scientific data

Check for updates

OPEN

DATA DESCRIPTOR

High resolution synthetic residential energy use profiles for the United States

Swapna Thorve[1,2] ✉, Young Yun Baek[1], Samarth Swarup[1], Henning Mortveit[1,3], Achla Marathe[1], Anil Vullikanti[1,2] & Madhav Marathe[1,2] ✉

Efficient energy consumption is crucial for achieving sustainable energy goals in the era of climate change and grid modernization. Thus, it is vital to understand how energy is consumed at finer resolutions such as household in order to plan demand-response events or analyze impacts of weather, electricity prices, electric vehicles, solar, and occupancy schedules on energy consumption. However, availability and access to detailed energy-use data, which would enable detailed studies, has been rare. In this paper, we release a unique, large-scale, digital-twin of residential energy-use dataset for the residential sector across the contiguous United States covering millions of households. The data comprise of hourly energy use profiles for synthetic households, disaggregated into Thermostatically Controlled Loads (TCL) and appliance use. The underlying framework is constructed using a bottom-up approach. Diverse open-source surveys and first principles models are used for end-use modeling. Extensive validation of the synthetic dataset has been conducted through comparisons with reported energy-use data. We present a detailed, open, high resolution, residential energy-use dataset for the United States.

Background & Summary

Modernization of the U.S. electric grid is occurring at a noteworthy rate due to the installation of new technologies within the grid such as smart meters. They enable two-way communication between the customer and utilities, providing information and granular control of power usage for individual households[1,2]. The grid is also witnessing rapid transformations due to increasing penetration of electric vehicles (EV) and distributed energy resources (DER) such as rooftop photovoltaics (PV), community solar, and wind energy. While this wave of modernization is beneficial, the electric grid is simultaneously facing a sharp increase in crisis situations as a result of climate change phenomena[3,4] such as extreme weather events and global warming. One example of extreme weather is the February 2021 North American cold wave that caused a tremendous strain on the power grid especially in Texas where millions lost power for days[5]. Another example is where global warming impacts household HVAC energy use. Although the rise of 1° to 2 °C in winter temperatures is expected to decrease heating requirements, a similar rise in summer temperatures is expected to increase cooling needs significantly[6].

In the face of these challenges, achieving sustainable energy goals has become paramount for maintaining a healthy grid. To this end, the research community is faced with important questions regarding reduction of carbon footprints[7–11], incentivizing DER adoption[12], studying benefits of building energy retrofit[9,13,14], integration of electric vehicles[15] and consumer behavior[16] in the grid, and mechanisms for designing electricity pricing[17,18] to create efficient residential consumption patterns. Answering many of these questions requires comprehensive knowledge of energy-use patterns, building stock, the structure of distribution networks, consumer behaviors, and so on. However, such exhaustive datasets are rarely freely available (or available at all) for research use, making it hard for the research community to pursue these endeavours[19]. Reasons for unavailability of such data range from privacy concerns to the lack of a system for making data available to researchers.

Most of the published energy use data are metered data, a result of longitudinal studies conducted by researchers (Table 1) with relatively small samples of households that may not be representative of the wider

[1]Network Systems Science and Advanced Computing, Biocomplexity Institute and Initiative, University of Virginia, Charlottesville, USA. [2]Department of Computer Science, University of Virginia, Charlottesville, USA. [3]Department of Engineering Systems and Environment, University of Virginia, Charlottesville, USA. ✉e-mail: st6ua@virginia.edu; marathe@virginia.edu

Authors/Dataset	Description
Klemanjak *et al.*[26,75]	A synthetic energy demand dataset was released for 21 appliances in Austria in 2020. Data collected from two households was used to train models and then appropriate noise was added for appliance start times and durations to mimic variations in actual consumption patterns.
Kolter *et al.*[76,77]	The Reference Energy Disaggregation Data Set (REDD) is published by MIT. The dataset contains high-frequency current/voltage waveform data of the power mains in households along with labeled circuits in the house.
Makonin *et al.*[78]	The Rainforest Automation Energy (RAE) dataset was published by Harvard in 2017. The dataset contains 1 Hz data (mains and sub-meters) from two residential houses.
Murray *et al.*[79,80]	Load measurements from 20 households of UK from a two year longitudinal study.
Pecan Street[22,23]	Labeled circuit data for households across major cities in the U.S. This is said to be the most comprehensive dis-aggregate energy data available for the U.S.
Rashid *et al.*[81,82]	The I-blend dataset has recorded minute-level consumption of all the buildings at an academic institute in India over a period of 52 months
Paige *et al.*[83,84]	The flEECe dataset provides energy data at a 1 Hz sampling rate for four circuits for six net-zero energy senior housing units in Virginia, USA for nine months
Shin *et al.*[85,86]	The first Korean dataset measuring appliance-level energy data was released in 2019 for 22 houses in Korea.
Kelly *et al.*[20,87]	Power demand is recorded from five houses UK houses at two levels – whole house and individual appliances. This dataset is referred to as the UK-Dale dataset. Two versions of this dataset have been released.
Anderson *et al.*[88,89]	Building-Level fUlly-labeled dataset for Electricity Disaggregation (BLUED) for one household in Pittsburg U.S. for one week. State transition of appliances are labeled and time-stamped, providing the necessary ground truth for the evaluation of NILM algorithms.
Barker *et al.*[90,91]	Electricity usage data is monitored every minute from nearly every plug load from 400 anonymous homes.
Beckel *et al.*[92]	Electricity consumption is monitored via smart plugs for six households in Switzerland over a period of 8 months.
Pereira *et al.*[93–95]	Power usage for 44 apartments and 6 homes in Portugal is collected for 264 days at 30 minute intervals. The advanced version of this dataset 'SustDataED2' dataset contains 96 days of aggregated and individual appliance consumption from one household in Portugal.
Monacchi *et al.*[96,97]	Common household devices are monitored for power consumption in Austria and Italy (GREEND dataset).
Pullinger *et al.*[98,99]	1-second electricity data is gathered over a period of 23 months from 255 UK homes (IDEAL household energy dataset).
Ruhnau *et al.*[100,101]	Synthetic national time series of heat demand that covers over 16 countries in the EU from 2008 to 2018.

Table 1. Energy-use datasets published in the residential sector.

geographical region and demographics. Some of these studies monitor households over a longer period of time (e.g. two years), however, the downside of such experiments is that it takes a considerable amount of time (e.g. participant consent, equipment setup, monitoring) and manual effort (e.g., data cleaning, imputing missing values) before such data is usable. Although these studies release energy data for free use, many of them limit publishing participant details (e.g. building characteristics and location, household level demographics). Participant details are usually withheld due to privacy reasons/participant consent, lack of information, or unavailability of these attributes in the free version of the data. Literature has attempted to address some of these issues by creating appropriate data structures for releasing appliance metadata information for households alongwith their energy use data[20,21]. However, we observe that many of the issues still persist in the U.S. context. One such example is the Pecan Street Dataport[22]. Pecan Street Inc[23]. is the largest publisher of energy-use data in the U.S. through their portal – *Dataport*. They collect energy-use data in California (CA), Texas (TX), New York (NY), and Colorado (CO). This is a potentially very useful data set. However, only a small sample (~25 households in CA and TX) of energy-use data is freely available for public use and do not contain sufficient (or any) demographic or building information.

A dataset synthesized over a larger spatial scope offers the opportunity to study regional and temporal differences in energy use while a smaller region dataset offers studying energy use patterns that may be particular to the region. Irrespective of spatial scope, small sample size makes it difficult to get a good representation of the population variation in the region (e.g. explaining/exploiting role of household demographics, behavior, and building characteristics in energy use). In addition to the spatial scope and number of samples, many of the datasets do not release sufficient (or any) participant details. Such limited data restricts the usage of these energy-use data for detailed practical analyses or studying scenario interventions and equity questions in the grid (e.g., which type of demographic and building stock is best suited for EV adoption, or how much carbon footprint can be reduced by retrofitting buildings). Thus, we observe that there is a general sparsity of large scale high resolution energy use datasets along with detailed metadata information at household level such as appliance ownership, building data, important demographic features.

We summarize key drawbacks of energy datasets for the U.S. as follows – limited spatial scope, small sample size, lack of sufficient household, appliance, & building metadata. Given these wide array of problems with the state-of-art energy-use data availability, we introduce synthetic energy use datasets that are able to address many of these issues. Synthetic data is defined as data generated by models that provide accurate statistical representations of the real world. Examples of such data for the smart grid are synthetic power distribution networks[24], energy consumption profiles for offices and commercial buildings[25] and for residential buildings[26–29]. Our work specifically addresses the data scarcity gap in energy use research for the U.S. residential sector. We propose a synthetic framework for modeling large-scale high resolution energy use data by integrating diverse datasets and end-use models for bottom-up dis-aggregate energy modeling. This results in a novel synthetic energy use dataset (i.e., a digital twin of household level energy demand) comprising hourly electrical energy demand profiles

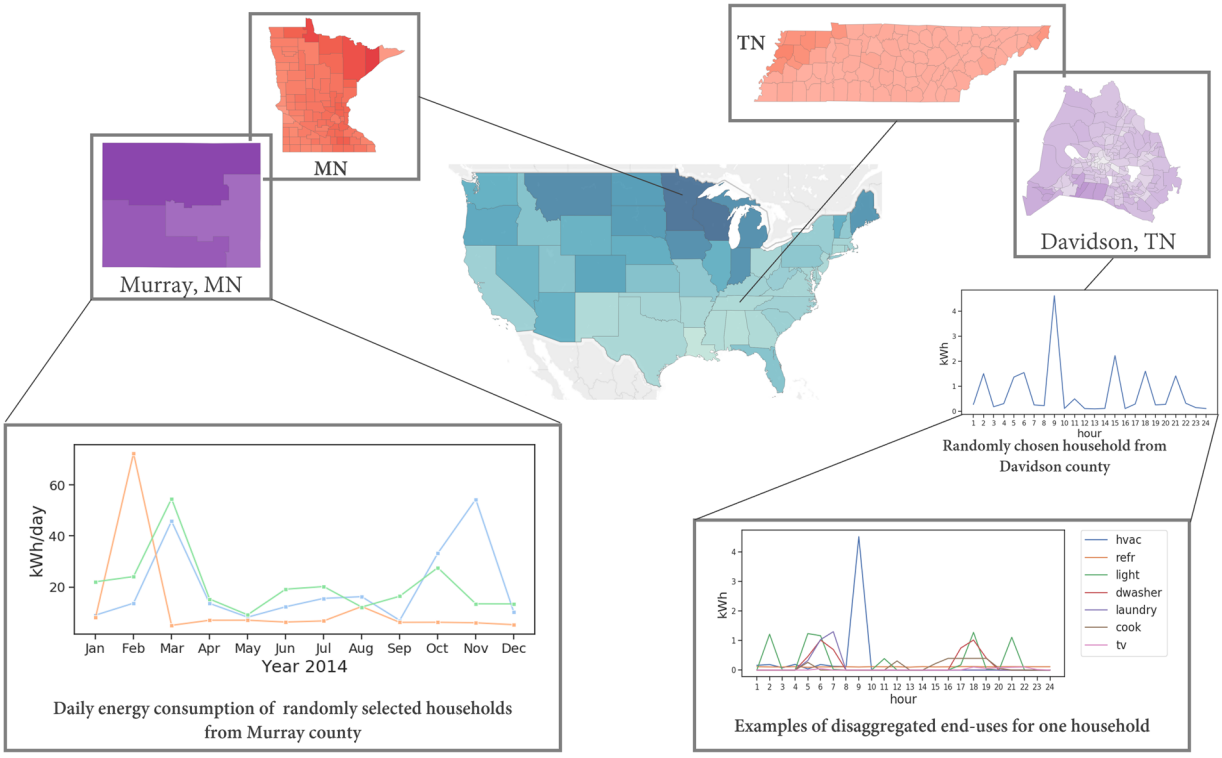


Fig. 1 Data overview. This figure shows examples of the spatio-temporal resolutions of multiple facets of the dis-aggregated synthetic energy demand data. The figure shows sample data at state, county, and household level at different temporal granularities. The data is generated for all households in the U.S.

for U.S. households. The total electrical energy use is published as a composition of eight primary end-uses in a household – heating/air-conditioning (HVAC), lighting, dishwashing, cooking, laundry (clothes washer and clothes dryer), refrigeration, hot water, and miscellaneous plug load (vacuuming, computer use, TV). A detailed data-intensive bottom-up framework is developed to generate synthetic energy-use profiles by integrating multiple open-source surveys and a synthetic population for the U.S[30]. A mixture of methods (stochastic, machine learning, physics-based engineering methods) is used to model different end-uses in all households that consume electricity as a primary fuel across the 48 contiguous states and Washington, D.C. in North America. To the best of our knowledge, this synthetic energy-use dataset is the first detailed, large-scale, freely available household-level electricity consumption behaviors dataset for the U.S. Our synthetic energy-use infrastructure is well-suited to solve the newer smart grid problems mentioned earlier. We publish the dis-aggregated energy use timeseries for all the synthetic households. The published data is representative of the U.S. households, provide household level metadata, and are a good representation of the real world energy use. Fig. 1 provides a graphic illustration of the synthesized residential energy demand digital twin.

Methods

This section describes the datasets and models employed to generate synthetic energy use time series at the household level, see Table 2. All notations used in the paper are described in Table 3.

The presented framework is composed of a synthetic representation of the U.S. population, regression models for surveys, and bottom-up energy use models. A synthetic population is composed of households and people in households. The synthetic households are generated using census surveys and statistical methods such that the synthetic population is *statistically similar* to the original population. An open-source version of the U.S. synthetic population – Synthetic Populations and Ecosystems of the World (SPEW)[30,31] is used in our framework. The SPEW synthetic population is comprised of demographic characteristics of synthetic households and synthetic individuals. The synthetic population is created using U.S. census data such as PUMS (Table 2) and statistical methods such as sampling and the Iterative Proportional Fitting (IPF) method[32].

