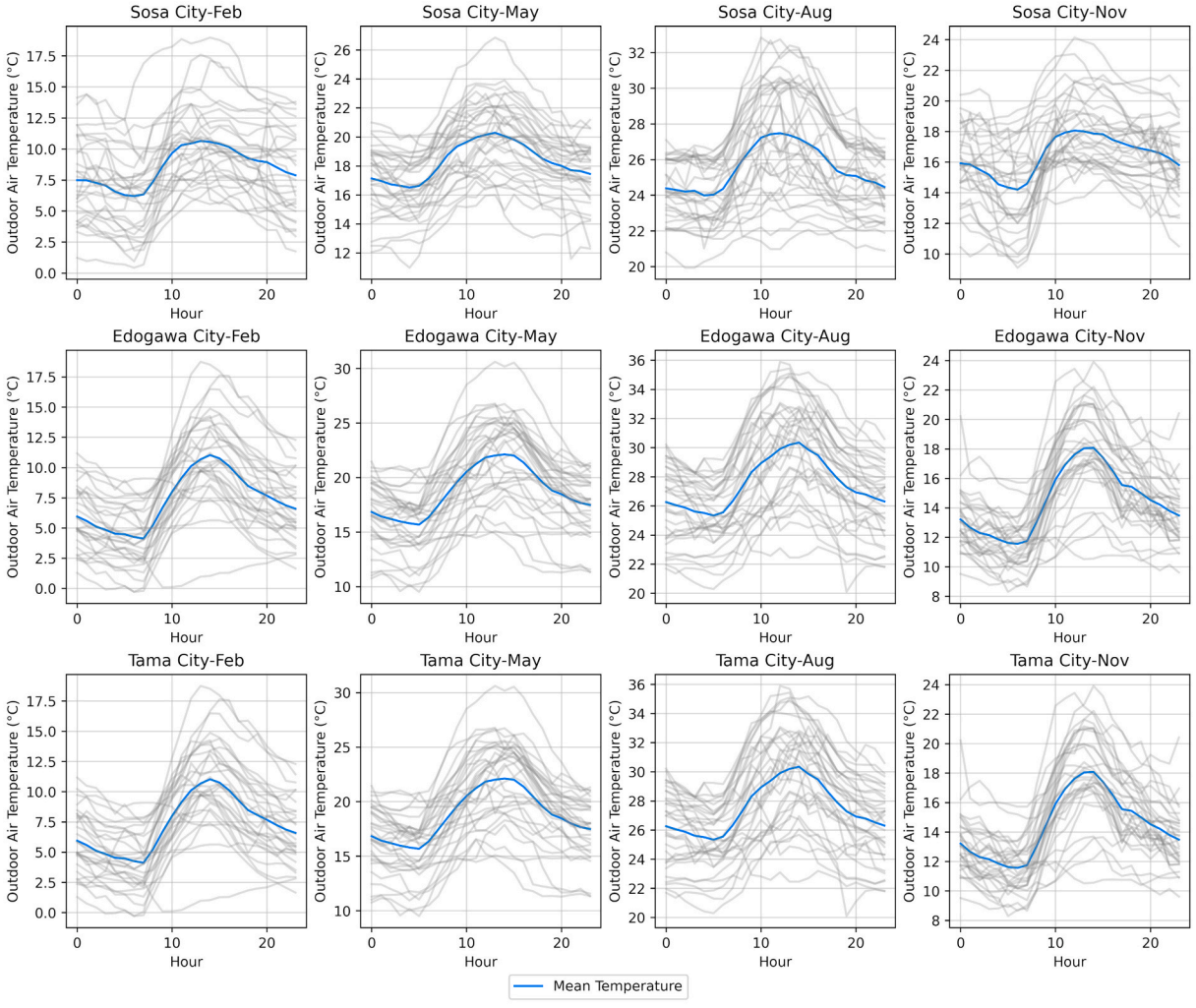



**Fig. A1.** Outdoor air temperature for four representative months of the seasonal cycle for each municipality. The blue line indicated the monthly mean diurnal signal. The varying range of the y axis between months and cities should be noted prior to comparison.

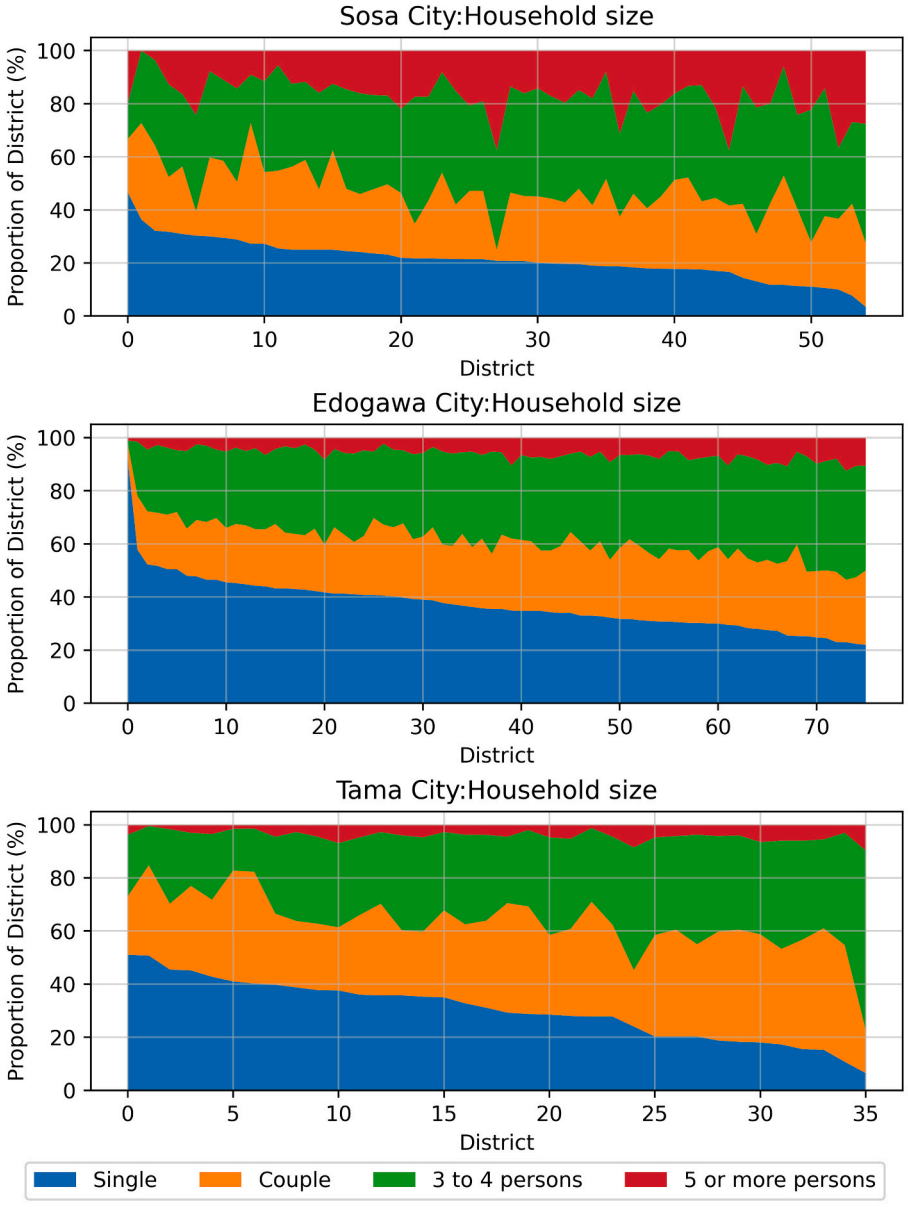



**Fig. A2.** Proportion of household sizes for each district within the case study cities. Sosa city had a noticeably larger household size compared to the denser Edogawa and Tama cities.

