A Method to Generate Heating and Cooling Schedules Based on Data from Connected Thermostats

Journal: Energy and Buildings

Tsuyoshi Ueno and Alan Meier
Lawrence Berkeley National Laboratory
January 2020

Abstract

200 words max. We'll write the abstract when we are finished

Highlights

4 - 6 highlights, max 90 characters
Highlight #1
Highlight #2
Highlight #3
Highlight #4
Highlight #5

1. Introduction

1.1. Simulation of Building Energy Consumption

Building energy simulation is now an accepted practice to provide a quantitative assessment for estimating energy consumption, compliance with building codes, and determining the size of key equipment. Simulation is also used to explore the impacts of design changes and, more recently, comfort and health implications. Researchers have steadily improved techniques to model heat transfer, equipment, and controls operation (Lomas et al. 1997; Li and Wen 2014). At the same time, measurements of actual weather conditions have been refined, both in accuracy and frequency. The result has been increasing accuracy in the models' estimates of building's energy consumption in actual conditions.

With the improving precision of modeling the performance of materials, equipment, and controls, the greatest uncertainty in predictions of a building's energy consumption is increasingly the indoor temperatures (Booten et al. 2017). Building scientists often arbitrarily select temperature schedules for case studies. Arbitrary temperatures make sense because the *differences* between two simulations are more important than their absolute results. This is the case when investigating the relative merits of different energy-saving technologies. However, simulations undertaken to satisfy policy objectives require indoor temperatures that reflect actual practices (Hendron and Engebrecht 2010) . These situations include building energy codes, health codes, and resilience, where the analyst must compare the costs of improvements to the value of energy savings or other benefits. A change of 0.5°C in the indoor temperature assumption can raise or lower a home's predicted heating or cooling use up to 10% (Booten et al. 2017). Using realistic temperatures—temperatures found in actual buildings—is therefore an important input to simulations (Seryak and Kissock 2003). A recent application is predicting the performance of heat pump water heaters placed inside the conditioned space. The efficiencies of these devices depend on the absolute ambient temperature (Amirirad, Kumar, and Fung 2018). The problem of obtaining realistic indoor temperatures and schedules becomes even more difficult when seeking to estimate regional or national benefits from improvements in building performance.

Actual temperature schedules inside buildings can be obtained either from direct measurements or from surveys. Each approach has advantages and limitations; these are summarized below.

Researchers have measured temperatures in individual buildings or groups of buildings for many decades. Notable studies have taken place in Japan (H. Yoshino et al. 2006) , China (Hiroshi Yoshino et al. 2006), United States (Roberts and Lay 2013), Ireland (Healy and Clinch 2002), and Sweden (Johansson, Bagge, and Lindstrii 2013). These measurements are typically collected in support of other goals, such as understanding thermal comfort, health effects, or

performance of building components. Temperature measurements have grown more detailed and extensive as the cost of sensors and data collection have declined. Temperature data collected through measurements are ideal for simulations because the researcher can understand the precise locations and frequency of measurements and then ensure that the simulation is consistent. The limitation of this approach is that most measurements are undertaken in small groups of buildings and for limited periods. Thus, measurements provide highly accurate temperature schedules but they are difficult to extrapolate to larger populations or even for the whole year.

Surveys are often used to collect temperature information for input to simulations. The surveys typically ask occupants to provide temperature settings in their homes during principal activities (sleeping, socializing, etc.). One of the most reputable and longest-running surveys in North America, is the U.S. Residential Energy Consumption Survey, RECS (EIA 2015). The survey is repeated every four years and, in 2015, about six thousand homes were surveyed. The sample is carefully selected to represent the entire stock of U.S. homes. RECS asks survey respondents to report just six indoor temperatures: when they are home, away, or sleeping for both winter and summer. The survey provides an excellent window into national heating and cooling habits, both lattitudenally and longitudinally. The survey results are far better than no information but leave considerable uncertainty in actual temperature preferences. Like all surveys, errors and inconsistencies can arise in self-reported temperatures and schedules. For example, the survey respondent may not be the person responsible for controlling the home's temperature. Each type of thermostat used to control the temperature—manual, programmable, Internet-connected, or no thermostat at all—has a different relationship between settings and actual temperatures. Changes in behavior during periods when the occupants are on vacation—often ten percent of the time—are not captured, too. In general, the data on temperatures and schedules derived from surveys are much less precise than the other inputs used in a building simulation.

Survey results are especially problematic for simulations because these responses must be translated into hourly, indoor temperatures. The researcher must further decide how to allocate the responses across weekends, holidays, and other situations.

In summary, both approaches to collecting temperature data have limitations and both cannot be used to accurately capture regionally representative temperature settings and schedules.

1.2. The Internet-Connected Thermostat

In about 2010 the first Internet-connected thermostats appeared were offered to consumers. These thermostats used an Internet connection (typically through Wifi) to communicate operating data to the thermostat vendor (in the "cloud") and to receive operating instructions from the vendor. The Internet connection enabled many new features to be offered to customers, such as control via smartphones and optimized operation of the homes' heating and

cooling systems. Now, in 2020, we estimate that about twenty million Internet-connected (or "communicating") thermostats have been installed in North American homes. This corresponds to roughly 15 percent of stock. About four million European homes have thermostats. The market appears to be growing at about fifteen percent per year in response to the new features the thermostats provide, incentives offered by utility companies, and the opportunity to more conveniently save energy.

Connected thermostats continuously transmit data to their vendors. Every five minutes, typical units transmit the following:

- Setpoint (or target) temperature
- Actual temperature in the home or zone
- Runtime of the heating or cooling system during the previous interval

Many models also detect motion, humidity, and detailed operating characteristics of the HVAC system. Recently, vendors have begun offering additional temperature sensors that can be placed in other zones to assist in more precise heating and cooling strategies.[1] The richness of the DYD data is revealed in Figure tlkGrape, a heat map of temperatures for one year. The vertical axis shows one week (5 minutes x 24 hours x 7 days), and the horizontal axis shows the week of one year (52 weeks). The difference between the daytime and nighttime temperatures appear as horizontal stripes (except on weekends). The seasonal transitions appear as one moves from right to left. Data gaps appear as white spaces.

[1] Note that most connected thermostats cannot link to electrical "smart meters" and are not capable of collecting concurrent energy consumption data.

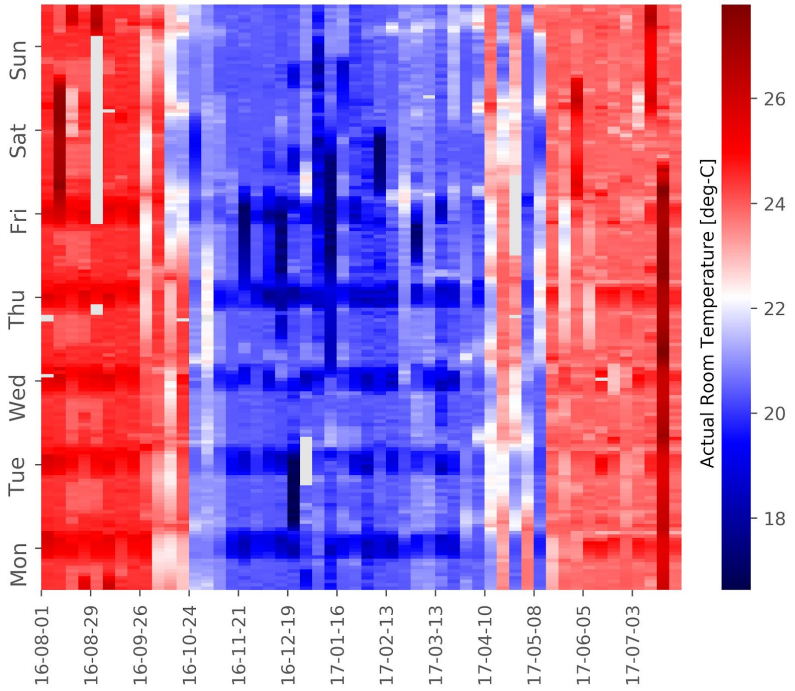


Figure tlkGrape. Temperatures in a typical DYD home for one year.

Runtimes of the home's heating and cooling system—more difficult to display—and other sensor outputs complement this temperature data.