The SPEW households are made of basic demographic (e.g., income, age) and locality information. Although the SPEW population is representative of the U.S. population on a finer spatial resolution, it is not equipped with energy and activity related information (e.g., building characteristics, time spent at home, number of cooking activities) necessary for estimating energy use at household level or person level. Building stock, energy and activity related information is collected by national surveys in the U.S. – Residential Energy Consumption Survey RECS[33] and American Time Use Survey ATUS[34] respectively. The basic synthetic population is augmented with energy and activity related attributes by building machine learning models. This augmentation is called as the *enrichment step*. The enriched synthetic population along with other freely available data sources can be used together as inputs to the energy use modeling framework. The energy use modeling framework has

Dataset	Description
American Time Use Survey (ATUS 2015)	ATUS provides nationally representative estimates of how, where, and with whom people in the U.S. spend their time, and is the only federal survey providing data on the full range of activities, from childcare to volunteering. This survey provides demographic information as well as information on energy-related activities[34]. 24-hour data is recorded for 5115 participants.
Synthetic Populations and Ecosystems of the World (SPEW)	SPEW[30,31] is a framework that produces synthetic populations for various countries. We used the open-sourced version of the synthetic population available for the U.S. constructed for the year 2013. The sampled base population is the byproduct of American Community Survey (ACS) Public Use Microdata Sample (PUMS) data. Statistical methods such as Simple Random Sampling (SRS) and Iterative Proportional Fitting (IPF)[102,103] are used to estimate joint distributions of population characteristics given their marginal distributions at a small geographic level (e.g. PUMA-level for the U.S.). Data records are available at household level for all of U.S. Descriptors are available for mapping records from PUMS data onto the base synthetic population.
Public Use Microdata Sample (PUMS 2013)	PUMS is a 5% representative sample for a larger region than block group referred to as a Public Use Microdata Area (PUMA)[104]. PUMAs are described by the Census as "a collection of counties or tracts within counties with more than 100,000 people". These statistical areas are defined for the circulation of PUMS data. PUMS contains individual records of the characteristics for a 5% sample of people and their households. One PUMS record is a complete Census record.
North American Land Data Assimilation System (NLDAS)	Hourly temperature data for North America. Data resolution is at 1/8th-degree grid over North America[105].
Residential Energy Consumption Survey (RECS 2015)	U.S. Energy Information Administration (EIA) Residential Energy Consumption Survey (RECS)[33] data is a national sample survey that collects energy-related data for housing units. For 2015, data was collected from 5,686 households to represent 118.2 million U.S. households. We use this dataset to obtain housing unit-specific information such as floor area, main heating fuel, fuel equipment, indoor temperature setting, presence of air conditioner, dishwasher, washer, dryer, refrigerator, water heater fuel, water heater size, water heater age, number of lighting units, etc,.
National Solar Radiation Database (NSRDB)	NREL provides solar radiation data for the U.S. We use hourly data that comes from the physics-based approach called the Physical Solar Model (PSM). Data is available for the U.S. for 1998–2014[106]. The GHI variable is used as an indicator of irradiance level in the lighting model. GHI is modeled solar radiation on a horizontal surface received from the sky. This is measured in $\frac{\text{watt}}{\text{meter}^2}$.
Miscellaneous	Appliance power and efficiencies, gallons of hot water required for activities, and any other input data required for models is drawn from surveys and data collected from ground and/or testing[50,51,62,64]

Table 2. List of primary datasets used for constructing the residential demand models.

six models for representing nine energy uses – HVAC, lighting, domestic hot-water, refrigerator, dishwasher, cooking, clothes washer, clothes dryer, and miscellaneous plug load such as TV, computer use, cleaning activities (e.g., vacuuming). The first subsection describes the modeling details of the *enrichment step* and the following subsection describes energy demand models.

Enrichment models. The enrichment models support creating comprehensive synthetic structures for calculating residential energy usage. This step is called as the *enrichment step*. Refer to Fig. 2 for a pictorial representation of the overview of the framework. Datasets used in this workflow are described in Table 2. Since the demographic features available in the synthetic population are not sufficient for computing energy usage, it is made richer by adding layers of information related to building stock and energy consumption from the RECS survey such as building characteristics, appliance ownership, and thermostat set-point behaviors. This mapping of features is made by building inference tree models. Activity schedules for a normative day of an ATUS survey respondent are attached to synthetic individual by building a multivariate random forest regression model. These models are described below.

The ATUS model. The ATUS data provides nationally representative surveys of people's activities in different location types such as childcare in or outside the house, time spent at work, laundry time at home, waiting times in hospital, and so on, see Table 2 for a description. The time-use diaries of the survey individuals can be attached to synthetic individuals by matching an appropriate survey individual to a synthetic individual. In our work, we consider *appropriate matching* based on amount of time a person spends in different location types such as home, work, school, shopping, and other miscellaneous locations. This seems a reasonable approach because we are interested in learning how an individual spends 24 hours of the day by categorizing the amount of time spent at important location types – for e.g., the time spent in different location types for a person works full-time is quite different than a house bound senior citizen or a college student. This rationale of assigning survey respondents to synthetic individuals is also presented in prior work by Lum *et al.*[35].

Random forest regression method is used to build a model that predicts the amount of time a person spends in locations types such as home, work, shopping, other, school, and trip counts during the day. Thus, six dependent variables are modeled – trip count during the day and time spent at each location type - home, work, shopping, other, school. Independent variables used to build the model are as follows – number of members in the household (hsize), number of children (nchild), age (age), working hours (wrkhrs), gender (gender), income modeled as a categorical variable (hinc2, hinc3), and binary variables such as an American citizen or not (nativity), worker or not (worker), owns home or not (ownhome), has a phone or not (tel), and race related variables such as if person is white, Hispanic, black, or Asian (white, hispanic, black, asian). Figure 3 shows example of feature importance for two dependent variables.

Once the model is trained on ATUS respondents, a synthetic person $P_{i,j}$ is randomly assigned a survey individual from the leaf nodes in the trained ensemble model. Thus, the result gives every synthetic individual a

Notation	Description
H_i	Household i drawn from the synthetic population
$P_{i,j}$	Synthetic household member j of household H_i
A_k	Respondent k from ATUS survey
S_l	Household l from RECS survey
Irr^{i}	Irradiance threshold for H_i. Drawn from a Normal distribution Normal(60,10)
$\langle O_{i,0}, \ldots, O_{i,t}, \ldots, O_{i,23}\rangle$	Occupancy time series of synthetic household i over 24 hours, $t \in \{0, 1, \ldots, 23\}$
$\langle \text{Irr}_0, \ldots, \text{Irr}_t, \ldots, \text{Irr}_{23}\rangle$	Hourly irradiance time series of a census tract for a given day in the year 2014
$\langle T_0^{\text{out}}, \ldots, T_t^{\text{out}}, \ldots, T_{23}^{\text{out}}\rangle$	Hourly temperature series of the outside environment for a given day (°F)
$\langle T_0^{\text{in}}, \ldots, T_t^{\text{in}}, \ldots, T_{23}^{\text{in}}\rangle$	Thermostat setpoint (°F)
η	Efficiency of the HVAC equipment and water heaters
$R^{\text{roof}}, R^{\text{wall}}$	Thermal resistance coefficient for roof and wall for different climate zones
T_v^{hot}	Temperature (°F) of hot water end-point category v,
	where $v \in$ {shower, bath, cwasher, dishwasher}
$T_{m,z}^{\text{cold}}$	Mains water temperature (°F) for month m and climate zone z
$d \in \text{D}$	End-use $d \in$ D where D = {hvac, h2o, light, refr, dwasher, cook, cwasher, cdryer, TV, computer, cleaning}
$\langle E_{i,0}^d, E_{i,t}^d, \ldots, E_{i,23}^d\rangle$	Hourly energy use profile of H_i for a end-use d and $t \in \{0, \ldots, 23\}$
E_i^d	Daily energy consumed over 24 hours by end-use d in household H_i. $E_i^d = \sum_{t=0}^{23} E_{i,t}^d$ and $d \in$ D and $t \in \{0, 1, \ldots, 23\}$
$\langle G_{i,0}^{\text{h2o}}, G_{i,t}^{\text{h2o}}, \ldots, G_{i,23}^{\text{h2o}}\rangle$	Hourly profile of hot-water use (gallons per hour) of H_i for a end-use h2o and $t \in \{0, \ldots, 23\}$. $G_{i,t}^{\text{h2o}} = \sum_{v \in \text{V}} G_{i,t,v}^{\text{h2o}}$ where V = {shower, bath, dishwasher, clotheswasher}
G_i^{h2o}	Daily amount of hot water consumed (in gallons) by a household H_i in a day. $G_i^{\text{h2o}} = \sum_{t=0}^{23} G_{i,t}^{\text{h2o}}$
$G_{i,v}^{\text{h2o}}$	Daily amount of water consumed (in gallons) by a household H_i in a day by an event v. $G_{i,v}^{\text{h2o}} = \sum_{t=0}^{23} G_{i,t,v}^{\text{h2o}}$

Table 3. Notations.

time-use diary. The energy-use models will extract home activities from a time-diary and also build a household-level occupancy schedule over the 24-hour duration, denoted as $\langle O_{i,0}, O_{i,1}, \ldots, O_{i,23}\rangle$. These are used as an input to the energy use models. Synthetic household member activity scheduling conflicts are handled in the activity model.

The RECS mapping model. The baseline synthetic population does not have any building structural characteristics and appliance ownership information. These salient features are important for modeling different categories of energy use and are available in the RECS survey. We overlay RECS household attributes onto a synthetic household by building multivariate conditional inference trees[36,37]. Conditional inference tree is a non-parametric class of regression trees that uses recursive partitioning of dependent variables based on the value of correlations. Four dependent variables are modeled – square footage of the dwelling, presence of laundry appliances, presence of air conditioner, presence of dishwasher. The independent variables are year in which the house was built, occupancy time of the current tenants, own or rent the residence, total number of rooms, income, number of refrigerators, number of members in the household, dwelling type, dwelling is located in urban or rural area, primary heating fuel type. The independent variables are common attributes between RECS survey records and synthetic household records. Conditional inference trees are trained on different census regions in the U.S. to tease out regional differences. A RECS household S_i is randomly selected from the appropriate leaf nodes of the conditional inference tree and assigned to the synthetic household H_i every time a new simulation is run. This dynamic assignment introduces stochasticity when the simulation is executed for same and/or different days.

Energy use modeling. The enriched synthetic population (i.e., the output of the *enrichment step*) enables encoding of behaviors (time spent in different energy related activities at home), normative attributes (e.g., square footage, age, income, gender), declarative attributes (e.g., individual activities as a sequence) and procedural attributes (e.g., behaviors capturing dependencies, interactions, frequency of performing activities) into the knowledge required for building energy use profiles[38]. The synthetic infrastructure is leveraged to build six energy use models (Fig. 2). Nine end-uses are synthesized for each household. These end-uses are divided into two parts – Thermostatically Controlled Loads (TCL) and appliance use. For a household i, nine end-uses published in the data are –

1. **HVAC** (E^{hvac}). This category includes heating and cooling electric load from central air conditioning during hot days and electric furnace/heater used during cold days. This is a TCL load.
2. **Domestic hot water use** (E^{h2o}). Energy consumed for heating water that is needed for personal grooming activities such as shower/bath, laundry activities such as using clothes washer, and dishwasher. This is a TCL load.
3. **Dishwasher** (E^{dwasher}). Energy used by dishwashers.

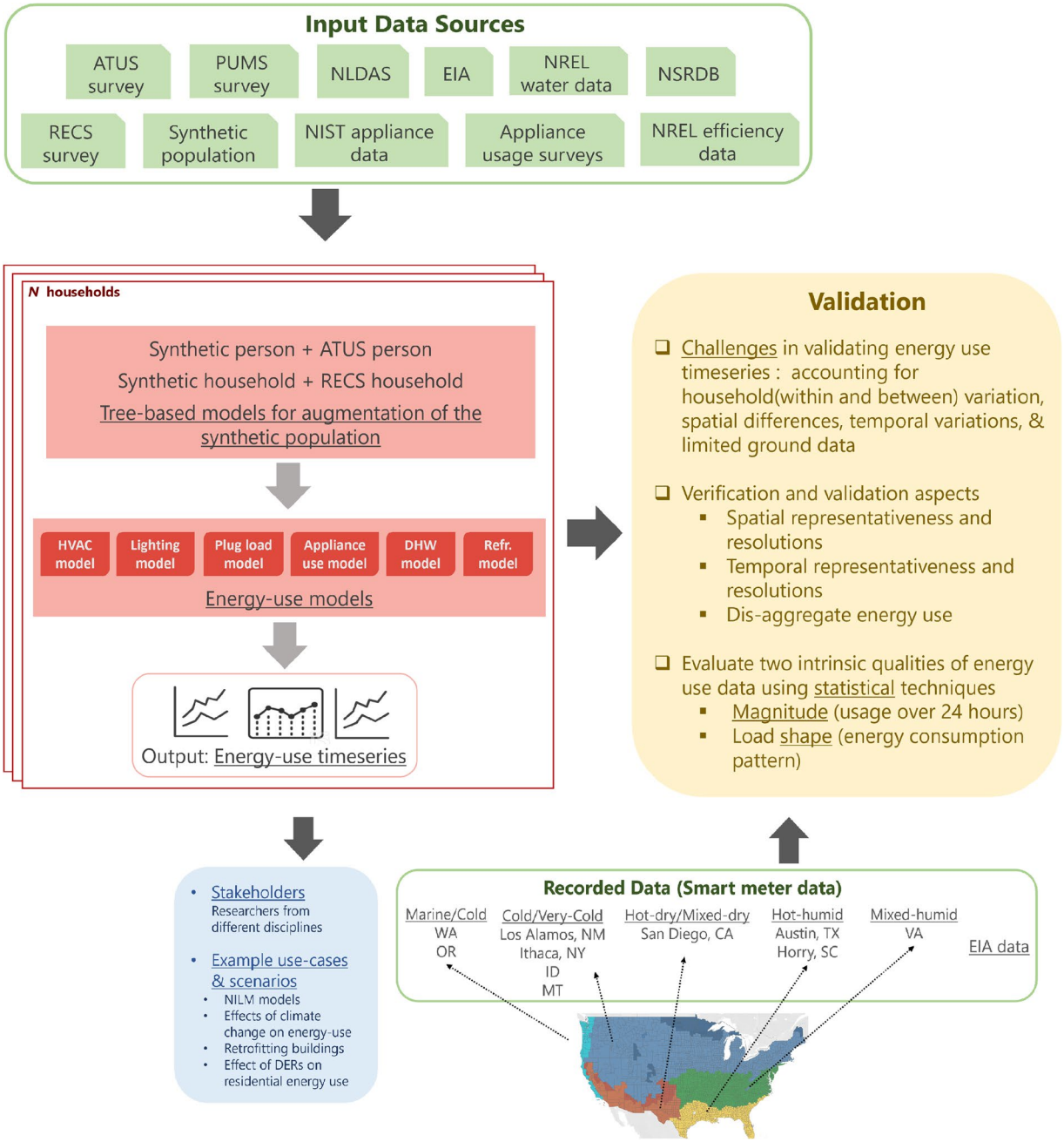


Fig. 2 Overview of the energy modeling infrastructure. Many different types of input data are used in the proposed modeling framework. These are shown at the top. For complete description of input datasets refer to Table 2. These datasets are input to different modeling components of the framework. Some datasets support augmentation of the synthetic population while others are input to the energy-use models. All the models are described in the *Methodology* section. The bottom rectangle describes the recorded data/smart meter data from different climate zones of the U.S. These datasets are used for validation of the synthetic energy-use timeseries. The validation block (yellow backdrop) describes three components of V&V - regional, magnitude, and structural/shape comparisons. This line of validation covers (**a**) different temporal aspects (hourly and daily), (**b**) spatial aspects in terms of regions and seasons, (**c**) diversity aspect of the large-scale synthetic data. The blue text refers to the V's of big data. Each colored block possesses the given V characteristic.

4. **Clothes Washer** ($E^{cwasher}$). Energy used by electric clothes washers.
5. **Clothes Dryer** (E^{cdyer}). Energy consumed by dryer.
6. **Cooking** (E^{cook}). Energy consumed by electric cooking range, oven, and other kitchen appliances such as coffee maker, microwave, toaster, etc.
7. **Miscellaneous plug load** (E^{misc}). This type of energy indicates plug load attributed to cleaning activities and electronic devices such as TV, computers, other smaller electronic gadgets.