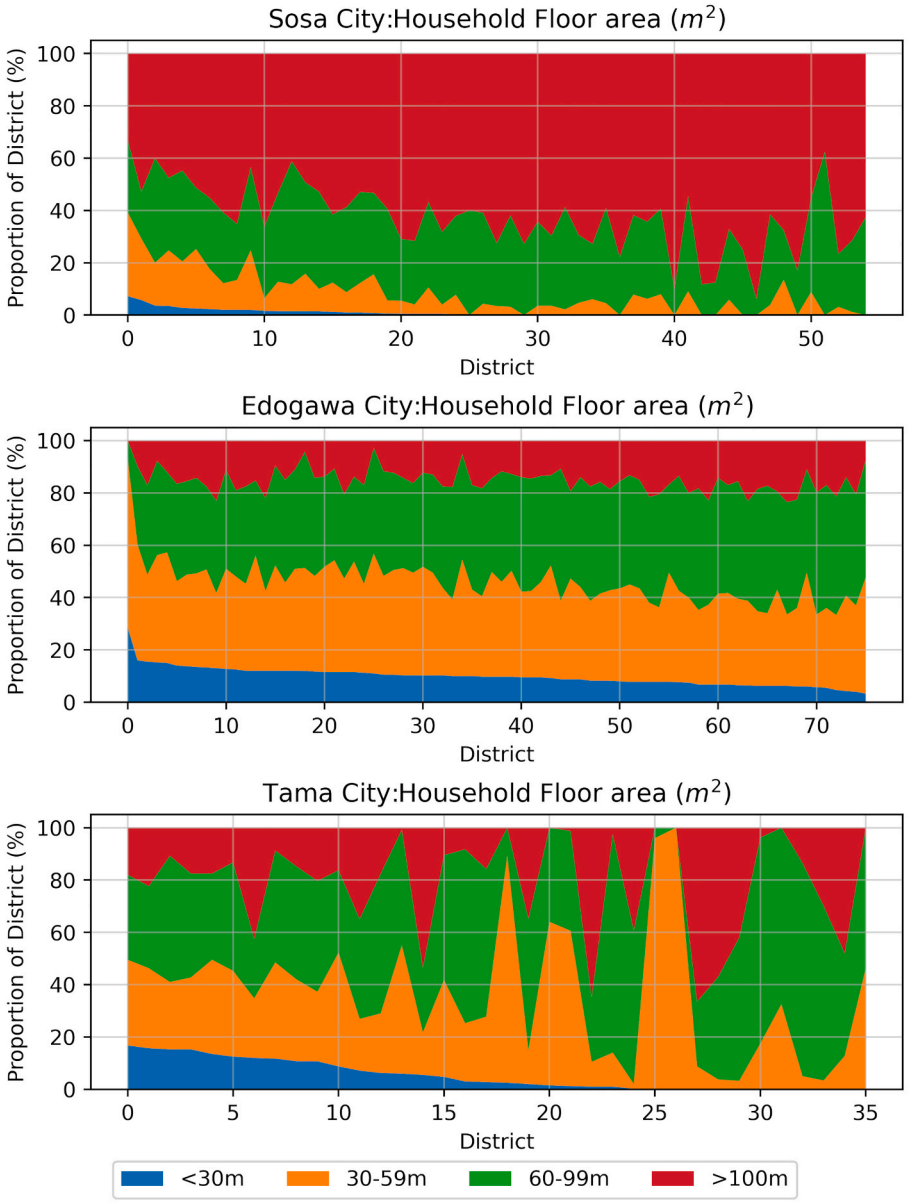



**Fig. A3.** Proportion of household floor area ($m^2$) derived from synthetic population data generated for each district. The larger floor areas in Sosa compared to Tama/Edogawa cities emphasize the difference in building stock between case study municipalities.

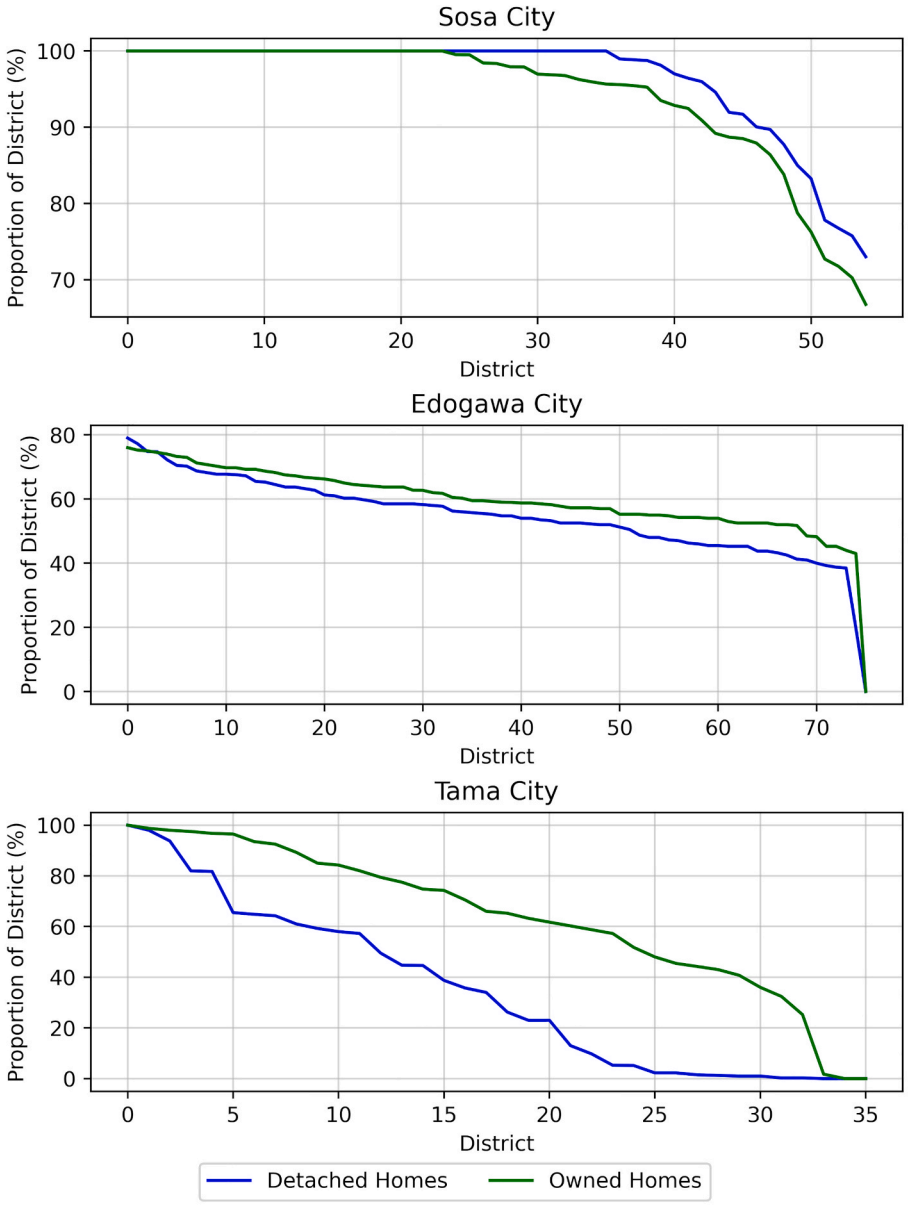



**Fig. A4.** Proportion of detached (blue) and owned (green) homes based on synthetic population data generated for each municipality.