Data from connected thermostats would appear to be an excellent source of temperature information. Unfortunately, most thermostat vendors have not shared this data in order to protect customer privacy (and possibly valuable market information). European (and other regions') data privacy laws may also prevent releasing this information. In at least two cases, vendors worked with researchers and provided them thermostat data. Booten et al. (2017) analyzed thermostat data from about 12,000 homes distributed across the United States. With it, they were able to estimate temperatures by climate region. Ge and Ho (2018) used thermostat data from 27,000 American homes to study the persistence of habits in consumers' temperature setting behavior. In both cases, however, the investigators had no additional information about the homes beyond their locations, which limited the scopes of their analyses.

In 2015 one thermostat vendor, ecobee, established an experimental program called "Donate Your Data (DYD)" where its customers could "donate" their data to researchers (Ecobee Inc. 2018). It further asked the "donors" to provide limited information about the household, including the city, the home's floor area, type of heating system, age of home, and the number of occupants. The customer names and all personally identifiable information were removed. Purchasers of new ecobee thermostats were offered the opportunity to "opt-in" at the time of registration. The program has attracted a growing number of participants. As of August 2018,

about 50,000 households have joined the DYD program by 2018. The trajectory of registrations, and the major geographical locations of the DYD participants is shown in Figure tlkCherry. The overwhelming majority of the thermostats are located in the United States.

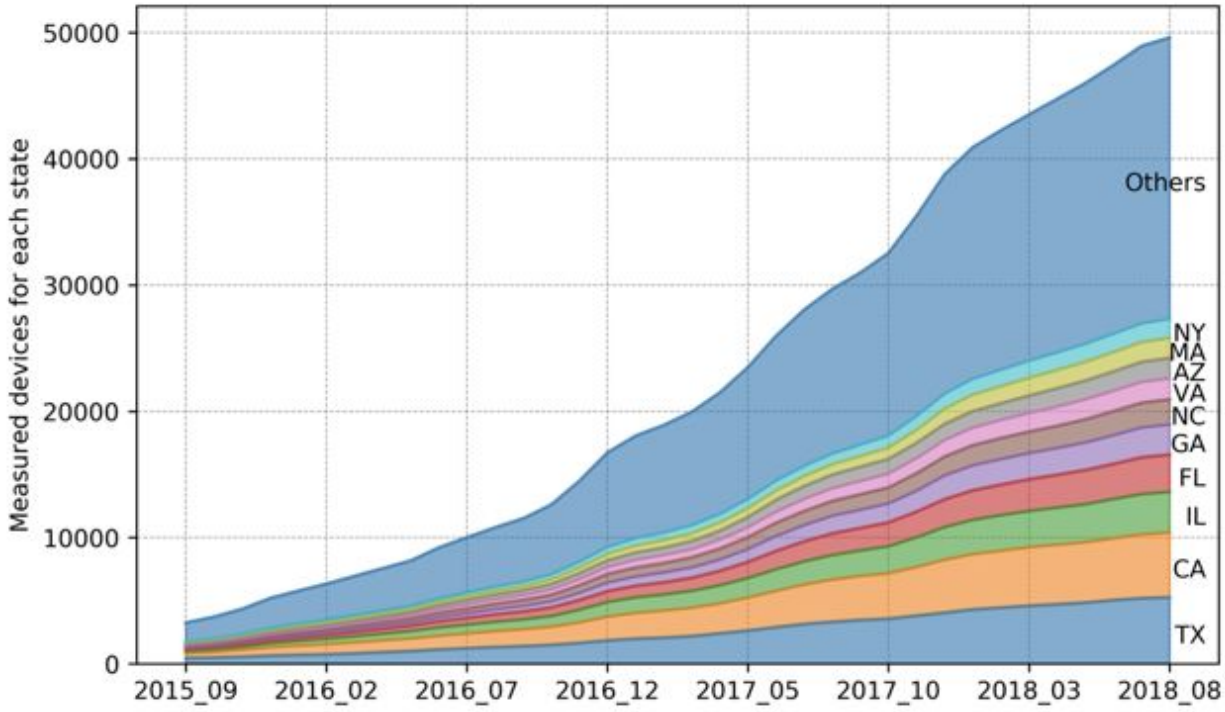


Figure tlkCherry. Growth of participants in the DYD program in the United States. The 2-letter codes refer to individual states and "Others" are the remaining states.)

Researchers have already begun to explore the DYD dataset and extract information about occupant behavior and peak demand (Meier et al. 2019), occupant temperature preferences (Huchuk, O'Brien, and Sanner 2018), estimating energy savings from thermostats (Daken, Meier, and Frazee 2016), and using the network to track power outages (Meier, Ueno, and Pritoni 2019). However, nobody has converted the DYD data into representative temperature schedules. In this paper, we present a method to convert actual temperatures recorded in DYD homes into data suitable for representative building simulations of American homes. We begin by comparing evaluating the representativeness of DYD homes. Then we present a method to convert DYD temperatures into a user-selected set of prototypes that capture the diversity in operating behaviors. Finally, we illustrate the tool with some examples.

2. How Representative are the DYD Homes?

Before developing representative operating schedules for American homes from DYD data, it is necessary to confirm that the participants in the DYD homes accurately reflect the diversity of homes in the United States as a whole. Each participant filled out a questionnaire. So as to not

discourage people from filling out questionnaires, ecobee avoided asking standard economic and demographic questions that could be easily compared to the census. We therefore relied on indirect methods of comparison described below.

Roughly 50,000 homes participated in the DYD program as of August 2018. While this is a large number, the sample suffers from obvious biases. The DYD homes are a triply self-selected sample. First, people buying ecobee thermostats require reliable broadband Internet connections (greater than 1 Megabyte/sec) and wifi networks in their homes. About 6% of American households lack broadband access. Most of those homes with inadequate connections are located in rural areas. Second, connected thermostats are still a new technology, so people who buy ecobee thermostats are probably early adopters and more technically proficient than the average. (This bias may be diluted somewhat by numerous utility programs subsidizing purchases.)[2] Finally, only a unique group of ecobee customers will choose to opt in to the DYD program and fill out the questionnaire. For all of the above reasons, the DYD sample is likely to not reflect the actual population and housing stock in the United States.

To understand the extent of this bias, we compared the DYD homes to the U.S. Department of Energy's Residential Energy Consumption Survey, RECS (see above). The RECS surveys only about 6000 homes, but the Department of Energy rigorously ensures that the homes accurately reflect the whole population. Our approach to exploring sample bias was to compare findings from similar questions in the DYD questionnaire and RECS.

According to RECS, about 63% of American households are detached single-family homes. In the DYD sample, roughly 63% are also single-family detached homes. However, the categories in the DYD questionnaire do not map directly into the RECS categories. About 18% of the DYD homes are in the self-described categories of "townhouse", "condominium", "rowhouse", and "semi-detached," compared to 6% in the single RECS category of "single-family attached". RECS estimates that about 26% of American households are apartments but only 5% of the DYD participants reported living in apartments. This bias towards single-family homes (detached and attached) is to be expected because ecobee thermostats are not compatible with most apartment heating and cooling systems. For that reason, we compared the DYD homes to RECS single-family homes (detached and attached).

We performed additional comparisons between DYD and RECS data, including: geographic distribution, floor area, number of occupants, type of heating system, and age of home. Three comparisons are presented graphically in Figure tlkApple, Figure tlkBanana, and Figure tlkOrange

[2] For competitive reasons, ecobee was not able to share with us the demographics of its customers.