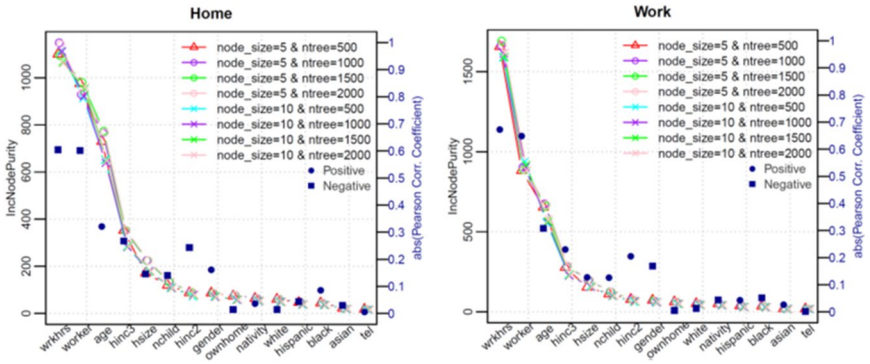


Fig. 3 Impurity-based feature importance and correlation. Each plot shows Gini importance of features for two dependent variables – home and work. The x-axis shows independent variables in order of importance based on *IncNodePurity*. The selection of the parameters for 'ntree' (number of decision trees) and 'node size' (minimum size of terminal nodes). Eight conditions are tested for the combination of the two parameters: ntree = 500, 1000, 1500, and 2000; node size = 5, and 10. The plots show robust results across the different conditions. According to the plots, the following five independent variables - wrkhrs; worker; age; hinc3; hsize mostly affect all the dependent variables. The right-hand y-axis shows the absolute Pearson Correlation Coefficient. The positive and negative coefficients are distinguished by blue dots and squares, respectively. Except wrkhrs; worker, all other independent variables weakly correlated with the dependent variables.

8. **Refrigeration** (E^{refr}). Energy consumed by refrigerators.
9. **Lighting** (E^{light}). Energy consumed by lighting units.

Table 3 describe the notations used in the methodology sections. The total energy summed over 24 hours (E_i^{total}) of a household i is given by the equations below –

$$E_i^{\text{total}} = E_i^{\text{TCL}} + E_i^{\text{appliances}} \tag{1a}$$

$$E_i^{\text{TCL}} = E_i^{\text{hvac}} + E_i^{\text{h2o}} \tag{1b}$$

$$E_i^{\text{appliances}} = E_i^{\text{dwahser}} + E_i^{\text{cook}} + E_i^{\text{cwasher}} + E_i^{\text{cdryer}} + E_i^{\text{light}} + E_i^{\text{refr}} + E_i^{\text{misc}} \tag{1c}$$

$$E_i^{\text{misc}} = E_i^{\text{tv}} + E_i^{\text{computer}} + E_i^{\text{cleaning}} \tag{1d}$$

HVAC model E^{hvac}. According to the U.S. Energy Information Administration (EIA)[39], HVAC is responsible for the highest proportion of energy consumption in households. The HVAC model calculates how much energy is required to maintain ambient/comfort temperature indoors. This is dependent on factors ranging from the area of the house, outdoor temperature, efficiency of HVAC equipment, and so on. Occupant behaviour of thermostat settings in different seasons and household occupancy during the day play an important role in understanding thermal comfort levels and how its effect on electricity consumption. Engineering and statistical approaches[40] are presented in the literature to simulate energy consumption of heaters/furnace and air conditioners[41–44]. We adopt the engineering based approach from Subbiah *et al.*[44] where the function of heating/cooling a household H_i at hourly intervals is defined as:

$$E_{i,t}^{\text{hvac}} = \frac{\Delta T}{\eta} \times \left(\frac{FloorArea_i}{R^{\text{roof}}} + \frac{WallArea_i}{R^{\text{wall}}} \right) \tag{2}$$

Here $E_{i,t}^{\text{hvac}}$ is the energy consumed by household H_i at the end of hour t in kWh by heating/cooling equipment to maintain thermal comfort. $FloorArea_i$ is the floor area and $WallArea_i$ is the wall area (extrapolated from floor area[44]) of H_i. The quantities R^{roof} and R^{wall} are R-values (insulation level) for households in different climate zones, while η is defined in Table 3. Next, ΔT is the absolute difference between T_t^{in} and T_t^{out}, and T_t^{in} is indoor thermostat temperature at hour t. The hourly outside temperature (T_t^{out}) is obtained from NOAA NLDAS data mentioned in Table 2. Efficiency and insulation data is obtained from guidelines published by EIA. All other household attributes are obtained from the enriched synthetic population. Depending upon occupancy patterns throughout the day, changes in thermostat behaviors are assigned to each household. Heating and cooling threshold temperatures for appliance on/off times are taken from the thermostat study published by NREL in 2017[45].

Event v	Range of T_v^{hot} (F)	Flow rate (gpm) μ, σ, distribution	Duration (minutes) μ, σ, distribution
Shower	[105, 116]	2.25, 0.68, Normal	7.81, 3.52, Normal
Bath	[105, 116]	4.40, 1.17, Normal	5.65, 2.09, Normal
Dishwasher	[120, 140]	1.39, 0.20, Normal	1.53, 0.41, LogNormal
Clothes washer	[60, 130]	2.20, 0.62, Normal	3.05, 1.62, Normal

Table 4. Hot water model characteristics.

Domestic Water Heating Model E^{h2o}. The EIA shows that 17%–32% of the household energy use is attributed to domestic hot water use (DHW). Literature shows models used for estimating hot water demand at multiple temporal resolutions – annual, daily, hourly, and minute intervals. One of the initial models for estimating load profiles of hot water demand was developed in 2001 by Jordan *et al.*[46] for a period of one year for temporal resolutions of 1 min, 6 min, and 1 hour. However, this work does not consider historical nor factual flow rates to determine how much hot water (gallons/day) is used by a household. A follow-up paper was developed for synthesizing water demand profiles for Switzerland[47] by calibrating this model using field data. A model to simulate yearly DHW event schedule for a single-family household was developed by Hendron *et al.*[48] from the National Renewable Energy Laboratory (NREL) in 2010. The simulator used two surveys that collected information about water demand in U.S. households for five categories: sink, bath, shower, clothes washer, and dishwasher. This model has been widely accepted in the literature. One recent example of the adaptation of Hendron's model is for simulating hot water demand in Canadian households[49]. The model is calibrated for survey data collected for Canada and appropriate adjustments are made with respect to Canadian lifestyles.

For our model, we use the distributions of duration and flow rates of activities involving hot water usage such as bath/shower, clothes washer, and dishwasher from Hendron *et al.* Note that duration and flow rates can take negative values (Table 4). The flow rate is capped to 0.05gpm and the duration is capped to 1 minute for any negative value[48]. Table 4 characterizes the average count of daily events, duration, and flow rates. The values of hot water temperature for different uses and the cold water inlet temperature are obtained from studies conducted by NREL in different regions of U.S[50–52]. An engineering based approach is used to estimate hot water usage[44,50] in household i for event v at time t

$$E_v^{\text{hot}} = \frac{G_{v,i,t}^{\text{hot}} \times \Delta T}{\eta} \times 0.00189, \quad \text{where}$$
$$G_{v,i,t}^{\text{hot}} = \text{duration}_v \times \text{flow\_rate}_v, \quad \text{and} \quad \Delta T = T_{m,z}^{\text{cold}} - T_v^{\text{hot}}. \tag{3}$$

The gallons of hot water $G_{v,i,t}^{\text{hot}}$ consumed by event v is computed as a product of flow_rate (gpm) and duration (minutes). Both these characteristics are drawn from distributions in Table 4. E_v^{hot} is the energy consumed by the event v to heat G_v^{hot} gallons of water. Last four entries in the Table 3 shows summation of multiple events occurring across the time horizon. Here η is the efficiency of the electric water heaters. Surveys conducted by NREL have shown that η is a complex function of storage capacity of water heater, type of water heater, age of water heater. No distributions are available for η in the current studies. Field data collected from NREL surveys[50–52] show that the efficiency varies anywhere between 80%–99%. Here 0.00189 $\left(\frac{\text{kWh}}{\text{gal}^\circ\text{F}}\right)$ is a conversion constant obtained from Subbiah *et al.*[44], and ΔT is the temperature difference (°F) between mains (inlet) water temperature $T_{m,z}^{\text{cold}}$ for a given month m in a climate zone z and the water temperature required for a particular end-point. The values for $T_{m,z}^{\text{cold}}$ and T_v^{hot} are obtained from NREL surveys[50,51]. Whenever the activity model detects the presence of an event v, we calculate the energy used by hot-water for the event using Eq. 3. Note that we compute hot water energy usage only for synthetic households having electric water heaters.

Lighting E^{light}. Lighting accounts for 5–10%[39] of the consumption with lighting usage in residential setting mainly characterized by outdoor lighting conditions and occupancy schedules in households[53]. A Markov-chain approach is adopted by Widen *et al.*[54] for modeling lighting demand in Swedish households using time use data in Sweden. A stochastic model is developed for residential lighting estimation for the city of Cordova in Spain by Palacios-Garcia[55] based on a model developed by Stokes *et al.*[56] using measured lighting data for 100 UK homes. Another stochastic model is developed by Richardson *et al.*[57] for UK households using time-use data and lighting data from the Energy Information Administration(EIA).

We build a stochastic model for lighting demand in U.S. dwellings by building on design concepts from work done by Richardson *et al.*[57], Stokes *et al.*[56], and Paatero & Lund *et al.*[58]. Richardson's model is particularly interesting since it supports important characteristics of light usage such as 'co-use' and 'relative weights'. The model uses the concept of 'co-use' of lighting, i.e., lighting in a dwelling is often shared by household members in the same space of the dwelling at the same time. The model also considers that all lighting units are not used at the same frequency (e.g. frequently occupied rooms such as kitchen space and living area will use more lighting than other rooms) and employs a weighting scheme to indicate relative usage.

Outdoor lighting conditions are modeled using irradiance time series. It is obtained from NSRDB described in Table 2. Hourly irradiance data is collected using the NSRDB API for the 365 days of the year 2014 at census tract resolution for the U.S. Thus, all synthetic households in a census tract use the same irradiance time series for a given day. The household level hourly occupancy profile $\langle O_{i,0}, O_{i,1}, \ldots, O_{i,23} \rangle$ is developed by examining

activities of awake synthetic household members of H_i at home. Presence of awake occupants in the dwelling support the decision making of light switch-on event. The distribution of lighting units in households are derived from the RECS survey. In general, distribution of lighting units of a H_i is taken from the matching S_i. Three types of lighting units are considered: incandescent, CFL, and LED. Power ratings of lighting unit categories are taken from a study conducted by the Bonneville Power Administration (U.S.) where lighting fixtures were analyzed for a sample of 161 Northwest residences[59]. For a given simulation day, we define an irradiance threshold ($\mathrm{Irr}^{\mathrm{i}}$) for a household H_i. It indicates that occupants may consider switching on lights when outdoor lighting is less than $\mathrm{Irr}^{\mathrm{i}}$. $\mathrm{Irr}^{\mathrm{i}}$ is sampled from a normal distribution[57] Normal(60, 10). All notations used in the model are described in Table 3. Annual lighting data for the U.S. is summarized for different household sizes from the RECS survey.

Literature shows that lighting usage increases by number of occupants in the household, however, the lighting usage does not double for every occupant added in the house. In order to simulate shared lighting usage, the concept of effective occupancy[57] of a household $\langle \hat{O}_{i,0}, \hat{O}_{i,t}, \ldots, \hat{O}_{i,23} \rangle$ is introduced. Effective occupancy ($\hat{O}_{i,t}$) is defined as a function of active occupancy ($O_{i,t}$). The values for effective occupancy are derived by scaling the annual lighting demand by household size such that the effective occupancy of a dwelling with one active occupant is one. The next step is to obtain the details of lighting units in a household. The proportion of lighting unit types are obtained from a RECS household S_l that matches H_i (RECS Model). Power ratings are attached to each lighting unit. In general, not all lighting units are used at the same frequency. This is observed in literature surveys such as DECADE report[60]. The frequency of usage of lighting units in households can be roughly modeled as a natural log curve[57], however, no formal methods have been presented in the literature due to lack of quantitative data. We use the natural log curve presented in Richardson *et al.*[57] to model the relative usage of a lighting unit. Once weights are assigned to lighting units, the probability of a switch-on event for every lighting unit is calculated at a regular time interval (in our case 1 hour). The probability of a switch-on event P_b^{on} of lighting unit b at hour t is calculated as

$$P_b^{\mathrm{on}} = \mathbb{I}_b \times b^{\mathrm{weight}} \times \hat{O}_{i,t} \times \gamma, \quad \text{where}$$
$$\mathbb{I}_b = \begin{cases} 1 & \text{irradiance threshold conditionis True for bulb } b \text{ at time } t \text{ if } \mathrm{Irr}_t \leq \mathrm{Irr}^{\mathrm{i}}, \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

Here b^{weight} is sampled from a natural logarithmic curve, γ is a calibration constant used to achieve the appropriate annual lighting consumption for the U.S., and $\hat{O}_{i,t}$ is the effective occupancy of H_i at time t. If a switch-on event occurs, then energy consumption is calculated for the respective lighting unit b. The lighting duration is picked randomly from the distribution described in Stokes *et al.*[56].

Refrigeration E^{refr}. The energy consumed by a refrigerator depends upon its size, age, ambient temperature, and several other factors as described in literature. They consume 3%–5% of the total residential energy usage. Shimoda *et al.*[42] show that the daily refrigerator consumption is affected by outside temperature, while Tsuji *et al.*[43] show a linear relationship between outside temperature and annual refrigerator demand. Both these work are done in context of refrigerators in Japan. The Lawrence Berkeley National Laboratory in California uses field metered energy use data from ~1500 refrigerators and freezers to develop a model that predicts annual usage of different freezer and refrigerator categories[61]. All of the above models collected relevant data from the field or utilized detailed surveys on refrigeration.

Our approach is to develop a regression model for predicting daily refrigerator usage (kWh/day) of a household (E_i^{refr}) as a function of outside environment temperature. The model is trained with the metered refrigerator usage data from Pecan Street Inc, where 30% of the total metered data is used for training and testing the model. The 30% data is obtained by conducting stratified sampling based on climate zones and daily average temperature bins. The dependent variable is the daily refrigerator usage E_i^{refr} in kWh/day for H_i. The independent variables are daily average temperature $\hat{T}^{\mathrm{out}}$ (°F) and categorical attributes indicating three major climate zones. The 24 hour load profile of a refrigerator $\langle E_{i,0}^{\mathrm{refr}}, E_{i,1}^{\mathrm{refr}}, \ldots, E_{i,23}^{\mathrm{refr}} \rangle$ is constructed from the daily usage, and the variation in the hourly usage of the refrigerator is modeled using a Guassian distribution. The refrigerator operates in an automated/standby mode, that is, occupant presence does not influence the energy consumption of this activity[43,44]. Thus, computing the 24 hour profile of the refrigerator by adding a small Gaussian noise to the hourly load can be considered acceptable. The validation section shows that addition of this noise creates good match to real data.