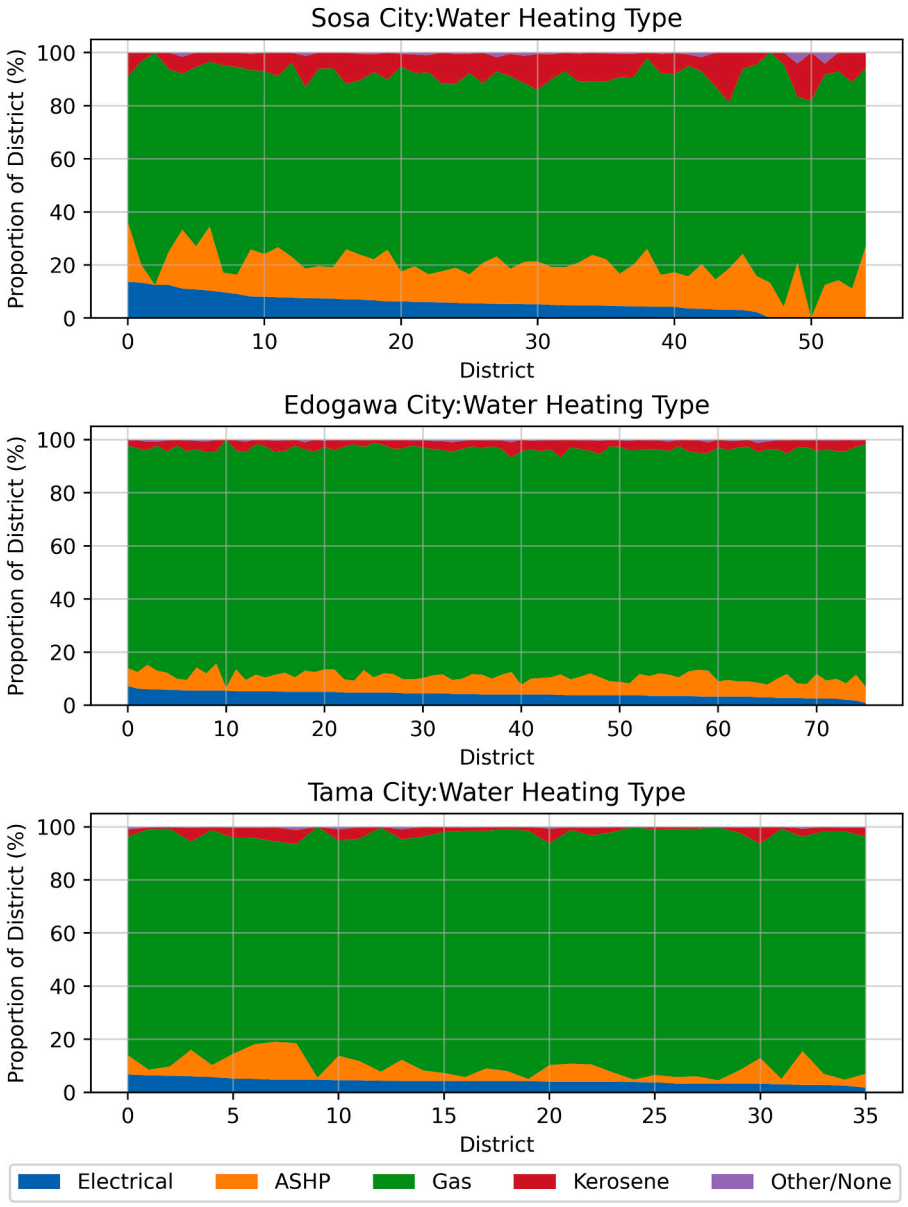



**Fig. A5.** Proportion of water heating system types based on synthetic population data generated for each municipality demonstrating the dominance of gas water heaters (green) in the case study areas. ASHP represented Air Source Heat Pumps.

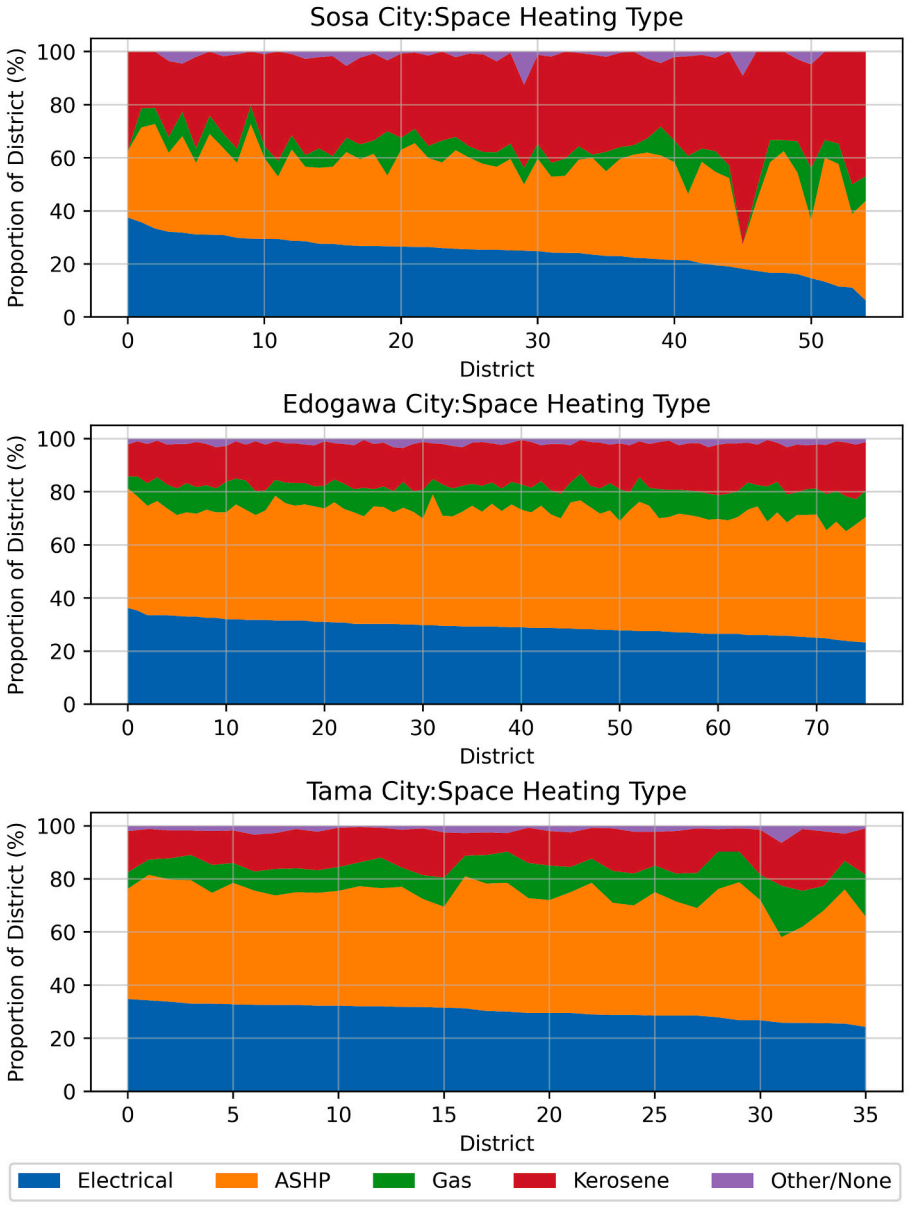



**Fig. A6.** Proportion of space heating system types based on synthetic population data generated for each municipality showing a mixture of space heating types led by Air Source Heat Pumps (ASHP, orange).

**Table A2**
Building parameters adopted by the TREES model used for simulating case study areas.

| Thermal resistance ($m^2$ K/W) of insulation material of exterior wall [49] | | |
|---|---|---|
| | Attached House | Detached House |
| 1980 Standard | 0.50 | 0.60 |
| 1992 Standard | 0.77 | 0.86 |
| 1999 Standard | 1.10 | 2.20 |
| Rated COP of RACs [40] | | |
| Cooling | 2.62–5.13 | |
| Heating | 3.19–5.69 | |
| Set Room Temperature (°C) | | |
| | Heating | Cooling |
| No insulation | 18 °C | 27 °C |
| 1985 Standard | 21 °C | |
| 1992 Standard | 22 °C | |
| 1999 Standard | | |
| Ventilation rate (times/hour) | | |
| No insulation | 3.0 | |
| 1985 Standard | 1.0 | |
| 1992 Standard | 0.5 | |
| 1999 Standard | | |

## Appendix B. Residential district filtering

Building data compiled from building point data contained the coordinates and floor area, grouped by use type, for each building. A spatial join aggregated building information attributed to each district. Summary statistics calculated for each district describe the sum of residential and total building floor area. The residential floor area ratio estimated the proportion of residential to total building floor area, providing a proxy of the influence of residential buildings in the district. Districts with residential floor area $< 50$ % of total building floor area were omitted from subsequent analysis at both the district and city levels. Fig. B.1 showed the distribution of floor areas by district relative to the threshold. This helped avoid areas with considerable influence from non-residential buildings which are outside the scope of the modelling.

