Pacific Northwest
NATIONAL LABORATORY

PNNL-38371

Artificial Intelligence for Enhancing Multiscale Analysis

Buildings Focus

September 2025

Ying Zhang
Rachel Hoesly
Milan Jain
Meredydd Evans

U.S. DEPARTMENT of ENERGY

Prepared for the U.S. Department of Energy
under Contract DE-AC05-76RL01830

DISCLAIMER

PACIFIC NORTHWEST NATIONAL LABORATORY
operated by
BATTELLE
for the
UNITED STATES DEPARTMENT OF ENERGY
under Contract DE-AC05-76RL01830

Printed in the United States of America

Available to DOE and DOE contractors from
the Office of Scientific and Technical Information,
P.O. Box 62, Oak Ridge, TN 37831-0062
www.osti.gov
ph: (865) 576-8401
fox: (865) 576-5728
email: reports@osti.gov

Available to the public from the National Technical Information Service
5301 Shawnee Rd., Alexandria, VA 22312
ph: (800) 553-NTIS (6847)
or (703) 605-6000
email: info@ntis.gov
Online ordering: http://www.ntis.gov

PNNL-38371

Artificial Intelligence for Enhancing Multiscale Analysis

Buildings Focus

September 2025

Ying Zhang
Rachel Hoesly
Milan Jain
Meredydd Evans

Prepared for
the U.S. Department of Energy
under Contract DE-AC05-76RL01830

Pacific Northwest National Laboratory
Richland, Washington 99354

Abstract

This project aims to develop multi-scale building energy data, potentially improving the representation of the U.S. buildings sector in GCAM-USA, an U.S.-focused human-energy-Earth systems model. Existing building energy datasets are typically limited to national or regional levels, which constrains the ability of models to capture fine-scale human-energy-Earth systems interactions and reduces their relevance for decision-making on issues such as energy security, resilience, and energy planning. By leveraging AI and advanced data integration methods, this work fuses multiple existing datasets to enhance the physical and geographic representation of both residential and commercial building energy use. So far, progress includes processing residential building data, designing the data structure for commercial buildings, and testing AI approaches for integrating datasets and addressing spatial-temporal gaps. This effort can not only advances GCAM-USA's capability in modeling the buildings sector but also supports broader DOE missions, such as developing digital testbeds, enhancing grid resilience analysis, and improving building–energy system modeling at decision-relevant scales.

Summary

Key building energy drivers such as household income, building type, age, and urban-suburb-rural context are currently missing or simplified in GCAM-USA but are being prioritized in this effort. For residential buildings, income and demographics strongly shape energy demand, while commercial demand is more closely tied to regional economic activity. AI-based integration allows harmonizing fragmented datasets like ACS, RECS, and CBECS, filling gaps where no single dataset provides complete spatial and temporal coverage. Early validation shows strong performance in capturing overall energy use, while highlighting opportunities to refine specific fuels such as propane and fuel oil. Future work will expand the use of advanced AI techniques to fill spatial and temporal gaps, incorporate richer building-level features, and apply explainability and uncertainty quantification methods to improve transparency and robustness of estimates. These developments will help establish a comprehensive AI-driven framework for building energy modeling that supports fine-scale, decision-relevant applications.

Acknowledgments

This research was supported by the Earth and Biological Sciences Directorate Mission Seed Investment, under the Laboratory Directed Research and Development (LDRD) Program at Pacific Northwest National Laboratory (PNNL). PNNL is a multi-program national laboratory operated for the U.S. Department of Energy (DOE) by Battelle Memorial Institute under Contract No. DE-AC05-76RL01830.

Acronyms and Abbreviations

ACS	American Community Survey
AHS	American Household Survey
AI	Artificial Intelligence
CBECS	Commercial Buildings Energy Consumption Survey
DHS	Department of Homeland Security
DOE	Department of Energy
GCAM	Global Change Analysis Model
ML	Machine Learning
NHTS	National Household Transportation Survey
PUMA	Public Use Microdata Area
PUMS	Public Use Micro Data Sample
RECS	Residential Energy Consumption Survey