Activity model $E^{\mathrm{appliances}}$. The energy consumption in a households that is attributed to appliance usage and plug load is 20%–26%. This energy is a result of the occupants' desires to perform activities such as taking baths, making hot meals, using the dishwasher, doing laundry, charging electronics such as TVs and computers, or using any other appliances that consume electricity. Equation 1b,c are used in this model. Based on the aforementioned end-uses, appliance usage behavior is characterized by[43] through operational mode of appliances, duration of operation, power consumption, limit on daily event occurrence, and saturation rate. Operational mode of appliances describes the functioning appliances and related behavior that can be categorized into three types: automatic (appliance use is independent of person), semi-automatic (appliance turned on by household member but turned off automatically), and manual (appliance turned off and on manually). The saturation rate can be used to determine the presence and/or penetration of certain appliances in households. Generally, the operational mode of appliances and saturation rate are deterministic in nature. However, parameters such as probability of activity occurrence, start time, duration, power consumption, and maximum occurrences vary from household to household and day to day. In general, some appliance usages can overlap and/or occur in parallel.

End-use	Relevant models	Our approach
HVAC	Muratori *et.al.*[41], Subbiah *et.al.*[44], Thorve *et.al.*[27], Tsuji *et.al.*[43]	Our model is based on the approach adopted in Subbiah *et.al.*[44] and Thorve *et.al.*[27]. These models were specific to Virginia state. The method employed in these works as well as ours is a physics model. This model is also documented in NREL Technical Reports. Additional details about thermostat settings, building characteristics such as insulation are obtained from RECS survey, EIA website, and NREL Technical Reports.
DHW	Maguire *et.al.*[50], Hendron *et.al.*[48], Thorve *et.al.*[27]	Hendron *et.al.*[48] and Maguire *et.al.*[50] present a general stochastic method to reproduce sample hot water draws based on two water usage surveys conducted in the U.S. The analyses concludes by reporting distributions related to hot water usage events such as showering, using dishwasher, and using clothes washer. Some of these results are summarized in Table 4 and used in our model. Hot and cold water temperatures for specific end-uses are obtained from NREL surveys. The above model does not consider the setting of specific household schedules. This context of household occupancy and occurrence of events is added to an existing model in literature presented in Thorve *et.al.*[27] in order to schedule these hot water usage events.
light	Richardson *et.al.*[57], Stokes *et.al.*[56], Paatero & Lund *et.al.*[58]	We mainly improve upon the stochastic lighting model developed for U.K. household by Richardson *et.al.* by adding context of U.S. households such as household size, household occupancy, annual lighting consumption in the U.S. for different household sizes, calibration of γ for U.S. households, and proportion of light bulbs in the U.S. households and their power ratings. The probability of switch-on event is modeled from Paatero & Lund *et.al.*[58] and Richardson *et.al.*[57]. Duration of switch-on event is taken from Stokes *et.al.*[56]. Power ratings for different categories of lighting units in U.S. is obtained from a study conducted by Bonneville Power Administration[59]. Proportion of lighting units in U.S. households and annual lighting consumption by household size is derived from RECS survey. Irradiance data for the U.S. is obtained from NREL.
refr	—	A linear regression model is developed to predict daily refrigerator usage for a household based on outside temperature and climate zones.
misc, act	Subbiah *et.al.*[44], Thorve *et.al.*[27], Tsuji *et.al.*[43]	All the three referenced models have inspired the design of activity models involving use of appliances. The actual activity occurrence is obtained from the individual/household occupancy schedule. Duration and power usage distributions of appliances is modeled from NIST datasets[62–64] and other datasets[65–68]. The start time is chosen randomly within the duration reported by ATUS individuals and the power ratings and duration of the activity/appliance is selected from the above mentioned distributions.

Table 5. Summary of referenced end-use modeling methods, including how these models are extended in this paper.

Table 6 outlines all the modeled activities and related appliances, their modes of operation, maximum allowed daily occurrences, activity duration, and power consumption. The distributions marked with an asterisk (*) denote that they are modeled by engineering judgement and/or other sources such as Energy Calculator (energyusecalculator.com). Power rating distributions for dishwashers are obtained from a survey conducted by NIST[62,63]. Power ratings and duration distributions for laundry appliances are derived from literature[27,44] and surveys[63]; power ratings for appliances in cook activity include electric ovens, microwaves, and electric cooktops (small- and large burners.) Power rating distributions for these appliances are derived from the NIST efficiency study[64], and durations of appliance usage are obtained from ATUS data, where the maximum limit for cooking activities is capped to three. Sample power ratings for TVs are observed from EnergyStar reports[65] and modeled using a normal distribution. The tv activity duration is modeled as a log-normal distribution after examining the ATUS survey data. Power ratings for computer use activity are derived from a small study conducted by EnergyStar[66]. Standard values for charging duration are used from reputed laptop manufacturers. Vacuum related data are obtained from EnergyStar vacuum report and a survey conducted by Electrolux covering 28,000 consumers from 23 countries including U.S.[67,68]. We assume that all households have vacuum cleaners. The usage frequency of vacuuming is 1–5 times per week[68] and the maximum number of daily occurrences is 1. Assuming Normal distribution for power ratings and duration of appliance usage is reasonable after examining rudimentary results from surveys/reports. The results of the hot water usage study conducted by NREL[48,52] as summarized in Table 4 show that most of the processes can be modeled as a Normal distribution.

The activity model simulates appliance usage based on activity indicators provided by ATUS when the occupant is present in the house. Considering the presence of appliance in each household (from matching RECS household) The time use diaries of adults in the synthetic population and frequency of occurrence of appliance usage such as dishwasher and laundry, and activities such as cooking are taken from RECS household. The activity model focuses on activities performed by an individual when at home. Similar to lighting, activities such as cooking, vacuuming, and leisure activities such as watching TV are shared by household members. A procedure is outlined below for generating household level activity sequence ActSeq_i. Let M be the number of adult members in the synthetic household. Then each household member $P_{i,j}$ has an activity sequence $\text{ActSeq}_{i,j}$. The goal is to find one household level activity sequence ActSeq_i composed of n activities (individual + shared appliance usage related activities) such that the sequence satisfies following constraints:

1. Each activity is performed when at least one occupant is home.
2. The limit on repeated usage is respected for each activity type.
3. Presence of appliance is considered for activities such as dishwasher, and laundry appliances.

Once the above constraints are satisfied, a start time is randomly selected for each activity from the activity duration reported by ATUS. The actual duration and power ratings for appliances used in different activities is chosen from Table 6. Table 5 provides an overview of all the energy (end-use) models in the framework.

Activity	Appliance	Mode	Max occ.	Duration (minutes)	Power (W)	Hot Water
dwasher	dishwasher	Semi-automatic	2	Normal (90, 30)*	Normal (900, 100)	Yes
cwasher	clothes washer	Semi-automatic	2	Normal (45, 20)*	Normal (400, 50)*	Yes
cdryer	clothes dryer	Semi-automatic	2	Normal (45, 20)*	Normal (2500, 200)*	No
cook	oven microwave cooktop (large) cooktop (small)	Manual/Semi-automatic	3	LogNormal (3, 0.96)	Normal (1426, 13.3) Normal (880, 14) Normal (213, 1.2) Normal (393, 3.1)	No
tv	television	Manual	—	LogNormal (4.24, 0.79)	Normal (120, 20)*	No
computer	desktop notebooks	Manual	—	Normal (90, 30)*	Normal (191.5, 32.7) Normal (60.5, 20.5)	No
cleaning	vacuum	Manual	1	Normal (30, 15)	Normal (1200, 300)	No

Table 6. Modeled activity and appliance usage behaviors.

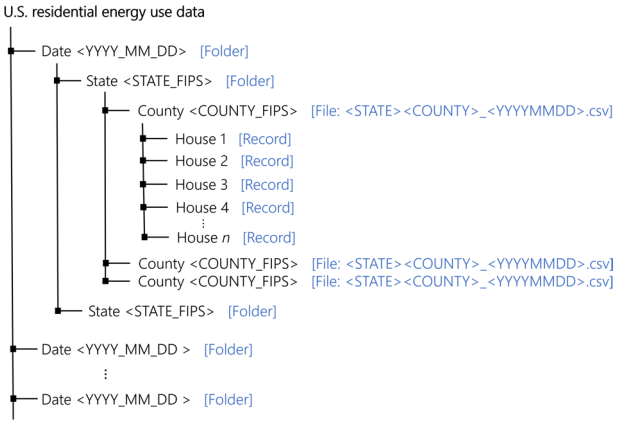


Fig. 4 Data organization. Dataset is available in the form of csv files. The files are organized by dates (temporal) and states (spatial). The blue text indicates the type (e.g. folder, file, record). The text within angular brackets denotes nomenclature templates of folders and files. A record csv file contains energy use data and metadata for a synthetic household in the SPEW population. There will be one file per county and date. One day generates several GBs of data.

attribute_name	description
hid	Synthetic household ID for the SPEW population
total_kwh_1 ,…. ,total_kwh_24	total energy use hourly profile for the particular household (in kWh)
hvac_kwh_1, …., hvac_kwh_24	heating/AC energy use hourly profile for the particular household (in kWh)
hoth2o_kwh_1, …., hoth2o_kwh_24	heating/AC energy use hourly profile for the particular household (in kWh)
refr_kwh_1, …., refr_kwh_24	refrigerator energy use hourly profile for the particular household (in kWh)
light_kwh_1, …., light_kwh_24	lighting energy use hourly profile for the particular household (in kWh)
misc_kwh_1, …., misc_kwh_24	miscellaneous device/appliance energy use hourly profile for the particular household (in kWh)
cook_wh_1, …., cook_wh_24	cooking appliances energy use hourly profile for the particular household (in Wh)
laundry_wh_1, …., laundry_wh_24	laundry appliances related energy use hourly profile for the particular household (in Wh)
dw_wh_1, …., dw_wh_24	dishwasher energy use hourly profile for the particular household (in Wh)

Fig. 5 Data Attributes. 24-hour dis-aggregated hourly household energy demand profiles are made available. 1–24 indicates the hour starting midnight. Eight end-use profiles are described (lines 3–10).

Data Records

The dataset for the entire year of 2014 for U.S. households is publicly available for download from the net.science repository through University of Virginia Dataverse[69]. The dataset is available in the form of csv files. It is organized in folders according to date and state. Figure 4 shows the hierarchy of data organization and file name templates. Each file corresponds to a U.S. county identifier and date. A county identifier is a FIPS code. FIPS codes are numbers which uniquely identify geographic areas by the U.S. census. A record in the file corresponds to a synthetic household. The record includes synthetic household metadata and energy data for that particular date. Attributes of the data record are shown in Fig. 5. All energy related data is in kWh. All the energy data is timestamped by local timezones in the country. A data header codebook is also included in the downloads. Note that, this work was reviewed by the University of Virginia's Institutional Review Board (IRB) and was determined to be exempt from board IRB approval, as this research project did not involve human subject research.

Climate	Location	Source	Year	Sample size	Area type	Resolution	Is open-source	Is data complete?	Is data dis-aggregated?
Hot-Humid	Austin,TX	Pecan Street	2018	25	Urban	15-min	Yes	Yes	Yes
Hot-Humid	Horry,SC	NRECA	2017	56000	Rural Semi-urban	Hourly	No	Yes	No
Mixed-Humid	Rappaha-nnock in VA	NRECA	2016	100	Rural	Hourly	No	Yes	No
Cold	Tompkins Cayuga in NY	Pecan Street	2019	25	Urban	15-min	Yes	No	Yes
Cold	Los Alamos in NM	Open data Dryad repository	2014	1600	Semi-urban	Hourly	No	Yes	No
Cold	MT	NEEA	2019	9	—	Hourly	Yes	No	Yes
Cold	ID	NEEA	2019	19	—	Hourly	Yes	No	Yes
Cold Marine	OR	NEEA	2019	102	—	Hourly	Yes	No	Yes
Cold Marine	WA	NEEA	2019	78	—	15-min	Yes	No	Yes
Hot-Dry/Mixed-Dry	San Diego in CA	Pecan Street	2014 2015 2016	25	Urban	15-min	Yes	No	Yes

Table 7. Datasets used for validation.

Technical Validation

Three studies are presented for validating the synthetic energy profiles. The first study quantifies the similarity between the real and synthetic energy use probability distributions using Jensen-Shannon and Hellinger distance. Comparisons are performed by end-use for real and synthetic data in all representative locations of the U.S. Strong similarities are observed for appliance use distributions between real and synthetic data as well as across spatial locations. TCL loads show differences in distributions across locations. The second study examines variations in the 24-hour energy use timeseries in real and synthetic data in all representative locations in the U.S. We uncover unique energy use patterns in the real and synthetic datasets and study similarities in patterns using unsupervised learning. We introduce two metrics in the process – coverage and closeness. The synthetic data has patterns similar to that of real data. The last study is focused on observing trends in the synthetic energy use in different representative locations in the U.S. We notice that the synthetic data is able to incorporate the effects of mixture of variables such as weather, irradiance, building attributes and demographic characteristics on household level energy usage. The study is a quick demonstration of energy use variability at multiple spatio-temporal levels in different end-uses.

The remaining V&V section is outlined as follows. First, we describe challenges in validating a large synthetic dataset for energy use. Then, we highlight the temporal and spatial resolutions of the data that are considered in the validation experiments. Next, ground truth datasets (real/recorded/actual data) used for evaluation are briefly described. This is followed by description of the experimental setup and results.

Validating the quality of the large-scale synthetic timeseries data for a sizeable region such as the U.S. is challenging, owing to the vast extent, diversity, and contrasting climates in the country. One of the challenges of validating an energy consumption timeseries at household level is the large variety and variability of the load patterns within and between households. In addition to external elements such as weather and building characteristics, consumer lifestyles and affordances play a vital role in shaping the demand such as a curve with morning peak, or a curve with a small afternoon peak and sharp evening peak. This leads to a big spectrum of variations and patterns in energy use. Thus, in-depth comparative analyses of synthetic data to actual data is required. However, it is conditioned on the availability of a reasonable amount of representative real data. Here, we employ real/recorded data such as load research data, end-use metering data, and smart meter data from ten locations in the country that are representative of the U.S. climate zones (Table 7). The availability of public smart meter data in the U.S. is limited, which may cause a potential skew towards the selected sample of households and may not be spatially representative. Thus, framing our understanding of validation in this context is important.

We address the quality of the synthetic energy consumption data on two intrinsic qualities of energy use data: magnitude (usage over 24 hours) and load shape (pattern of consumption). Magnitude and load shape can be examined across the temporal (hour/day/month/year) and spatial (household/census tract/city/county/state/climate zones) axes. Thus, the verification and validation (V&V) process covers:

- *Spatial representativeness and resolutions*. Due to limited availability of real data, we define spatial representativeness by choosing atleast one location in each climate zone in the U.S. to carry out validation experiments. The major climate zones[70] in the contiguous United States are as follows: (*i*) marine, (*ii*) hot-dry/mixed-dry, (*iii*) hot-humid, (*iv*) mixed-humid, and (*v*) cold/very-cold. Comparisons are then performed at household and city/county resolutions.
- *Temporal representativeness and resolutions*. Temporal representativeness is studied by observing similarities between real and synthetic hourly demand profiles. Furthermore, daily and seasonal energy usage is studied for different locations.