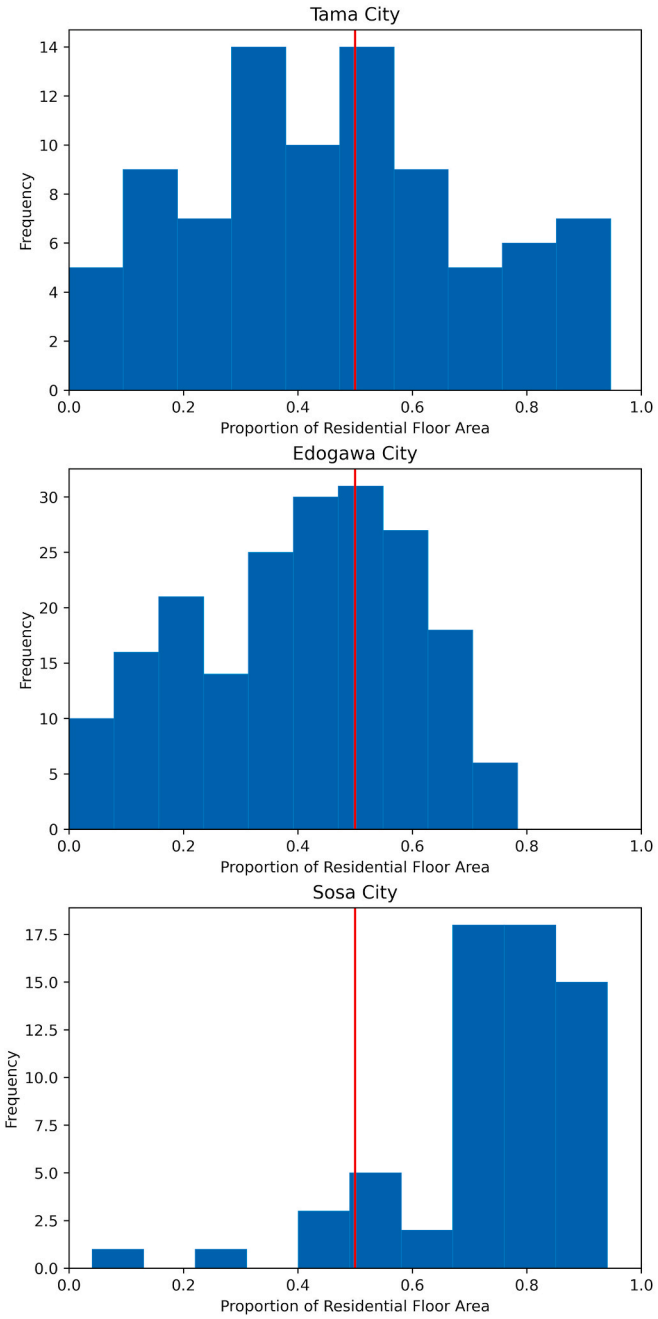



**Fig. B1.** Histogram of the ratio of residential floor area to total building area for each city, based on district level aggregation, compared against the threshold of 0.5 (red vertical line) used to filter districts for subsequent analysis.

## Appendix C. Temperature-time supplementary figures- counts and histograms

The KS-test compared the distribution of CVRMSE (Fig. C.1) and NMBE (Fig. C.2) for all city-day type pair against the assumption that the binned results were randomly distributed in T-T. Comparison of the distribution of CVRMSE in temperature–time bins revealed an underlying bias in all cases except for Tama City- Holiday. On the other hand, all city-day pairs passed the KS-test for NMBE based on T-T bins.

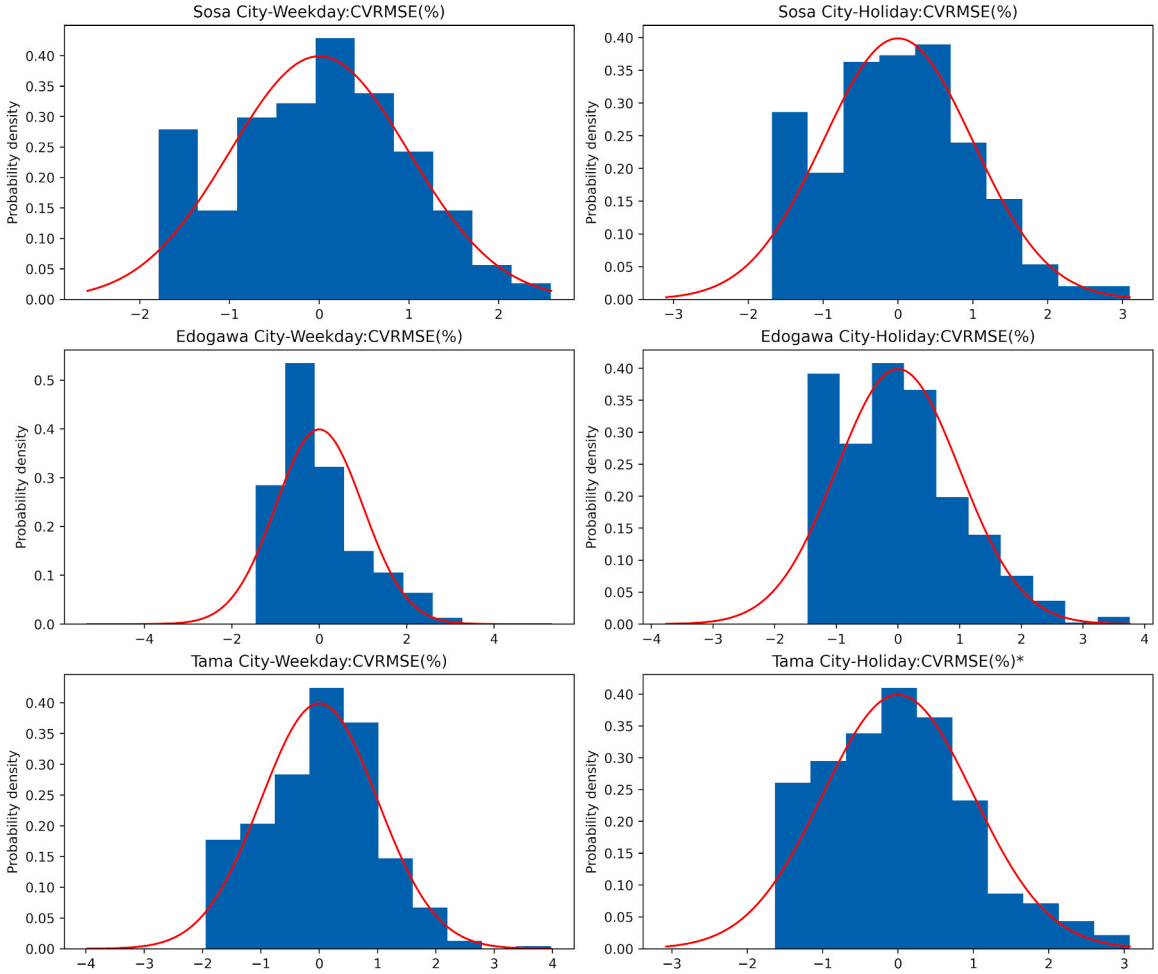



**Fig. C1.** Histogram of standardized CVRMSE based on temperaturetime bins compared against a random normal distribution (red line). Values with a * indicate that the CVRMSE was randomly distributed based on the KS-test at a 0.05 alpha level.