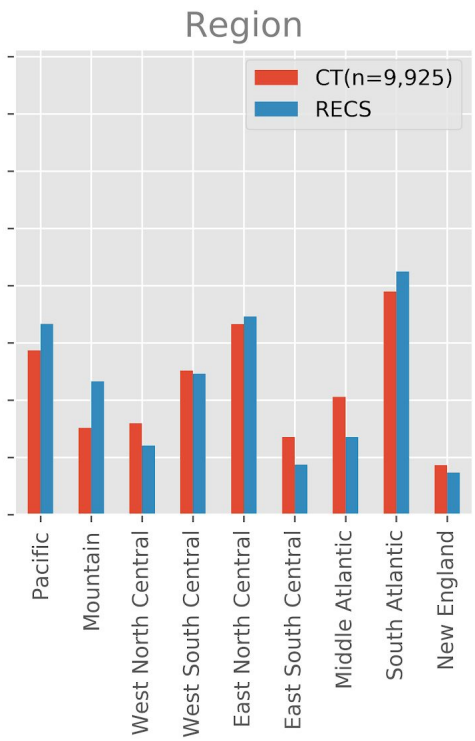


Figure tlkApple. Geographical distribution of DYD participants compared to RECS

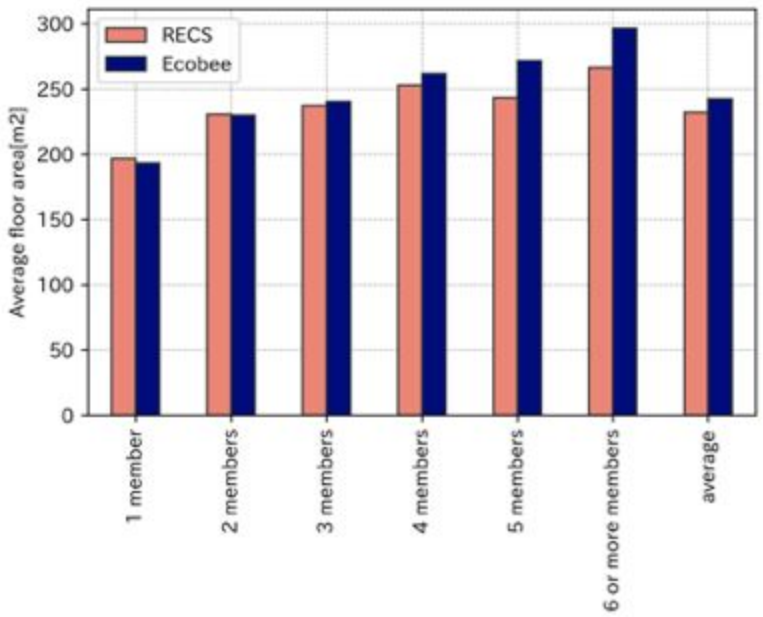


Figure tlkBanana. Distribution of floor areas with respect to number of occupants for DYD participants and RECS

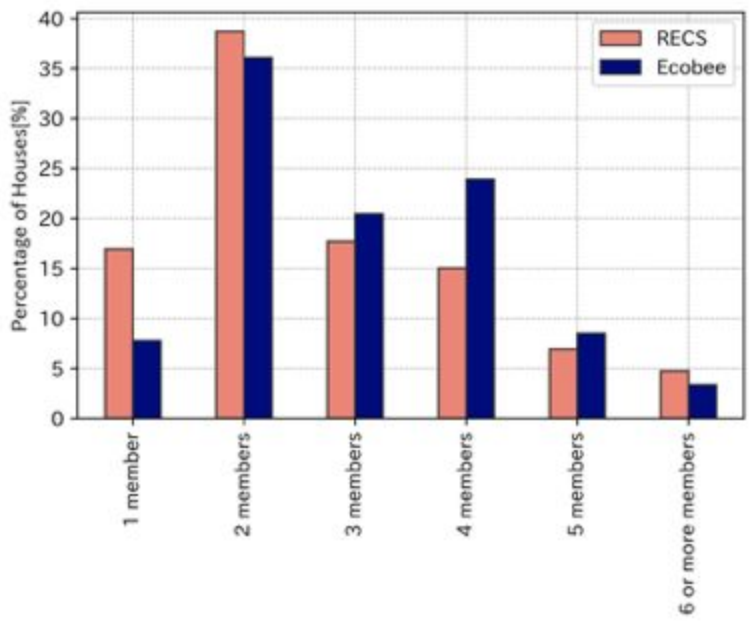


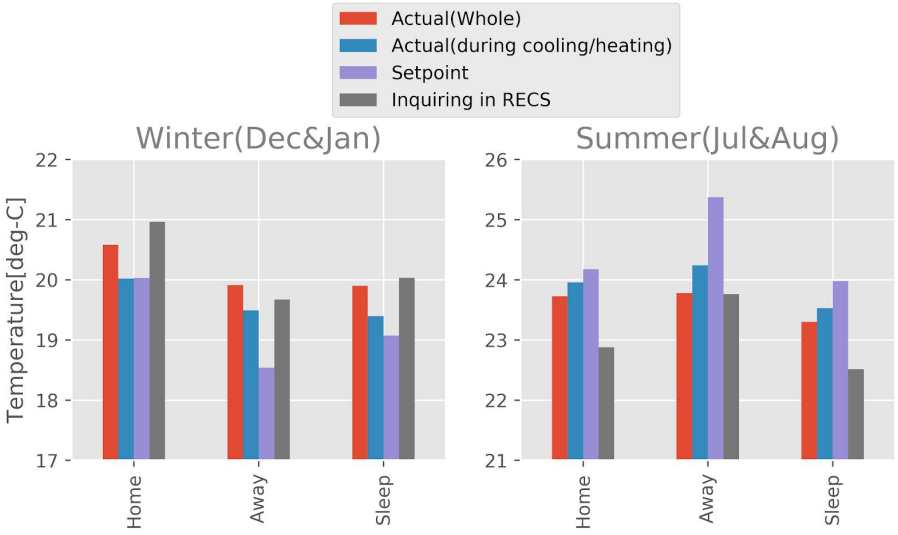
Figure tlkOrange. Distribution of number of occupants in homes for DYD participants and single-family detached homes in the RECS sample

There are some differences between the groups but fewer than one might expect. The DYD homes are geographically distributed roughly the same as RECS (Figure tlkApple) There are relatively fewer DYD homes than RECS homes in the Mountain region and more in the mid-Atlantic region, but the overall differences are small. Figure tlkBanana shows the floor areas in the different regions. The DYD and RECS homes are nearly the same size—DYD average floor area is only 4% larger. The relationships between the number of occupants and floor area are also similar. Figure tlkOrange shows the distribution of occupants in DYD and RECS homes. With the exception single-occupant homes, the number of occupants in the two groups are similar. For example, 35%–40% of the homes in both groups have two occupants.

The age distribution of homes in the two groups is also similar (data not shown). The heating systems differ because the ecobee thermostat is not fully compatible with electric resistance heating systems and heat pumps (data not shown).

It is also possible to compare measured temperatures in the DYD homes to temperatures reported by the occupants in RECS homes. The RECS survey asks occupants to report temperature settings while at home, sleeping, and away for both winter and summer. In this comparison, the median value of the responses was used. Ecobee adopted the same terms for its primary settings (or modes): Home, Sleep, and Away. Unlike RECS, ecobee collects both the setpoint (that is, the desired temperature) and the actual temperature. These may differ because of periods when the actual temperature floats above the setpoint or “smart recovery” is

enabled. The results of the comparison are shown in Figure tlkLemon. The Figure also displays the average temperatures for the whole heating and cooling seasons.



2.0.1. Figure tlkLemon. Comparisons of temperatures in DYD and RECS households

The temperatures follow expected behaviors, that is, in the winter temperatures are highest (warmest) when occupants are at home and lowest (coolest) when they are away and reversed in the summer. The impacts of floating temperatures can be observed by comparing the actual and setpoint DYD temperatures. In the winter the actual temperatures are slightly higher than the setpoints during Away and Sleep periods. In the summer, the DYD Away setpoint is significantly higher than the actual temperature, possibly because it captures cooler periods when no air conditioning was needed.

The RECS respondents report significantly higher (warmer) setpoints during the winter than measured setpoints in the DYD homes. This trend applies for Home, Away, and Sleep periods. The relationship continues during summer, that is, RECS setpoints are higher (warmer) than those measured in DYD homes. In general the RECS occupants appear to set their thermostats so that they are less comfortable—colder in the winter and warmer in the summer—than occupants of the DYD homes. It is not clear if this is a difference in behaviors or an artifact of the data collection techniques. The two groups have more similar temperatures when the DYD temperature (rather than the setpoint) is compared to the RECS values.

In summary, the DYD homes are not perfectly representative of the stock of single-family homes but they are reasonably similar with respect to location, floor area, number of occupants, and age of the homes. It is still possible that the occupants of the DYD homes differ greatly with respect to income or education, but there is no evidence suggesting this.

3. A Method for Creating Representative Temperature Schedules

3.1. Technical Approach

No single temperature schedule can represent the wide range of temperatures and schedules. Simulations of a home's energy use based on average conditions are likely to be highly misleading. They would, for example, not capture homes operated under extreme conditions, where energy consumption might be especially high. One solution is to construct a set of schedules that capture this diversity. Ten temperature schedules would more effectively capture this diversity (and a million would be even better). The technical challenge, however, is determining the correct weighting for the different temperature schedules so that the combinations of the simulated homes reflect the national situation. The DYD data provides the necessary information to create sets of representative temperature schedules. The method of generating representative prototype temperature schedules is described below.