- *Dis-aggregate energy use*. Note that we publish dis-aggregated energy use data at household level. Thus, a finer level of evaluation such as an energy use sub-type (e.g. HVAC, cooking, etc,.) is possible at various temporal and spatial levels.

All the real datasets used in the V&V process are listed in Table 7. Recorded datasets are obtained from Pecan Street Dataport[23], Northwest Energy Efficiency Alliance (NEEA)[71], National Rural Electric Cooperative Association (NRECA). The Los Alamos dataset is obtained from a public data sharing repository Dryad[72]. Unfortunately, we do not have any metadata about households (e.g. household size, dwelling type, etc) in these datasets. The datasets only have energy use timeseries.

Three studies are presented to cover temporal, spatial, and dis-aggregate nature of the synthetic time-series:

I. Comparing real and synthetic end-use energy usage (magnitude)
II. Comparing real and synthetic energy use patterns (shape/structure)
III. Observing differences and similarities in synthetic energy use data in spatially representative locations

I. Comparing real and synthetic end-use energy usage (magnitude). In this experiment, distributions of synthetic and real daily end-use data are compared using statistical metrics. One way of comparing these distributions is by measuring distance between the real and synthetic end-use distributions. Many metrics can be used to perform this task (e.g., Kullback–Leibler divergence (KL), the Hellinger distance, total variation distance (TVD), the Wasserstein metric, the Jensen-Shannon divergence (JS), and the Kolmogorov–Smirnov statistic (KS)). Klemenjak *et al.*[26] use JS distance and Hellinger distance as examples to compare distributions of appliance energy use between different datasets. A similar method is implemented in this section using the JS distance and the Hellinger distance metric. In our case, computing the distances between daily end use distributions allows us to perform regional comparisons as well as comparisons between real and synthetic datasets.

The Jensen-Shannon distance is the square root of the Jensen-Shannon divergence[73]. The range of this metric ranges between [0, 1] where 0 implies the distributions are similar. We prefer JS divergence over KL divergence since it is a symmetric measure. If P and Q are two probability vectors, then the JS distance JS(P, Q) is given by

$$\mathrm{JS}(P,\,Q) = \sqrt{\frac{\mathrm{KL}(P||M) + \mathrm{KL}(Q||M)}{2}}, \tag{5}$$

where M is the pointwise mean of P and Q and KL is the Kullback-Leibler divergence. To supplement our study, we use Hellinger distance as a second metric to quantify the similarity between two probability distributions. Hellinger distance is also a symmetric measure. Its range of values is [0, 1] with 0 encoding that the distributions are similar. The Hellinger distance of two probability vectors P and Q is denoted by H(P, Q) and defined as

$$\mathrm{H}(P,\,Q) = \frac{1}{\sqrt{2}}\sqrt{\sum_{i=1}^{k}(\sqrt{p_i} - \sqrt{q_i})^2}, \tag{6}$$

where k is the length of the vectors, and p_i, q_i are the i^{th} elements of the vectors P and Q, respectively.

Daily end-use energy usage (e.g. E_i^{hvac}) at household level are compared in the real and synthetic data for every location specified in Fig. 6. Vectors P and Q denote values in a single end-use for two datasets. Figure 6a–c list JS distances and Fig. 6d–f list Hellinger distances for selected end-uses (HVAC, refrigerator, cooking appliances). Each matrix represents distances between two energy usage distributions for an end-use. The row and column headers represent different data-sources and different regions and each cell represents the probability distribution similarity/distance value in the form of heatmap where the bar shows the range of the values on a continuous scale.

The JS and Hellinger distance tables for end-uses show strong similarities (the distance is close to zero). Furthermore, within each matrix three types of comparisons are performed. We compute similarity between end-use distributions for different regions within synthetic data, different regions within real data, and different regions in different data sources (namely real and synthetic data). For appliance usage (e.g. cooking), the distributions are quite similar across regions and data-sources. This supports findings from Fig. 11 that there exists significant similarities between different regions for synthetic daily energy consumption of different appliances. For HVAC end-use, it is observed that the distributions grow apart between regions for both – synthetic and real data sources. This is particularly true due to the strong association of HVAC with outdoor/environment temperature conditions and the time span for which these temperature conditions prevail (e.g., warmer temperatures are observed for a longer time in Texas (TX)).

II. Comparing energy use patterns (load shape/structural similarity). In this section, the synthetic energy use timeseries are evaluated using the concepts of diversity, coverage, and closeness. The diversity in energy use patterns is captured by segmenting the normalized timeseries $\langle \bar{e}_0, \ldots, \bar{e}_{23} \rangle$ using unsupervised learning techniques such as clustering. This is followed by studying *coverage* in terms of what percentage of synthetic timeseries population is represented in the real timeseries population and vice versa. Thus, coverage is used to measure diversity. However, learning only coverage is not sufficient. It is necessary to measure the accuracy of the matches found. Hence, we introduce the *closeness* metric. It studies how close (e.g. *dist*(*i, j*) are the synthetic and real data points.

Fig. 6 Left column: Jensen-Shannon distance matrices, Right column: Hellinger distance matrices. Each of the column shows Jensen-Shannon distance and Hellinger distance matrices between end-use probability distributions. Each matrix represents distances between two energy usage distributions for a particular enduse (e.g. HVAC, refrigerator, cooking). The row and column headers of the matrix represent different data-sources and different regions and each cell represents the probability distribution similarity/distance value in the form of heatmap, where the bar shows the range of the values on a continuous scale.

Let $\mathcal{R}$ and $\mathcal{S}$ be the set of load shapes of real and synthetic energy use timeseries. Let $K_{\mathcal{R}}$ be the number of unique load shapes (segments/patterns/clusters) found in set $\mathcal{R}$. Then, we define the *coverage*($\mathcal{S}$) as a ratio

$$\begin{aligned} \textit{coverage}(\mathcal{S}) &= \frac{\text{Number of unique shapes in } \mathcal{R} \text{ that contain atleast one data point from } \mathcal{S}}{\text{Number of unique shapes in } \mathcal{R}} \\ &= \frac{1}{K_{\mathcal{R}}} \times \sum_{b=1}^{K_{\mathcal{R}}} \mathbb{I}_b \quad \text{where} \\ \mathbb{I}_b &= \begin{cases} 1 & \text{if cluster } b \text{ contains atleast one time series } j \in \mathcal{S} \\ 0 & \text{otherwise.} \end{cases}, \end{aligned} \tag{7}$$

Thus, *coverage*($\mathcal{S}$) reflects the degree to which samples from set $\mathcal{S}$ cover the patterns in set $\mathcal{R}$. Similarly, if K_S is the number of unique segments in set $\mathcal{S}$, then, *coverage*($\mathcal{R}$) reflects the the percentage of unique patterns in set $\mathcal{S}$ covered by data points in set $\mathcal{R}$. Coverage is bounded between 0 and 1. Figure 13b shows *coverage*($\mathcal{S}$) and *coverage*($\mathcal{R}$) as K varies.

To measure closeness we calculate distance of individual timeseries to it's respective cluster center/representative. If $K_{\mathcal{R}}$ is the number of clusters in set $\mathcal{R}$, then, the *closeness*($\mathcal{S}, \mathcal{R}$) of set $\mathcal{S}$ to set $\mathcal{R}$ is measured by comparing the distributions of distances of individual timeseries $i \in \mathcal{R}$ and $j \in \mathcal{S}$ in each cluster $c \in K_{\mathcal{R}}$ to the respective center/representative timeseries of the cluster. Figure 13b illustrates the schematic of building the distance distributions. Let $P_{\mathcal{R}}$ and $P_{\mathcal{S}}$ denote the probability vectors of distances of sets $\mathcal{R}$ and $\mathcal{S}$ respectively. To measure the degree of closeness, we compare the two probability distributions using Hellinger distance $\mathrm{H}(P_{\mathcal{R}}, P_{\mathcal{S}})$ (Eq. 6). If distributions $P_{\mathcal{R}}$ and $P_{\mathcal{R}}$ are similar, then we say that set $\mathcal{S}$ is close to set $\mathcal{R}$.

$$closeness(\mathcal{S}, \mathcal{R}) = \mathrm{H}(P_{\mathcal{R}}, P_{\mathcal{S}}) \tag{8}$$

Closeness is bounded between 0 and 1. 0 implies that the two sets are close. Note that closeness is not a symmetric metric i.e. $closeness(\mathcal{S}, \mathcal{R}) \neq closeness(\mathcal{R}, \mathcal{S})$. Figure 13b describes the variation in similarity score of the probability with different number of segments K.

Now, we briefly describe the experimental setup. Two cases are considered to examine coverage, closeness and robustness of cluster groupings (k). For each case the energy use timeseries is normalized resulting in a *load shape* $\langle \bar{e}_0, \ldots, \bar{e}_{23} \rangle$. We choose normalization by total consumption (Eq. 9) in order to consider pronounced effects of peak-load in the profile. Household preferences or lifestyles can be typically captured by one or more load shapes[74], hence we choose this representation for uncovering patterns in the data. Thus, every $i \in \mathcal{R}$ and $j \in \mathcal{S}$ are normalized energy use vectors of length 24.

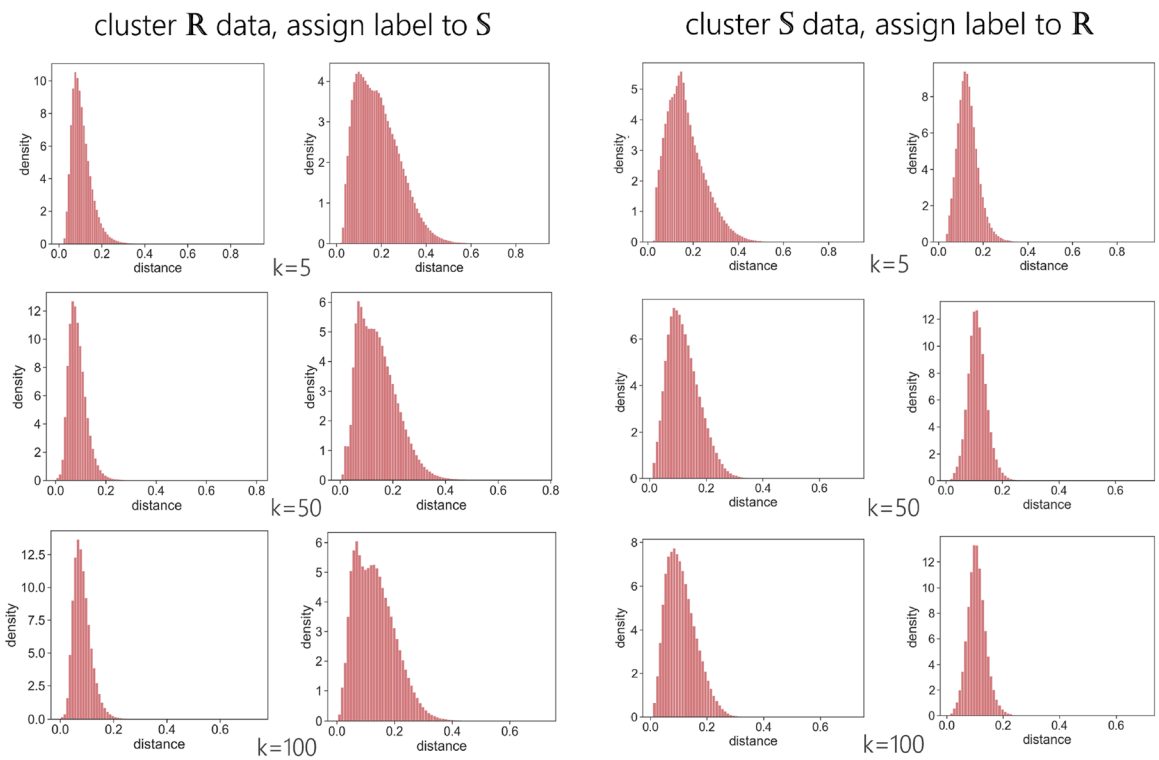
$$\bar{e}_t = \frac{e_t}{E^{\text{total}}}, \quad \text{where } E^{\text{total}} = \sum_{t=0}^{23} e_t \tag{9}$$

In the first case (Case 1), we generate $K_{\mathcal{R}}$ patterns from set $\mathcal{R}$ by clustering the real normalized energy use vectors using k-means clustering algorithm with Euclidean distance. This is followed by assigning a cluster label $k \in K_{\mathcal{R}}$ to each synthetic energy use timeseries $j \in \mathcal{S}$. Let c_k be the center/representation vector of group k. Then, $j \in \mathcal{S}$ is assigned to the cluster whose cluster center distance is minimum from j and is given by $min\Big(dist(j, c_0), \ldots, dist\big(j, c_{K_{\mathcal{R}}}\big)\Big)$. Then, we calculate the coverage of synthetic data *coverage*($\mathcal{S}$) and closeness of synthetic data to real data among all clusters as *closeness*($\mathcal{S}, \mathcal{R}$). In Case 2, we generate K_S clusters from set $\mathcal{S}$ (synthetic data) by segmenting the normalized energy use vectors using k-means clustering algorithm with Euclidean distance. This is followed by assigning a cluster label $k \in K_S$ to each real energy use timeseries $i \in \mathcal{R}$. i is assigned to the cluster whose cluster center distance is minimum from i and is given by $min_{\forall k \in K_S} dist(i, c_k)$. Then, we calculate the coverage of real data in synthetic groups *coverage*($\mathcal{R}$) and closeness of real data and synthetic data among all synthetic clusters as *closeness*($\mathcal{R}, \mathcal{S}$).

Results of both the cases are summarized in Fig. 8. A 100% *coverage* is observed in both the cases for different values of k. Observations for *closeness* metric are interesting. The Hellinger distance is close to zero in all the scenarios, however there is a slight uptake in the value as k increases. We inspect this further in Fig. 7. Figure 7 shows histograms of distances of real data points and synthetic data points from their assigned cluster center. In case 1, the distribution of distances of synthetic data points is slightly broader than the distribution of distances of real data points for all k. Thus, we see a distance for *closeness*($\mathcal{R}, \mathcal{S}$) in Fig. 8c. As k increases it is observed that some individual clusters have a broad and/or bimodal distance distribution indicating that there are data points that are very close to the cluster center while a few are far away. This difference is apparent as the number of clusters increases.