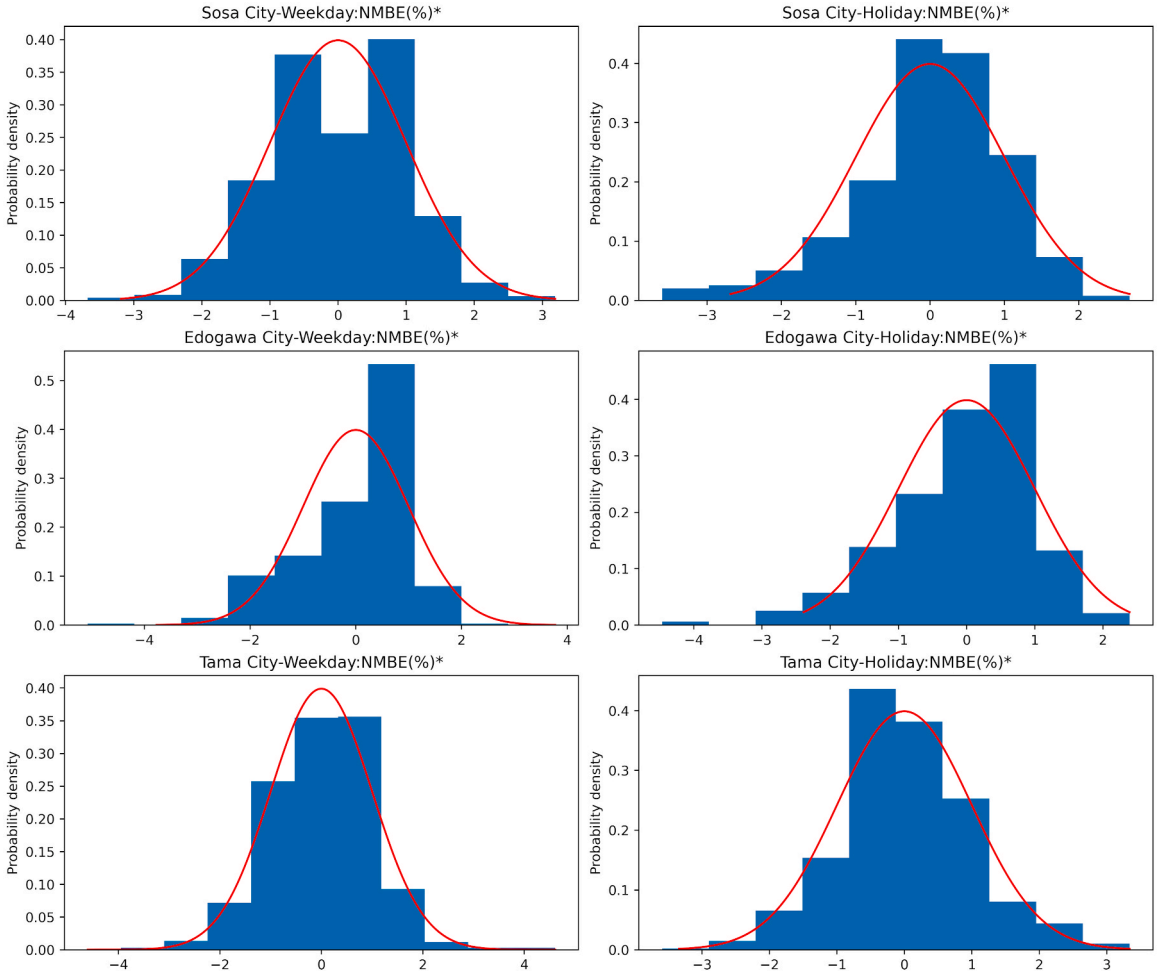



**Fig. C2.** Histogram of standardized NMBE based on temperaturetime bins compared against a random normal distribution (red line). Values with * indicate that the NMBE was randomly distributed based on the KS-test at a 0.05 alpha level.

Fig. C.3 Plotted the number of hourly data points in each temperature–time bin showing no clear patterns between the data count and the patterns in the CVRMSE and NMBE plots. This demonstrates that the CVRMSE and NMBE values appear to track the qualitative interpretation of temperature–time plots rather than being a relic of unequal data binning.

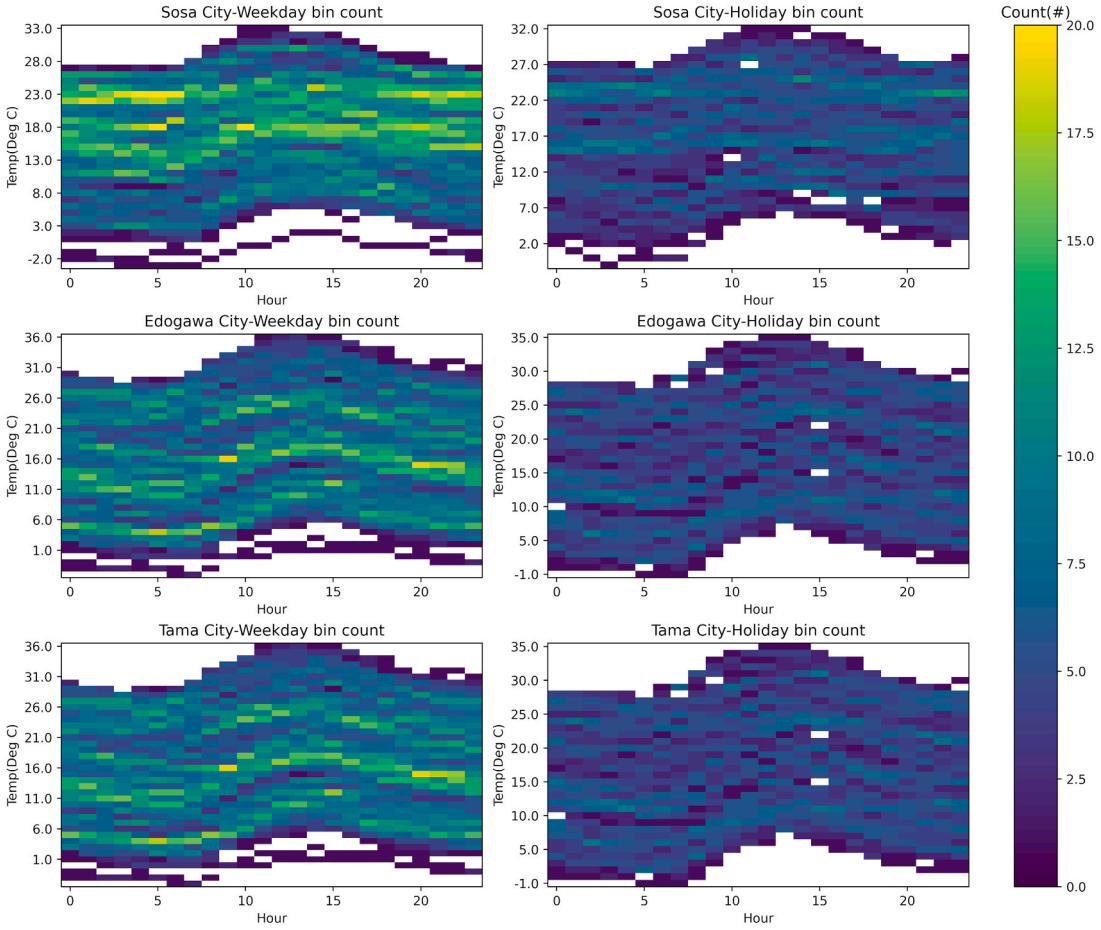



**Fig. C3.** Temperature-time bin counts highlighting the most frequently occurring bins for hourly data.

## Appendix D. Hourly heating and cooling elasticity metric comparison against CVRMSE

Heating and cooling specific CVRMSE revealed the relationship between elasticity and CVRMSE. Heating and cooling hourly data separated based on the TRF heating and cooling thresholds from measured data for each hour. CVRMSE was calculated for each hour annually for heating and cooling separately. Comparison of the CVRMSE calculated for only heating (Fig. D.1) or cooling (Fig. D.2) hourly data binned by hour and the hourly aggregated differences in measured and simulated elasticity revealed no tangible correlation between these metrics. This demonstrates that elasticity and CVRMSE based on energy consumption are not necessarily related.