Contents

Abstract ii
Summary iii
Acknowledgments iv
Acronyms and Abbreviations v
1.0 Introduction 1
2.0 Key Drivers and Data 2
2.1 Residential Buildings Sector 4
2.2 Commercial Buildings Sector 5
3.0 Methodology 6
3.1 Bayesian Multivariate Regression of Household Energy Use 6
3.2 Posterior Inference 7
3.3 Weighted Aggregation to PUMA Level 8
4.0 Evaluation 9
4.1 Data Preprocessing 9
4.1.1 Housing Level Estimates: Link fusionACS to ACS 9
4.1.2 Handling Implicates in fusionACS 9
4.1.3 Feature Engineering 10
4.1.4 Data Preparation for Model Training and Validation 10
4.2 Model Validation: Housing-Level Energy Estimation by Fuel Type 10
4.2.1 Predictive Performance 11
4.2.2 Posterior Diagnostics and Residual Dependencies 11
4.2.3 State-Level Aggregation ..... **Error! Bookmark not defined.**
4.2.4 Overall Assessment 12
5.0 Future Work 13
6.0 References 14

Figures

Figure 1 Model prediction performance as measured through RMSE 11
Figure 2 Residual heatmap correlation (mean) 12

Tables

Table 1. Building service (supplysector), fuel (subsector), and technologies in GCAM-USA 2
Table 2. Model predictive performance in estimating log-normal energy consumption by fuel type as measured through RSME 10

1.0 Introduction

This project aims to integrate buildings data at multiple scales in order to build capabilities in modeling the buildings sector in the open-source Global Change Analysis Model (GCAM). Traditionally, energy-related buildings data are only available at the national or regional level, which limits our ability to accurately and quickly model human-energy-Earth systems at finer scales, reducing the relevance for GCAM as a decision making and planning tool around energy security related issues. This work would build our ability to model human-energy-Earth systems at multiple scales by leveraging Artificial Intelligence (AI) tools to fuse existing datasets that can be used to improve the physical and geographical representation of buildings energy consumptions in GCAM-USA, a version of GCAM focusing on the U.S. with more detail in the buildings sector.

In addition to working toward better buildings sector representation in GCAM-USA, the cohesive dataset developed in this project would be useful to other missions of the Department of Energy (DOE) or Department of Homeland Security (DHS). For example, the Office of Science's Biological and Environmental Research program is considering creating Digital Testbeds to integrate multiple different types of human and Earth science modeling at varying scales. This dataset supports digital testbeds by providing high-resolution, multi-scale information on building energy use, allowing models to more accurately capture the interactions between human systems and the environment. By leveraging AI or Machine Learning (ML) to fuse disparate datasets, it enhances the realism, scalability, and decision-relevance of testbed experiments, enabling more robust evaluation of energy scenarios. Furthermore, the proposal can provide useful input to the Office of Energy Efficiency and Renewable Energy by allowing them to access modeling of buildings integrated with other parts of the energy and Earth system at fine scales. For the Office of Electricity, this capacity could be useful in connecting one of the largest electricity loads (buildings) with power system models at scales relevant to resilience questions.

2.0 Key Drivers and Data

The buildings sector in GCAM-USA includes a detailed structure representing residential and commercial building energy use across services, fuels, and technologies. However, key parameters that drive the dynamics and heterogeneity of building energy service demand, such as income (for residential buildings), urban/suburban/rural disaggregation, building type, and building age, are either simplified or absent in the GCAM-USA core version. This work prioritizes enhancing GCAM-USA's capability in more accurately modeling the buildings sector, particularly through improvements of the representation of key drivers in building energy use. This work also works to increase the geographical detail of underlying GCAM-USA building data, to enable swift breakouts of sub-state GCAM-USA in the future.

In summary, GCAM-USA version 8.3, the latest version as of today, models two building types (residential and commercial), where 14 residential energy services and 10 commercial services are included for each state. Each energy service is associated with different fuel and technology choices (e.g., standard vs. high-efficiency gas furnace vs. electric heat pump for residential heating service). See Table 1 below. Different choices are associated with different technology characteristics (e.g., cost, efficiency, and lifetime). There is also building shell conductance being modeled over years, which is related to the building age, a parameter currently missing in GCAM-USA. The current version of GCAM-USA does not incorporate consumer-specific characteristics, such as income levels, although extensions exist in development model branches.