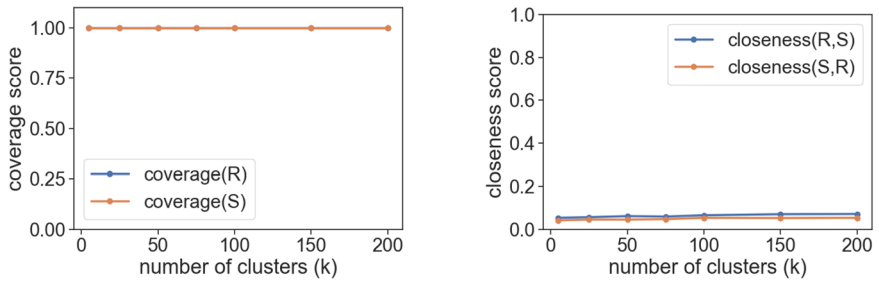
The goal of this V&V exercise was to verify if the diversity and trends of the real energy use profiles are replicated in the synthetic energy use profiles. Due to a biased and skewed sample of the real energy use data, it is challenging to perform validation of synthetic data. Some of the characteristics of the real datasets that hinder the implementation of using existing evaluation metrics *as is* are mentioned below. No supporting information of the real households is available (e.g. household size, dwelling type, square footage, indoor thermostat setting). We have shown that all of these factors are extremely important in the generation of household demand at a given time. Some of households in the real data may also be participants in demand-response programs resulting in unique load shapes due to shifting demand/reducing peak demand that may not be found in households not participating in DR programs (e.g. synthetic data). The real datasets are collected for different years for each region. The data are incomplete for some regions (e.g. San Diego samples do not have lighting data). The sample size (number of households) is highly skewed. It varies from 9 households in Montana to 56000 households for Horry,SC. Thus, it is important to note that $|\mathcal{R}| << |\mathcal{S}|$ (e.g. the number of households simulated in our framework for Washington state is far greater than that of 78 households in real data for Washington state.) All of these observations are summarized in Table 7.

III. Observing differences and similarities in synthetic energy use data in spatially representative locations. This empirical study uses only the synthetic data to conduct a comparative regional analyses to examine similarities and dissimilarities between energy use for different end-uses. We observe the spatio-temporal patterns and variations in different end-uses with respect to environmental elements such as



(a) Case 1: *closeness*($\mathcal{R}, \mathcal{S}$) **(b)** Case 2: *closeness*($\mathcal{S}, \mathcal{R}$)

Fig. 7 Example of *closeness* in different cases with varying k. Figures show the distances of data points from sets $\mathcal{R}$ and $\mathcal{S}$ to their respective cluster center. (**a**) demonstrates histograms of distances for different k. The plot on left is for real data points and on right is for synthetic data points. Then, we calculate *closeness*($\mathcal{R}$, $\mathcal{S}$) using Hellinger distance (corresponds blue line in Fig. 8c). For $k = 5$ a bimodal pattern is observed in distances for synthetic data points which tends to diminish as the number of clusters k increases. Figure b shows histograms of distances for different k for case 2. The plot on left is for synthetic data points and on right is for real data points. *closeness*($\mathcal{S}$, $\mathcal{R}$) is calculated using Hellinger distance (corresponds to orange line in Fig. 8c).

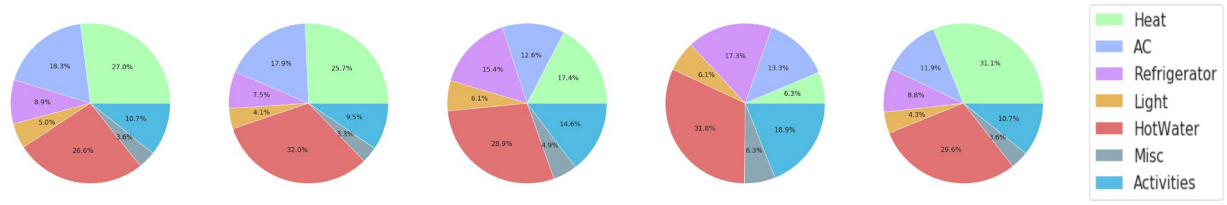


(a) Coverage **(b)** Closeness

Fig. 8 Summary of the two case scenarios. Orange color is denoted for findings of case 1 where we cluster real data set $\mathcal{R}$ and assign a cluster label to synthetic data set $\mathcal{S}$. Blue color is denoted for findings of case 1 where we cluster synthetic data set $\mathcal{S}$ and assign a cluster label to real data set $\mathcal{R}$. (**a**) illustrates 100% coverage in both cases even as k varies. This means that, in each case at least one data point belongs to every cluster for a given k. (**b**) shows the closeness between the two distance vectors: distance of real data points in a cluster to its respective centroid and distance of synthetic data points in a cluster to its respective centroid. Closeness is given by the Hellinger distance which suggests that a value of 0 signifies that the two distributions are similar. The value of distances is close to 0 for all values of k in both the cases. However, an upward trend is observed as k increases. Overall we see the robustness of results w.r.t. k.

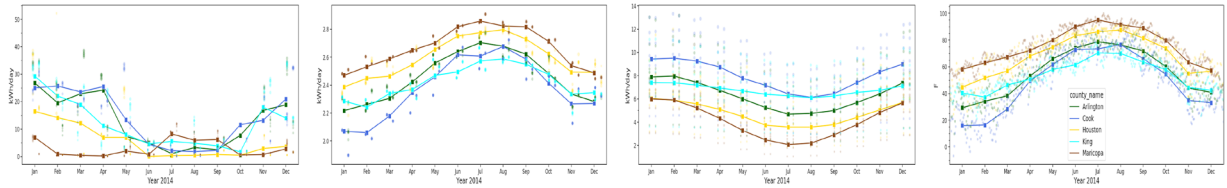
irradiance and temperature as well as demographic and structural characteristics of the households. The selected target locations are spatially representative of different climate zones of the U.S.:

Arlington, VA; Cook County, IL; Houston County, TX; Maricopa County, AZ; King County, WA



(a) Arlington, VA **(b)** Cook, IL **(c)** Houston, TX **(d)** Maricopa, AZ **(e)** King, WA **(f)** Legend

Fig. 9 Composition of synthetic electric consumption in the representative target locations. Heating and cooling constitute the majority part of the residential electric consumption. Refrigerators consume slightly higher energy in hotter regions such as Maricopa and Houston. Activities such as dishwashing, laundry, and cooking represents between 8–17% for different regions. Lighting and water heating have a consistent proportion of consumption across all locations. The proportions bear similarities with data published by EIA.



(a) HVAC energy use **(b)** Refrigerator energy use **(c)** Hot water energy use **(d)** Outside temperature

Fig. 10 Monthly synthetic energy use changes in end-uses such as HVAC, refrigerator, domestic hot water w.r.t. temperature. The above line charts monthly energy use changes in end-uses such as HVAC, refrigerator, domestic hot water w.r.t. outside temperature. The line chart shows average daily consumption over all households in the target regions. The scatter plot in the background describes average daily consumption for an end-use for sampled days color coded by location. The size of the markers denotes the standard deviation of the end-use consumption. Legend: Arlington, VA (green); Cook County, IL (blue); Houston County, TX (yellow); Maricopa County, AZ (brown); King County, WA (cyan).

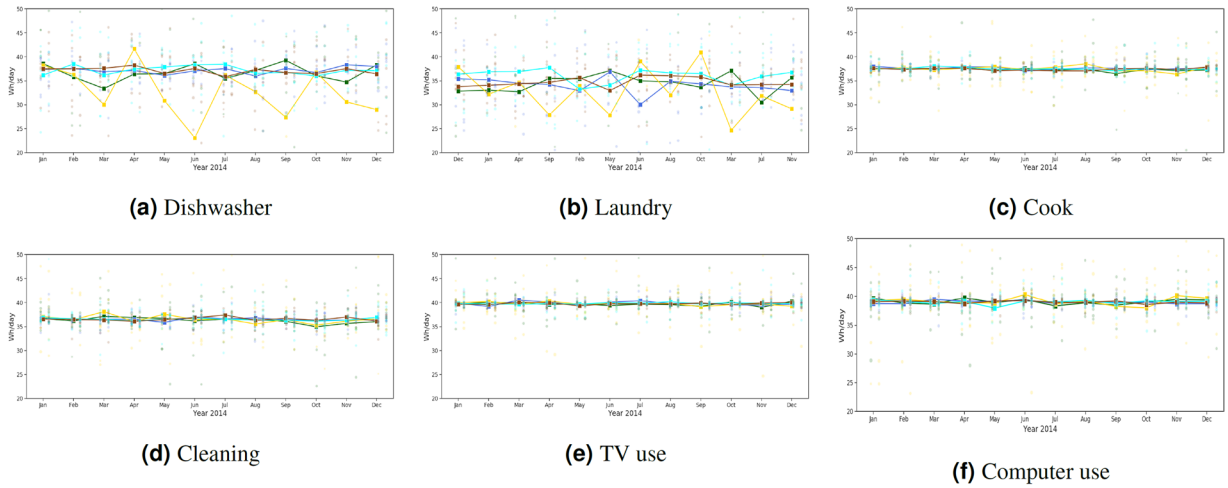
The composition of electric consumption by end-uses is shown in the form of pie diagrams in Fig. 9. EIA reports the shares of the major end-uses as follows: DHW 17–32%, lighting 5–10%, refrigerator 3–5%, activities/appliances 20–26%, space heating 25–47%, and air conditioning 5–10%. In general, the percentages of major end-use categories lie in the ranges similar to those reported by EIA. HVAC has a dominant share in the energy consumption in households as compared to usage of appliances and/or other activities.

Seasonal energy use variations for HVAC, refrigerator, and hot water is captured in Fig. 10. The plot shows variation in daily average energy use of the four end-uses on a monthly basis alongwith temperature across the year 2014. Refrigerator energy use increases slightly with temperature while energy used to heat water decreases with increase in temperature.

Electricity usage for heating water is the lowest during summer months for all locations (Fig. 10c). In particular, regions from hot-humid and hot-dry climate zones consume the least amount of energy. This observation stems from the relation between $E^{\mathrm{h2o,v}}$ and $T^{\mathrm{cold}}_{m,z}$ described in Eq. 3. The water inlet temperature ($T^{\mathrm{cold}}_{m,z}$) differs across temporal as well as spatial scale and is dependent on outside environment temperatures[50] (Details in Appendix). Figure 13 shows plots describing relation between household size and the number of gallons of hot water consumed and energy required to heat water. Note that, we consider only electric water heaters in this work.

Figure 10a shows that the HVAC consumption varies significantly throughout the year. HVAC use is higher in hot-dry areas in summer as compared to other regions possibly due to higher temperatures. Structural characteristics such as dwelling size (square footage), insulation quality, age and efficiency of HVAC equipment also affect household HVAC consumption. Another important variable that drives HVAC consumption is indoor thermostat behavior which is related to household occupants' behavior/actions. In this work, indoor thermostat temperatures are set constant throughout the day. Insulation quality is not monitored in households (due to lack of data). We assume that the dwelling is well-insulated and the insulation values are implemented according to the DOE standards for the respective climate zones. In Fig. 12a we show effect of square footage (conditioned space) of a dwelling on hvac energy use. In general, we observe that as the conditioned space in the dwelling increases, the HVAC consumption increases.

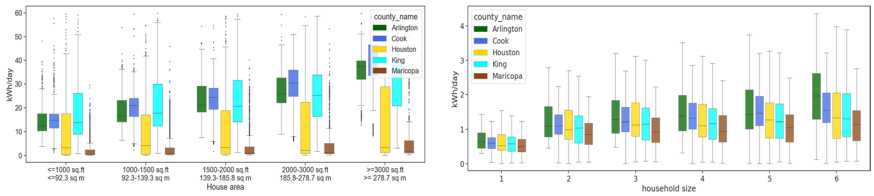
Lighting energy-use varies by seasons in all regions as irradiance levels change with weather events and seasons. Figure 14b shows average irradiance time series for the target locations. The corresponding lighting usage is shown in Fig. 14a. As an example, we look at monthly irrandiance profiles across 24 hours in Virginia for the year 2014 (Fig. 14d). The corresponding monthly lighting energy use time series is shown in Fig. 14c. Example of lighting consumption w.r.t. household size is explored in Fig. 12b.



(a) Dishwasher **(b)** Laundry **(c)** Cook

(d) Cleaning **(e)** TV use **(f)** Computer use

Fig. 11 Synthetic appliance energy use variation in target locations throughout the year. The line charts show variation in daily energy consumption for different appliance energy use throughout the year averaged by month. The lines depict average daily consumption over all households in the target region. The scatter plot in the background describes average daily consumption for an end-use for sampled days color coded by location. The size of the markers denotes the standard deviation of the end-use consumption. There are noticeable similarities in appliance-usage throughout all locations indicating that people in different parts of the country use appliances in a similar style. This is a reasonable observation since day-to-day activities such as cooking and cleaning will occur in all households. Their usage pattern may change during the day, but the total energy consumed by the appliance at the end of the day is similar. Arlington, VA (green); Cook County, IL (blue); Houston County, TX (yellow); Maricopa County, AZ (brown); King County, WA (cyan).



(a) HVAC vs. house area (i.e. floor area) **(b)** Lighting vs. household size

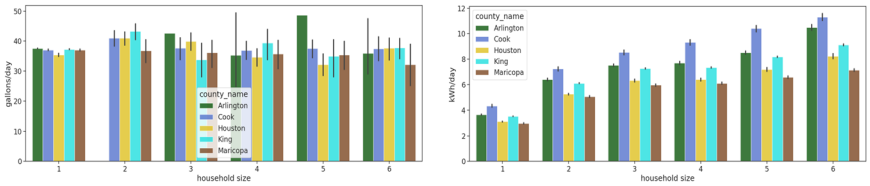
Fig. 12 (**a**) Synthetic HVAC use and house area (i.e. floor area). Boxplot comparing daily HVAC consumption in a winter day for the selected target locations by house area (i.e. floor area). The x-axis groups floor area of houses in five bins denoted in two units sq. ft (ft^2) and sq m (m^2). The bins are as follows: $\leq$1000 ft^2, 1000 - 1500 ft^2, 1500 - 2000 ft^2, 2000 - 3000 ft^2, $\geq$3000 ft^2. It is observed that as floor area of the house increases HVAC consumption increases in all regions. Winter temperatures are relatively moderate in AZ and TX, thus, the HVAC consumption is less as compared to other regions. (**b**) Synthetic lighting use and household size. Lighting consumption increases as household size increases. Household size indicates number of members in a household.

Figure 11 shows the breakdown of appliance usage for different appliances and electronic devices. Both figures show a line chart indicating average daily consumption for the month. The scatter plot in the background describes average daily consumption for an end-use for sampled days color coded by location, where the size of the markers denotes the standard deviation of the end-use consumption. It is observed that appliance usage in activities such as cooking, dishwashing, performing laundry, watching TV, using computer, and cleaning are fairly similar in different regions. The above comment is intuitively true since appliance use duration and their ratings may not vary across regions. However, the occurrence timing throughout the day may vary from house to house depending upon occupant schedules irrespective of which geographic regions they belong to.

Usage Notes

In order to analyze the dataset, researchers can use any programming languages such as Python, Java, Matlab, or R. As described in the 'Data Records' section, the files are stored in csv format, so most of the file reading functions in the above languages can support reading/accessing the dataset. Next, we discuss the potential applications of the released synthetic data. We also highlight important challenges and limitations of this work.