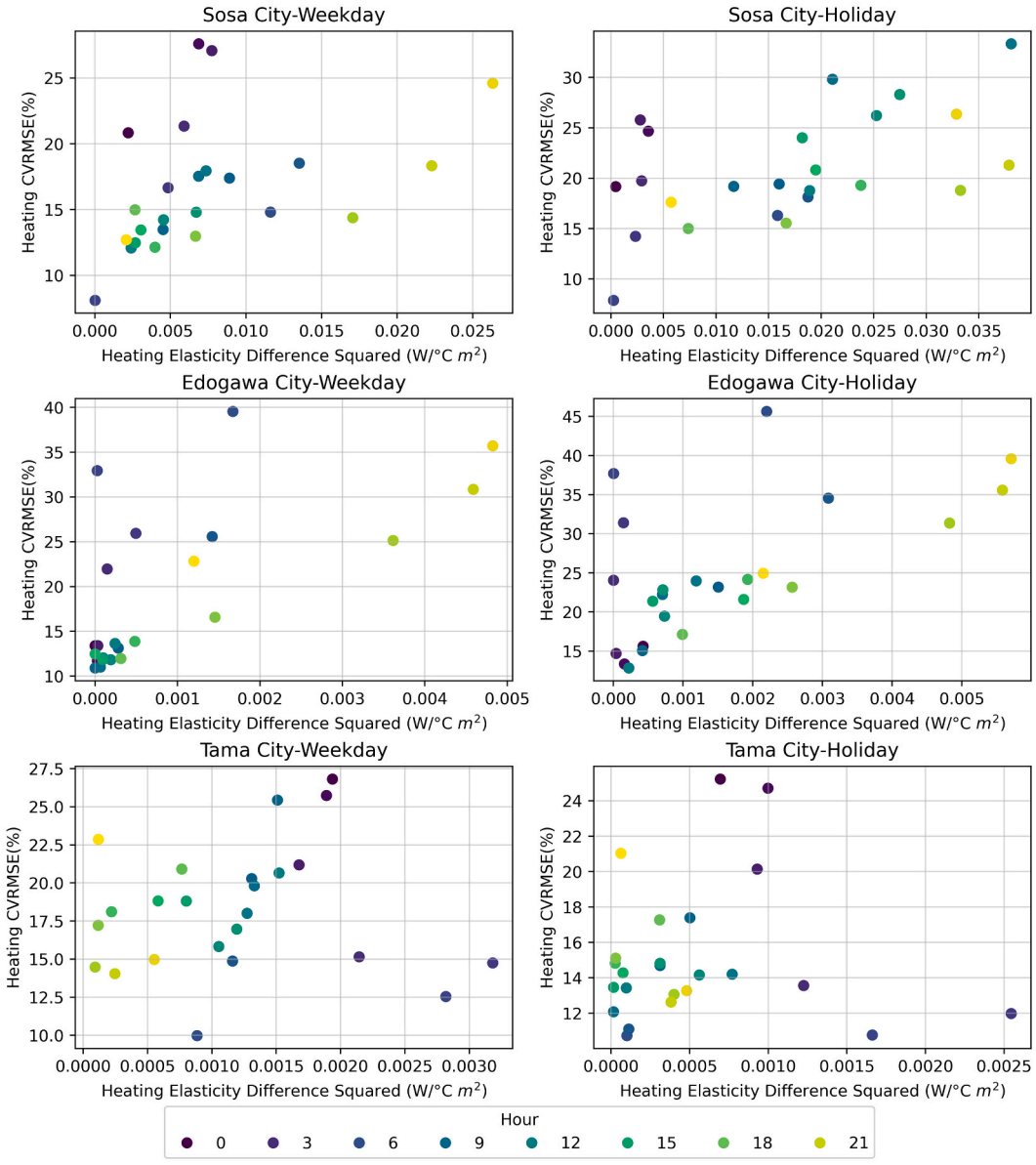



**Fig. D1.** CVRMSE (%) of electricity demand for heating hours, defined as hourly data with temperatures below the heating threshold for hourly measured data, compared against the difference in measured and simulated heating elasticity.

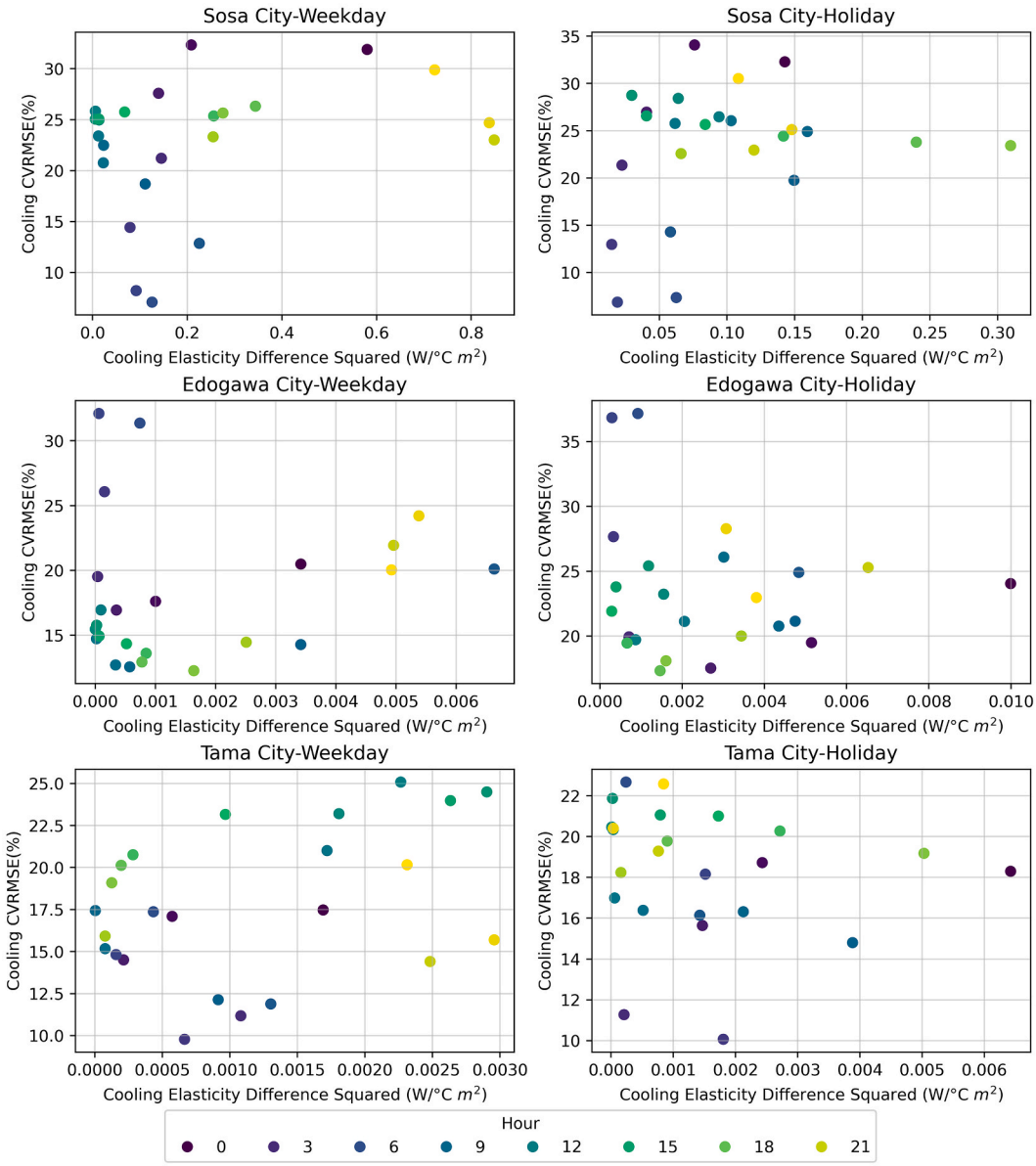



**Fig. D2.** CVRMSE (%) of electricity demand for cooling hours, defined as hourly data with temperatures above the cooling threshold for hourly measured data, compared against the difference in measured and simulated heating elasticity.