Table 1. Building service (supplysector), fuel (subsector), and technologies in GCAM-USA

supplysector	subsector	technology
resid heating	biomass	wood furnace
resid heating	coal	coal furnace
resid heating	gas	gas furnace
resid heating	gas	gas furnace hi-eff
resid heating	electricity	electric furnace
resid heating	electricity	electric heat pump
resid heating	refined liquids	fuel furnace
resid heating	refined liquids	fuel furnace hi-eff
resid cooling	electricity	air conditioning
resid cooling	electricity	air conditioning hi-eff
resid hot water	gas	gas water heater
resid hot water	gas	gas water heater hi-eff
resid hot water	electricity	electric resistance water heater
resid hot water	electricity	electric resistance water heater hi-eff
resid hot water	electricity	electric heat pump water heater
resid hot water	refined liquids	fuel water heater

resid hot water	refined liquids	fuel water heater hi-eff
resid lighting	electricity	incandescent
resid lighting	electricity	fluorescent
resid lighting	electricity	solid state
resid refrigerators	electricity	refrigerator
resid refrigerators	electricity	refrigerator hi-eff
resid freezers	electricity	freezer
resid freezers	electricity	freezer hi-eff
resid dishwashers	electricity	dishwasher
resid dishwashers	electricity	dishwasher hi-eff
resid cooking	electricity	electric oven
resid cooking	gas	gas oven
resid cooking	gas	gas oven hi-eff
resid cooking	refined liquids	lpg oven
resid cooking	refined liquids	lpg oven hi-eff
resid clothes dryers	electricity	clothes dryer
resid clothes dryers	electricity	clothes dryer hi-eff
resid clothes dryers	gas	clothes dryer
resid clothes washers	electricity	clothes washer
resid clothes washers	electricity	clothes washer hi-eff
resid televisions	electricity	electricity
resid computers	electricity	electricity
resid furnace fans	electricity	electricity
resid other	gas	gas
resid other	electricity	electricity
resid other	refined liquids	refined liquids
comm heating	biomass	wood furnace
comm heating	coal	coal furnace
comm heating	gas	gas furnace
comm heating	gas	gas furnace hi-eff
comm heating	electricity	electric furnace
comm heating	electricity	electric heat pump
comm heating	refined liquids	fuel furnace
comm cooling	gas	gas cooling
comm cooling	electricity	air conditioning

comm cooling	electricity	air conditioning hi-eff
comm hot water	gas	gas water heater
comm hot water	gas	gas water heater hi-eff
comm hot water	electricity	electric resistance water heater
comm hot water	electricity	electric heat pump water heater
comm hot water	refined liquids	fuel water heater
comm ventilation	electricity	ventilation
comm ventilation	electricity	ventilation hi-eff
comm cooking	gas	gas range
comm cooking	gas	gas range hi-eff
comm cooking	electricity	electric range
comm cooking	electricity	electric range hi-eff
comm lighting	electricity	incandescent
comm lighting	electricity	fluorescent
comm lighting	electricity	solid state
comm refrigeration	electricity	refrigeration
comm refrigeration	electricity	refrigeration hi-eff
comm office	electricity	office equipment
comm other	gas	gas
comm other	electricity	electricity
comm other	refined liquids	refined liquids
comm non-building	electricity	electricity

2.1 Residential Buildings Sector

To better capture residential building energy use, we identify several key drivers (Berrill et al. 2021), including energy use by technology, floorspace, building age (related to shell conductance parameter), income, building type (multifamily vs single family vs large apt buildings), area type (urban vs suburb vs rural identification). Although GCAM-USA's buildings sector operates at the state level, we aim to develop a dataset flexible in the scale at which we integrate parameters. It is worth noting that income deciles are implemented in the residential buildings sector in a model branch, where higher income generally leads to higher energy demand; however, technology choices are not linked to income levels. Therefore, future improvements, out of the scope of this LDRD, could include not only leveraging the income deciles in determining energy demand, but also refining behavioral parameters linked to income levels (e.g., resistance to electrification in low-income areas). Additionally, we could improve the shell conductance parameter based on building age data to better capture spatial and temporal variations in shell conductance and thus the thermal energy demand response.

The dataset we plan to develop will provide annual values over the historical years, aligning with observed data. This dataset will provide energy use by service and fuel (as does the current

GCAM-USA dataset) as well as more detailed geography, building-type, income, and urban-rural designation (if not noted by the geographic area). In addition, it would include estimates of statistical significance/error to support data needs in modeling and analysis.