Our approach to generating representative temperature schedules builds upon patterns observed in the DYD data. These data enable us to identify the variables that strongly affect temperatures and schedules. As described earlier, ecobee thermostats divide the day into three technical modes: Home, Away, and Sleep. The frequencies of these modes at each hour were calculated for every hour. These frequencies were calculated separately for weekdays and weekends because the distributions are so different (see Figure tlkLime). Annual data were used to calculate the frequencies.[3]

[3] Frequencies based on monthly (rather than annual) temperatures could be calculated, but this would require much more computation.

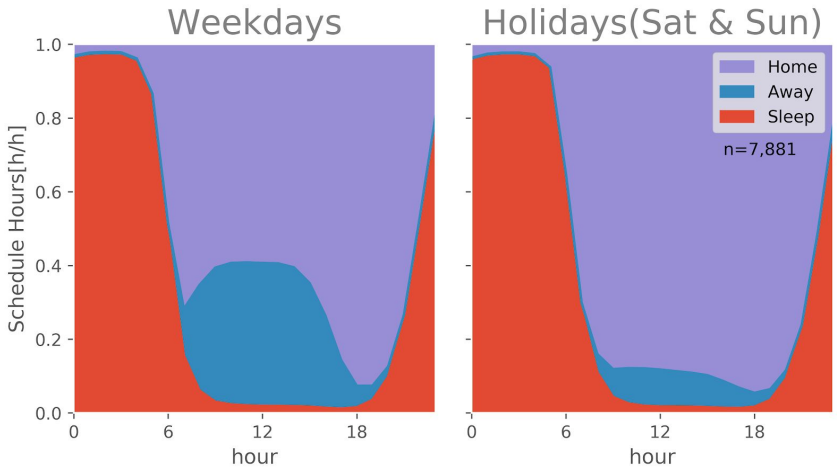


Figure tlkLime. Frequencies of occurrence of Home, Away, and Sleep for weekdays and weekends.

The figures show that on weekdays, about 95% of the homes are in Sleep mode until about 5 AM and then drops rapidly to a minimum near noon. Meanwhile, the fraction of homes in Away mode climbs sharply after 6 AM to almost 40% at noon. The maximum fraction of homes in Home mode occurs about 18:00. On weekends, the Sleep mode extends about one hour later and the fraction of homes in Away mode is much less than half that of weekdays.

The number of occupants also affects the time the house resides in each mode. Figure tlkPersimmon show the impact of the number of occupants on the occurrence of Home mode. Not surprisingly, the fraction of homes in Home mode increases with the number of occupants. This phenomenon is especially strong near 14:00 on weekdays, where single-occupancy homes are 0.5 while 6-person homes are 0.75. These differences almost vanish on weekends. The DYD data set was large enough to examine regional variations in occupancy; however, no significant differences were found.

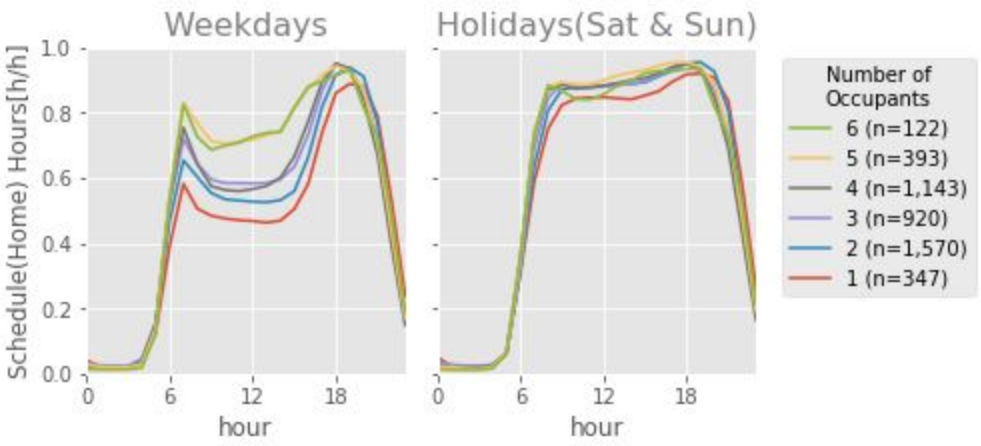
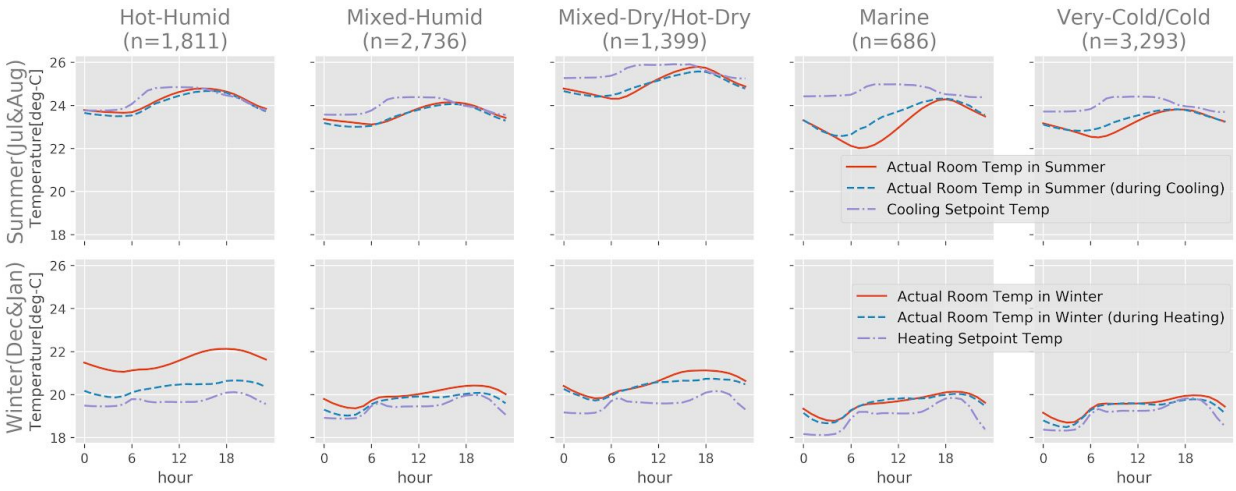


Figure tlkPersimmon. Influence on number of occupants on fraction of time in Home mode.

Figure tlkLemon summarizes the hourly setpoints and temperatures for the entire country. However, the richness of the DYD dataset allows further disaggregation of temperatures into five separate climate zones defined by Building America for its prototypes (see Figure tlkMelon). Variations between the climate zones are easily observed. For example, during the summer the setpoints in the Mixed-Dry/Hot-Dry regions are significantly higher (warmer) than in the Hot-Humid regions. During the winter, the setpoints show less variation; however, the homes in the Marine region have lower nighttime setpoints.



3.1.3. Figure tlkMelon. Room temperatures for each climate zone

The above analyses identified several variables that affect a home's heating and cooling energy consumption—temperatures, schedules, number of occupants, and climate zone—and should be taken into account when simulating a home's HVAC use. The DYD dataset makes it possible to quantify the frequency of occurrence of these factors. In the following sections, a method is described to generate an arbitrary number of typical schedules and temperatures for inputs to simulation models.

3.2. Generating Typical Temperatures and Schedules for Simulation Model Inputs

A program was written to generate typical temperature setpoints and schedules for use in simulation models based on the DYD data. For example, if a user wishes to represent the entire range of residential temperatures and schedules in the United States with six input files for their simulations, what should they be? This method can provide 1 - 40 input files. The logic behind the procedure is described below.

The program consists of two major procedures: a method for generating temperature setpoints and a method for generating the typical schedules. The procedure to generate setpoint temperatures is shown as a flow chart in Figure tklNashi. First, the DYD data must be organized for simple computation. For each home, the distribution of the setpoint temperatures is acquired for each season (Summer/Winter), mode (Home/Sleep/Away), and climate zone (five separate zones + all zones), and loaded into a database. These data are similar to the “setpoint” temperatures shown in Figure tlkLemon, but now assembled for each home.

Before generating setpoint temperatures, the user must specify the "number of desired samples (N)", that is, the number of input files to be generated. The generator then outputs a setpoint temperature schedule that can be used as an input for simulation of any climate zone, season, or schedule. The empty box in the flow chart represents the end of the loop for XXX.