## Data availability

The authors do not have permission to share smartmeter data but model results data are available upon request.

## References


[1] M.C. Georgiadou, T. Hacking, P. Guthrie, A conceptual framework for future-proofing the energy performance of buildings, Energy Policy 47 (2012) 145–155, https://doi.org/10.1016/j.enpol.2012.04.039.

[2] Task Force on Climate-related Financial Disclosures, Final Report: Recommendations of the Task Force on Climate-related Financial Disclosures, 2017. https://www.fsb-tcfd.org/recommendations/.

[3] T. Hong, J. Malik, A. Krelling, W. O'Brien, K. Sun, R. Lamberts, M. Wei, Ten questions concerning thermal resilience of buildings and occupants for climate adaptation, Build. Environ. 244 (2023) 110806, https://doi.org/10.1016/j.buildenv.2023.110806.

[4] F. Soares, M.C. Silva, I. Azevedo, Urban decarbonization policies and strategies: a sectoral review, Renew. Sustain. Energy Rev. 215 (2025) 115617, https://doi.org/10.1016/j.rser.2025.115617.

[5] C. Wang, J. Song, D. Shi, J.L. Reyna, H. Horsey, S. Feron, Y. Zhou, Z. Ouyang, Y. Li, R.B. Jackson, Impacts of climate change, population growth, and power sector decarbonization on urban building energy use, Nat. Commun. 14 (2023) 6434, https://doi.org/10.1038/s41467-023-41458-5.

[6] K. Honjo, M. Fujii, Impacts of demographic, meteorological, and economic changes on household CO2 emissions in the 47 prefectures of Japan, Reg. Sci. Policy Pract. 6 (2014) 13–31, https://doi.org/10.1111/rsp3.12013.

[7] M. Santamouris, C. Cartalis, A. Synnefa, D. Kolokotsa, On the impact of urban heat island and global warming on the power demand and electricity consumption of buildings—A review, Energy Build. 98 (2015) 119–124, https://doi.org/10.1016/j.enbuild.2014.09.052.

[8] Y.Q. Ang, Z.M. Berzolla, C.F. Reinhart, From concept to application: a review of use cases in urban building energy modeling, Appl. Energy 279 (2020) 115738, https://doi.org/10.1016/j.apenergy.2020.115738.

[9] L. Dahlström, T. Broström, J. Widén, Advancing urban building energy modelling through new model components and applications: a review, Energy Build. 266 (2022) 112099, https://doi.org/10.1016/j.enbuild.2022.112099.

[10] Y. Li, H. Feng, Integrating urban building energy modeling (UBEM) and urban-building environmental impact assessment (UB-EIA) for sustainable urban development: a comprehensive review, Renew. Sustain. Energy Rev. 213 (2025) 115471, https://doi.org/10.1016/j.rser.2025.115471.

[11] D. Kong, A. Cheshmehzangi, Z. Zhang, S.P. Ardakani, T. Gu, Urban building energy modeling (UBEM): a systematic review of challenges and opportunities, Energy Effic. 16 (2023) 69, https://doi.org/10.1007/s12053-023-10147-z.

[12] Y.Q. Ang, Z. Berzolla, C. Reinhart, Smart meter-based archetypes for socioeconomically sensitive urban building energy modeling, Build. Environ. 246 (2023) 110991, https://doi.org/10.1016/j.buildenv.2023.110991.

[13] Y. Kusumoto, R. Delage, T. Nakata, Machine learning application for estimating electricity demand by municipality, Energy 296 (2024) 131138, https://doi.org/10.1016/j.energy.2024.131138.

[14] M. Ferrando, F. Causone, T. Hong, Y. Chen, Urban building energy modeling (UBEM) tools: a state-of-the-art review of bottom-up physics-based approaches, Sustain. Cities Soc. 62 (2020) 102408, https://doi.org/10.1016/j.scs.2020.102408.

[15] M. Hekkenberg, H.C. Moll, A.J.M.S. Uiterkamp, Dynamic temperature dependence patterns in future energy demand models in the context of climate change, Energy 34 (2009) 1797–1806, https://doi.org/10.1016/j.energy.2009.07.037.

[16] L.G. Swan, V.I. Ugursal, Modeling of end-use energy consumption in the residential sector: a review of modeling techniques, Renew. Sustain. Energy Rev. 13 (2009) 1819–1835, https://doi.org/10.1016/j.rser.2008.09.033.

[17] D. Papantonis, V. Stavrakas, D. Tzani, A. Flamos, Towards decarbonisation or lock-in to natural gas? A bottom-up modelling analysis of the energy transition ambiguity in the residential sector by 2050, Energy Convers. Manag. 324 (2025) 119235, https://doi.org/10.1016/j.enconman.2024.119235.

[18] Y. Shimoda, Y. Yamaguchi, Y. Iwafune, K. Hidaka, A. Meier, Y. Yagita, H. Kawamoto, S. Nishikiori, Energy demand science for a decarbonized society in the context of the residential sector, Renew. Sustain. Energy Rev. 132 (2020) 110051, https://doi.org/10.1016/j.rser.2020.110051.

[19] P. Kastner, T. Dogan, Towards auto-calibrated UBEM using readily available, underutilized urban data: a case study for Ithaca, NY, Energy Build. 317 (2024) 114286, https://doi.org/10.1016/j.enbuild.2024.114286.

[20] Y. Yamaguchi, Y. Shoda, S. Yoshizawa, T. Imai, U. Perwez, Y. Shimoda, Y. Hayashi, Feasibility assessment of net zero-energy transformation of building stock using integrated synthetic population, building stock, and power distribution network framework, Appl. Energy 333 (2023) 120568, https://doi.org/10.1016/j.apenergy.2022.120568.

[21] U. Ali, M.H. Shamsi, C. Hoare, E. Mangina, J. O'Donnell, Review of urban building energy modeling (UBEM) approaches, methods and tools using qualitative and quantitative analysis, Energy Build. 246 (2021) 111073, https://doi.org/10.1016/j.enbuild.2021.111073.

[22] L. Lefort, R. Bonabe de Rougé, P. Schetelat, T. Berthou, P. Riederer, B. Duplessis, E. Peirano, Development of a methodology of validation for urban building energy models and application to French residential consumption, in: Edenbourg (on line), United Kingdom, 2020. https://doi.org/hal-03202728.

[23] A. Oraiopoulos, B. Howard, On the accuracy of urban building energy modelling, Renew. Sustain. Energy Rev. 158 (2022) 111976, https://doi.org/10.1016/j.rser.2021.111976.

[24] Z. Deng, K. Javanroodi, V.M. Nik, Y. Chen, Using urban building energy modeling to quantify the energy performance of residential buildings under climate change, Build. Simul. 16 (2023) 1629–1643, https://doi.org/10.1007/s12273-023-1032-2.

[25] S. Erba, F. Causone, R. Armani, The effect of weather datasets on building energy simulation outputs, Energy Procedia 134 (2017) 545–554, https://doi.org/10.1016/j.egypro.2017.09.561.