To develop this dataset, we will leverage a combination of U.S. datasets with national coverage and AI-based data integration methods. The open-source fusionACS dataset (Ummel et al. 2024), which integrates the American Community Survey (ACS) Public Use Micro Data Sample (PUMS) (US Census Bureau 2016) with the Residential Energy Consumption Survey (RECS) (US Department of Energy 2018) and others, enables us to examine energy use and fuel/technology choice patterns across income groups at fine spatial scale (Public Use Microdata Area or PUMA level). This dataset will be used to evaluate how income affects technology adoption and energy intensity in residential buildings. We will also compare overlapping variables such as floorspace and building age per building type across datasets of RECS for matching geographies and years to assess consistency and fill in gaps. To integrate these datasets over time and across spatial levels (PUMA, state, and census region), we will design AI-based data integration methodology to perform spatial and temporal interpolation and extrapolation and generate consistent and comprehensive estimates of these key variables. The harmonized dataset will serve as inputs to refine GCAM-USA's residential buildings types.

2.2 Commercial Buildings Sector

The approach to developing the dataset for commercial building energy use follows that for the residential buildings described above, with some important distinctions. First, the spatial resolution for commercial building energy use is expected to be coarser than for residential buildings (e.g., PUMA level), because the backbone dataset, Commercial Buildings Energy Consumption Survey or CBECS (US Department of Energy 2022), provides statistically robust estimates only at the census division level. To address this, we plan to integrate CBECS with ModelAmerica (New et al. 2021) data and apply AI-based methods to refine the spatial resolution where possible. Second, while household income is a strong determinant of residential energy demand, it is not a meaningful driver for commercial buildings. Instead, indicators such as regional GDP or economic activity levels will be used to represent commercial activity and associated energy consumption. Finally, commercial building energy consumption is generally less sensitive than residential demand to fluctuations in energy prices. To better understand these dynamics, we will apply sensitivity analysis to evaluate the relative importance of input variables on commercial energy demand outcomes.

3.0 Methodology

Estimating household and building energy use by fuel type at fine geographic scales is inherently difficult because the variables most critical to energy demand, such as floor area, building type, vintage, and primary heating fuel, are never jointly observed in a single dataset at national scale. For instance, ACS provides rich demographic and socioeconomic information but contains only coarse measures of building characteristics. On the other hand, energy surveys such as RECS and CBECS capture end-use and fuel consumption in detail but are limited in sample size and lack geographic coverage below census region or division. Existing datasets (ACS, RECS, CBECS, ResStock, ComStock, ModelAmerica) independently capture demographics, buildings' characteristics, and employment statistics, none of these datasets can provide estimates of building floor area and building type by fuel type and other relevant features.

AI offers a powerful opportunity to fill these gaps by learning from the joint distributions of known features, such as location, building age, land use, or demographics, to generate plausible estimates of missing attributes like building type or floor area. By learning from the overlapping margins across disparate datasets, AI models can infer plausible values for unobserved combinations of building and household features. In addition to that, these models offer an added advantage of capturing the temporal dynamics, a key element needed for GCAM-USA modeling. In comparison to traditional statistical models, these AI models can capture non-linear dynamics and possess the ability to integrate multimodal data (such as satellite imagery, parcel data, and urban morphology features) to infer building typologies or use statistical relationships from well-surveyed regions to impute floor area in less-documented areas.

Our approach leverages a Bayesian multivariate regression framework that unifies demographic, socioeconomic, and building characteristics at the household level and then aggregates these estimates to the PUMA scale.

3.1 Bayesian Multivariate Regression of Household Energy Use

Let $i \in \{1, \dots, N\}$ index households, and $j \in \{1, \dots, 4\}$ index fuel types: natural gas (NG), fuel oil (FO), liquefied petroleum gas/propane (LP), and electricity (EL). For each household i, predictors: $\mathbf{x}_i \in \mathbb{R}^K$ is a vector of household and building level covariates (income, building age, structure type, household size, etc.); and observed y_{ij} is the annual energy use (Btu) for fuel j. Since residential energy use is highly skewed, we transform the target variable by taking the log normal.

$$z_{ij} = \log(1 + y_{ij})$$

We posit a Bayesian multivariate normal regression model:

$$\mathbf{z}_i = \begin{bmatrix} z_{i1} \\ z_{i2} \\ z_{i3} \\ z_{i4} \end{bmatrix} \sim \mathcal{N}_4(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}),$$

with mean

$$\boldsymbol{\mu}_i = \alpha + \mathbf{B}^{\mathrm{T}}\mathbf{x}_i,$$

where,

- $\boldsymbol{\alpha} \in \mathbb{R}^4$ are intercepts for each fuel,
- $\mathbf{B} \in \mathbb{R}^{K \times 4}$ are regression coefficients linking predictors to each fuel,
- $\boldsymbol{\Sigma} \in \mathbb{R}^{4 \times 4}$ is the residual covariance matrix capturing correlations across fuels.