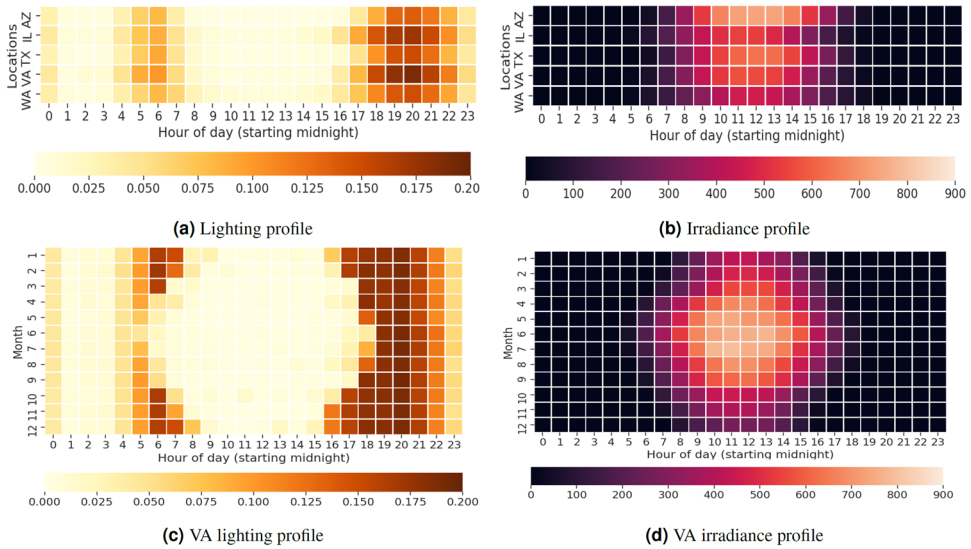
Applicability and benefits of the dataset. We are releasing a comprehensive household level dataset for energy use. In addition to the household level disaggregated energy use data, household composition is also included from census data. This work was reviewed by the University of Virginia's Institutional Review Board (IRB) and was determined to be exempt from board IRB approval, as this research project did not involve human



(a) Hot water (gallons) vs. household size

(b) Hot water (energy) vs. household size

Fig. 13 Synthetic hot water usage and energy vs. synthetic household size. Household size indicates number of household members. The clustered bar charts show the amount of hot water consumed (in gallons in (**a**)) and corresponding energy usage in (**b**) according to household size in a winter day. The vertical black line on each bar shows the variation. Water usage and its variation increases with household size. The amount of energy for hot water end-use increases with household size and differs by region.



(a) Lighting profile

(b) Irradiance profile

(c) VA lighting profile

(d) VA irradiance profile

Fig. 14 Heatmap depicting relation between hourly synthetic lighting usage and hourly irradiance. (**a**) shows average annual 24-hour lighting profiles of representative target locations. (**b**) shows average annual 24-hour irradiance profile of representative target locations. (**c**) and (**d**) present the variation in lighting usage and corresponding irradiance profiles at monthly level for Arlington, VA. (**c**) presents lighting consumption variation throughout the day in different months across the year. (**d**) shows variation in monthly irradiance profile. The units of measurements for energy usage is kWh and irradiance is Watts/m^2. The lighting energy use is inversely proportional to the irradiance. The energy usage is higher in evening and night hours when the occupant is active in the dwelling. The average lighting and irradiance profiles show regional differences in irradiance availability and subsequent lighting energy usage. The VA profiles show that the day light is available for longer durations leading to lower lighting energy consumption as compared to winter.

subject research. The dataset can be effectively employed in various applications such as NILM (non-intrusive load monitoring), load profile analyses for observing similarities/differences between end use consumption of different regions and seasons, evaluating effects of retrofits in buildings, studying effects of temperature rise in different regions, and so on. In addition, this data can also be used for energy model calibration, occupant behavior evaluation, implementing demand response strategies and policy interventions. The dataset can be especially leveraged in training deep learning models where massive amount data is appreciated. Such models can be used for real-time residential demand forecasting. The dataset released are essentially time-series along with categorical and numerical attributes. Thus, any statistical tool or programming language can be used to analyze them. Study III in the 'Technical Validation' study illustrates examples of the possible uses of the dataset.

Challenges and limitations. The use of synthetic residential energy demand data has its pros and cons. National scale hourly synthetic data can be used to carry out national and even potentially international policy analysis. The spatio-temporal variability allows one to access important emerging questions related to energy equity, fairness and accessibility at a fine scale. A systems level approach can be taken to vexing questions outlined in the 2030 Intergovernmental Panel on Climate Change (IPCC) goals. On the other hand, synthetic data sets have their limitations as well. For instance, the fine-scale variability (minutes level as well as weekly variation) of usage amongst households cannot be captured easily in such synthetic data sets. Additionally, the behavior exhibited by any single synthetic family might be biased by the data used for synthesis. Thus, any insight generated from high resolution analyses should be considered carefully.

An important challenge in developing the realistic synthetic residential load profiles at a national scale and at a high spatio-temporal resolution is to find appropriate datasets for representing different types of climates, demographics, appliances, and activity patterns. Accessibility and availability of all the above information from legitimate sources is crucial to maintain trustworthiness in the resulting models. A robust and extensible infrastructure is developed to synthesize diverse data sources into detailed information structure at various spatial resolutions (e.g. combining household level data with climate zone related data such as insulation values). The infrastructure consists of methods to compose multiple models and data sets. The overall time to generate the synthetic data was reduced by using high performance computing capabilities.

Some of the limitations of our work are discussed. The current synthetic data does not include power consumption by electric vehicles and energy generation via renewable generation (e.g. solar panel, wind). The ATUS data is available for a normative day for individuals. Thus, activity and appliance related demands are generated for a normative day with minor variations coming from the activity model. Hence, our synthetic data might not be able to capture daily activity variation appropriately (e.g. as observed in real-time smart metering). This can be challenging to work with especially when studying demand response scenarios. The building envelop considered for a synthetic household is simplified due to lack of information needed to represent a large population group, thus limiting our ability to employ state-of-the-art and sophisticated building modeling techniques. (e.g. we use a simple HVAC physics based model to generate heating and cooling related energy demand).

Concluding remarks. The paper describes a bottom up approach to generate large-scale digital twin data of dis-aggregated residential energy use hourly timeseries for the residential sector at household resolution across the contiguous United States for millions of households. The approach integrates diverse open-source surveys and datasets, where the end-use models are developed by either extending well-established methods or by building new models. Extensive validation of the synthetic datasets is conducted using real/recorded energy-use data across spatial and temporal resolutions.

Code availability

Programming languages such as Python 3 and Java 8 are used for modeling, analyzing, and developing the framework. The code is deposited in the repository[69] alongwith the dataset.

Received: 26 May 2022; Accepted: 15 December 2022;
Published online: 06 February 2023

References

1. Hart, D. G. Using AMI to realize the Smart Grid. *2008 IEEE Power and Energy Society General Meeting - Conversion and Delivery of Electrical Energy in the 21st Century* 1–2, https://doi.org/10.1109/PES.2008.4596961 (2008).
2. Mohassel, R. R., Fung, A. S., Mohammadi, F. & Raahemifar, K. A survey on advanced metering infrastructure and its application in Smart Grids. *2014 IEEE 27th Canadian Conference on Electrical and Computer Engineering (CCECE)* 1–8, https://doi.org/10.1109/CCECE.2014.6901102 (2014).
3. Hailegiorgis, A., Crooks, A. & Cioffi-Revilla, C. An agent-based model of rural households& adaptation to climate change. *Journal of Artificial Societies and Social Simulation* **21**, 4, https://doi.org/10.18564/jasss.3812 (2018).
4. Auffhammer, M., Baylis, P. & Hausman, C. H. Climate change is projected to have severe impacts on the frequency and intensity of peak electricity demand across the united states. *Proceedings of the National Academy of Sciences* **114**, 1886–1891, https://doi.org/10.1073/pnas.1613193114 (2017).
5. Busby, J. W. *et al.* Cascading risks: Understanding the 2021 winter blackout in Texas. *Energy Research & Social Science* **77**, 102106, https://doi.org/10.1016/j.erss.2021.102106 (2021).
6. Petri, Y. & Caldeira, K. Impacts of global warming on residential heating and cooling degree-days in the united states. *Scientific Reports* **5**, 12427, https://doi.org/10.1038/srep12427 (2015).
7. Goldstein, B., Gounaridis, D. & Newell, J. P. The carbon footprint of household energy use in the United States. *Proceedings of the National Academy of Sciences* **117**, 19122–19130, https://doi.org/10.1073/pnas.1922205117 (2020).
8. National Academies of Sciences, Engineering, and Medicine. *Accelerating Decarbonization of the U.S. Energy System* (The National Academies Press, Washington, DC, 2021).
9. Gillingham, K. T., Huang, P., Buehler, C., Peccia, J. & Gentner, D. R. The climate and health benefits from intensive building energy efficiency improvements. *Science Advances* **7**, eabg0947, https://doi.org/10.1126/sciadv.abg0947 (2021).
10. Berrill, P., Gillingham, K. T. & Hertwich, E. G. Drivers of change in US residential energy consumption and greenhouse gas emissions, 1990–2015. *Environmental Research Letters* **16**, 034045, https://doi.org/10.1088/1748-9326/abe325 (2021).
11. Berrill, P., Gillingham, K. T. & Hertwich, E. G. Linking housing policy, housing typology, and residential energy demand in the united states. *Environmental Science & Technology* **55**, 2224–2233, https://doi.org/10.1021/acs.est.0c05696 (2021).
12. Kassakian, J. *et al.* The Future of the Electric Grid: An Interdisciplinary MIT Study. *Massachusetts Institute of Technology, MIT Energy Initiative* (2011).
13. Deb, C., Dai, Z. & Schlueter, A. A machine learning-based framework for cost-optimal building retrofit. *Applied Energy* **294**, 116990, https://doi.org/10.1016/j.apenergy.2021.116990 (2021).
14. Nutkiewicz, A., Choi, B. & Jain, R. K. Exploring the influence of urban context on building energy retrofit performance: A hybrid simulation and data-driven approach. *Advances in Applied Energy* **3**, 100038, https://doi.org/10.1016/j.adapen.2021.100038 (2021).
15. Muratori, M. Impact of uncoordinated plug-in electric vehicle charging on residential power demand. *Nature Energy* **3**, 193–201, https://doi.org/10.1038/s41560-017-0074-z (2018).
16. Mahdavi, A. *et al.* The Role of Occupants in Buildings' Energy Performance Gap: Myth or Reality? *Sustainability* **13**, https://doi.org/10.3390/su13063146 (2021).
17. Tanaka, K., Wilson, C. & Managi, S. Impact of feed-in tariffs on electricity consumption. *Environmental Economics and Policy Studies* https://doi.org/10.1007/s10018-021-00306-w (2021).
18. Tsaousoglou, G., Efthymiopoulos, N., Makris, P. & Varvarigos, E. Personalized real time pricing for efficient and fair demand response in energy cooperatives and highly competitive flexibility markets. *Journal of Modern Power Systems and Clean Energy* **7**, 151–162, https://doi.org/10.1007/s40565-018-0426-0 (2019).
19. National Academies of Sciences, Engineering, and Medicine. *Analytic Research Foundations for the Next-Generation Electric Grid*. (The National Academies Press. Washington, DC., 2016).