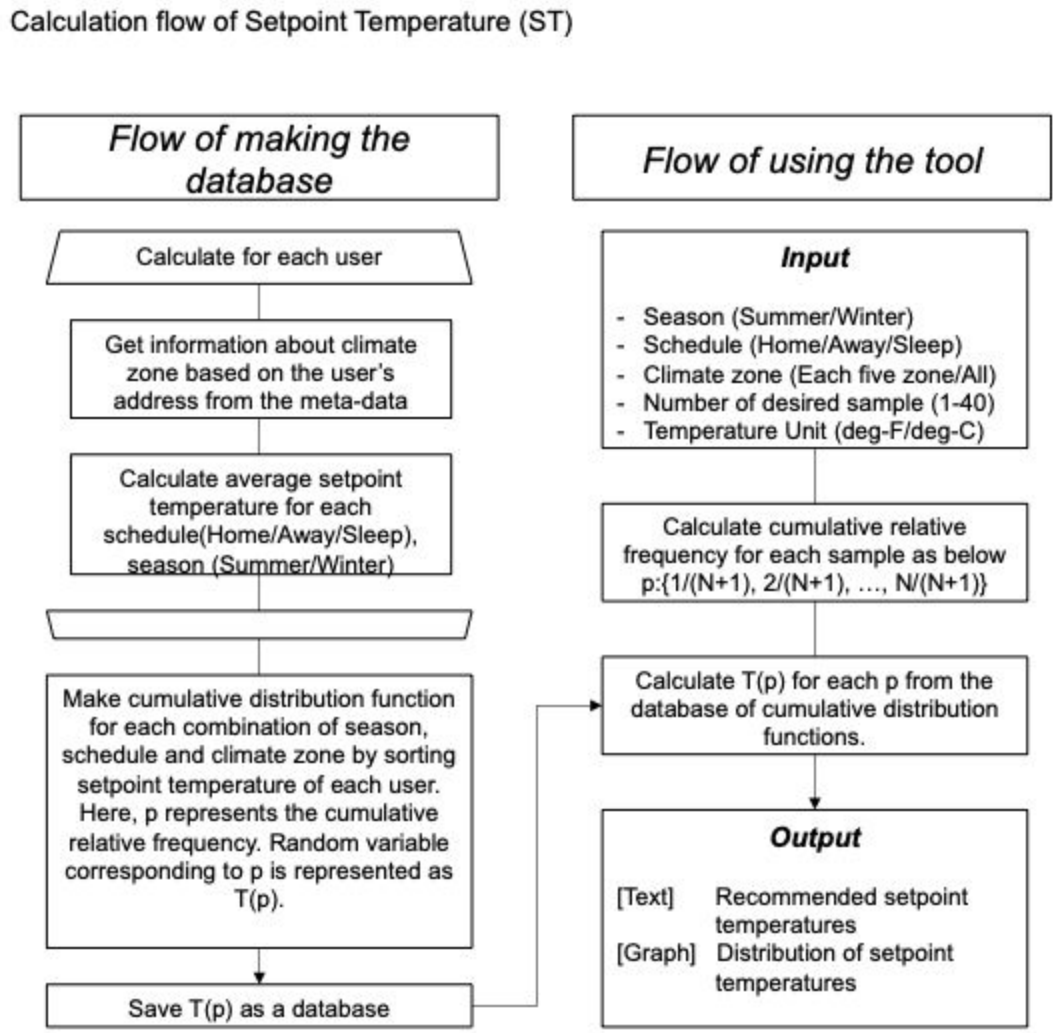


Figure tklNashi. Flow chart showing procedure to generate setpoints

The calculation method is straightforward. For each parameter, the setpoint temperature in the database is sorted in ascending order to create a distribution function T(p). The set of setpoint temperatures represented as follows is output,

T (k / (N + 1)) for k in 1,2, ..., N <equation 1>

That is, the entire distribution is divided into (N + 1) digits, and the value of the delimiter is output. When N = 1, T (0.5): the median of all distributions is output, and when N = 4, four values of T(0.2), T(0.4), T(0.6), and T(0.8) are output. Figure tlkMango illustrates the setpoints calculated from this procedure for N=4 and Figure tlkPlum illustrates the setpoints calculated for N=20.

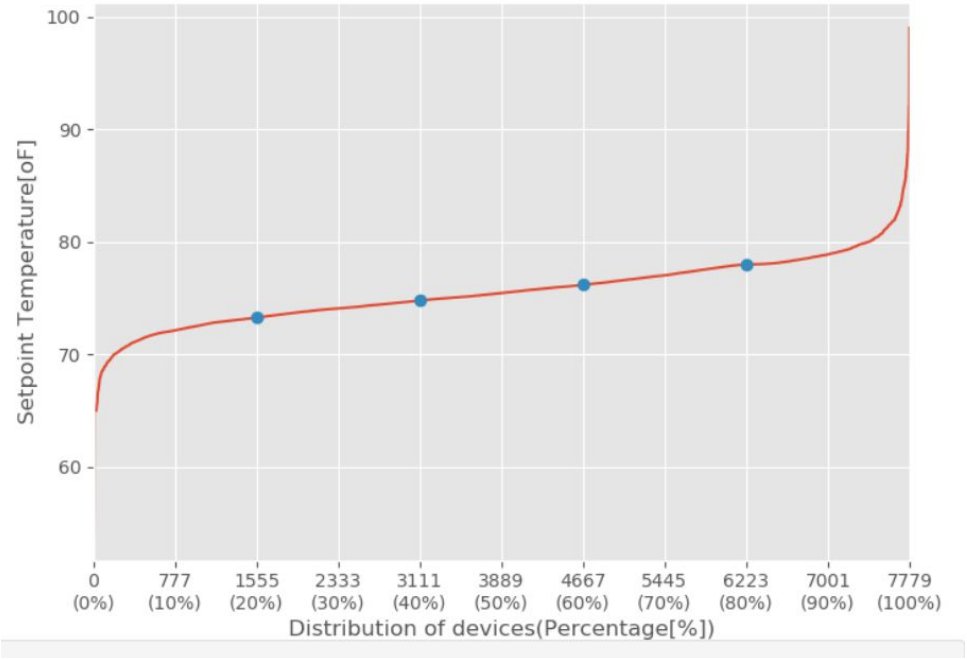


Figure tlkMango. **Distribution of setpoints in Home mode for all climate zones in summer with N = 4.**

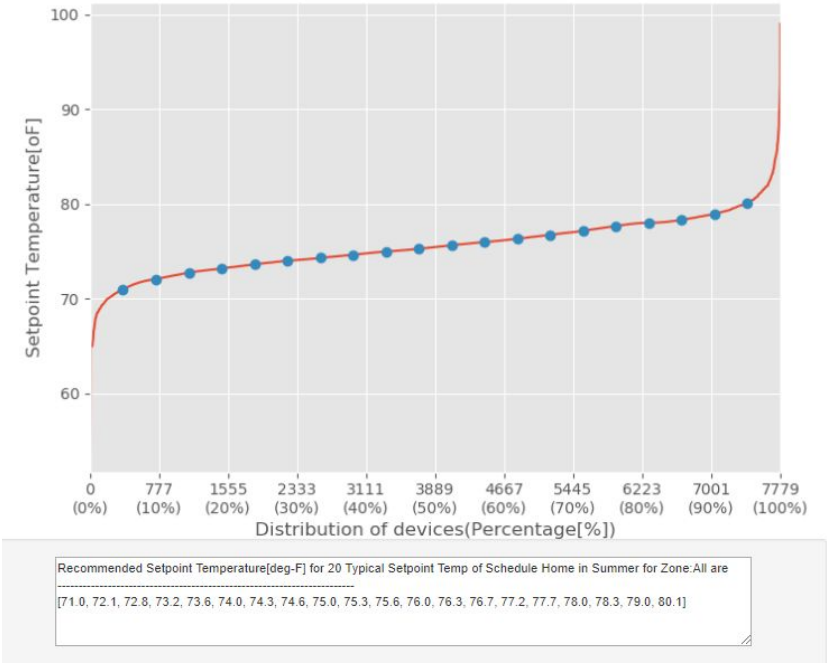


Figure tlkPlum. Distribution of setpoints in Home mode for all climate zones in summer with N = 20.

Figure tlkPlum shows the recommended setpoints for use in simulations during the summer while the thermostats are in Home mode when 20 files have been selected. The distribution is interpreted as follows: roughly 10% of the homes have setpoints below 72°F, 50% of the homes have setpoints below 75°—the median setpoint—and 10% of the homes have setpoints above 79°. Similar distributions can be generated for other modes and, if desired, specific climate zones. It is interesting to observe that about 80% of setpoints lie between 72° and 79°.

The advantage of increasing N appears in the extremes; the fraction of homes with either very low or high setpoints are explicitly captured. These homes, for example, might be vulnerable to moisture problems.