We assign weakly informative priors to regression coefficients to stabilize estimation without imposing overly restrictive assumptions

$$\alpha_j \sim \mathcal{N}(0, 2^2), \qquad B_{kj} \sim \mathcal{N}(0, 1^2)$$

for all predictors k and fuels j.

Weakly informative priors help prevent pathological estimates in high-dimensional settings (e.g., extremely large positive or negative coefficients that are inconsistent with plausible household energy use), while avoiding the rigidity of strongly informative priors. These priors reflect the belief that most effects are likely small to moderate on the log-energy scale, but they still allow the data to dominate inference when strong evidence is present.

For the covariance, we use an LKJ prior:

$$\Sigma = \mathbf{L}\mathbf{L}^{\mathrm{T}}, \qquad \mathbf{L} \sim \mathrm{LKJCholeskyCov}(\eta = 2, \sigma_j \sim \mathrm{HalfNormal}(1))$$

which yields marginal standard deviations σ_j and a correlation matrix with weakly informative prior toward independence. The covariance matrix Σ allows residuals across fuels (NG, FO, LP, EL) to be correlated: households that consume more electricity may systematically consume less natural gas (substitution) or more (complementarity).

The covariance matrix must be symmetric, positive definite, and estimated from the noisy data, and if put naïve priors directly on the covariance elements, we risk invalid or unstable covariance matrices. The LKJ prior is a principled distribution over correlation matrices and ensures that any draw is a valid correlation matrix and the prior shape can be tuned with a single parameter η. An LKJ prior with $\eta = 1$ places a uniform distribution over all possible correlation, treating strong positive, strong negative, and near-zero correlation as equal likely. Values of $\eta > 1$ mildly favor correlations closer to zero (independence), while values of $\eta < 1$ favor extreme correlations near ± 1. We set the LKJ prior parameter to $\eta = 2$ and by doing so we express a weak prior belief that independence is slightly more plausible than strong correlations, while still allowing the data to reveal strong cross-fuel relationships if present. This aligns well with domain knowledge: some fuels may act as substitutes (e.g., natural gas v/s electricity), while others are only weakly related (e.g., propane and electricity).

3.2 Posterior Inference

Once trained, we obtain posterior samples of $\{\alpha, B, \Sigma\}$ using Hamiltonian Monte Carlo (NUTS) as implemented in PyMC. For each posterior draw s, we compute household-level posterior predictive means:

$$\hat{y}_{ij}^{(s)} = \exp\left(\mu_{ij}^{(s)}\right) - 1,$$

where $\mu_{ij}^{(s)}$ is the sampled regression mean on the log scale. Additionally, we generate full posterior draws from the multivariate normal likelihood to incorporate residual variation.

3.3 Weighted Aggregation to PUMA Level

Each ACS household record carries a survey weight w_i indicating the number of households it represents in the population. Let $g(i)$ denote the PUMA (and optionally building category) group to which household i belongs. For each group G and fuel j, we compute weighted posterior totals:

$$T_{G_j}^{(s)} = \sum_{i \in G} w_i \hat{y}_{ij}^{(s)}$$

The posterior distribution of T_{Gj} across draws *s* provides point estimates (posterior mean or median) and confidence intervals (e.g., 10^{th} – 90^{th} percentiles) for annual fuel-specific energy use at the PUMA level, further stratified by building category when required.

4.0 Evaluation

Our work builds upon the work of fusionACS and ResStock/ComStock – efforts focused on enriching household, demographic, and socioeconomic characteristics captured in ACS with household-level energy consumption. In this section, we describe the data preprocessing steps, the experimental setup, and validation of our multivariate energy consumption prediction model at the housing level. The study is conducted for the state of WA, with results aggregated to the PUMA level and validated against state-level estimates of energy use by fuel type.