20. Kelly, J. & Knottenbelt, W. The UK-DALE dataset, domestic appliance-level electricity demand and whole-house demand from five UK homes. *Scientific Data* **2**, https://doi.org/10.1038/sdata.2015.7 (2015).
21. Kelly, J. & Knottenbelt, W. Metadata for energy disaggregation. *2014 IEEE 38th International Computer Software and Applications Conference Workshops* https://doi.org/10.1109/compsacw.2014.97 (2014).
22. Webber, M. Pecan Street Dataport. *Pecan Street Inc*. https://www.pecanstreet.org/dataport/ (2013).
23. Nagasawa, K. *et al*. Data Management for a Large-Scale Smart Grid Demonstration Project in Austin, Texas. *ASME 2012 6th International Conference on Energy Sustainability* (2013).
24. Meyur, R. *et al*. Creating Realistic Power Distribution Networks using Interdependent Road Infrastructure. *2020 IEEE International Conference on Big Data (Big Data)* 1226–1235, https://doi.org/10.1109/BigData50022.2020.9377959 (2020).
25. Li, H. AlphaBuilding Synthetic Dataset. *Lawrence Berkeley National Laboratory* (2021).
26. Klemenjak, C., Kovatsch, C., Herold, M. & Elmenreich, W. A synthetic energy dataset for non-intrusive load monitoring in households. *Scientific Data* **7**, https://doi.org/10.1038/s41597-020-0434-6 (2020).
27. Thorve, S. *et al*. Simulating residential energy demand in urban and rural areas. *Winter Simulation Conference* (2018).
28. Roth, J., Martin, A., Miller, C. & Jain, R. K. Syncity: Using open data to create a synthetic city of hourly building energy estimates by integrating data-driven and physics-based methods. *Applied Energy* **280**, 115981, https://doi.org/10.1016/j.apenergy.2020.115981 (2020).
29. Tong, K., Nagpure, A. & Ramaswami, A. All urban areas energy use data across 640 districts in India for the year 2011. *Scientific Data* **8**, https://doi.org/10.1038/s41597-021-00853-7 (2021).
30. Bill, E., Shannon, G., Lee, R., & Sam, V. Synthetic populations and ecosystems of the world. Tech. Rep., Department of Statistics, Carnegie Mellon University http://stat.cmu.edu/~spew/assets/spew_documentation.pdf (2017).
31. Gallagher, S., Richardson, L. F., Ventura, S. L. & Eddy, W. F. SPEW: Synthetic Populations and Ecosystems of the World. *Journal of Computational and Graphical Statistics* **27**, 773–784, https://doi.org/10.1080/10618600.2018.1442342 (2018).
32. Beckman, R. J., Baggerly, K. A. & McKay, M. D. Creating synthetic baseline populations. *Transportation Research Part A:Policy and Practice* **30**(6), 415–429, https://doi.org/10.1016/0965-8564(96)00004-3 (1996).
33. United States Energy Information Administration. 2015 RECS Survey Data, https://www.eia.gov/consumption/residential/data/2015/. Accessed: Nov, 2017 (2015).
34. ATUS Survey. U.S. Bureau of Labor Statistics: American Time Use Survey, https://www.bls.gov/tus/datafiles_2015.htm Accessed: Mar, 2018 (2015).
35. Lum, K., Chungbaek, Y., Eubank, S. & Marathe, M. A Two-stage, Fitted Values Approach to Activity Matching. *International Journal of Transportation* **4**, 41–56 (2016).
36. Torsten, H., Kurt, H. & Achim, Z. *ctree: Conditional Inference Trees*. R package version 1.3-5 (2006).
37. Hothorn, T., Hornik, K. & Zeileis, A. Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics* **15**, 651–674, https://doi.org/10.1198/106186006X133933 (2006).
38. Barrett, C. L., Johnson, J. & Marathe, M. High Performance Synthetic Information Environments: An Integrating Architecture in the Age of Pervasive Data and Computing: Big Data (Ubiquity Symposium). *Ubiquity* **2018**, 1:1–1:11, https://doi.org/10.1145/3158342 (2018).
39. EIA. U.s. energy information administration (2020).
40. Swan, L. G. & Ugursal, V. I. Modeling of end-use energy consumption in the residential sector: A review of modeling techniques. *Renewable and Sustainable Energy Reviews* **13**, 1819–1835, https://doi.org/10.1016/j.rser.2008.09.033 (2009).
41. Muratori, M., Roberts, M. C., Sioshansi, R., Marano, V. & Rizzoni, G. A highly resolved modeling technique to simulate residential power demand. *Applied Energy* **107**, 465–473, https://doi.org/10.1016/j.apenergy.2013.02.057 (2013).
42. Shimoda, Y., Asahi, T., Taniguchi, A. & Mizuno, M. Evaluation of city-scale impact of residential energy conservation measures using the detailed end-use simulation model. *Energy* **32**, 1617–1633, https://doi.org/10.1016/j.energy.2007.01.007 (2007).
43. Kiichiro, T., Fuminori, S., Tsuyoshi, U., Osamu, S. & Takehiko, M. Bottom-Up Simulation Model for Estimating End-Use Energy Demand Profiles in Residential Houses. *Proceedings from ACEEE Summer Studies on Energy Efficiency in Buildings* (2004).
44. Subbiah, R., Pal, A., Nordberg, E. K., Marathe, A. & Marathe, M. V. Energy Demand Model for Residential Sector: A First Principles Approach. *IEEE Transactions on Sustainable Energy* **8**, 1215–1224, https://doi.org/10.1109/TSTE.2017.2669990 (2017).
45. Chuck, B. *et al*. Residential Indoor Temperature Study. *National Renewable Energy Laboratory. Technical Report NREL/TP-5500-68019* (2017).
46. Ulrike, J. & Klaus, V. Realistic Domestic Hot-Water Profiles in Different Time Scales. *Universität Marburg* 1–18 (2001).
47. de Santiago, J., Rodriguez-Villalón, O. & Sicre, B. The generation of domestic hot water load profiles in swiss residential buildings through statistical predictions. *Energy and Buildings* **141**, 341–348, https://doi.org/10.1016/j.enbuild.2017.02.045 (2017).
48. Bob, H., Jay, B. & Greg, B. Tool for Generating Realistic Residential Hot Water Event Schedules. *SimBuild Conference* (2010).
49. Rouleau, J., Ramallo-González, A. P., Gosselin, L., Blanchet, P. & Natarajan, S. A unified probabilistic model for predicting occupancy, domestic hot water use and electricity use in residential buildings. *Energy and Buildings* **202**, 109375, https://doi.org/10.1016/j.enbuild.2019.109375 (2019).
50. Jeff, M., Xia, F. & Eric, W. Comparison of Advanced Residential Water Heating Technologies in the United States. *National Renewable Energy Laboratory Technical Reports* (2013).
51. Wiehagen, J. & Sikora, J. Performance Comparison of Residential Hot Water Systems. *National Renewable Energy Laboratory Reports* (2003).
52. Hendron, R. Building America Research Benchmark Definition, Technical Report NREL/TP-550-44816. *National Renewable Energy Laboratory Reports* (2008).
53. Capasso, A., Grattieri, W., Lamedica, R. & Prudenzi, A. A bottom-up approach to residential load modeling. *IEEE Transactions on Power Systems* **9**, 957–964, https://doi.org/10.1109/59.317650 (1994).
54. Widén, J., Nilsson, A. M. & Wäckelgård, E. A combined Markov-chain and bottom-up approach to modelling of domestic lighting demand. *Energy and Buildings* **41**, 1001–1012, https://doi.org/10.1016/j.enbuild.2009.05.002 (2009).
55. Palacios-Garcia, E. *et al*. Stochastic model for lighting's electricity consumption in the residential sector. *Impact of energy saving actions. Energy and Buildings* **89**, 245–259, https://doi.org/10.1016/j.enbuild.2014.12.028 (2015).
56. Stokes, M., Rylatt, M. & Lomas, K. A simple model of domestic lighting demand. *Energy and Buildings* **36**, 103–116, https://doi.org/10.1016/j.enbuild.2003.10.007 (2004).
57. Richardson, I., Thomson, M., Infield, D. & Delahunty, A. Domestic lighting: A high-resolution energy demand model. *Energy and Buildings* **41**, 781–789, https://doi.org/10.1016/j.enbuild.2009.02.010 (2009).
58. Paatero, J. V. & Lund, P. D. A model for generating household electricity load profiles. *International Journal of Energy Research* **30**, 273–290, https://doi.org/10.1002/er.1136 (2006).
59. Tribwell, L. S. & Lerman, D. I. Baseline Residential Lighting Energy Use Study. *American Council for an Energy-Efficient Economy (ACEEE)* (1996).
60. Boardman, B. *et al*. DECADE - Domestic Equipment and Carbon Dioxide Emissions. *Energy and Environment Programme Environmental Change Unit University of Oxford* (1995).
61. Greenblatt, J., Hopkins, A., Letschert, V. & Blasnik, M. Energy use of US residential refrigerators and freezers: function derivation based on household and climate characteristics. *Energy Analysis and Environmental Impacts Department Environmental Energy Technologies Division Lawrence Berkeley National Laboratory* (2012).

62. Castro, N. S., Bowman, J. & Twigg, B. The New U.S. Department of Energy Dishwasher Test Procedure: Development and First Results. *National Institute of Standards & Technology* (2005).
63. Christopher, I., Natascha Milesi, F. & Michael A. G. Consumer Use of Dishwashers, Clothes Washers, and Dryers: Data Needs and Availability. NIST Technical Note 1696, Mechanical Systems and Control Group Building Environment Division Engineering *Laboratory, Department of Energy* (2011).
64. Nabinger, S. J. Evaluation of Kitchen Cooking Appliance efficiency Test Procedures. *National Institute of Standards and Technology, U.S. Department of Commerce* (1999).
65. EnergyStar. *Product Retrospective: TVs. ENERGY STAR Report* (2021).
66. EnergyStar. ENERGY STAR Program Requirements for Computers. *ENERGY STAR Report* (2010).
67. Palmstedt, P. Vacuum Cleaners. *ENERGY STAR Market @AND@ Industry Scoping Report* (2011).
68. Palmstedt, P. Electrolux Global Vacuuming Survey 2013 Report. *Electrolux* (2013).
69. Thorve, S., Mortveit, H. & Marathe, M. Household-level disaggregated hourly synthetic residential energy use profiles for the United States. *University of Virginia Dataverse* https://doi.org/10.18130/V3/VJUZSH (2022).
70. Michael, B. C., Theresa, G. L., C., P. C., Marye, H. & Kathi, R. High-Performance Home Technologies: Guide to Determining Climate Regions by County. *Pacific Northwest National Laboratory* **7.3**, 1–50, https://www.energy.gov/eere/buildings/downloads/building-america-best-practices-series-volume-73-guide-determining-climate (2015).
71. Residential building stock assessment (rbsa) metering data, northwest energy efficiency alliance. https://neea.org/data/residential-building-stock-assessment. Accessed: 2022-03-23.
72. Souza, V., Estrada, T., Bashir, A. & Mueen, A. LADPU Smart Meter Data. *Dryad* https://doi.org/10.5061/dryad.m0cfxpp2c (2020).
73. Lin, J. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory* **37**, 145–151, https://doi.org/10.1109/18.61115 (1991).
74. Kwac, J., Flora, J. & Rajagopal, R. Household energy consumption segmentation using hourly data. *IEEE Transactions on Smart Grid* **5**, 420–430, https://doi.org/10.1109/TSG.2013.2278477 (2014).
75. Klemenjak, C., Kovatsch, C., Herold, M. & Elmenreich, W. SynD: A Synthetic Energy Dataset for Non-Intrusive Load Monitoring in Households. *figshare* https://doi.org/10.6084/m9.figshare.c.4716179 (2020).
76. Kolter, J. Z. & Johnson, M. J. REDD: A Public Data Set for Energy Disaggregation Research. *SustKDD workshop on Data Mining Applications in Sustainability* (2011).
77. Kolter, J. Z. & Johnson, M. J. REDD: The Reference Energy Disaggregation Data Set. *MIT Initial REDD Release, Version 1.0* http://redd.csail.mit.edu/ (2011).
78. Makonin, S., Wang, Z. J. & Tumpach, C. RAE: the rainforest automation energy dataset for smart grid meter data analysis. *CoRR* **abs/1705.05767**, http://arxiv.org/abs/1705.05767 (2017).
79. Murray, D., Stankovic, L. & Stankovic, V. An electrical load measurements dataset of United Kingdom households from a two-year longitudinal study. *Scientific Data* **4**, https://doi.org/10.1038/sdata.2016.122 (2017).
80. Murray, D., Stankovic, L. & Stankovic, V. REFIT: Electrical Load Measurements (Cleaned). *University of Strathclyde* https://doi.org/10.15129/9ab14b0e-19ac-4279-938f-27f643078cec (2015).
81. Rashid, H., Singh, P. & Singh, A. I-BLEND, a campus-scale commercial and residential buildings electrical energy dataset. *Scientific Data* **6**, https://doi.org/10.1038/sdata.2019.15 (2019).
82. Rashid, H., Singh, P. & Singh, A. I-BLEND, a campus-scale commercial and residential buildings electrical energy dataset. *figshare* https://doi.org/10.6084/m9.figshare.c.3893581 (2019).
83. Paige, F., Agee, P. & Jazizadeh, F. flEECe, an energy use and occupant behavior dataset for net-zero energy affordable senior residential buildings. *Scientific Data* **6**, https://doi.org/10.1038/s41597-019-0275-3 (2019).
84. Paige, F. & Agee, P. flEECe, an Energy Use and Occupant Behavior Dataset for Net Zero Energy Affordable Senior Residential Buildings. *Open Science Framework* https://doi.org/10.17605/OSF.IO/2AX9D (2019).
85. Shin, C. *et al*. The ENERTALK dataset, 15 Hz electricity consumption data from 22 houses in Korea. *Scientific Data* **6**, https://doi.org/10.1038/s41597-019-0212-5 (2019).
86. Shin, C. *et al*. The ENERTALK dataset, 15 Hz electricity consumption data from 22 houses in Korea. *figshare* https://doi.org/10.6084/m9.figshare.c.4502780 (2019).
87. Kelly, J. & Knottenbelt, W. The UK-DALE dataset. *UKERC Energy Data Centre* https://doi.org/10.5286/UKERC.EDC.000002 (2015).
88. Anderson, K., Ocneanu, A., Carlson, D. R., Rowe, A. G. & Bergés, M. BLUED: A Fully Labeled Public Dataset for Event-Based Non-Intrusive Load Monitoring Research. *Proceedings of the 2nd KDD Workshop on Data Mining Applications in Sustainability* (2012).
89. Anderson, K. Dataset Name: Building-Level fUlly labeled Electricity Disaggregation dataset (BLUED). *github* https://tokhub.github.io/dbecd/links/Blued.html (2011).
90. Barker, S. *et al*. An Open Data Set and Tools for Enabling Research in Sustainable Homes. *Proceedings of the 1st KDD Workshop on Data Mining Applications in Sustainability (SustKDD)* (2012).
91. Barker, S. UMass Smart* Dataset - 2017 release. *UMassTraceRepository* https://traces.cs.umass.edu/index.php/smart/smart (2017).
92. Beckel, C., Kleiminger, W., Cicchetti, R., Staake, T. & Santini, S. The ECO Data Set and the Performance of Non-Intrusive Load Monitoring Algorithms. *Proceedings of the 1st ACM Conference on Embedded Systems for Energy-Efficient Buildings* 80–89, https://doi.org/10.1145/2674061.2674064 (2014).
93. Pereira, L., Quintal, F., Gonçalves, R. & Nunes, N. SustData: A Public Dataset for ICT4S Electric Energy Research. *ICT4S* (2014).
94. Pereira, L. SustData: A Public Dataset for ICT4S Electric Energy Research. *Open Science Framework* https://osf.io/2ac8q/ (2021).
95. Pereira, L., Costa, D. & Ribeiro, M. A residential labeled dataset for smart meter data analytics. *Scientific Data* **9**, 134, https://doi.org/10.1038/s41597-022-01252-2 (2022).
96. Monacchi, A., Egarter, D., Elmenreich, W., D'Alessandro, S. & Tonello, A. M. GREEND: An energy consumption dataset of households in Italy and Austria. *2014 IEEE International Conference on Smart Grid Communications (SmartGridComm)* 511–516, https://doi.org/10.1109/SmartGridComm.2014.7007698 (2014).
97. Monacchi, A., Egarter, D., Elmenreich, W., D'Alessandro, S. & Tonello, A. M. GREEND: An energy consumption dataset of households in Italy and Austria. *Duke Energy Initiative Lakeside Labs* https://energy.duke.edu/content/greend-electrical-energy-dataset (2021).
98. Pullinger, M. *et al*. The IDEAL household energy dataset, elecstricity, gas, contextual sensor data and survey data for 255 UK homes. *Scientific Data* **8**, 146, https://doi.org/10.1038/s41597-021-00921-y (2021).
99. Goddard, N. *et al*. The IDEAL Household Energy Dataset. *Edinburgh DataShare* https://doi.org/10.7488/ds/2836 (2021).
100. Ruhnau, O., Hirth, L. & Praktiknjo, A. Time series of heat demand and heat pump efficiency for energy system modeling. *Scientific Data* **6**, https://doi.org/10.1038/s41597-019-0199-y (2019).
101. Ruhnau, O. When2Heat Heating Profiles. *Open Power System Data* https://doi.org/10.25832/when2heat/2019-08-06 (2019).
102. Deming, W. E. & Stephan, F. F. On a Least Squares Adjustment of a Sampled Frequency Table When the Expected Marginal Tables are Known. *Annals Math. Stats* **11**, 427–444 (1940).
103. Fienberg, S. E. An Iterative Procedure for Estimation in Contingency Tables. *The Annals of Mathematical Statistics* **41**, 907–917 (1970).

104. Public Use Microdata Sample (PUMS). PUMS Documentation, https://www.census.gov/programs-surveys/acs/microdata/documentation.2013.html. Accessed: Nov, 2017 (2013).
105. Land Data Assimilation System. North American Land Data Assimilation System (NLDAS) Climate Data, https://ldas.gsfc.nasa.gov/nldas/. Accessed: Mar, 2018 (2016).
106. National Renewable Energy Laboratory (NREL). *National Solar Radiation Database (NSRDB)*, https://nsrdb.nrel.gov/data-sets/us-data. Accessed: Nov, 2020 (2014).

Acknowledgements

We thank the anonymous reviewers for their very helpful comments that helped us improve the manuscript. We thank members of National Rural Electric Cooperative Association (NRECA) for providing validation data for Rappahannock county, Virginia and Horry county, South Carolina. This work is partially supported by University of Virginia Strategic Investment Fund award number SIF160, NSF EAGER CMMI-1745207, NSF Grant OAC-1916805, and NSF BIGDATA IIS-1633028.

Author contributions

S.T. collected data for the models, developed and implemented the modeling framework and all the individual energy use models, prepared the manuscript; Y.Y.B. implemented the ATUS model, read and edited the paper; H.M and S.S. helped with model development, data collection, read and edited the manuscript, provided guidance in writing Background and Summary; M.M. worked on model development, validation, manuscript preparation, provided feedback for the Methodology and Validation section; A.M. read and edited the manuscript and helped with validation; A.V. edited the manuscript and helped with validation. All authors participated in writing and reviewing the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to S.T. or M.M.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.