The second component of the tool generates schedules. The methodology is summarized in Figure tklIchigo. First, the DYD schedule data must be organized for simple computation. This organization is identical to the setpoint temperatures, that is, for each home, the distribution of the setpoint temperatures is acquired for each season (Summer/Winter), mode (Home/Sleep/Away), and climate zone (five separate zones + all zones), and loaded into a database. Next, the average number of hours [h/h] of each schedule of the day of week (Weekday/Holiday) and every hour (0 - 23 o'clock) for all target households is acquired, and a database is created for each number of occupants(1 - 6 persons + whole). In addition to the above parameters, the generator also has a "number of desired sample(N)" of schedules to be acquired as an input, and outputs a schedule that can be used as an input of simulation for any day of the week or number of occupants. The empty boxes represent the end of the loop for each "hour", "day of week", "user", etc.

Clustering techniques are applied to identify the representative schedules. The K-Means method was used to generate the groups based on the value. Then the schedules are generated by calculating averaged probability for each group and each hour, as described in the flowchart. The number of schedules to be acquired (N) is calculated as the number of clusters, and a schedule with the maximum number of hours of schedule for each group/time is taken as the output at that group / time. K-Means is implemented using KMeans of the scikit-learn/cluster module of Python 2.7, and the initial value of the module are used for parameters other than the number of clusters.

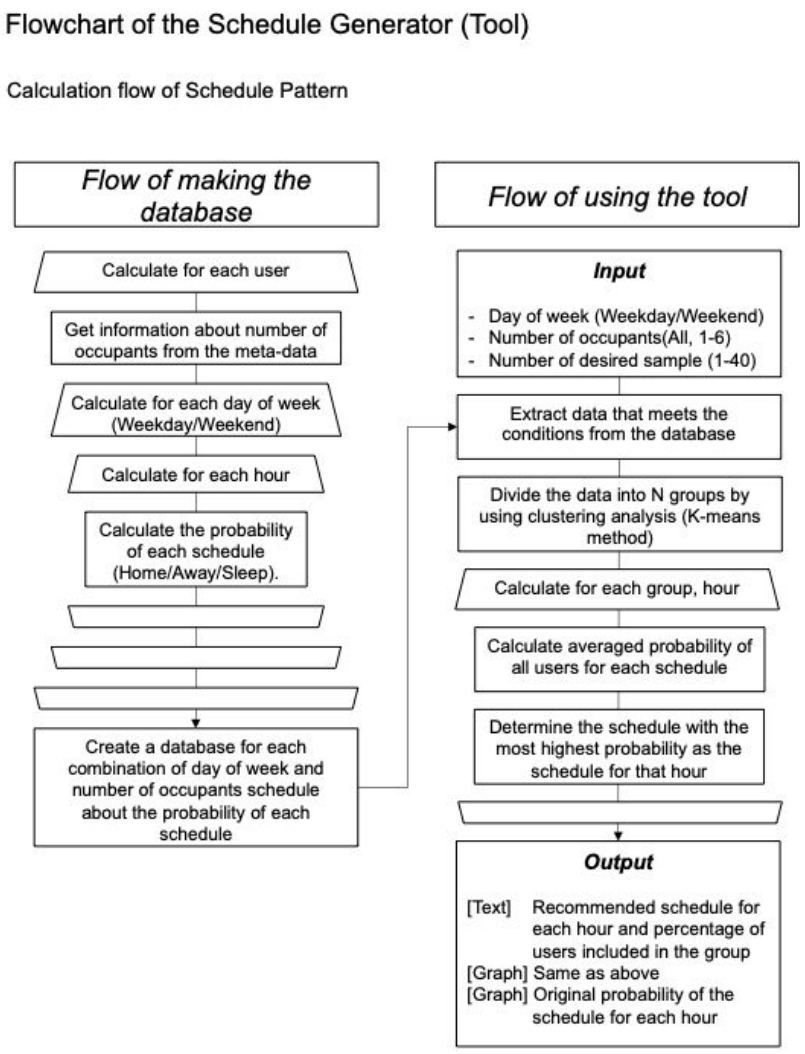


Figure tklIchigo. Flowchart of the logic used to generate schedules

Figure tlkBlueberry illustrates the output for N=4. Table tlkWalnut displays the results in tabular form when four schedules are selected to represent the national housing stock. In two of the schedules for N=4, there are no Away periods. These schedules with no Away time (that is,

somebody is always at home) represent about 57% of the homes.

Figure tlkBlueberry. Schedules for weekdays, all occupants, and all regions for N = 4.

Recommended 4 Typical Schedule for Weekday, Number of Occupants: All are

S:Sleep, H:Home,A:Away

[Hour],0,1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23

1 (34.6%),S,S,S,S,S,S,S,S,H,H,H,H,H,H,H,H,H,H,H,H,H,H,H,S

2 (24.2%),S,S,S,S,S,S,S,S,H,H,H,H,H,H,H,H,H,H,H,H,H,S,S,S

3 (21.8%),S,S,S,S,S,S,S,H,H,A,A,A,A,A,A,A,A,A,A,A,H,H,H,H,S,S

4 (19.4%),S,S,S,S,S,S,S,H,H,A,A,A,A,A,A,A,A,A,H,H,H,H,H,H,S,S

Table TlkWalnut. Recommended modes when four schedules are selected (N=4).

For N=10, even more diverse schedules appear. Figure tlkMango illustrates the output for N=10 and Table tlkPecan displays the results in tabular form. For example, 3.5% of the homes have essentially all of the non-sleeping hours in Away mode during weekdays.

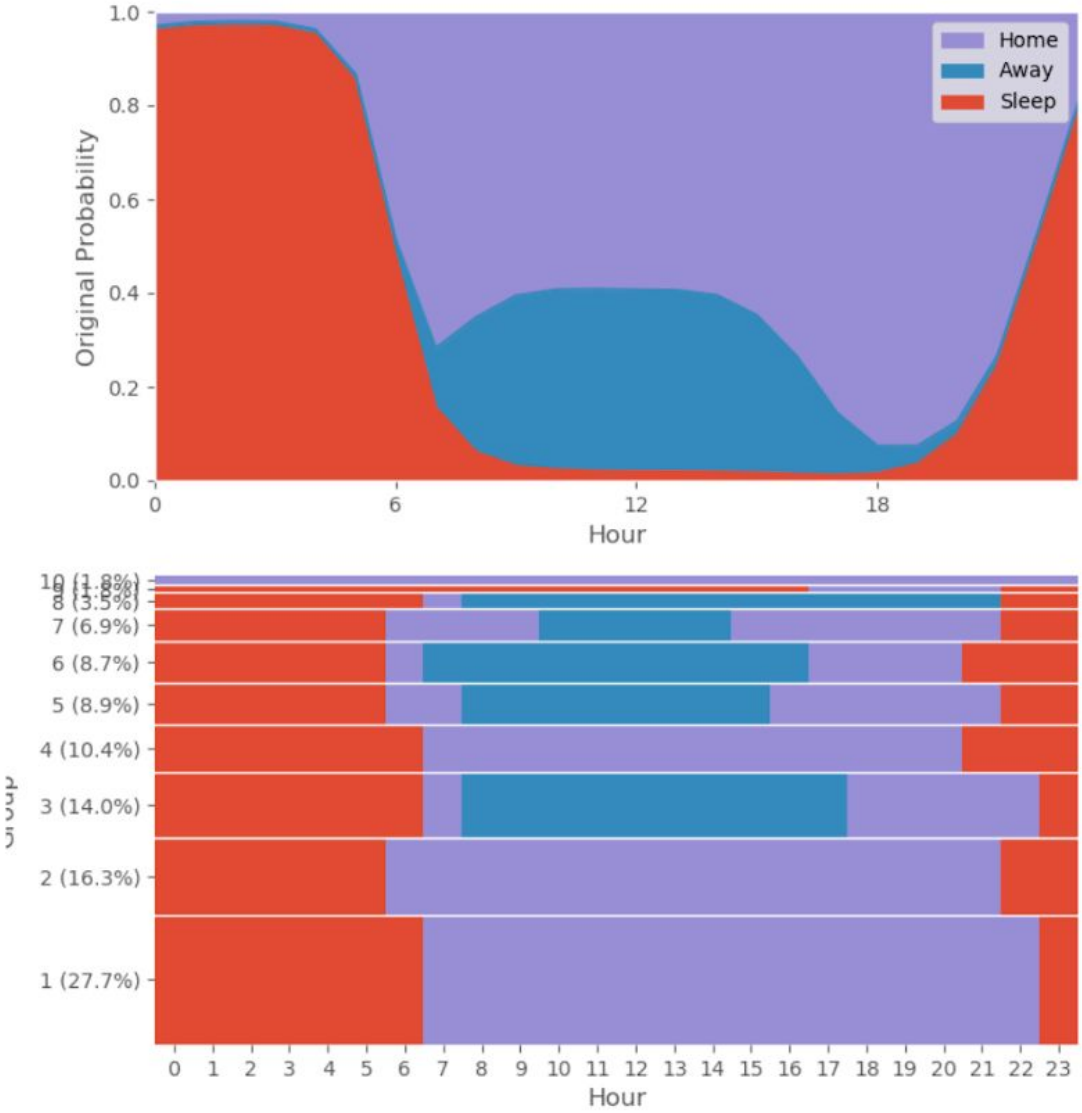


Figure tlkMango. Schedules for weekdays, all occupants, and all regions for N = 10.