4.1 Data Preprocessing

fusionACS uses ACS microdata as the recipient backbone and fuses in variables from donor surveys such as RECS, American Household Survey (AHS), National Household Transportation Survey (NHTS), and the Consumer Expenditure Survey. This enriches ACS records with fuel consumption, expenditures, and related attributes. By contrast, ResStock and ComStock are large-scale, physics-based building stock models developed by NREL to characterize the U.S. residential and commercial building sectors. They integrate survey data, building physics, and regional parameters to generate high-resolution estimates of energy use, retrofit potential, and technology adoption.

4.1.1 Housing Level Estimates: Link fusionACS to ACS

fusionACS includes a persistent household identifier, ACS house ID (*acs_hid*), that links fusionACS record to the recipient dataset (ACS 5Y 2011-2015) microdata. Because ACS 5Y products pool multiple years, the same household may appear with estimates for different years. Since our objective is to produce household-level energy estimates, we index records by *acs_hid* and retain the most recent estimate within 2011-2015.

Also, ACS microdata for 2011 refers PUMA 2000 boundaries, while 2012 onward uses PUMA 2010 boundaries. To provide consistent geography, we standardize all households to PUMA 2010. For 2011 records (originally linked to PUMA 2000), we use GeoPandas to map each household to the best-matching PUMA 2010 polygon by intersecting PUMA 2000 and PUMA 2010 shapefiles, then assigning the area-dominant PUMA 2010. This produces a single, consistent PUMA10 key for aggregation and reporting.

4.1.2 Handling Implicates in fusionACS

fusionACS provides 40 implicates per household to reflect fusion uncertainty. In this study, we take the average across implicates to create a single working value per household. In future versions, we will treat implicates as draws from a posterior predictive distribution and propagate them through through model fitting and PUMA aggregation to produce principled uncertainty intervales that reflect both model parameters and fusion uncertainty.

After these steps, each row in the modeling frame represents a unique household (identified by acs_hid) with (i) its most recent estimates (2011-2015), (ii) a single implicate – averaged for fused variable, and (iii) a harmonized PUMA10 identifier. ACS household weights (WGTP) are retained to scale each sampled household to the population. In modeling, households are treated as independent records and treat WGTP as a post-modeling scaling factor.

4.1.3 Feature Engineering

Finally, we harmonize categorical variables and derive numeric features that better represent building stock characteristics.

- **Building type (BLD)** from ACS is collapsed into three categories – *single-family*, *multi-family*, and *others* – with non-response coded as *unknown*.
- **Number of units** in multi-family structures is approximated using midpoints for categorical ranges (e.g., 3-4 apartments → 4 units, 20-49 apartments → 35 units).
- **Year built (YBL)** from ACS is mapped to a representative construction year using midpoints for categorical ranges (when an exact year was unavailable), then converted to building age (AGE) by subtracting from the reference year 2025.
- **Heating fuel (HFL)** is collapsed into four categories – *natural gas*, *propane*, *electricity*, and *oil/kerosene*. All other fuels are grouped as *others*; non-response is coded as *unknown*.
- Continuous housing features such as **number of bedrooms (BDSP)**, **number of rooms (RMSP)**, and **number of persons in family (NPF)** were preserved as numeric inputs.
- **Household income (HINCP)** is inflation-adjusted to constant 2015 dollars using ACS's adjustment factor (ADJINC).

4.1.4 Data Preparation for Model Training and Validation

To prepare the data for model training, we begin by splitting the dataset into randomly sampled 70:30 training and testing split. The training set container 63,589 observations and the test set 27,253 observations. The split is performed with stratification based on BLD category and HFL category, ensuring that the relative distribution of these categorical groups is maintained across both subsets. Stratification helps prevent bias by making sure that the test set is representative of the overall population, rather than disproportionately containing more or fewer samples from certain building or household fuel categories. For preprocessing, we apply one-hot encoding to categorical variables and numeric features are standardized using *Standard Scaler*, which transforms each numeric feature to have zero mean and unit variance.

4.2 Model Validation: Housing-Level Energy Estimation by Fuel Type

We trained the model using MCMC sampling with two chains and 2000 posterior draws after a 2000-step tuning phase.