[26] ASHRAE, ASHRAE Guideline 14-2014: Measurement of Energy, Demand, and Water Savings, 2014.

[27] J. Yuan, Z. Jiao, X. Xiao, K. Emura, C. Farnham, Impact of future climate change on energy consumption in residential buildings: a case study for representative cities in Japan, Energy Rep. 11 (2024) 1675–1692, https://doi.org/10.1016/j.egyr.2024.01.042.

[28] W. Hu, Y. Scholz, M. Yeligeti, Y. Deng, P. Jochem, Future electricity demand for Europe: unraveling the dynamics of the temperature response function, Appl. Energy (2024).

[29] W. Hu, Y. Scholz, M. Yeligeti, Y. Deng, P. Jochem, Future electricity demand for Europe: unraveling the dynamics of the temperature response function, Appl. Energy 368 (2024) 123387, https://doi.org/10.1016/j.apenergy.2024.123387.

[30] K. Nakajima, Y. Takane, S. Fukuba, K. Yamaguchi, Y. Kikegawa, Urban electricity–temperature relationships in the Tokyo Metropolitan Area, Energy Build. 256 (2022) 111729, https://doi.org/10.1016/j.enbuild.2021.111729.

[31] G. Ruiz, C. Bandera, Validation of calibrated energy models: common errors, Energies 10 (2017) 1587, https://doi.org/10.3390/en10101587.

[32] Y. Wang, J.M. Bielicki, Acclimation and the response of hourly electricity loads to meteorological variables, Energy 142 (2018) 473–485, https://doi.org/10.1016/j.energy.2017.10.037.

[33] P. Virtanen, R. Gommers, T.E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S.J. Van Der Walt, M. Brett, J. Wilson, K.J. Millman, N. Mayorov, A.R.J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E.W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E.A. Quintero, C.R. Harris, A.M. Archibald, A.H. Ribeiro, F. Pedregosa, P. Van Mulbregt, SciPy 1.0 Contributors, A. Vijaykumar, A. P. Bardelli, A. Rothberg, A. Hilboll, A. Kloeckner, A. Scopatz, A. Lee, A. Rokem, C. N. Woods, C. Fulton, C. Masson, C. Häggström, C. Fitzgerald, D.A. Nicholson, D. R. Hagen, D.V. Pasechnik, E. Olivetti, E. Martin, E. Wieser, F. Silva, F. Lenders, F. Wilhelm, G. Young, G.A. Price, G.-L. Ingold, G.E. Allen, G.R. Lee, H. Audren, I. Probst, J.P. Dietrich, J. Silterra, J.T. Webber, J. Slavič, J. Nothman, J. Buchner, J. Kulick, J.L. Schönberger, J.V. De Miranda Cardoso, J. Reimer, J. Harrington, J.L. C. Rodríguez, J. Nunez-Iglesias, J. Kuczynski, K. Tritz, M. Thoma, M. Newville, M. Kümmerer, M. Bolingbroke, M. Tartre, M. Pak, N.J. Smith, N. Nowaczyk, N. Shebanov, O. Pavlyk, P.A. Brodtkorb, P. Lee, R.T. McGibbon, R. Feldbauer, S. Lewis, S. Tygier, S. Sievert, S. Vigna, S. Peterson, S. More, T. Pudlik, T. Oshima, T.J. Pingel, T.P. Robitaille, T. Spura, T.R. Jones, T. Cera, T. Leslie, T. Zito, T. Krauss, U. Upadhyay, Y.O. Halchenko, Y. Vázquez-Baeza, SciPy 1.0: fundamental algorithms for scientific computing in Python, Nat. Methods 17 (2020) 261–272, https://doi.org/10.1038/s41592-019-0686-2.

[34] M. Chen, G.A. Ban-Weiss, K.T. Sanders, The role of household level electricity data in improving estimates of the impacts of climate on building electricity use, Energy Build. 180 (2018) 146–158, https://doi.org/10.1016/j.enbuild.2018.09.012.

[35] S. Choi, S. Yoon, Change-point model-based clustering for urban building energy analysis, Renew. Sustain. Energy Rev. 199 (2024) 114514, https://doi.org/10.1016/j.rser.2024.114514.

[36] T. Ihara, Y. Genchi, T. Sato, K. Yamaguchi, Y. Endo, City-block-scale sensitivity of electricity consumption to air temperature and air humidity in business districts of Tokyo, Japan, Energy 33 (2008) 1634–1645, https://doi.org/10.1016/j.energy.2008.06.005.

[37] Y. Hiruta, L. Gao, S. Ashina, A novel method for acquiring rigorous temperature response functions for electricity demand at a regional scale, Sci. Total Environ. 819 (2022) 152893, https://doi.org/10.1016/j.scitotenv.2021.152893.

[38] C.F. Jekel, G. Venter. pwlf: A Python Library for Fitting 1D Continuous Piecewise Linear Functions, 2019.

[39] Y. Shimoda, M. Sugiyama, R. Nishimoto, T. Momonoki, Evaluating decarbonization scenarios and energy management requirement for the residential sector in Japan through bottom-up simulations of energy end-use demand in 2050, Appl. Energy 303 (2021) 117510, https://doi.org/10.1016/j.apenergy.2021.117510.

[40] A. Taniguchi-Matsuoka, Y. Shimoda, M. Sugiyama, Y. Kurokawa, H. Matoba, T. Yamasaki, T. Morikuni, Y. Yamaguchi, Evaluating Japan's national greenhouse gas reduction policy using a bottom-up residential end-use energy simulation model, Appl. Energy 279 (2020) 115792, https://doi.org/10.1016/j.apenergy.2020.115792.

[41] Y. Yamaguchi, Y. Shimoda, A stochastic model to predict occupants' activities at home for community-/urban-scale energy demand modelling, J. Build. Perform. Simul. 10 (2017) 565–581, https://doi.org/10.1080/19401493.2017.1336255.

[42] Y. Shimoda, T. Asahi, A. Taniguchi, M. Mizuno, Evaluation of city-scale impact of residential energy conservation measures using the detailed end-use simulation model, Energy 32 (2007) 1617–1633, https://doi.org/10.1016/j.energy.2007.01.007.

[43] A. Taniguchi, T. Inoue, M. Otsuki, Y. Yamaguchi, Y. Shimoda, A. Takami, K. Hanaoka, Estimation of the contribution of the residential sector to summer peak demand reduction in Japan using an energy end-use simulation model, Energy Build. 112 (2016) 80–92, https://doi.org/10.1016/j.enbuild.2015.11.064.

[44] J. Butzbaugh, R. Hosbach, A. Meier, Miscellaneous electric loads: characterization and energy savings potential, Energy Build. 241 (2021) 110892, https://doi.org/10.1016/j.enbuild.2021.110892.

[45] Intended Nationally Determined Contributions (INDC): Greenhouse Gas Emission Reduction Target in FY2030, Minist. Foreign Aff. Jpn. (n.d.). https://www.mofa.go.jp/ic/ch/page1we_000104.html (accessed January 24, 2025).

[46] Ministry of the Environment, Japan's National Greenhouse Gas Emissions and Removals in FY2023 (Executive Summary), (2025). https://www.env.go.jp/en/press/press_04099.html.

[47] T. Hong, J. Langevin, K. Sun, Building simulation: ten challenges, Build. Simul. 11 (2018) 871–898, https://doi.org/10.1007/s12273-018-0444-x.

[48] J. Cao, M. Li, R. Zhang, M. Wang, An efficient climate index for reflecting cooling energy consumption: cooling degree days based on wet bulb temperature, Meteorol. Appl. 28 (2021) e2005.

[49] Y. Shimoda, Y. Yamaguchi, T. Okamura, A. Taniguchi, Y. Yamaguchi, Prediction of greenhouse gas reduction potential in Japanese residential sector by residential energy end-use model, Appl. Energy 87 (2010) 1944–1952, https://doi.org/10.1016/j.apenergy.2009.10.021.