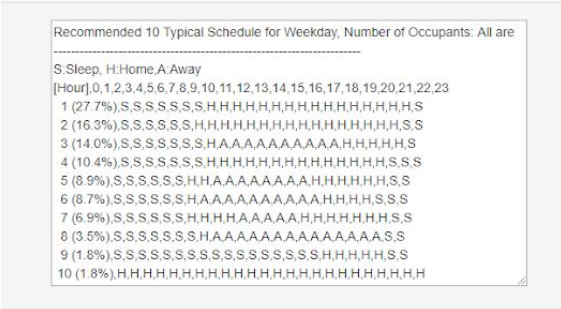


Table tlkPecan displays the results in tabular form when ten schedules are selected to represent the national housing stock.

The power of this method and the underlying data is illustrated in the example below. In this case, six schedules were generated for each climate zone. Separate schedules were generated for weekdays and weekends (and holidays). A further distinction was made between homes with a single occupant and those with four occupants. Tables tlkCashew and tlkPeanut show the temperatures, the schedules, and the fraction of housing stock represented by each schedule. In this case, six prototypes (N=6) were generated for each climate zone. Table tlkCashew shows the temperatures for the prototypes in each climate zone, mode, and season. Note that the temperatures of prototypes in the same climate zone differ by as much as 3°C for the same mode.

	Season:	Summer			Winter		
	Schedule:	Sleep	Away	Home	Sleep	Away	Home
Climate zone	Sample number						
Hot-Humid	1	22.7	24.4	23.5	18.5	17.9	19.5
	2	23.7	25.4	24.4	19.7	18.9	20.3
	3	25.0	26.5	25.3	20.7	19.8	21.2
Mixed-Humid	1	22.3	23.8	22.9	17.9	17.5	19.3
	2	23.4	25.0	23.7	19.1	18.5	20.0
	3	24.7	26.5	24.7	20.1	19.4	20.9
Mixed-Dry/Hot-Dry	1	23.9	25.3	24.3	17.9	17.2	19.3
	2	25.2	26.5	25.3	19.4	18.5	20.3
	3	26.4	27.9	26.1	20.6	19.9	21.3
Marine	1	23.3	24.8	23.5	16.8	17.0	19.1
	2	25.0	26.5	24.8	18.5	18.2	19.9
	3	26.6	27.7	25.6	19.7	19.4	20.6
Very-Cold/Cold	1	22.5	24.1	22.9	17.3	17.2	19.1
	2	23.7	25.4	23.9	18.8	18.4	20.0
	3	25.1	26.9	25.0	19.9	19.4	20.8

Table tlkCashew. Temperature settings for the prototypes in each climate zone, mode, and season.

Table tlkPeanut shows the hour-by-hour schedules for the modes and the percentage of DYD homes represented by the prototype. Separate timings are also generated for weekdays/holidays and for the number of occupants (one or four). The percentages attributed to each prototype vary widely. For example in Weekday (1 occupant), most of the homes are represented by two prototypes about 5.5% of homes are represented by sample number 3. Sample 3 also has a complex schedule because it has two periods while in Away mode.

Day of week	Number of occupants	Sample number	Percentage	Sleep	Away	Home
Weekday	1	1	50.7	0-6, 22-23	8-16	7, 17-21
		2	43.8	0-6, 23	-	7-22
		3	5.5	0-12	13-15, 19-23	16-18
	4	1	43.4	0-5, 22-23	8-15	6-7, 16-21
		2	29	0-5, 23	-	6-22
		3	27.6	0-5, 21-23	-	6-20
Holiday	1	1	50.1	0-5, 23	-	6-22
		2	34.3	0-7, 23	-	6-22
		3	15.6	0-7, 22-23	8-17	18-21
	4	1	47.1	0-6, 23	-	7-22
		2	42.8	0-6, 21-23	-	7-20
		3	10.1	0-6, 22-23	9-16	7-8, 17-21

Table tlkPeanut. Hour-by-hour schedules for the modes and the percentage of DYD homes represented by the prototype. Separate schedules and percentages are also generated for weekdays/holidays and for the number of occupants.

This information is sufficient to create temperature schedules for each prototype and to weight the resulting simulations so as to create a national average heating and cooling energy consumption.

4. A Web-Based Schedule Generator

The above examples were generated for homes located in all climate regions, with all numbers of occupants, during weekdays. Other schedules can be generated for specific climate zones, days of the week, number of occupants, and floor area. However, each schedule requires access to the sorted data as described in Figures tklNashi and tklIchigo. To enable wider access to the results, we developed a web-based tool to generate temperature schedules. Figure TlkPapaya is a screenshot of the user interface.

>>> we need a pretty picture of the user interface
Figure TlkPapaya. A screenshot of the user interface for the web-based schedule generator

The user can specify the number of schedules (up to N = 40), xxxx, and xxxxx. The tool returns graphical displays of the results and tables similar to those presented earlier. These results are suitable for input into schedules for building energy simulation models. The tool makes it possible to quickly identify temperature schedules that may cause unusual energy consumption or performance issues and estimate the fractions of homes falling into those categories.

The tool is available to researchers upon request to the authors. A future version will be made public after the next installment of DYD data has been incorporated. (The next installment will result in a roughly 10-fold increase in the number of homes.)

5. Discussion

The DYD database gives insights into the temperature preferences and schedules in homes that were never before available. Before this, national estimates could only be formed from surveys based on guesses by occupants, with a few temperatures representing behavior through a season and in different types of occupancy. In contrast, the DYD data is based on actual measurements in thousands of homes taken every five minutes. It therefore represents a transformation of our knowledge of heating and cooling preferences from point values to patterns and cycles. This information enables more realistic simulations of American heating and cooling behavior, leading to more accurate estimates of energy consumption and savings. The information can also improve government and utility recommendations for energy-saving thermostat settings. The DYD database has applications not directly related to temperatures, too, such as HVAC sizing or improving estimates of energy consumption of heat pump water heaters.

It is essential to understand the DYD's limitations before generalizing the findings to the entire U.S. housing stock. First, the overwhelming majority of participants are single-family homes. Second, the database contains relatively few homes equipped with heat pumps. Several sources of bias in the participants were also identified, such as self-selection and early-adoption. The participants provided some socio-demographic information but not income, precise location, and other key indicators. Nevertheless, the DYD homes were surprisingly similar to the single-family homes in the Residential Energy Consumption Survey with respect to location, floor area, and number of occupants.

A final limitation is the absence of homes where both temperature and energy data are available. This is mostly an institutional problem—thermostat vendors and utility companies

refuse to share their data—but is also understandable to protect privacy and security. The failure to share energy and temperature data makes it impossible to perform some of the most fundamental explorations, such as the relationship between indoor temperatures and energy use.

Another unknown factor is the manner in which people use their thermostats. We cannot exclude the possibility that DYD participants heat and cool their homes differently than other homes because their thermostats have additional features. One unique feature is remote control (via smartphone or the web), which gives DYD participants the ability to pre-heat or pre-cool their homes. Another feature is that ecobee can adjust temperatures and schedules to reduce HVAC energy consumption (if allowed by the participant). Finally, we have no direct information that the participants are correctly operating their thermostats and are satisfied with the thermostat's performance. Errors in programming thermostat operation have been documented in a large fraction of homes (Meier et al. 2011). However, we have indirect evidence of satisfaction: the participants maintained settings for long periods and kept their thermostats for several years.

The DYD data gives insights into conditions that depart from the average. For example, it is possible to estimate the fraction of homes maintained below 16°C in the winter or cooled to above 28°C in the summer. About XXX% of the homes are continuously occupied (in Home or Sleep modes). The DYD data also reveal that homes are vacant (in Away mode) roughly XX% of the time, which is a feature not always included in simulations.