Table 2. Model predictive performance in estimating log-normal energy consumption by fuel type as measured through RSME

	Train	Test
Natural Gas	4.88	4.85
Oil/Petroleum	3.27	3.29

Kerosene Oil	3.74	3.75
Electricity	0.44	0.44

4.2.1 Predictive Performance

Table 2 summarize the root mean squared error (RMSE) which captured the absolute magnitude of prediction errors. On the training set, RMSE values ranged from 0.44 (Electricity) to 4.88 (Natural Gas). Oil/Petroleum and Kerosene Oil fell between these extremes. The close alignment of train and test performance indicates that the model generalizes well to unseen data.

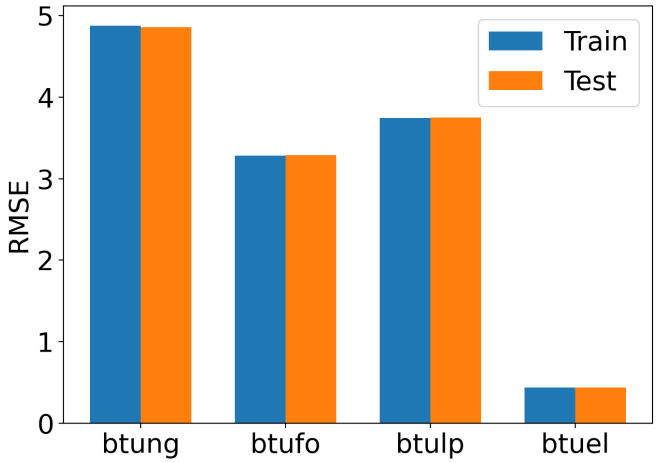

Figure 1 Model prediction performance as measured through RMSE

4.2.2 Posterior Diagnostics and Residual Dependencies

Posterior convergence was assessed using the Gelman-Rubin $\hat{R}$ statistic, effective sample size (ESS), and the number of divergent transitions. The maximum $\hat{R}$ observed was 1.0, with effective sample sizes for both bulk and tail distributions exceeded 4,500 and no divergent transitions were recorded. These diagnostics indicated that the sampler mixed well, providing reliable posterior uncertainty estimates.

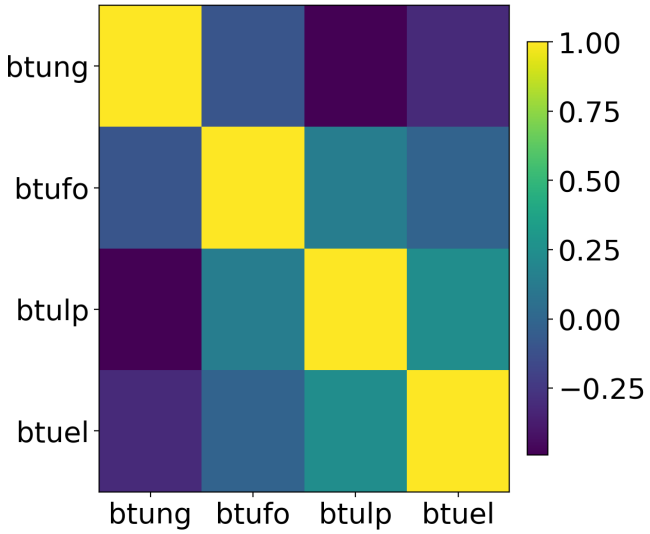

Figure 2 Residual heatmap correlation (mean)

Analysis of residual correlations revealed some remaining structure across predictors and fuels (see Figure 2). The correlation matrix shows the degree to which pairs of variables move together. Each cell in the matrix contains a correlation coefficient, which ranges from -1 to +1. A value close to +1 indicates that the two variables tend to increase or decrease together (positive association), while a value close to -1 indicates that one variable tends to decrease when the other increases (negative association). Values near 0 suggest little or no linear relationship between the variables.

In this study, the correlation matrix is used to examine the residuals – the differences between observed and predicted energy consumption. If the model captures all systematic patterns, the residuals should behave like random noise, and the correlations across different fuels or predictors should be close to zero. However, if we observe blocks of higher positive or negative correlations, it suggests that certain relationships remain unexplained by the model. For example, strong residual correlation between two fuels might indicate that household using those fuels share unmodeled characteristics (such as building envelope type or heating equipment efficiency) that systematically affect energy use.