The DYD database is expected to keep growing and exceed 100,000 participants in 2021. This will provide much more detailed insights into temperature behaviors. Newer thermostats are often equipped with multiple temperature sensors, so researchers can explore the intra-home temperature variations. Unfortunately, the value of a larger sample will be constrained by the poor meta-data about the participants. So an important goal will be to improve the quality of information about the occupants—floor area, demographics, etc.—to complement the rich temperature and HVAC operation data.

This paper explored only the temperature aspects of the DYD data. The Connected Thermostats also collect runtimes of the HVAC equipment. Here, too, completely new insights into residential heating and cooling operation can be obtained. Operational data will help verify the performance of HVAC systems in simulation models in ways that were never before possible. These results will be reported in other communications.

6. Conclusions

A new type of thermostat, which is connected to the Internet, collects temperature and operating data every five minutes from millions of homes in North America and in a growing number in

other countries. This paper explores the application of this data to simulations of energy residential building energy use. The goal is to create more realistic temperatures and schedules in the simulation models than those used today. The approach assumes that a portfolio of simulations, each capturing one set of temperatures and schedules, will provide more insights than a single simulation with average temperatures and schedules. The analysis relies on a unique dataset: the owners of ecobee thermostats who opted to share their thermostat's performance information with researchers through the Donate Your Data Program. At the time of this study, over XXX thousand homes were in the dataset. The DYD data is based on actual measurements in thousands of homes taken every five minutes and represents a transformation of our knowledge of heating and cooling preferences from a few point values to detailed patterns and cycles.

The first step was to determine if the homes in the program were representative of the stock of homes in the United States. A series of comparisons were made between the limited meta-data available from the participants and a national survey of representative homes. The DYD dataset generally matched the survey results for single-family homes with respect to location, floor area, and other characteristics. Thus, we concluded the DYD homes were reasonably representative of the U.S. single-family homes.

A method was developed to generate temperature schedules based on the DYD data. The goal was to create a flexible program that could generate 1 - 40 different temperature schedules for simulations. The program generates distributions of indoor temperatures in each of the three operating modes (Home, Sleep, and Away) and under different conditions, such as season, day of week, and number of occupants. The user must select the number of simulations desired. The program then searches for the temperatures that best reflect the shape of the distribution for the desired number of simulations. Next, the program generates distributions of time that the homes spend in each operating mode, both with respect to actual time of day and the durations. The program then searches for the schedules that best reflect the shape of the distribution for the number of simulations selected. The program outputs hourly temperature profiles, suitable for inputs to building energy simulation programs. The program also calculates the fraction of housing stock for which each profile applies. Thus, the user can weight the results of each simulation to estimate average heating or cooling energy consumption for the entire stock of homes.

The program can also identify the fraction of homes operated with less-common temperatures or schedules. These situations are difficult to capture when simulations only use average conditions yet may be important because they may be associated with unique technical or health problems.

The Donate Your Data database has important limitations—notably the absence of linked energy consumption data—but this study shows the unexpected sources and applications of big data and the insights that these analyses can give into technical, health, and behavioral issues. Further insights are likely as the dataset grows and other characteristics are investigated.

Acknowledgments

The authors deeply appreciate the cooperation of ecobee, Inc. and its anonymous customers who participated in the Donate Your Data program. This work was supported by the U.S. Environmental Protection Agency and the Assistant Secretary for Energy Efficiency and Renewable Energy, Office of Building Technology, U.S. Department of Energy under Contract No. DE-AC02-05CH11231. TU was supported by the Central Research Institute of the Electric Power Industry (Japan).

References

Amirirad, Afarin, Rakesh Kumar, and Alan S. Fung. 2018. “Performance Characterization of an Indoor Air Source Heat Pump Water Heater for Residential Applications in Canada.” *International Journal of Energy Research* 42 (3): 1316–27. https://doi.org/10.1002/er.3932.

Booten, Chuck, Joseph Robertson, Dane Christensen, Mike Heaney, David Brown, Paul Norton, and Chris Smith. 2017. “Residential Indoor Temperature Study.” National Renewable Energy Lab.(NREL), Golden, CO (United States).

Daken, Abigail, Alan Meier, and Douglas Frazee. 2016. “Do Internet-Connected Thermostats Save Energy?” In *ACEEE 2016 Summer Study on Energy Efficiency in Buildings*. Pacific Grove, Calif.: American Council for An Energy Efficient Economy (Washington, D.C.).

Ecobee Inc. 2018. “Donate Your Data.” Ecobee. 2018. https://www.ecobee.com/donateyourdata/.

EIA. 2015. “Residential Energy Consumption Survey (RECS).” Washington, D.C.: Energy Information Administration. https://www.eia.gov/consumption/residential/.

Ge, Qi, and Benjamin Ho. 2018. “Energy Use and Temperature Habituation: Evidence from High Frequency Thermostat Usage Data.” *Economic Inquiry*. https://doi.org/10.1111/ecin.12744.

Healy, John D., and J. Peter Clinch. 2002. “Fuel Poverty, Thermal Comfort and Occupancy: Results of a National Household-Survey in Ireland.” *Applied Energy* 73 (3): 329–43. https://doi.org/10.1016/S0306-2619(02)00115-0.

Hendron, R, and C Engebrecht. 2010. “Building America House Simulation Protocols.” DOE/GO-102010-3141. Golden, CO: National Renewable Energy Laboratory.

Huchuk, Brent, William O’Brien, and Scott Sanner. 2018. “A Longitudinal Study of Thermostat Behaviors Based on Climate, Seasonal, and Energy Price Considerations Using Connected Thermostat Data.” *Building and Environment* 139 (July): 199–210. https://doi.org/10.1016/j.buildenv.2018.05.003.

Johansson, Dennis, Hans Bagge, and Lotti Lindstrii. 2013. “User Related Energy Use in Buildings--Results From Two Years of Measurement of Household Electricity in 1300 Apartments in Sweden.” *ASHRAE Transactions* 119 (2).

Li, Xiwang, and Jin Wen. 2014. “Review of Building Energy Modeling for Control and

Operation." *Renewable and Sustainable Energy Reviews* 37 (September): 517–37. https://doi.org/10.1016/j.rser.2014.05.056.
Lomas, K. J., H. Eppel, C. J. Martin, and D. P. Bloomfield. 1997. "Empirical Validation of Building Energy Simulation Programs." *Energy and Buildings* 26 (3): 253–75. https://doi.org/10.1016/S0378-7788(97)00007-8.
Meier, Alan, Cecilia Aragon, Therese Peffer, Daniel Perry, and Marco Pritoni. 2011. "Usability of Residential Thermostats: Preliminary Investigations." *Building and Environment* 46 (10): 1891–98. https://doi.org/10.1016/j.buildenv.2011.03.009.
Meier, Alan, Tsuyoshi Ueno, and Marco Pritoni. 2019. "Using Data from Connected Thermostats to Track Large Power Outages in the United States." *Applied Energy* 256 (December): 113940. https://doi.org/10.1016/j.apenergy.2019.113940.
Meier, Alan, Tsuyoshi Ueno, Leo Rainer, Marco Pritoni, Abigail Daken, and Dan Baldewicz. 2019. "What Can Connected Thermostats Tell Us about American Heating and Cooling Habits?" In *ECEEE 2019 Summer Study*. Hyères, France: European Council for an Energy Efficient Economy.
Roberts, David, and Kerylyn Lay. 2013. "Variability in Measured Space Temperatures in 60 Homes." NREL/TP-5500-58059. Golden, CO: National Renewable Energy Laboratory.
Seryak, John, and Kelly Kissock. 2003. "Occupancy and Behavioral Affects on Residential Energy Use." In *Proceedings of the Solar Conference*, 717–22. American Solar Energy Society; American Institute of Architects.
Yoshino, H., J. C. Xie, T. Mitamura, T. Chiba, H. Sugawara, K. Hasegawa, K. Genjo, and S. Murakami. 2006. "A Two Year Measurement of Energy Consumption and Indoor Temperature of 13 Houses in a Cold Climatic Region of Japan." *Journal of Asian Architecture and Building Engineering* 5 (2): 361–68. https://doi.org/10.3130/jaabe.5.361.
Yoshino, Hiroshi, Yasuko Yoshino, Qingyuan Zhang, Akashi Mochida, Nianping Li, Zhenhai Li, and Hiroyuki Miyasaka. 2006. "Indoor Thermal Environment and Energy Saving for Urban Residential Buildings in China." *Energy and Buildings*, Energy and Environment of Residential Buildings in China, 38 (11): 1308–19. https://doi.org/10.1016/j.enbuild.2006.04.006.