Heatmaps of residual correlations indicated that dependencies were most pronounced for specific fuel categories, particularly fuel oil and propane, suggesting that the current model although captures broad-scale variation, it does not fully explain heterogeneity in all fuel-use pattern. Incorporating additional building-level features (e.g., insulation, roof type) or hierarchical structures may help reduce residual dependencies.

4.2.3 Overall Assessment

Overall, the evaluation of the prediction model highlighted that the model was able to capture broad variation in household energy consumption and achieved excellent posterior convergence. Electricity was modeled with relatively low proportional errors, while other energy sources remain more difficult due to variability in usage and limited representation in the dataset. The absence of convergence issues strengthens confidence in inference, but residual structure points to opportunities for extending the model specification.

5.0 Future Work

Ongoing efforts focus on processing residential building datasets and finalizing the data structure design for commercial buildings. Future work can include explorations of various AI methods for integrating the datasets and filling spatial-temporal data gaps. A comprehensive and detailed framework of how to use AI methods to model building energy consumptions can be designed in the future, including data processing, model training, sensitivity analysis, and performance evaluation. Future work can also include exploration of statistical significance of the estimates across varying geographies (e.g., PUMAS, counties).

Specifically, within AI, multiple extensions are possible. An important area is addressing spatial and temporal gaps in the data. The existing model relies on a limited set of features (present in ACS) to predict energy consumption by fuel type; however, prior studies have shown that additional building-level characteristics, such as building envelope type, roof type, insulation, are equally important drivers of consumption patterns, which are unfortunately not captured in the ACS. Moreover, GCAM-USA requires estimates over time, which add a temporal dimension to the prediction challenge. Generative models, including variational autoencoders or diffusion-based approaches, can provide much more realistic imputations by leveraging correlations across geography and time to fill these gaps. Such methods could leverage multiple diverse datasets, ranging from household surveys and building stock models to remote sensing and climate records, to produce spatially and temporally complete estimates.

Advanced representation learning techniques, such as graph neural networks or knowledge graph embeddings, could be employed to explicitly capture relationships between building characteristics, energy use, and contextual factors. By modeling these interdependencies, it becomes possible to move beyond treating each building as an isolated unit and instead represent the broader system of connections (e.g., shared infrastructure, neighborhood-level demographics, and climate influences). This richer representation would allow one to identify clusters of buildings with similar consumption profiles and detect structural vulnerabilities.

Beyond predictions, future efforts could also focus on designing a comprehensive modeling framework that incorporate explainability methods to interpret model drivers. Providing transparent reasoning behind estimates, such as identifying whether household income, building age, or equipment type drives observed energy use, would help improve feature engineering to further improve the model. Lastly, uncertainty quantification through Bayesian deep learning or ensemble techniques will be essential to assess robustness, particularly when generalizing to underrepresented building types or regions. By attaching credible intervals to forecasts, decision-makers can explicitly account for risk and variability, reducing the chance of overconfidence.

6.0 References

Berrill, Peter, Kenneth T Gillingham, and Edgar G Hertwich. 2021. “Drivers of Change in US Residential Energy Consumption and Greenhouse Gas Emissions, 1990–2015.” *Environmental Research Letters* 16 (3): 034045. https://doi.org/10.1088/1748-9326/abe325.

New, Joshua, Mark Adams, Anne Berres, Brett Bass, and Nicholas Clinton. 2021. “Model America.” April 14. https://doi.org/10.13139/ORNLNCCS/1774134.

Ummel, Kevin, Miguel Poblete-Cazenave, Karthik Akkiraju, et al. 2024. “Multidimensional Well-Being of US Households at a Fine Spatial Scale Using Fused Household Surveys.” *Scientific Data* 11 (1): 142. https://doi.org/10.1038/s41597-023-02788-7.

US Census Bureau. 2016. “2015 American Community Survey 5-Year Estimates Public Use Microdata Sample.” Washington D.C. Census.gov.

US Department of Energy. 2018. “2015 Residential Energy Consumption Survey (RECS) Microdata.” December. https://www.eia.gov/consumption/residential/.

US Department of Energy. 2022. “2018 Commercial Building Energy Consumption Survey Microdata.” December. https://www.eia.gov/consumption/commercial/data/2018/index.php?view=microdata.

Pacific Northwest National Laboratory

902 Battelle Boulevard
P.O. Box 999
Richland, WA 99354

1-888-375-PNNL (7665)

www.pnnl.gov